



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

CENSOS
 **2021**

Estudo de viabilidade para os Censos 2021

Interligação das diferentes bases de dados
provenientes de fontes administrativas, no âmbito
do novo modelo censitário para 2021
(Relatório QUAR*)

*Quadro de Avaliação e Responsabilização

30 de junho de 2015

Versão 1.4

Controlo de versões

Nome ficheiro	Versão	Data	Principais alterações
C2021EV_RELATÓRIO_INTER LIGAÇÃO_FONTES_ADMINIS TRATIVAS_20150324_V1.0.do CX	1.0	2015-03-24	(VERSAO 1.0)
C2021EV_RELATÓRIO_INTER LIGAÇÃO_FONTES_ADMINIS TRATIVAS_20150326_V1.1.do CX	1.1	2015-03-26	(VERSAO 1.1)
C2021EV_RELATÓRIO_INTER LIGAÇÃO_FONTES_ADMINIS TRATIVAS_20150327_V1.2.do CX	1.2	2015-03-27	(VERSAO 1.2)
C2021EV_RELATÓRIO_INTER LIGAÇÃO_FONTES_ADMINIS TRATIVAS_20150330_V1.3.do CX	1.3	2015-03-30	(VERSAO 1.3)
C2021EV_RELATÓRIO_INTER LIGAÇÃO_FONTES_ADMINIS TRATIVAS_20150630_V1.4.do CX	1.4	2015-06-30	(VERSAO 1.4)

Índice

Sumário executivo.....	5
Enquadramento	7
1. Trabalhos de preparação com a CNPD.....	8
2. Acesso a ficheiros administrativos e processo de transmissão	9
2.1. Articulação institucional.....	9
2.2. Processo de encriptação e abreviatura de descritivos.....	10
2.3. Processo de transmissão de dados ao INE	11
3. Análise, tratamento e normalização das bases de dados	12
3.1. ADSE	13
3.2. AT	16
3.3. CGA.....	16
3.4. DGEEC.....	19
3.5. GEP	21
3.6. IEPF.....	24
3.7. IISS	27
3.8. IRN/BDIC.....	29
3.9. SEF	32
3.10. Síntese da informação disponível e principais dificuldades.....	35
4. Ligação entre CENSOS2011 e ficheiros administrativos	44
4.1. Trabalhos preparatórios nas várias bases de dados	45
4.2. Identificação de problemas encontrados nos dados	47
4.3. Comparação CENSOS2011 e BDIC2010.....	48
4.4. Comparação CENSOS2011 e BDIC2014.....	55
4.5. Comparação CENSOS2011 e Segurança Social 2013.....	58
4.6. Comparação usando o Quality Stage	61
4.7. Síntese dos principais resultados	63
4.8. Caracterização dos registos que não foi possível ligar	65
5. Avaliação da qualidade do processo de interligação das bases de dados	70
5.1. Como determinar a qualidade das ligações?	70

5.2.	Exatidão das ligações	72
5.3.	Resultados – Indicadores de Qualidade	73
5.4.	Outros indicadores da qualidade das ligações.....	75
5.5.	Considerações finais sobre a qualidade do processo	78
6.	Conclusões e trabalho futuro	79

Sumário executivo

O presente relatório sintetiza os primeiros resultados do programa de trabalho conjunto entre o GC 2021 e o DMSI, no âmbito da ligação de ficheiros e da infraestrutura informacional de suporte ao estudo de viabilidade para um novo modelo censitário para 2021.

Num quadro de mudança de um modelo censitário que passe pela utilização de fontes administrativas, os processos de ligação entre registos assumem um papel fundamental, nomeadamente para a construção de uma Base de População Residente (BPR). A constituição de uma base de dados do tipo censitário a partir de informação dispersa por diferentes fontes administrativas, depende da capacidade de integração dos registos das diversas fontes, pelo que a eficácia dos métodos de *matching* de registos são um fator decisivo no processo.

Face à inexistência de um identificador comum às diferentes fontes administrativas, e em algumas situações, às baixas taxas de preenchimento dos identificadores que possibilitam algumas ligações, são utilizados processos de *matching* através de chaves de atributos respeitantes a características das unidades estatísticas em análise. Por outro lado, e tendo também como objetivo o eventual aproveitamento de alguma informação administrativa para substituição de dados recolhidos por inquérito, a integração dessa informação passará sempre pela utilização de chaves de *matching* não numéricas.

Neste sentido, foi realizado um exercício de ligação entre os Censos 2011 e os ficheiros administrativos BDIC (2010 e 2014) e Segurança Social (2013), através da utilização de 12 chaves de ligação. As chaves de ligação utilizadas combinaram de forma diferenciada informação relativa a várias características dos indivíduos, nomeadamente, os primeiros 3 caracteres do primeiro nome, 3 primeiros ou últimos caracteres do último nome, sexo, data de nascimento, freguesia/município de residência, naturalidade e estado civil.

Da ligação entre os Censos 2011 e os ficheiros administrativos utilizados foi atribuído NIC a 9 654 666 registos censitários, o que corresponde a uma taxa de 91,1%. Refira-se que a grande maioria dos registos foi ligada na primeira etapa de *matching*, em que foi garantida a igualdade de 8 atributos comuns.

A qualidade do processo de *matching* foi avaliada através da exatidão das ligações e da estimação dos falsos positivos. Os resultados obtidos apontam para uma taxa de

exatidão das ligações de 95%, ou seja, pares de registos que pertencem ao mesmo indivíduo.

Com a manutenção das características dos ficheiros e com a eventual possibilidade de serem colmatadas parte das limitações associadas a alguns atributos, a par de um maior aperfeiçoamento dos processos e técnicas de *matching*, poderão estar reunidas as condições para assegurar a interligação de ficheiros, a um nível que permita a construção de uma base de população com base em informação administrativa.

O presente relatório sintetiza os trabalhos desenvolvidos, até ao momento, no que respeita à ligação das diferentes bases de dados. Neste sentido, o relatório aborda os seguintes pontos:

No ponto 1 – *Trabalhos de preparação com a CNPD*, refere-se a criação do enquadramento legal necessário para acesso por parte do INE aos ficheiros administrativos.

No ponto 2 – *Acesso a ficheiros administrativos e processo transmissão*, evidencia-se a articulação institucional entre o INE e as entidades detentoras da informação administrativa e apresenta-se o sistema de encriptação dos dados e o processo de transmissão.

No ponto 3 – *Análise, tratamento e normalização das bases de dados*, são analisadas as diferentes fontes administrativas na componente de metainformação e indicadores de qualidade. Apresenta-se a normalização efetuada aos dados e os principais resultados da ligação entre as várias bases através dos identificadores numéricos NIC, NISS e NIF.

No ponto 4 – *Ligação entre os Censos 2011 e ficheiros administrativos*, apresenta-se o processo de *matching* entre os Censos, a BDIC e a Segurança Social e os respetivos resultados.

No ponto 5 - *Avaliação da qualidade do processo de interligação das bases de dados*, são apresentados os resultados da avaliação da qualidade do processo de matching realizado.

No ponto 6 apresentam-se as conclusões e o trabalho a desenvolver no futuro.

Enquadramento

O presente relatório constitui um primeiro passo para o estudo e aplicação de processos de *matching* de microdados, indispensável para a utilização de ficheiros administrativos para fins estatísticos. O *matching* de microdados assume um papel decisivo no quadro do estudo de viabilidade em curso no INE, relativo ao modelo censitário a aplicar em 2021.

Face à inexistência, em Portugal, de um identificador comum às diferentes fontes administrativas (a Constituição da República impede a existência de um número único do cidadão), revelam-se decisivos os processos de *matching* através de chaves de atributos alfanuméricos, nos quais são utilizadas características das unidades estatísticas em análise.

Os ficheiros administrativos já analisados possuem, em regra, apenas um ou dois dos principais números de identificação utilizados em Portugal: Nº de Identificação Civil (NIC), Nº de Identificação Fiscal (NIF) e Nº de Identificação da Segurança Social (NISS). Por outro lado, nalguns casos, verificam-se elevadas taxas de não preenchimento destes identificadores numéricos.

O desenvolvimento deste estudo tem como *objetivos*:

- Permitir a ligação entre os ficheiros administrativos disponibilizados por várias entidades da Administração Pública e os ficheiros provenientes de inquéritos estatísticos, designadamente a base dos Censos 2011.
- Apoiar a operacionalização da metodologia de constituição de uma Base de População Residente (BPR).
- Preparar a progressiva substituição de informação recolhida através de inquéritos por informação administrativa.

O aprofundamento metodológico nestas matérias requer investimento e conhecimento especializado. Países com processos já consolidados de alteração do modelo censitário, como os países nórdicos, investem continuamente nesta área. Em países como a Espanha, que utilizou em 2011 um modelo censitário misto, o *matching* entre ficheiros, foi uma peça fundamental no processo de mudança.

1. Trabalhos de preparação com a CNPD

A legislação dos Censos 2011, Decreto-Lei 226/2009, de 14 de setembro, enquadrava legalmente o acesso, por parte do INE, aos registos individuais dos ficheiros administrativos. Com a conclusão dessa operação estatística, o INE, por deliberação da CNPD, deixou de poder aceder aos dados individualizados necessários à realização do estudo de viabilidade.

Neste sentido, foi necessário a criação de condições legais para acesso aos registos administrativos, fator decisivo para o avanço do estudo de viabilidade para o novo modelo censitário.

A par da Lei que regula o Sistema Estatístico Nacional, Lei nº 22/2008 de 13 de maio, foi necessária a articulação com a Comissão Nacional de Proteção de Dados (CNPD) para a produção de uma deliberação que possibilitasse o acesso à informação administrativa necessária à realização do estudo de viabilidade.

Neste contexto, foi criado um mecanismo de colaboração entre o INE e a CNPD de modo a encontrar as soluções técnicas que permitissem o acesso à informação detida por vários organismos da administração pública, tendo como pressuposto que os dados individuais não sejam suscetíveis de identificação.

A Deliberação da CNPD nº 929/2014, datada de 11 de junho, conjugada com a Lei do SEN, conferem a moldura legal necessária e indispensável, para esta fase do projeto onde se pretende avaliar os contributos das diferentes fontes administrativas. Em momento posterior e, tendo em conta os resultados do estudo de viabilidade, a Lei de enquadramento dos Censos 2021, deverá incorporar também os princípios subjacentes na referida deliberação.

Importa contudo sublinhar que a deliberação da CNPD impõe algumas restrições no acesso à informação administrativa, nomeadamente no acesso ao nome dos indivíduos e à respetiva morada. No nome apenas se encontram disponíveis as 3 primeiras letras do 1º nome e as 3 últimas letras do último nome, e quanto à morada apenas se dispõe de informação sobre a Localidade e o Código Postal. Todos os identificadores numéricos (NIC, NIF, NISS) foram encriptados pelas entidades responsáveis pelos dados, recebendo o INE os ficheiros com a informação já anonimizada através de um canal seguro.

2. Acesso a ficheiros administrativos e processo de transmissão

2.1. Articulação institucional

A articulação com as várias entidades envolvidas é um fator decisivo no desenvolvimento de um projeto desta natureza.

Neste sentido, na sequência da deliberação da CNPD foi enviada à tutela do INE (Secretário de Estado para a Modernização Administrativa), em setembro de 2014, uma proposta de Resolução do Conselho de Ministros que visa fomentar as condições de cooperação e articulação do INE com as várias entidades envolvidas neste estudo.

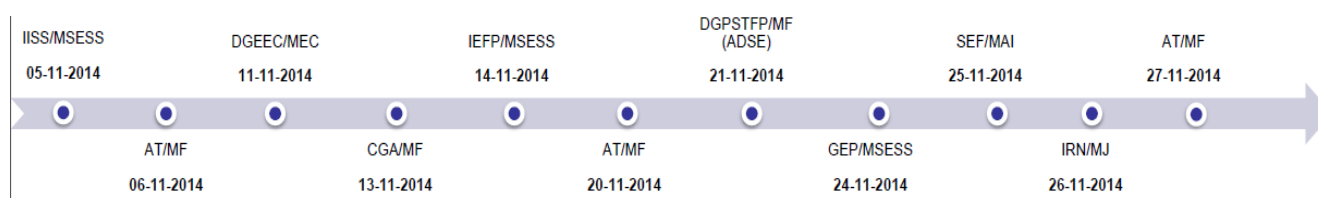
Tendo em conta a informação necessária ao estudo, as entidades incluídas foram, nesta fase, as seguintes:

- [IRN] Instituto dos Registos e do Notariado (Ministério da Justiça)
- [SEF] Serviço de Estrangeiros e Fronteiras (Ministério da Administração Interna)
- [DGEEC] Direção-Geral de Estatísticas da Educação e Ciência (Ministério da Educação e Ciência)
- [IISS] Instituto de Informática da Segurança Social (Ministério da Solidariedade, Emprego e Segurança Social)
- [GEP] Gabinete de Estratégia e Planeamento (Ministério da Solidariedade, Emprego e Segurança Social)
- [IEFP] Instituto do Emprego e Formação Profissional (Ministério da Solidariedade, Emprego e Segurança Social)
- [AT] Autoridade Tributária (Ministério das Finanças)
- [CGA] Caixa Geral de Aposentações (Ministério das Finanças)
- [ADSE] Direção-Geral de Proteção Social aos Trabalhadores em Funções Públicas (Ministério das Finanças)
- [SRERH] Secretaria Regional de Educação e Recursos Humanos da R.A. da Madeira (Governo Regional da Madeira)
- [IEM] Instituto de Emprego da Madeira (Governo Regional da Madeira)
- [SREC] Secretaria Regional da Educação e Cultura dos Açores
- [DREQP] Direção Regional do Emprego e Qualificação Profissional

Em paralelo, e durante o mês de outubro de 2014, foram encetados contactos bilaterais com todas as entidades, através do envio de uma carta explicando o estudo que se pretende desenvolver, respetivo enquadramento legal e solicitando a indicação de um interlocutor para o projeto.

A cooperação por parte dos organismos detentores dos ficheiros administrativos foi determinante no avanço do estudo de viabilidade e permitiu que durante o mês de novembro fosse possível ao INE reunir com todas as entidades (ver figura 1 abaixo). Nestas reuniões foi definido o conteúdo informacional necessário ao estudo de viabilidade, as condições de tratamento e de transmissão da informação.

Figura 1 – Cronograma das reuniões entre o INE e as várias entidades



O clima de grande cooperação e empenhamento revelado pelas diversas entidades, permitiu que ainda no mês de dezembro de 2014 fossem recebidos os primeiros ficheiros.

Neste momento foi recebida praticamente toda a informação solicitada, estando apenas em falta a informação proveniente da Autoridade Tributária.

2.2. Processo de encriptação e abreviatura de descritivos

O INE desenvolveu uma aplicação que permite a cada entidade realizar a encriptação, abreviatura e transmissão ao INE das variáveis a disponibilizar no âmbito da deliberação da CNPD nº929/2014, de 11/06/2014.

Esta aplicação de Codificação de Dados Administrativos (CDA) permite que aos identificadores numéricos – número de identificação civil (NIC), número de identificação fiscal (NIF) e número de identificação de segurança social (NISS) seja aplicado na fonte um processo de encriptação de dados por aplicação de um *hash* com um algoritmo

SHA256, recebendo o INE os ficheiros com a informação anonimizada. A aplicação permite ainda a abreviatura do campo Nome, seleccionando as 3 primeiras letras do primeiro nome e as 3 últimas letras do último nome.

Através de uma interface gráfica com o utilizador ou por linha de comando, é possível seleccionar o ficheiro a tratar, previamente produzido. É ainda possível a identificação da estrutura do ficheiro, bem como os campos e operações a realizar: encriptação, abreviatura ou cópia simples. A CDA permite a visualização do tempo de execução, bem como o resumo e ocorrências durante o processamento. Após o processamento bem-sucedido, o ficheiro é compactado e mediante a introdução de utilizador e palavra-chave - previamente disponibilizadas a cada uma das entidades pode efetuar-se a transferência segura para a infraestrutura do INE. No final, é indicado ao utilizador o sucesso da transmissão efetuada.

2.3. Processo de transmissão de dados ao INE

As transmissões de dados das fontes administrativas são efetuadas em canal de comunicação seguro, com utilização de credenciais de autenticação.

A transmissão do ficheiro é preferencialmente realizada através da utilização da aplicação CDA, que centraliza num único processo e repositório centralizado a informação recebida no INE.

Ao interlocutor indicado por cada entidade são fornecidas as respetivas credenciais de acesso: utilizador e palavra-chave. Estas credenciais são utilizadas na aplicação de encriptação aquando da transmissão do ficheiro para o INE. Aquando da utilização da aplicação CDA, é enviada uma notificação para uma lista de distribuição de contactos de correio eletrónico dos responsáveis pela receção dos ficheiros.

Para este processo de transmissão e armazenamento, o INE disponibiliza uma infraestrutura segura nos seus próprios servidores, com uma área específica para cada entidade, acessível através de *hyper-text transport protocol secure (https)* a um endereço *universal resource locator (url)*, mediante a utilização de um código de utilizador e palavra-chave previamente referidos.

A palavra-chave pode ser alterada no *url*, <https://cloud.ine.pt> com as credenciais fornecidas.

3. Análise, tratamento e normalização das bases de dados

A análise e o tratamento de informação proveniente de várias fontes administrativas coloca alguns desafios, quer pela variedade de estruturas de dados, das codificações e de formatos existentes, quer pela especificidade de algum tipo de informação.

No sentido de se compatibilizar a informação das diversas fontes foi definido um conjunto de regras de uniformização de variáveis, cujo objetivo foi reduzir as diferenças nas codificações e nos formatos utilizados.

Esta uniformização passou pela própria designação das variáveis e pela respetiva codificação dos dados. Sempre que possível, foram utilizadas as mesmas tabelas de descodificação nas variáveis disponibilizadas pelas várias fontes (Exemplo: Códigos para a representação dos nomes dos países Norma internacional - *Iso alpha2*; Código da divisão administrativa em vigor, Sexo, Estado Civil, etc.).

Esta normalização possibilitou a análise e comparação da informação das diversas fontes de forma mais eficaz.

No tratamento e normalização dos dados das fontes administrativas, desenvolveu-se um conjunto de processos que resultaram na construção de uma ou mais tabelas, da respetiva fonte, para cada ano. Neste sentido, foi adicionado, para todas as fontes, o campo ANO, de forma a identificar claramente, o ano que os respetivos dados correspondem.

Relativamente às variáveis, todas as fontes têm associada pelo menos uma variável de identificação, NIC, NISS, NIF ou Nome, variáveis estas que nem sempre se encontram preenchidas na totalidade. A este propósito, refira-se que para o cálculo da taxa de preenchimento, sempre que possível, foram excluídos, para além dos registos nulos, os registos inválidos.

Destacam-se alguns dos campos alvo de tratamento / normalização:

- Distrito, Município e Freguesia (de residência ou naturalidade): Foi aplicada a tabela de divisão administrativa atualmente em vigor. Nas fontes em que os dados incluíam apenas os códigos, foram associadas as respetivas designações. Em

algumas fontes apenas estava disponível a designação da unidade administrativa, pelo que houve necessidade de ser associado o respetivo código;

- País (de residência, naturalidade, nacionalidade): Foi aplicada a tabela de *Códigos para a representação dos nomes dos países Norma internacional - Iso alpha2*, o que originou a conversão de códigos utilizados em algumas fontes. Em algumas situações existiam apenas os códigos pelo que tiveram que ser associadas as respetivas designações;
- Datas: A todas as datas existentes nas fontes, nomeadamente data de nascimento, data de validade, data de emissão, etc., foi aplicado o mesmo formato. As várias datas foram transformadas em variáveis numéricas de forma a maximizar a sua pesquisa e ordenação;
- Nome: Retirou-se a acentuação e foi transformado em maiúsculas;
- Sexo: Foi recodificado em algumas fontes, de forma a possuir códigos comuns e uma tabela de descodificação única;
- Estado Civil: Foi recodificado em algumas fontes, de forma a possuir códigos comuns e uma tabela de descodificação única;
- Código Postal: Variável dividida em duas CP4 e CP3, a primeira corresponde aos primeiros quatro dígitos do código postal e a segunda aos restantes três dígitos;
- Tipos de documentos de identificação: Campo recodificado em algumas fontes, de forma a possuir códigos comuns e uma tabela de descodificação única.

Para comodidade de análise, quando numa fonte existia uma variável codificada e foi disponibilizada a tabela utilizada, foi adicionada sempre que possível uma nova coluna com a sua descodificação.

3.1. ADSE

A ADSE (Direção-Geral de Proteção Social aos Trabalhadores em Funções Públicas) é um Serviço Integrado do Ministério das Finanças e da Administração Pública, que tem a responsabilidade de assegurar a proteção aos beneficiários nos domínios da promoção da saúde, prevenção da doença, tratamento e reabilitação.

A informação disponibilizada pela ADSE revela-se importante na tentativa de caracterização dos trabalhadores da administração pública, área que apresenta algumas lacunas ao nível de dados individualizados.

Os dados da ADSE abrangem:

- Os trabalhadores com relação jurídica de emprego público da administração central, regional e local, desde que estejam inscritos na Caixa Geral de Aposentações ou na Segurança Social, e não beneficiem, como titulares, de outro subsistema de saúde integrado na Administração Pública;
- O pessoal docente do ensino particular e cooperativo, desde que para o efeito seja celebrado um acordo com a ADSE ou outro pessoal contemplado na legislação
- Os aposentados que não sejam abrangidos por qualquer outro subsistema de saúde integrado na Administração Pública;
- Os dependentes dos titulares: cônjuges, descendentes e ascendentes ou equiparados.

Para a ADSE estão carregados os anos de 2012 e 2013.

Foram detetados alguns registos anómalos: para o ano de 2012 a variável Número CGA apresenta alguns registos não encriptados; não foram considerados 350 085 registos com o mesmo número de ADSE e que não possuíam qualquer tipo de informação adicional.

A ficha técnica que se apresenta em seguida sintetiza a informação disponibilizada por esta fonte.

Metainformação	
Anos disponíveis e datas de receção	Universo
	Beneficiários da ADSE (Titulares - trabalhadores no ativo ou aposentados - e Familiares ou equiparados) nomeadamente: os trabalhadores com relação jurídica de emprego público da administração central, regional e local, inscritos na Caixa Geral de Aposentações ou na Segurança Social, e não beneficiem de outro subsistema de saúde integrado na Administração Pública; o pessoal docente do ensino particular e cooperativo, desde celebrado acordo com a ADSE; os aposentados que não sejam abrangidos por qualquer outro subsistema de saúde integrado na Administração Pública; os dependentes dos titulares: cônjuges, descendentes e ascendentes ou equiparados.
Âmbito geográfico	Variáveis disponíveis
Portugal	Nº ADSE, NIC, NIF, NCGA, sexo, nome, data de nascimento, estado civil, parentesco, país distrito, município, freguesia de residência, localidade, código postal, data de inscrição, serviço, movimento despesas reembolso: regime livre e convencionado.
Unidade de observação	
Indivíduo	

Total de registos		Registos únicos	
2012	2013	2012	2013
1 377 353	1 290 784	1 377 353	1 290 784

Taxas de matching	
ADSE2012	Sem matching (problemas no NIC encriptado em tratamento)
ADSE2013	Sem matching (problemas no NIC encriptado em tratamento)

Indicadores de qualidade			
Nome da variável	Designação da variável	Taxa de preenchimento (%)	
		2012	2013
NUB_ADSE	Número de identificação da ADSE	100,0	100,0
NIC	Número de identificação civil	67,5	66,5
NIF	Número de identificação fiscal	64,8	65,3
NCGA	Número de CGA	-	39,5
SEXO	Sexo	100,0	100,0
DTNASC	Data de nascimento	100,0	100,0
NOME	Nome	100,0	100,0
PAIS_RES_COD	País de residência	100,0	100,0
DT_RESID_COD	Distrito de residência	99,9	99,9
MN_RESID_COD	Município de residência	99,5	99,5
FR_RESID_COD	Freguesia de residência	37,6	34,6
MORADA_LOCALIDADE	Localidade	100,0	100,0
CP4	4 primeiros dígitos do código postal	100,0	100,0
CP3	3 últimos dígitos do código postal	85,5	86,7
ESTCIVIL	Estado civil	100,0	100,0
PARENTESCO	Situação parentesco na ADSE (Titular, filho, cônjuge, ascendente,...)	100,0	100,0
DT_INS_ADSE	Data de inscrição na ADSE	100,0	100,0
NOME_SERVIÇO	Nome serviço/entidade	100,0	100,0
MOVDESP_RL2012	Movimento de despesas Reembolso 2012 (Regime livre)	100,0	100,0
MOVDESP_RC2012	Movimento de despesas Reembolso 2012 (Regime convencionado)	100,0	100,0

3.2. AT

Foi solicitada à AT informação sobre o cadastro dos contribuintes, o IRS e o IMI, que ainda não foi rececionada no INE.

3.3. CGA

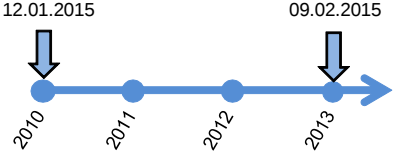
A Caixa Geral de Aposentações (CGA) é a instituição de previdência que tem a seu cargo a gestão do regime de segurança social dos funcionários públicos e trabalhadores equiparados, admitidos até 31 de dezembro de 2005, em matéria de pensões de aposentação, de reforma, de sobrevivência e outras de natureza especial, designadamente, pensões de preço de sangue e pensões por serviços excepcionais prestados ao País.

Os dados da CGA apresentam especial interesse como fonte complementar ao ficheiro da Segurança Social, quer na caracterização das pessoas que já não se encontram no ativo, mas também dos funcionários da administração pública com data de admissão anterior a 2006.

A informação disponibilizada pela CGA abrange os indivíduos subscritores (funcionários públicos e agentes administrativos civis e militares da Administração Pública Central, Local e Regional, professores do ensino particular e cooperativo e trabalhadores de algumas empresas públicas e sociedades anónimas de capitais públicos), os reformados da função pública e os pensionistas de sobrevivência de funcionários públicos ou equiparados.

No ficheiro da CGA constavam duplicados ao nível do número de subscritor, consequência do mesmo indivíduo poder acumular diferentes situações: por exemplo ser pensionista de sobrevivência ou subscritor e outra situação. Neste sentido foi implementado um conjunto de regras de modo a assegurar que, para cada número de subscritor, exista apenas um registo. Na definição das regras deu-se prevalência às situações de atividade económica, de acordo com as regras habitualmente utilizadas em termos censitários.

Nesta fonte estão disponíveis dois anos: 2010 e 2013. Para ambos os anos as variáveis Nome e número CGA encontram-se preenchidas na totalidade. A ficha que se apresenta de seguida sintetiza os principais indicadores relativos a esta fonte administrativa.

Metainformação	
Anos disponíveis e datas de receção 	Universo Indivíduos subscritores (funcionários públicos e agentes administrativos civis e militares da Administração Pública Central, Local e Regional, professores do ensino particular e cooperativo e trabalhadores de algumas empresas públicas e sociedades anónimas de capitais públicos), os reformados da função pública e os pensionistas de sobrevivência de funcionários públicos ou equiparados.
Âmbito geográfico Portugal	Variáveis disponíveis ANO, NUMERO DE SUBSCRITOR, NIC, NIF, NOME, SEXO, DATA DE NASCIMENTO, CÓDIGO POSTAL, DISTRITO E MUNICÍPIO DE RESIDÊNCIA, SITUAÇÃO NA CGA, ORIGEM, MINISTÉRIO, SERVIÇO, MORADA DO SERVIÇO
Unidade de observação Indivíduo	

Total de registos		Registos únicos	
2010	2013	2010	2013
1 130 790	1 085 872	1 103 980	1 080 116

Taxas de matching	
	BDIC NIC
CGA 2010	94,8%
CGA 2013	96,1%

Indicadores de qualidade			
Nome da variável	Designação da variável	Taxa de preenchimento (%)	
		2010	2013
ANO	Ano	100,0%	100,0%
NUMERO_SUBSCR	Número de subscritor	100,0%	100,0%
NOME	Nome	100,0%	100,0%
BI_00	Bilhete de identidade	71,4%	77,0%
NIF	Número de identificação fiscal	84,6%	81,9%
CP4	Código postal	50,8%	57,6%
CP3	Código postal auxiliar	50,8%	57,6%
DT_RESID_COD	Código do Distrito de Residência	49,8%	56,6%
MN_RESID_COD	Código do Município de Residência	49,8%	56,6%
SEXO	Sexo	100,0%	100,0%
DTNASC	Data de Nascimento	100,0%	100,0%
SITUACAO	Situação	100,0%	100,0%
ORIGEM	Origem	81,2%	79,2%
MINISTERIO	Ministério	81,2%	79,2%
SERVICO	Serviço	81,9%	82,3%
RUA_SERV	Rua do Serviço	81,4%	81,8%
LOCALIDADE_SERV	Localidade do Serviço	39,0%	37,3%
CP4_SERV	Código de postal do Serviço	81,9%	82,3%
CP3_SERV	Código postal auxiliar do Serviço	81,9%	82,3%

3.4. DGEEC

A Direção Geral de Estatísticas da Educação e Ciência (DGEEC) é a entidade responsável pela recolha, monitorização, tratamento, produção e divulgação de informação estatística nas áreas de intervenção do Ministério da Educação.

A informação rececionada abrange os alunos matriculados/inscritos nos diversos níveis de ensino público e privado: Ensino Básico, Secundário e Superior. Os dados disponibilizados, para além de permitirem caracterizar a população que frequenta o sistema de ensino, são fundamentais para a validação da população residente para fins censitários, atendendo a que produzem indícios de residência muito consistentes.

Dado que a DGEEC é responsável apenas pela informação relativa ao continente, para os níveis de ensino básico e secundário, foi necessário contactar as entidades das regiões autónomas. A Secretaria Regional de Educação e Recursos Humanos da Região Autónoma da Madeira disponibilizou a informação solicitada. No entanto, para a Região Autónoma dos Açores não estão atualmente disponíveis dados individualizados dos alunos, não se prevendo que esta situação se altere a curto/médio prazo. Foi necessário compatibilizar a informação proveniente das duas entidades, de forma a integrarem uma mesma tabela de dados. Essa integração já foi realizada para o ano letivo de 2013/2014, estando ainda em tratamento a integração para o ano letivo de 2011/2012.

Para esta fonte estão carregados os anos de 2011 e 2013. A obtenção da base de dados final resultou do carregamento de diferentes tabelas relativas às escolas privadas, escolas públicas e ao ensino superior, que continham estruturas diferentes, pelo que foi necessário estabelecer uma estrutura e codificação comuns.

A ficha técnica, que a seguir se apresenta, sintetiza a informação disponibilizada por esta fonte.

Metainformação	
Anos disponíveis e datas de receção	Universo
25.03.2014 ↓ 2011/12 17.12.2014 23.01.2015 ↓ 2013/14	Alunos matriculados nos diferentes graus e modalidades de ensino (ensino básico, secundário e superior). Os níveis de ensino creche e pré-escolar apenas estão disponíveis para a Madeira. Foram enviados em 2011, 7 ficheiros distintos e em 2013, 5 tabelas distintas (4 do Continente e 1 da Madeira)
Portugal continental e R.A. Madeira	Variáveis disponíveis
Unidade de observação	ANO E ANO LETIVO, NIC, NISS, TIPO E NÚMERO DO DOCUMENTO DE IDENTIFICAÇÃO, NOME, SEXO, DATA DE NASCIMENTO, NACIONALIDADE, CÓDIGO POSTAL, DISTRITO, MUNICÍPIO E FREGUESIA DE RESIDÊNCIA, CICLO E MODALIDADE DE ENSINO, CURSO, ANO E SITUAÇÃO, MUNICÍPIO E FREGUESIA DA ESCOLA
Indivíduo	

Total de registos	
2011 *	2013
1 965 842	1 706 779

Taxas de matching	
	BDIC NIC
DGEEC2011	88,0%
DGEEC/DEM2013	98,3%

Indicadores de qualidade			
Nome da variável	Designação da variável	Taxa de preenchimento (%)	
		2011	2013
ANO	Ano	100,0%	100,0%
A0_ANOLETIVO	Ano letivo	100,0%	100,0%
NOME	Nome	100,0%	100,0%
A2_TIPODOCID	Tipo de documento de identificação	98,4%	100,0%
A2_NUMDOCID	Número de identificação	98,6%	100,0%
A2_NUMDOCID_CHKDIGIT	Número de controlo do documento de identificação	-	64,7%
NISS	Número de identificação da Segurança Social	-	62,5%
PAIS_NAC_COD	País de Nacionalidade	99,9%	100,0%
DT_NASC	Data de nascimento	100,0%	100,0%
SEXO	Sexo	100,0%	100,0%
CP4	Código postal	73,4%	66,1%
CP3	Código postal auxiliar	58,7%	64,7%
FR_RESID_COD	Freguesia de residência	65,1%	71,6%
MN_RESID_COD	Município de residência	84,6%	92,0%
DT_RESID_COD	Distrito de residência	84,9%	92,0%
A10_PLANOESTUDOS	Plano de Estudos	-	75,9%
A10_ENSINO	Nível de Ensino	100,0%	100,0%
A10_CICLO	Ciclo de Ensino	71,7%	98,8%
A12_MODALIDADE	Modalidade de Ensino	78,3%	78,4%
A12_CURSO	Curso	47,7%	97,6%
A13_ANOTIPO	Ano de Ensino	90,4%	78,6%
A16_SITUACAO	Situação de aprovação	93,8%	78,9%
A20_CODFREGUESIAESC OLA	Freguesia da Escola	2,8%	78,9%
A20_CODCONCELHOESC OLA	Município da Escola	2,8%	78,9%

3.5. GEP

Os Quadros de Pessoal, Anexo A do Relatório Único, são disponibilizados pelo Gabinete de Estratégia e Planeamento do Ministério da Solidariedade, Emprego e Segurança Social (GEP) e resultam da obrigação anual, a cargo dos empregadores, do envio de informação sobre a atividade social da empresa.

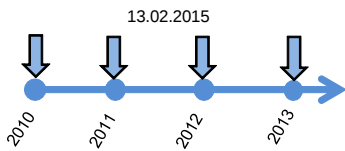
Na perspetiva do aproveitamento de informação administrativa para efeitos censitários, a informação disponibilizada nos Quadros de Pessoal funciona como informação complementar aos dados da Segurança Social na área do emprego e mercado de trabalho, possibilitando uma caracterização mais alargada do grupo dos trabalhadores com contrato individual de trabalho.

Nos Quadros de Pessoal estão incluídas as pessoas empregadas com contrato individual de trabalho, em empresas com pelo menos uma pessoa ao serviço e abrangidas pelo código do trabalho. Estão fora do âmbito dos Quadros de Pessoal os trabalhadores por conta própria, os trabalhadores do serviço doméstico, os trabalhadores da administração pública, dos institutos públicos e os que possuem regime de contrato de trabalho em funções públicas.

Foi necessário proceder ao tratamento da informação disponibilizada nos Quadros de Pessoal, no sentido de assegurar apenas um registo por indivíduo. Dado que existem pessoas que trabalham em mais do que uma empresa, e tendo em conta que para efeitos censitários apenas se caracteriza o “emprego principal”, os registos foram selecionados de acordo com o maior número de horas de trabalho na semana, critério idêntico ao utilizado habitualmente nos censos da população.

Desta fonte estão carregados quatro anos: 2010, 2011, 2012 e 2013. As variáveis relativas ao NISS e ao Nome encontram-se totalmente preenchidas, sendo que a variável do nome apresenta, por vezes, alguns caracteres especiais que não pertencem à lista padrão definida.

Apresenta-se de seguida a ficha técnica relativa aos Quadros de Pessoal.

Metainformação							
Anos disponíveis e datas de receção				Universo			
				Pessoas empregadas com contrato individual de trabalho, em empresas com pelo menos uma pessoa ao serviço e abrangidas pelo código do trabalho (não estão incluídos os trabalhadores por conta própria, os trabalhadores do serviço doméstico, os trabalhadores da administração pública, dos institutos públicos e os que possuem regime de contrato de trabalho em funções públicas).			
Âmbito geográfico				Variáveis disponíveis			
Portugal				ANO, NISS, NOME, SEXO, DATA DE NASCIMENTO, NACIONALIDADE, REGIME DE REFORMA APLICADO, DATA DE ADMISSÃO, ANTIGUIDADE NA EMPRESA, CATEGORIA PROFISSIONAL, SITUAÇÃO NA PROFISSÃO, TIPO DE CONTRATO, REGIME DE DURAÇÃO DO TRABALHO, HABILITAÇÕES LITERÁRIAS, NÍVEL DE QUALIFICAÇÃO, PROFISSÃO, PERÍODO DE TRABALHO SEMANAL, NIF EMPRESA, MORADA DA EMPRESA (DISTRITO, MUNICÍPIO, FREGUESIA, LOCALIDADE E CÓDIGO POSTAL) NATUREZA JURÍDICA, ATIVIDADE ECONÓMICA DA EMPRESA, NÚMERO DE PESSOAS AO SERVIÇO EMPRESA, MORADA DO ESTABELECIMENTO (DISTRITO, MUNICÍPIO, FREGUESIA, LOCALIDADE E CÓDIGO POSTAL) ATIVIDADE ECONÓMICA DO ESTABELECIMENTO, NÚMERO DE PESSOAS AO SERVIÇO ESTABELECIMENTO.			
Unidade de observação							
Indivíduo							
Total de registos				Registos únicos			
2010	2011	2012	2013	2010	2011	2012	2013
2 843 391	2 798 081	2 617 252	2 610 344	2 780 487	2 736 659	2 562 816	2 554 641
Taxas de matching							
				SS NISS			
QP2010				94,8%			
QP2011				96,9%			
QP2012				96,1%			
QP2013				98,1%			

Indicadores de qualidade					
Nome da variável	Designação da variável	Taxa de preenchimento (%)			
		2010	2011	2012	2013
NISS	Número de identificação da Segurança Social	98,8%	98,9%	98,9%	99,0%
NOME	Nome	100,0%	100,0%	100,0%	100,0%
REG_REFORMA	Regime de reforma aplicado	100,0%	100,0%	100,0%	100,0%
PAIS_NAC_COD	País de Nacionalidade	100,0%	100,0%	100,0%	100,0%
SEXO	Sexo	100,0%	100,0%	100,0%	100,0%
DTNASC	Data de nascimento	99,8%	99,9%	99,9%	99,9%
DT_ADM	Data de Admissão	100,0%	100,0%	100,0%	100,0%
ANTIG	Antiguidade na empresa	100,0%	100,0%	100,0%	100,0%
CTPRO	Categoria Profissional	100,0%	100,0%	100,0%	100,0%
SITPRO	Situação na profissão	100,0%	100,0%	100,0%	100,0%
TIPO_CONTR	Tipo de contrato	93,8%	93,6%	93,5%	93,5%
REG_DUR	Regime de duração do trabalho	93,8%	93,6%	93,5%	93,5%
HABIL	Habilitações literárias	100,0%	100,0%	100,0%	100,0%
NQUAL1	Nível de qualificações	100,0%	100,0%	100,0%	100,0%
PROF	Profissão - Classificação portuguesa de profissões	100,0%	100,0%	100,0%	100,0%
PNT	Período normal de trabalho semanal	100,0%	93,6%	93,5%	93,5%
NIF	Número de identificação fiscal da empresa	100,0%	100,0%	100,0%	100,0%
MORADA_EMPRESA	Morada da empresa	100,0%	100,0%	100,0%	100,0%
LOCAL_EMPRESA	Localidade da empresa	100,0%	100,0%	100,0%	100,0%
CP4	Código postal	100,0%	100,0%	100,0%	100,0%
CP3	Código postal auxiliar	100,0%	100,0%	100,0%	100,0%
DESCRICAO_POSTAL	Localidade postal da empresa	100,0%	99,8%	100,0%	100,0%
NATJU	Natureza jurídica	100,0%	100,0%	100,0%	100,0%
DTEMP	Distrito da empresa	100,0%	100,0%	100,0%	100,0%
IDEMP	Município da empresa	100,0%	100,0%	100,0%	100,0%
FREGEMP	Freguesia da empresa	100,0%	100,0%	100,0%	100,0%
CAEMP	Atividade económica da empresa	100,0%	100,0%	100,0%	100,0%
PEMP	Número de pessoas ao serviço na empresa (31 outubro)	100,0%	100,0%	100,0%	100,0%
MORADA_ESTAB	Morada do estabelecimento	100,0%	100,0%	100,0%	100,0%
LOCAL_ESTAB	Localidade do estabelecimento	100,0%	100,0%	100,0%	100,0%
CP4_ESTAB	Código postal do estabelecimento	100,0%	100,0%	100,0%	100,0%
CP3_ESTAB	Código postal auxiliar do estabelecimento	100,0%	100,0%	100,0%	100,0%
DESC_POST_ESTAB	Localidade postal do estabelecimento	99,9%	99,8%	100,0%	100,0%
DTEST	Distrito do estabelecimento	100,0%	100,0%	100,0%	100,0%
IDEST	Município do estabelecimento	100,0%	100,0%	100,0%	100,0%
FREGEST	Freguesia do estabelecimento	100,0%	100,0%	100,0%	100,0%
CAEST	Atividade económica do estabelecimento	100,0%	100,0%	100,0%	100,0%
PEST	Número de pessoas ao serviço do estabelecimento (31 outubro)	100,0%	100,0%	100,0%	100,0%

3.6. IEF

O Instituto de Emprego e Formação Profissional (IEFP) é o serviço público de emprego, cuja missão é promover a criação e a qualidade do emprego e combater o desemprego, através da execução de políticas ativas de emprego.

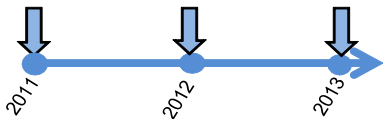
A informação disponibilizada pelo IEF afigura-se importante no sentido da caracterização da população desempregada, mas também se revela muito útil para aferir os indícios de residência em território nacional.

Dado que o IEF é responsável apenas pela informação do desemprego no continente, foi necessário contactar também os serviços de emprego das regiões autónomas da Madeira e dos Açores. O Instituto de Emprego da Madeira (IEM) já disponibilizou a informação solicitada. No entanto, ainda não foi possível receber os dados relativos à Região Autónoma dos Açores.

Foi necessário compatibilizar a informação proveniente do IEF e do IEM, de forma a integrarem uma única tabela de dados, uma vez que utilizavam estruturas de dados e codificações distintas.

Desta fonte estão carregados os anos 2011, 2012 e 2013. De referir, que as variáveis relativas ao Documento de identificação e Nome estão totalmente preenchidas.

Apresenta-se de seguida a ficha com a caracterização da fonte IEF.

Metainformação					
Anos disponíveis e datas de receção			Universo		
23.12.2014 (IEM) 05.01.2015 (IEFP) 			Indivíduos desempregados inscritos no centro de emprego		
Âmbito geográfico			Variáveis disponíveis		
Portugal continental e Região Autónoma da Madeira			ANO, ANO_MES, BI, NIF, NISS, NOME, LOCALIDADE, NÚMERO DE UTENTE, CODIGO POSTAL, TIPO DOCUMENTO IDENTIFICAÇÃO, DISTRITO, MUNICIPIO E FREGUESIA DE RESIDÊNCIA, SEXO, DATA DE NASCIMENTO E ESTADO CIVIL, PAIS NACIONALIDADE, HABILITACAO E AREA CURSO, CATEGORIA, DATA DESEMPREGO, DATA E MOTIVO DE INSCRICAO, PROFISSÃO E ATIVIDADE ECONÓMICA, TEMPO INSCRICAO E DE NÃO INSCRIÇÃO		
Unidade de observação					
Indivíduo					

Total de registos			Registos únicos		
2011	2012	2013	2011	2012	2013
702 215	853 817	897 258	702 215	853 817	897 258

Taxas de matching				
	SS NIF	SS NISS	BDIC NIC	
IEFP e IEM 2011	71,6%	70,0%	72,8%	
IEFP e IEM 2012	72,8%	72,1%	77,2%	
IEFP e IEM 2013	67,1%	67,0%	83,5%	

Indicadores de qualidade				
Nome da variável	Designação da variável	Taxa de preenchimento (%)		
		2011	2012	2013
ANO	Ano	100,0%	100,0%	100,0%
ANO_MES	Ano e mês	96,5%	96,6%	96,8%
BI	Número de identificação civil	100,0%	100,0%	100,0%
NIF	Número de identificação fiscal	97,9%	98,8%	99,3%
NISS	Número de identificação Segurança Social	97,2%	98,1%	98,6%
NOME	Nome	100,0%	100,0%	100,0%
LOCALIDADE	Localidade	88,7%	87,6%	86,9%
UTENTE_ID	Número de utente	3,5%	3,4%	3,2%
CP4	Código postal	100,0%	100,0%	100,0%
CP3	Código postal auxiliar	96,8%	97,6%	97,9%
DCPOSTAL	Designação do Código Postal	100,0%	100,0%	100,0%
TIPO_DOC_ID	Tipo de documento de identificação	100,0%	100,0%	100,0%
DT_RESID_COD	Distrito de Residência	100,0%	100,0%	100,0%
MN_RESID_COD	Município de Residência	100,0%	100,0%	100,0%
FR_RESID_COD	Freguesia de Residência	99,5%	99,5%	99,4%
SEXO	Sexo	100,0%	100,0%	100,0%
DTNASC	Data de nascimento	100,0%	100,0%	100,0%
PAIS_NAC_COD	Código do país de nacionalidade	100,0%	100,0%	100,0%
ESTCIVIL	Estado civil	100,0%	100,0%	100,0%
HABILITACAO	Habilitação	100,0%	100,0%	100,0%
AREA_CURSO	Curso (apenas para habilitações superiores)	11,2%	13,4%	14,9%
DATA_DESEMPREGO	Data de início de contagem da situação de desemprego	96,5%	96,6%	96,8%
CATEGORIA	Categoria pedido de emprego	100,0%	100,0%	100,0%
DATA_INSCRICAO	Data da última inscrição para emprego	100,0%	100,0%	100,0%
MOTIVO_INSCRICAO	Motivo de inscrição do pedido de emprego	100,0%	100,0%	100,0%
INT_TEMPO_INSCRICAO	Contagem do tempo de inscrição	100,0%	100,0%	100,0%
INT_TEMPO_NAO_INSC	Contagem do tempo em que não esteve inscrito (só para reinscrições)	0,1%	0,2%	0,1%
CAE_ANTERIOR	Código de Atividade Económica anterior	92,7%	91,4%	88,8%
CNP_ANTERIOR	Código Nacional de Profissões anterior (até 2013)	89,7%	91,8%	90,0%
PROFISSAO_ANT_CPP_APARTIR 2014	Código Profissão anterior (Classificação Portuguesa de Profissões: a partir de 2014)	3,5%	3,4%	3,2%

3.7. IISS

A informação fornecida pelo Instituto de Informática da Segurança Social (IISS) revela-se fundamental não só pelo número de indivíduos que abrange, mas também pelo facto de agregar numa única base de dados os 3 identificadores numéricos mais relevantes: NISS - Número de Identificação da Segurança Social, NIC - Número de Identificação Civil e NIF - Número de Identificação Fiscal. Esta última característica faz deste ficheiro uma plataforma crucial na articulação da informação proveniente das várias entidades.

A Segurança Social disponibilizou informação sobre todas as qualificações ativas que a pessoa possuía à data de referência (31 de Dezembro do ano de referência), o que originou que cada pessoa pudesse ter que mais do que uma qualificação - situação existente para quase 2 milhões de registos.

Não sendo um ficheiro para fins estatísticos, a informação da Segurança Social foi tratada no sentido de fornecer dados estatísticos, com possibilidade de serem utilizados para efeitos censitários. Assim, no sentido de assegurar a existência de apenas uma qualificação para cada indivíduo, foi implementado um conjunto de regras, cujo princípio geral foi o da prevalência das qualificações relacionadas com atividade económica sobre as situações de reforma, apoio social, etc, princípio equivalente ao utilizado para a informação de carácter censitário.

Posteriormente foi construída uma tabela de correspondência, que permitiu transformar/aproximar as qualificações da segurança social, em modalidades da variável censitária condição perante a atividade económica, criando-se 4 grandes categorias: “população empregada”, “população desempregada”, “população reformada” e “outra situação” (onde se incluiu por exemplo os indivíduos que recebem Rendimento Social de Inserção ou outro tipo de apoios não passíveis de classificação nas categorias anteriores).

Para a fonte IISS foram carregados os anos de 2011 e 2013, sendo que, para ambos os anos, o NISS, é a única variável de identificação totalmente preenchida. Importa salientar que para o ano de 2011 a variável NOME não se encontra disponível.

A ficha seguinte sintetiza os principais indicadores relativos à fonte Segurança Social.

Metainformação			
Anos disponíveis e datas de receção		Universo	
		Indivíduos que têm uma qualificação ativa, no sentido de um enquadramento no sistema de segurança social. Abrange todas as pessoas singulares que mantêm uma inscrição ativa na segurança social, independentemente do tipo de qualificação: trabalhador por conta de outrem, trabalhador independente, pensionista de velhice, pensionista de invalidez, rendimento social de inserção,....	
Âmbito geográfico		Variáveis disponíveis	
Portugal		ANO, SEXO, DATA DE NASCIMENTO, ESTADO CIVIL, NÚMERO E TIPO DOCUMENTO IDENTIFICAÇÃO, NIF, NISS, PAIS, MUNICÍPIO E FREGUESIA DE NATURALIDADE, DISTRITO, MUNICÍPIO, FREGUESIA, ARTÉRIA E CÓDIGO POSTAL DE RESIDÊNCIA, NACIONALIDADE, DATA DE INÍCIO, DE FIM E TIPO DE QUALIFICAÇÃO, MOTIVO DE FIM, REGIME, NOME E NISS DA ENTIDADE EMPREGADORA, MORADA, DISTRITO, CONCELHO, FREGUESIA E MORADA DA ENTIDADE EMPREGADORA	
Unidade de observação			
Indivíduo			

Total de registos		Registos únicos	
2011	2013	2011	2013
9 062 616	9 058 682	7 209 027	7 124 233

Taxas de matching	
	BDIC NIC
SS2011	95,7%
SS2013	98,4%

Indicadores de qualidade			
Nome da variável	Designação da variável	Taxa de preenchimento (%)	
		2011	2013
ANO	Ano a que os dados dizem respeito	100,0%	100,0%
NISS	Número de identificação da Segurança Social	100,0%	100,0%
TIPO_DOC_CIVIL	Designação do documento de identificação	69,9%	79,4%
BI	Número de identificação Civil	73,9%	79,3%
NIF	Número de identificação Fiscal	96,4%	97,6%
SEXO	Sexo	100,0%	100,0%
DTNASC	Data de nascimento	99,9%	99,9%
NOME	Nome	-	100,0%
PAIS_NAC_COD	País de Nacionalidade	86,6%	90,5%
PAIS_NAT_COD	País de Naturalidade	100,0%	100,0%
MN_NAT_COD	Município de Naturalidade	87,7%	87,0%
FR_NAT_COD	Freguesia de Naturalidade	87,7%	87,0%
DT_RESID_COD	Distrito de residência	97,8%	97,4%
MN_RESID_COD	Município de residência	97,8%	97,4%
FR_RESID_COD	Freguesia de residência	97,8%	97,4%
ARTERIA_RES	Rua	98,6%	98,1%
LOCALIDADE_RES	Localidade	98,6%	98,1%
CP4	Código postal	86,4%	90,0%
CP3	Código postal auxiliar	87,1%	90,8%
ESTCIVIL	Estado Civil	100,0%	100,0%
TIPO_QLF	Tipo de qualificação	100,0%	100,0%
DATA_INICIO_QLF	Data de início da qualificação	100,0%	100,0%
DATA_FIM_QLF	Data de fim da qualificação	2,9%	3,4%
MOTIVO_FIM	Motivo do fim da qualificação	2,9%	3,4%
REGIME	Regime de enquadramento na Segurança Social	64,7%	63,6%
NISS_EE*	NISS da Entidade empregadora	100,0%	100,0%
NIPC*	Número de identificação de pessoa coletiva	-	95,0%
FIRMA_DENOMINACAO*	Nome da entidade empregadora	94,5%	95,1%
ARTERIA_EE*	Morada da entidade empregadora	94,5%	95,0%
CP4_EE*	Código postal da entidade empregadora	94,5%	89,7%
CP3_EE*	Código postal auxiliar da entidade empregadora	94,5%	89,7%
LOCALIDADE_EE*	Localidade da entidade empregadora	94,5%	95,0%
COD_DISTrito_EE*	Distrito da entidade empregadora	94,5%	95,0%
COD_CONCELHO_EE*	Município da entidade empregadora	94,5%	95,0%
COD_FREGUESIA_EE_ORIG*	Freguesia de origem da entidade empregadora	94,5%	95,0%
COD_FREGUESIA_EE*	Freguesia da entidade empregadora	94,5%	95,0%

*Considerando qualificações DOMINANTES: MOE (MEMBRO DE ORGÃO ESTATUTÁRIO) e TCO (TRAB.POR CONTA DE OUTREM)

3.8. IRN/BDIC

A Base de Identificação Civil (BDIC) é a fonte administrativa de referência para a contabilização da população portuguesa e respetiva caracterização das variáveis demográficas básicas.

A BDIC é da responsabilidade do Instituto dos Registos e do Notariado (IRN) e é alimentada pelo SIRIC - Sistema de Informação do Registo Civil - que atribuía até 2007 o Bilhete de Identidade (de acordo com a Lei nº 32/99) e a partir de 2007, de acordo com a Lei nº7/2007, o Cartão do Cidadão.

Esta fonte é constituída por todos os cidadãos de nacionalidade portuguesa, residentes em Portugal ou no estrangeiro. Desde 2007, as crianças à nascença passaram a ser registadas na BDIC. Face ao objetivo estatístico de constituição de uma base de população residente, a BDIC apresenta dois tipos de limitações:

- Âmbito: O universo da BDIC abrange apenas a população de nacionalidade portuguesa (os cidadãos brasileiros são em número residual), não inclui os cidadãos de nacionalidade estrangeira residentes em Portugal
- Conceito diferente entre população registada (BDIC) e população residente (Censos)

Não obstante estas limitações, a BDIC constitui um dos pilares para a constituição de uma base de dados de população residente, que implica obrigatoriamente a ligação entre registos provenientes das diversas fontes.

Para esta fonte foram carregados os anos de 2010 e 2014. As variáveis Nome e Número do documento de identificação estão totalmente preenchidas.

Apresenta-se de seguida a ficha de caracterização da BDIC.

Metainformação			
Anos disponíveis e datas de receção		Universo	
		2010: Cidadãos de nacionalidade portuguesa residentes em Portugal (11 565 714) ou no estrangeiro (1 323 280). 2014: Indivíduos de nacionalidade Portuguesa e brasileiros com estatuto de porto seguro com morada de residência em Portugal ou morada desconhecida (inclui 11 884 913 indivíduos vivos e 2 489 817 falecidos).	
Âmbito geográfico		Variáveis disponíveis	
Portugal		ANO, NIC, NOME, SEXO, DATA DE NASCIMENTO, CÓDIGO POSTAL, LOCALIDADE, DISTRITO, MUNICÍPIO e FREGUESIA DE RESIDÊNCIA, PAÍS DE NACIONALIDADE, PAÍS, MUNICÍPIO E FREGUESIA DE NATURALIDADE, ESTADO CIVIL, DATA DE EMISSÃO, DATA DE VALIDADE, ESTADO DO REGISTO	
Unidade de observação			
Indivíduo			
Total de registos recebidos		Registos de residentes em Portugal	
2010	2014	2010	2014
12 888 994	14 374 730	11 565 714	11 884 913

Indicadores de qualidade			
Nome da variável	Designação da variável	Taxa de preenchimento (%)	
		2010	2014
ANO	Ano a que os dados dizem respeito	100,0%	100,0%
DT_RESID_COD	Código de distrito	100,0%	98,8%
MN_RESID_COD	Código de município	100,0%	98,8%
FR_RESID_COD	Código de freguesia	100,0%	98,8%
NOME	Nome	100,0%	100,0%
BI	Número de identificação Civil	100,0%	100,0%
SEXO	Sexo	100,0%	100,0%
DTNASC	Data de nascimento	100,0%	100,0%
PAIS_NAC_COD	País de Nacionalidade	100,0%	100,0%
PAIS_NAT_COD	País de Naturalidade	100,0%	100,0%
MN_NAT_COD	Município de Naturalidade	100,0%	100,0%
FR_NAT_COD	Freguesia de Naturalidade	100,0%	100,0%
ESTCIVIL	Estado Civil	100,0%	100,0%
CP_LOCAL_POSTAL	Localidade postal	0,0%	72,3%
CP4	Código postal	0,0%	72,3%
CP3	Código postal auxiliar	0,0%	72,3%
DATA_EMISSAO	Data de emissão	100,0%	100,0%
DATA_VALIDADE	Data de validade	0,0%	87,5%
ESTADO	Vivo / óbito	0,0%	100,0%

3.9. SEF

O Serviço de Estrangeiros e Fronteiras (SEF) é o organismo responsável pelo registo da informação relativa à entrada, permanência e afastamento de cidadãos estrangeiros em território Nacional.

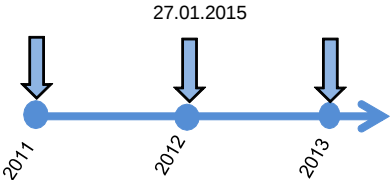
A informação proveniente do SEF é essencial para completar a população residente em Portugal, uma vez que a informação constante da BDIC abrange apenas a população com nacionalidade portuguesa.

O ficheiro do SEF contém informação sobre os indivíduos de nacionalidade estrangeira com título de residência válido. Pode permitir o recenseamento, em termos de cobertura e de algumas variáveis de conteúdo, da população estrangeira residente legalmente em Portugal.

O SEF disponibilizou ainda alguma informação sobre a constituição dos agregados familiares da população com títulos de residência, que ainda não foi possível analisar.

Da fonte SEF foram carregados três anos, 2011, 2012 e 2013. As variáveis relativas ao nome e ao número de registo no SEF são as únicas que se encontram totalmente preenchidas.

Apresentam-se, em seguida, os principais indicadores relativos a esta fonte.

Metainformação					
Anos disponíveis e datas de receção			Universo		
			Indivíduos de nacionalidade estrangeira com título de residência válido		
Âmbito geográfico			Variáveis disponíveis		
Portugal			ANO, RSN, NIF, NISS, NOME, SEXO, DATA E LOCAL DE NASCIMENTO, ESTADO CIVIL, PAÍS NACIONALIDADE E NATURALIDADE, PAÍS ÚLTIMA RESIDÊNCIA, MORADA, CÓDIGO POSTAL, ZONA POSTAL, LOCALIDADE, DISTRITO E MUNICÍPIO DE RESIDÊNCIA, Nº, TIPO E DATA DE VALIDADE DA AUTORIZAÇÃO DE RESIDÊNCIA, SITUAÇÃO, TIPO E GRUPO PROFISSIONAL, PROFISSÃO, HABILITAÇÕES		
Unidade de observação			Indivíduo		
Total de registos			Registos únicos		
2011	2012	2013	2011	2012	2013
434 708	414 610	398 268	434 708	414 610	398 268
Taxas de matching					
		SS NISS		SS NIF	
SEF2011		69,3%		59,8%	
SEF2012		69,4%		58,5%	
SEF2013		67,1%		56,7%	

Indicadores de qualidade				
Nome da variável	Designação da variável	Taxa de preenchimento (%)		
		2011	2012	2013
ANO	Ano a que os dados dizem respeito	100,0%	100,0%	100,0%
RSN	Número interno de processo	100,0%	100,0%	100,0%
NIF	Número de Identificação Fiscal	57,8%	59,9%	61,4%
NISS	Número de Identificação da Segurança Social	49,9%	50,6%	51,1%
NUMERO_AR	Número de autorização de residência	99,9%	99,9%	100,0%
TIPO_AR	Tipo de autorização de residência	100,0%	100,0%	100,0%
DATA_VALIDADE	Data de validade da autorização de residência	100,0%	100,0%	100,0%
NOME	Nome	100,0%	100,0%	100,0%
SEXO	Sexo	100,0%	100,0%	100,0%
ESTCIVIL	Estado civil	100,0%	100,0%	100,0%
DTNASC	Data de nascimento	100,0%	100,0%	100,0%
LOCAL_NASCIMENTO	Local de nascimento	92,2%	91,8%	91,2%
PAIS_NAC_COD	País de Nacionalidade	100,0%	100,0%	100,0%
PAIS_NAT_COD	País de Naturalidade	100,0%	100,0%	100,0%
DT_RESID_COD	Distrito de residência	100,0%	100,0%	100,0%
MN_RESID_COD	Município de residência	100,0%	100,0%	100,0%
CP4	Código postal	99,7%	99,8%	99,9%
CP3	Código postal auxiliar	99,7%	99,8%	99,9%
LOCALIDADE	Localidade	99,7%	99,8%	99,9%
MORADA	Morada	99,7%	99,8%	99,9%
PAIS_ULTIMA_RESID	País de residência anterior	27,6%	42,6%	53,1%
PROFISSAO_CNP	Profissão	13,3%	14,1%	14,2%
TIPO_OCUPACAO	Tipo de ocupação	36,8%	51,7%	61,9%
SITUACAO_PROFISSIONAL	Situação profissional	36,8%	51,7%	61,9%
GRUPO_PROFISSIONAL	Grupo profissional	28,7%	35,4%	40,2%
HABILITACOES	Localidade	53,8%	59,1%	63,6%

3.10. Síntese da informação disponível e principais dificuldades

O quadro 1 seguinte sintetiza a informação atualmente disponível para as diferentes entidades.

Quadro 1 – Síntese das bases de dados disponíveis e respetivos anos

Base de dados	2010	2011	2012	2013	2014
ADSE - Direção-Geral de Proteção Social aos Trabalhadores em Funções Públicas			X	X	
CGA - Caixa Geral de Aposentações	X			X	
Educação - Direção-Geral de Estatísticas da Educação e Ciência e Direção Educação da Madeira		X		X	
Quadros de pessoal - Gabinete de Estratégia e Planeamento	X	X	X	X	
IEFP- Instituto do Emprego e Formação Profissional e Instituto de Emprego da Madeira		X	X	X	
Segurança Social - Instituto de Informática		X		X	
BDIC - Instituto dos Registos e do Notariado	X				X
SEF - Serviço de Estrangeiros e Fronteiras		X	X	X	

A integração dos diferentes ficheiros administrativos é uma condição essencial no âmbito do estudo de viabilidade para o novo modelo censitário, nomeadamente para a constituição de uma base de população.

A constituição de uma base de dados de população residente implica obrigatoriamente a ligação, ou *record linkage*, entre os registos provenientes dos diferentes ficheiros.

As ligações através de identificadores numéricos são mais fáceis de concretizar e, de uma forma geral, conferem maior grau de certeza e qualidade ao processo de interligação entre ficheiros. O quadro 2 resume a situação dos ficheiros quanto à taxa de preenchimento dos principais identificadores numéricos.

Quadro 2 - Existência de identificador numérico e respetivo grau de preenchimento

Identificadores numéricos	Base de dados disponíveis								Não disponíveis	
	BDIC (2014)	SEF (2013)	SS (2013)	CGA (2013)	QP (2013)	Educação (2013)	IEFP (2013)	ADSE (2013)	IRS	IMI
Nº de registos	11,8 M	398 mil	7,1 M	1,0 M	2,6 M	1,7 M	897 mil	1,3 M	x	x
NIC_N.º Identificação Civil	100%	-	76%	77%	-	91%	95%	66%	-	-
NIF_Nº Identificação Fiscal	-	61%	98%	82%	-	-	99%	65%	x%	-
NISS_N.º Identificação Segurança Social	-	51%	100%	-	99%	62%	98%	-	-	-
NIF_Nº Identificação Fiscal do proprietário	-	-	-	-	-	-	-	-	-	x%
NIF_Nº Identificação Fiscal da empresa	-	-	100%	-	100%	-	-	-	-	-

O identificador numérico mais representado é o NIC, que consta em 6 dos 8 ficheiros analisados, variando a sua taxa de preenchimento entre 100% na BDIC e os 66% na ADSE.

O NISS consta também em 5 das bases de dados disponíveis no INE, com taxas de preenchimento entre os 100% na Segurança Social e os 51% no ficheiro do SEF.

O NIF consta em 5 dos 8 ficheiros analisados, destacando-se o IEFP com 99% dos registos preenchidos, assim como os 98% da Segurança Social.

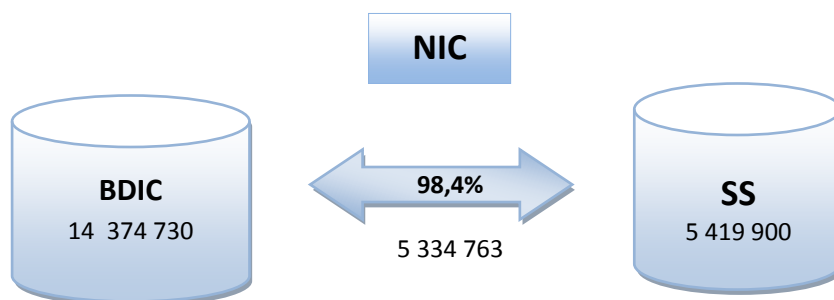
Atendendo às taxas de preenchimento, a ligação entre os vários ficheiros só poderá ser concretizada parcialmente através dos identificadores numéricos, pelo que tem vindo a ser testada a utilização de outros atributos nas chaves de *matching*. De entre esses atributos destacam-se: Nome (mais concretamente as 3 primeiras letras do primeiro Nome e as três últimas letras do último nome), sexo, data de nascimento, freguesia/município de residência, estado civil, naturalidade, etc.

De seguida apresentam-se os principais resultados do *matching* através de identificadores numéricos entre as várias bases de dados disponíveis.

Matching entre as diversas fontes através de NIC

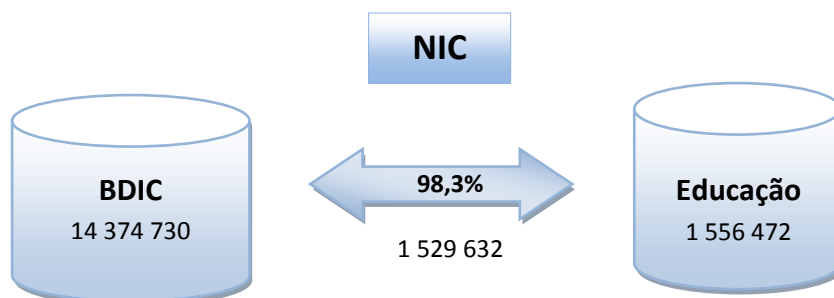
SS2013 e BDIC 2014

Dos 5 419 900 NIC's únicos existentes no Ficheiro da Segurança Social 5 334 763 fazem *matching* com a BDIC 2014, o que corresponde a uma taxa de **98,4%**.



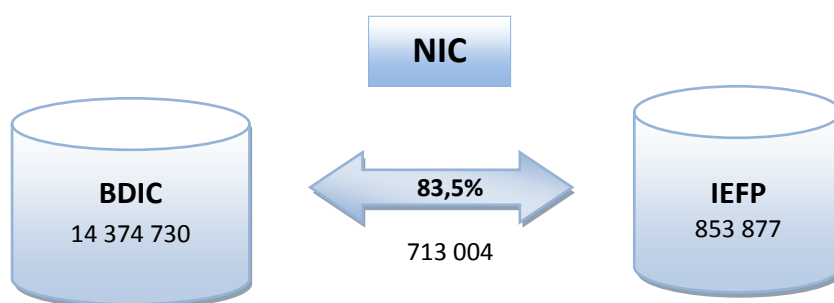
Educação 2013 e BDIC 2014

Cruzando os NIC's presentes no ficheiro da educação verifica-se que **98,3%** têm correspondência na BDIC 2014.



IEFP 2013 e BDIC 2014

Dos 853 877 NIC's únicos existentes na base de dados do IEPF têm correspondência na BDIC 713 004, o que corresponde a uma taxa de *matching* de **83,5%**.



CGA 2013 e BDIC 2014

No *matching* entre os ficheiros da CGA e da BDIC foram detetados alguns problemas no código NIC encriptado, pelo que os valores de *matching* para esta fonte não estão ainda disponíveis.

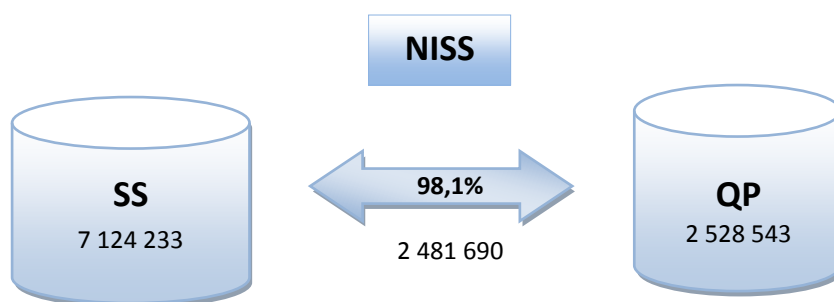
ADSE 2013 e BDIC 2014

No *matching* entre a base de dados da ADSE e da BDIC foram detetados alguns problemas no código NIC encriptado, pelo que os valores de *matching* para esta fonte não estão ainda disponíveis.

Matching entre as diversas fontes através de NISS

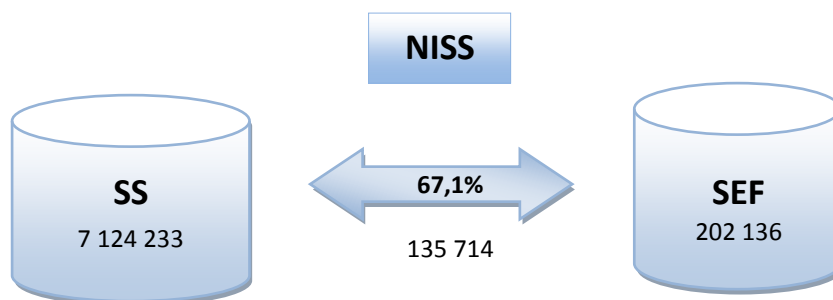
Quadros de pessoal 2013 e SS2013

Dos 2 528 543 NISS únicos existentes na base de dados dos Quadros de Pessoal fazem *matching* com a Segurança Social 2 481 690 com a o que corresponde a uma taxa de 98,1%.



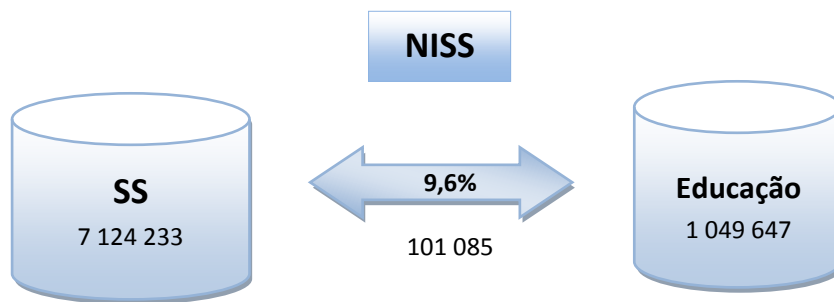
SEF 2013 e SS2013

Dos 202 136 NISS's únicos existentes no Ficheiro SEF fazem *matching* com a Segurança Social 135 714 com a o que corresponde a uma taxa de 67,1%.



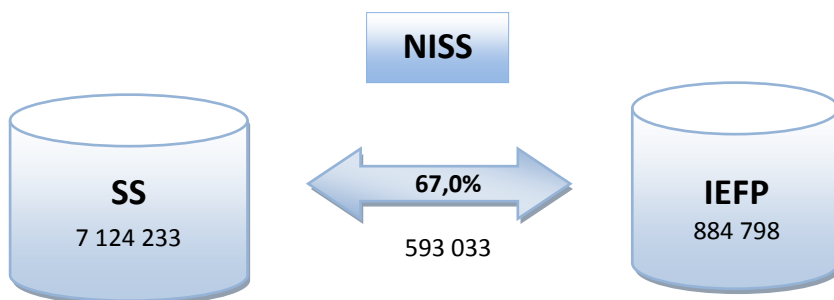
Educação 2013 e SS2013

Cruzando os NISS presentes no ficheiro da educação verifica-se que apenas **9,6%** têm correspondência na base de dados da SS 2013. O valor da taxa de *matching* está relacionado com o facto destes dois ficheiros terem uma interseção de universos muito baixa.



IEFP e SS2013

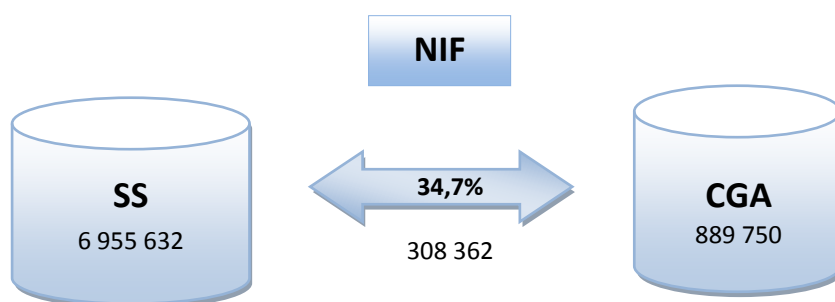
Dos 884 798 NISS's únicos existentes na fonte IEPF fazem *matching* com a Segurança Social 593 033, o que corresponde a uma taxa de 67%.



Matching entre as diversas fontes através de NIF

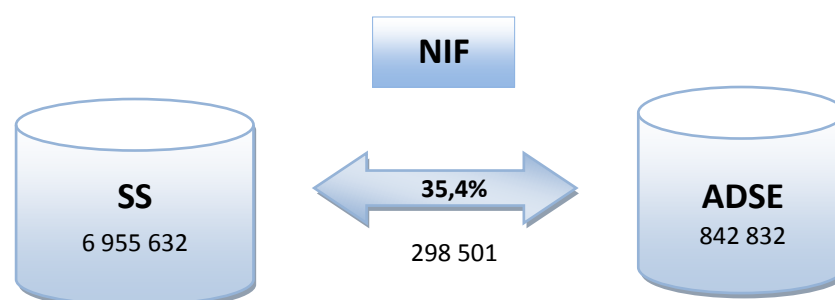
GGA 2013 e SS2013

Dos 889 750 NIF's únicos existentes na base de dados da CGA fazem matching com a Segurança Social 308 362 o que corresponde a uma taxa de 34,7%. Este valor justifica-se na medida em que os universos destas duas fontes são complementares, uma vez que grande parte dos beneficiários da CGA não integra o regime da Segurança Social.



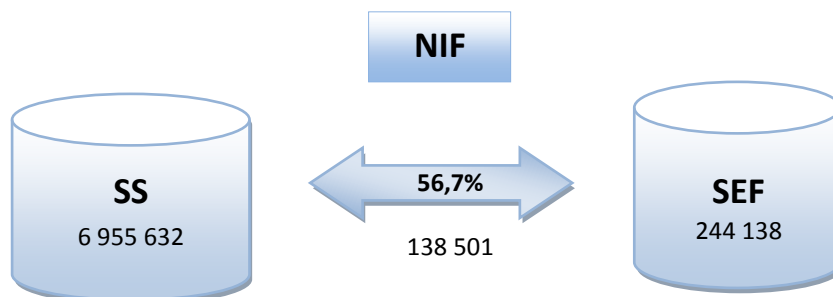
ADSE 2013 e SS2013

Dos 842 832 NIF's existentes na ADSE fazem *matching* com a Segurança Social apenas 298 501, o que se traduz numa taxa de 35,4%. Este valor decorre da própria complementaridade do universo das duas bases de dados, dado que muitos dos beneficiários da ADSE não integram o regime da Segurança Social mas o da CGA.



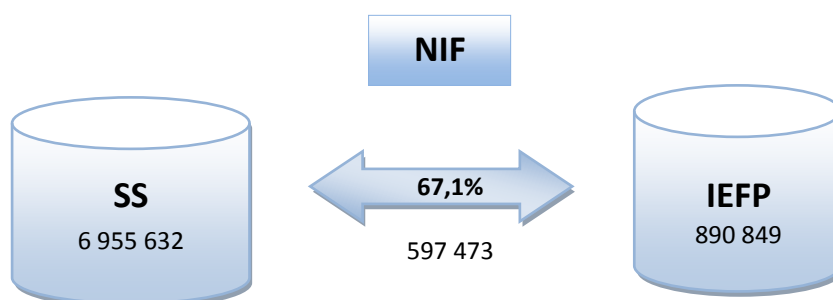
SEF 2013 e SS2013

Dos 244 138 NIF's únicos existentes no ficheiro do SEF têm correspondência na Segurança Social 138 501, o que corresponde a uma taxa de *matching* 56,7%.



IEFP 2013 e SS2013

Dos 890 849 NIF's existentes no Ficheiro IEFP fazem *matching* com a base de dados da Segurança Social 597 473 com a o que corresponde a uma taxa de 67,1%.



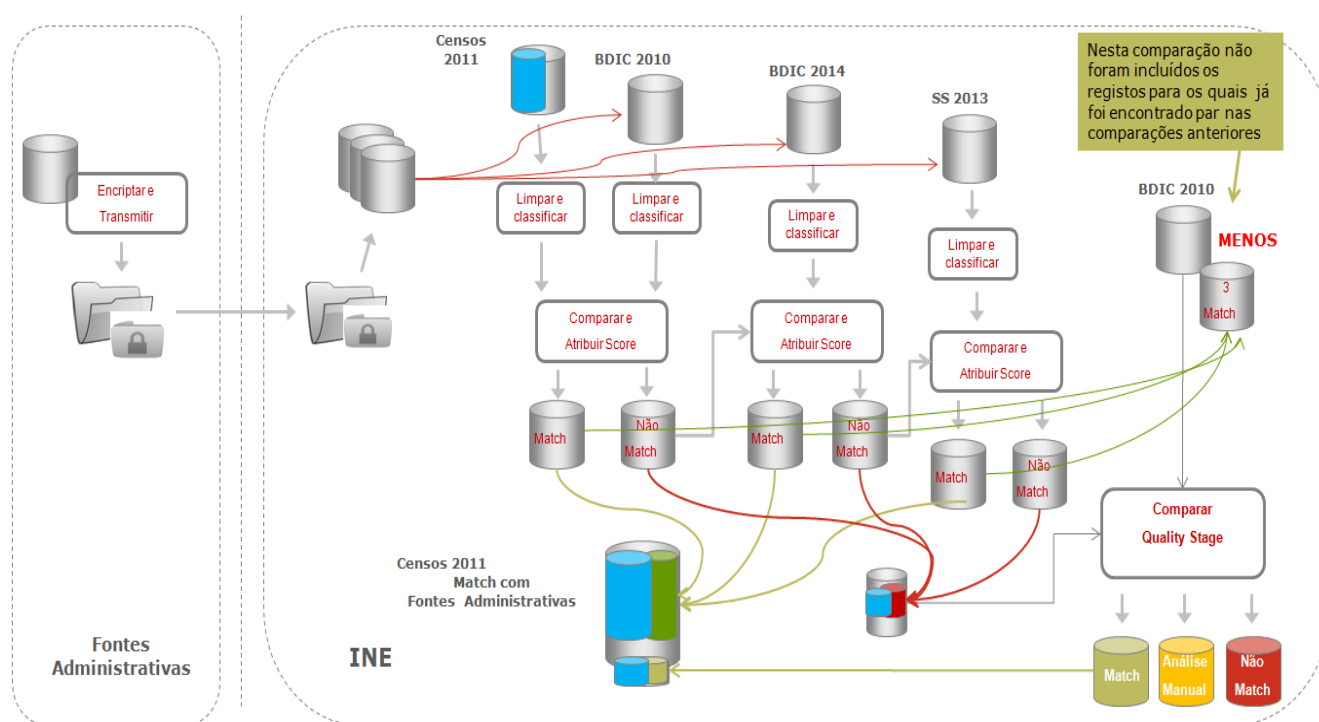
Finalmente referem-se algumas das dificuldades encontradas na organização das bases de dados das diversas fontes administrativas. Estas relacionam-se com:

- A descodificação dos códigos geográficos utilizados pelas diversas entidades. A utilização de códigos de distritos, municípios e freguesias, que não os que se encontram atualmente em vigor gera problemas no tratamento destas variáveis;
- A não existência de regras na recolha e preenchimento do identificador NIC. A forma como é recolhido e registado pelas várias entidades é diversa: há entidades que recolhem o NIC com o *check digit* agregado, outras recolhem o NIC sempre com 8 caracteres incluindo os zeros à esquerda, noutras não há qualquer tipo de validação do identificador recolhido. Estas situações são problemáticas uma vez que vão originar códigos encriptados distintos, deixando de ser possível identificar um mesmo indivíduo nos diferentes ficheiros. Deste modo afigura-se essencial acautelar a harmonização na recolha deste identificador.
- A atribuição, na fonte, de número de identificação fictício (tipo 99999999 ou 00000000) quando as entidades não recolhem o NIC. Esta situação impossibilita a distinção e a posterior ligação destes registos;
- Diversidade de estruturas, de formatos e de codificações utilizados em variáveis que são comuns a diferentes fontes;
- Identificadores numéricos com fracas taxas de preenchimento em algumas fontes;
- A não aplicação, na fonte, de regras de validação em alguns dos campos recolhidos pelas entidades.

4. Ligação entre CENSOS2011 e ficheiros administrativos

A metodologia de construção de uma base de população residente assenta em processos de ligação, ou *record linkage*, de registos correspondentes ao mesmo indivíduo que existem em simultâneo em diversas bases de dados, neste caso disponibilizadas por várias fontes administrativas. Como hipótese inicial, se um registo com certos fatores ou atributos (entre outros: nome, sexo, data de nascimento, freguesia de residência, ...) existe em várias bases de dados, então esse registo é único e representa um mesmo indivíduo. Nesta fase os trabalhos dizem respeito apenas à identificação de atributos comuns através de chaves de ligação. Depois de identificados os registos comuns, serão colocados na base de dados resultante os identificadores numéricos que lhe forem associados. A figura 2 ilustra este processo.

Figura 2 – Interligação de ficheiros provenientes de fontes administrativas



Existem várias possibilidades de ligação com metodologias e técnicas associadas, desde logo técnicas determinísticas e estocásticas. As técnicas determinísticas envolvem comparações exatas através de chaves de ligação, neste caso, entre pares de registos de bases de dados diferentes. As técnicas estocásticas ou probabilísticas utilizam indícios ou possíveis semelhanças (associadas a um grau de incerteza) entre pares de registos. Em ambos os casos, o que se pretende é identificar pares de registos que correspondem a um mesmo indivíduo, ou que fazem *matching*.

Esgotados todos os processos de *matching* automático, será necessário definir um plano de *matching* manual com critérios de procura exaustivos, em que através de observação direta se identifiquem registos comuns nas fontes consideradas.

Neste relatório, apresentam-se e analisam-se os resultados obtidos relativos aos dois primeiros tipos de *matching*.

Foram inicialmente realizados trabalhos de ligação determinística entre as várias bases de dados, tentando obter por este método o maior número de resultados possível. Para o efeito foram utilizados diversos campos ou atributos conjugados, até um máximo de 8 num total de 12 disponíveis. Para os registos em que não foi possível encontrar par foi já aplicada uma comparação probabilística ou *fuzzy* utilizando diferentes elementos de comparação, através do *software Quality Stage*.

4.1. Trabalhos preparatórios nas várias bases de dados

Trabalhos preparatórios nos dados dos Censos 2011

Foi criada uma tabela de informação dos Censos 2011 específica para o processo de interligação com as outras bases de dados administrativas. Essa tabela contém dados sobre as características dos indivíduos e informação alfabética de nome e morada dos indivíduos.

Foram analisados os conteúdos das diversas colunas envolvidas sob o ponto de vista da interligação de ficheiros. Decorrente desta análise foram efetuadas transformações de conteúdo. Nos campos alfabéticos foram removidos todos os caracteres que não constam de uma lista padrão. Nomes iniciados por “M<espaço>” passaram a “MARIA”.

As moradas, já separadas em partículas, foram objeto de normalização de abreviaturas para o tipo de via e designação do prefixo de edifício, tendo em atenção as abreviaturas definidas nos dicionários do processo já utilizados no INE para a partição de moradas.

Nos dados originais dos Censos 2011 foram criadas mais duas colunas adicionais codificadas segundo a nova geografia da CAOP2013 que foram utilizadas para comparar os dados dos Censos 2011 com os registos da BDIC2014, e os de todas as outras fontes que tenham dados por freguesia posteriores a 2013.

Trabalhos preparatórios nos dados das fontes administrativas BDIC2010, BDIC 2014 e SS 2013

Foram construídas tabelas específicas para apoio da comparação dos dados das fontes administrativas. Aqui foram realizados os seguintes processos de limpeza e normalização:

Na coluna NOME foram retiradas todas as acentuações e caracteres não pertencentes à lista padrão (p. ex. *.V=) e retiradas as "" (pelicas).

Foi feita a deteção de caracteres que podem ser substituídos por letras, de uma forma consistente (na coluna NOME "-AR" foi transformado em "MAR" e "-NT" em "ANT").

Foi realizado um teste com um primeiro processo de ligação entre ficheiros, que permitiu uma análise inicial e forneceu informação preparatória para o desenho do processo de ligação a implementar. Após verificação manual de alguns dos registos não comuns, foram encontradas situações de:

- Não *match*
- *Match* com números de BI duplicados
- *Match* com identificador de indivíduo dos censos duplicados

Estes foram alguns dos problemas identificados e analisados que seria necessário analisar também no resultado do processo a automatizar.

4.2. Identificação de problemas encontrados nos dados

Da análise efetuada aos diversos ficheiros das fontes administrativas, foi necessário realizar vários processos de limpeza e tratamento da informação. A maioria foi comum às diversas fontes, indicam-se os mais significativos para cada delas.

C2011

Foram encontrados registos que estão assinalados como estrangeiros e têm preenchido o número de bilhete de identidade.

BDIC2010

Alguns registos na base de dados da BDIC não têm morada, não sendo possível fazer o *match* através destes elementos.

A base de dados da BDIC não tem código postal.

Vários códigos de freguesia da BDIC tiveram que ser convertidos para os códigos usados nos Censos 2011.

Existem códigos de freguesia, de município de residência e de naturalidade que não foi possível converter. Existem datas de nascimento na BDIC não válidas que não foram consideradas.

BDIC2014

Foram encontradas lacunas no preenchimento de variáveis, por exemplo nos códigos postais, na versão dos códigos de município ou freguesia utilizados e nas datas não validadas.

Verificou-se a existência de números de identificação repetidos em registos associados a indivíduos diferentes. Estes registos não foram considerados.

4.3. Comparação CENSOS2011 e BDIC2010

Para a comparação das bases de dados dos Censos e da BDIC seguiu-se uma estratégia de comparação, através de verificação de igualdade de chaves, entre os campos disponíveis em cada uma das bases. Sobre os registos que não foi possível encontrar uma correspondência através dessa comparação foram desenvolvidos procedimentos no *software Quality Stage* (QS) para a comparação das moradas. Numa primeira abordagem, optou-se por fazer doze níveis de comparações, iniciando na mais restritiva e terminando na menos restritiva.

Em cada etapa, encontram-se os registos com match sendo-lhes atribuído um *score* decrescente. O valor de *score* foi atribuído com base na chave de ligação utilizada. Foram usadas nas primeiras fases o conjunto máximo de elementos para a chave de ligação, onde se convencionou atribuir o máximo de pontuação à ligação encontrada. Às ligações conseguidas através das combinações de chaves utilizadas nas fases seguintes foram atribuídas pontuações decrescentes.

A atribuição de um *score* à ligação de cada fase obedeceu a uma lógica de confiança nas chaves utilizadas nessa fase. Como a opção tomada foi utilizar campos de ligação mais restritivos em fases mais iniciais e depois alargar essa restrição em fases posteriores, o *score* atribuído foi diminuindo em cada uma das fases.

Não foram realizados testes que confirmem estes pressupostos. O valor atribuído é meramente classificativo de ponto de vista da ordenação, não significando proporcionalidade entre os valores atribuídos.

De referir que em cada fonte, foram utilizadas chaves de ligação diferentes em cada fase (por motivos de diferenças estruturais de cada fonte). Por conseguinte, não é possível comparar o número de ligações para cada fase entre diferentes fontes.

Os registos sem correspondência, em cada etapa, são analisados nas etapas seguintes.

1º Etapa: Match com o maior número de campos possível:

Nesta etapa efetuou-se o match com os seguintes campos:

- FREGUESIA RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ÚLTIMOS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO
- NATURALIDADE
- ESTADO CIVIL

2º Etapa: Match com os seguintes campos (retirou-se “primeiros 3 carateres do último nome”)

- FREGUESIA RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- ÚLTIMOS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO
- NATURALIDADE
- ESTADO CIVIL

3º Etapa: Match com os seguintes campos (retirou-se o dia da data de nascimento e acrescentou-se o “primeiros 3 carateres do último nome”):

- FREGUESIA RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ÚLTIMOS 3 CARACTERES DO ÚLTIMO NOME
- ANO E MÊS DA DATA NASCIMENTO
- NATURALIDADE
- ESTADO CIVIL

4º Etapa: Match com os seguintes campos (retirou-se “primeiros 3 carateres do último nome” e o estado civil e acrescentou-se o dia da data de nascimento):

- FREGUESIA RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- ÚLTIMOS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO
- NATURALIDADE

5º Etapa: Match com os seguintes campos (retirou-se a naturalidade):

- FREGUESIA RESIDENCIA
- SEXO

- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- ÚLTIMOS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO

6º Etapa: Match com os seguintes campos (retirou-se os “últimos 3 carateres do último nome”):

- FREGUESIA RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- DATA NASCIMENTO

7º Etapa: Match com os seguintes campos (acrescentou-se os “primeiros 3 carateres do último nome” e “últimos 3 carateres do último nome” e subiu-se o nível para município):

- MUNICIPIO RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ÚLTIMOS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO

8º Etapa: Match com os seguintes campos (acrescentou-se “últimos 3 carateres do primeiro nome” e naturalidade e subiu-se o nível para Portugal):

- PORTUGAL
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- ÚLTIMOS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ÚLTIMOS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO
- NATURALIDADE

9º Etapa: Match com os seguintes campos (retirou-se os “primeiros 3 carateres do último nome” e naturalidade):

- PORTUGAL
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- ÚLTIMOS 3 CARACTERES DO PRIMEIRO NOME
- ÚLTIMOS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO

10º Etapa: Match com os seguintes campos (acrescentou-se os “primeiros 3 caracteres do último nome”, a freguesia, e data de nascimento dos censos variou entre +/-10 anos)

- FREGUESIA RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- ÚLTIMOS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ÚLTIMOS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO (+/- 10 anos)

11º Etapa: Match com os seguintes campos (acrescentou-se a naturalidade, subiu-se o nível para município e data de nascimento dos censos variou entre +/-2 anos)

- MUNICIPIO RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- ÚLTIMOS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ÚLTIMOS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO (+/- 2 anos)
- NATURALIDADE

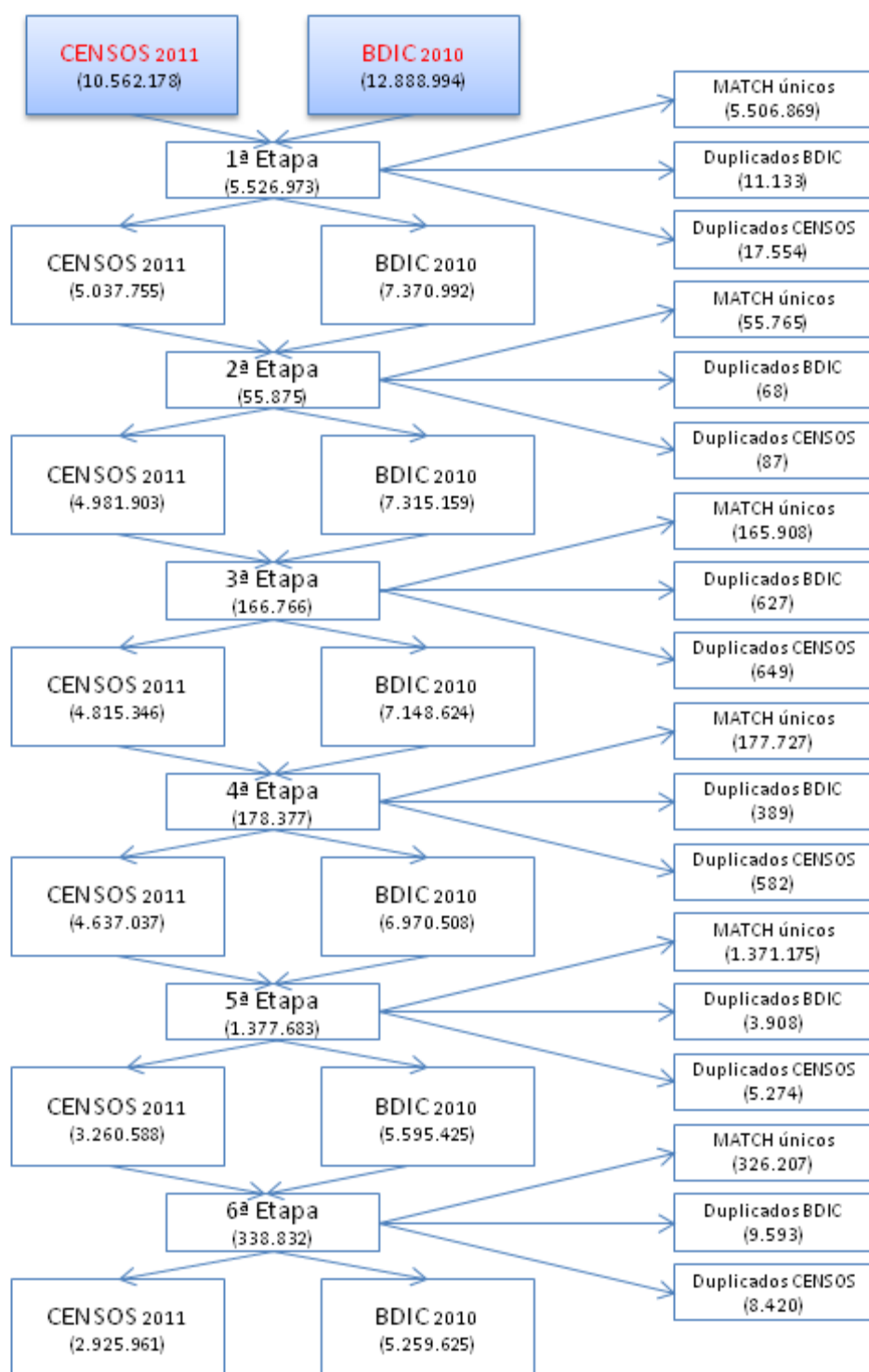
12º Etapa: Match com os seguintes campos (retirou-se o município)

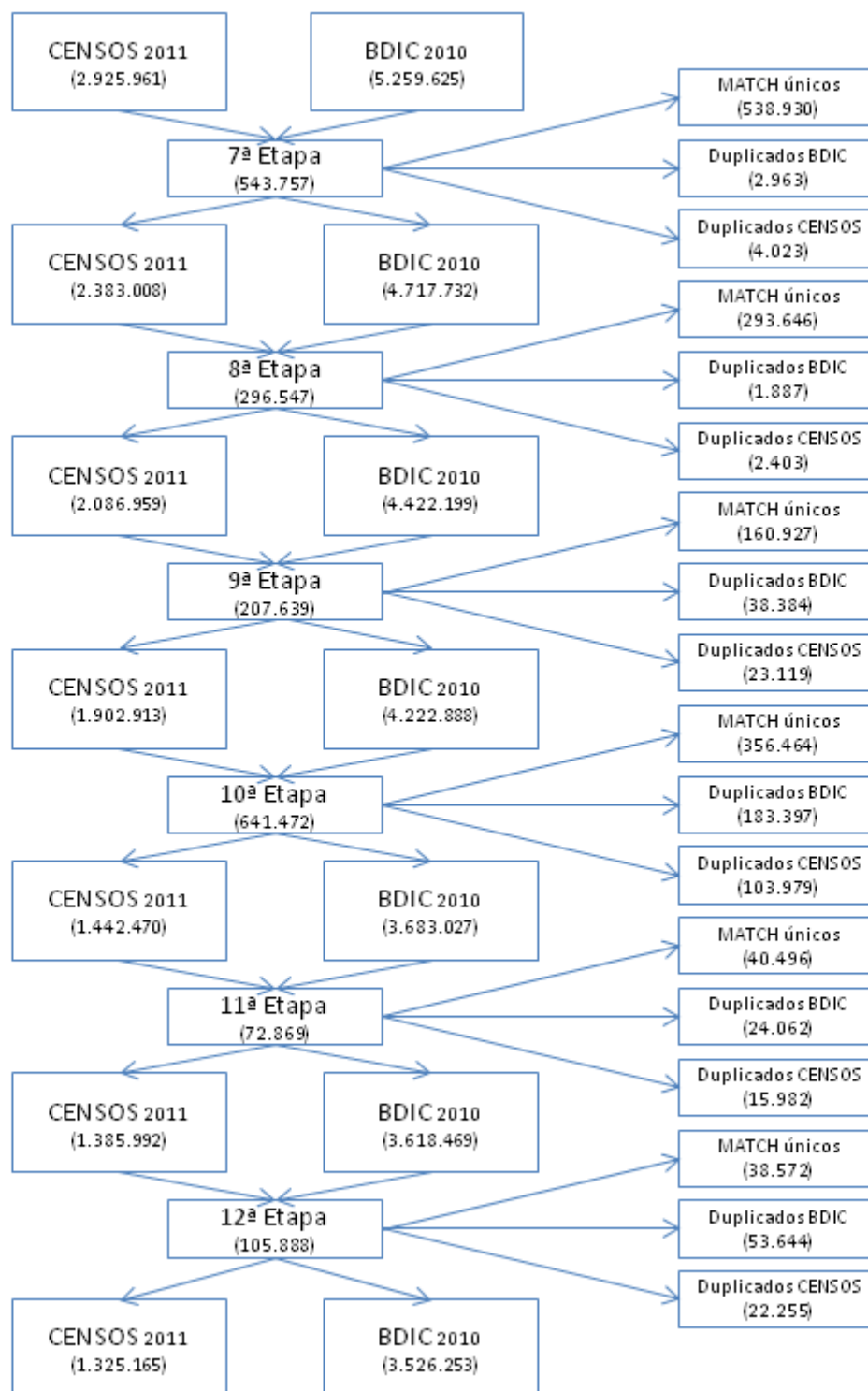
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- ÚLTIMOS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ÚLTIMOS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO (+/- 2 anos)
- NATURALIDADE

No final destas 12 etapas obtiveram-se os resultados apresentados no quadro 3.

	Nº pares ligados	Duplicados BDIC	Duplicados Censos	NDuplicados BDIC	NDuplicados Censos
1ª Etapa	5.506.869	8.935	2.543	2.198	15.011
2ª Etapa	55.765	42	23	26	64
3ª Etapa	165.908	229	209	398	440
4ª Etapa	177.727	261	68	128	514
5ª Etapa	1.371.175	2.589	1.225	1.319	4.049
6ª Etapa	326.207	2.966	4.046	6.627	4.374
7ª Etapa	538.930	1.849	796	1.114	3.227
8ª Etapa	293.646	1.009	494	878	1.909
9ª Etapa	160.927	7.457	16.645	30.927	6.474
10ª Etapa	356.464	68.777	77.476	114.620	26.503
11ª Etapa	40.496	5.800	8.721	18.262	7.261
12ª Etapa	38.572	10.554	16.001	43.090	6.254
Total:	9.032.686	110.468	128.247	219.587	76.080

O esquema seguinte representa o processo, dividido em cada uma das etapas e os resultados obtidos em cada uma delas.





Como se pode verificar, da análise do esquema anterior, restam 1.529.492 registos dos Censos (204.327 são duplicados) e 3.856.308 registos da BDIC (330.055 são duplicados e mais de 1 milhão são relativos a portugueses residentes no estrangeiro) sem correspondência ou que não fizeram *matching*.

No final deste processo foram obtidos os seguintes resultados:

- *Match* entre os Censos e a BDIC_2010
- Registos dos Censos sem *match*
- Registos da BDIC 2010 sem *match*

4.4. Comparação CENSOS2011 e BDIC2014

Na comparação das bases de dados dos Censos e BDIC 2014 foram usados apenas os registos sem *match* da comparação com a BDIC 2010 (1.529.492 registos).

Numa primeira abordagem, optou-se por fazer dez comparações, iniciando na mais restritiva e terminando na menos restritiva.

Em cada etapa, encontram-se os registos com *match* sendo-lhes atribuído um *score* decrescente. Os restantes registos (sem correspondência) são analisados na etapa seguinte, de uma forma sucessiva.

1º Etapa: Match com o maior número de campos possível

Nesta etapa efetuou-se o match com os seguintes campos:

- FREGUESIA RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO
- NATURALIDADE

2º Etapa: Match com o maior número de campos possível

Nesta etapa efetuou-se o match com os seguintes campos:

- FREGUESIA RESIDENCIA CAOP13
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO
- NATURALIDADE

3º Etapa: Match com os seguintes campos (retirou-se a naturalidade)

- FREGUESIA RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- ÚLTIMOS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO

4º Etapa: Match com os seguintes campos (retirou-se a naturalidade)

- FREGUESIA RESIDENCIA CAOP13
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- ÚLTIMOS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO

5º Etapa: Match com os seguintes campos (subiu-se o nível para município e acrescentou-se a naturalidade)

- MUNICIPIO RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ANO E MÊS DA DATA NASCIMENTO
- NATURALIDADE

6º Etapa: Match com os seguintes campos (subiu-se o nível para município e acrescentou-se a naturalidade)

- MUNICIPIO RESIDENCIA CAOP13
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ANO E MÊS DA DATA NASCIMENTO
- NATURALIDADE

7º Etapa: Match com os seguintes campos (retirou-se o dia da data de nascimento e usou-se a freguesia):

- FREGUESIA RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ANO E MÊS DA DATA NASCIMENTO
- NATURALIDADE

8º Etapa: Match com os seguintes campos (retirou-se o dia da data de nascimento e usou-se a freguesia):

- FREGUESIA RESIDENCIA CAOP13
- SEXO

- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ANO E MÊS DA DATA NASCIMENTO
- NATURALIDADE

9º Etapa: Match com os seguintes campos (o ano da data de nascimento varia entre +/- 10 anos)

- FREGUESIA RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ANO DA DATA NASCIMENTO (+/- 10)
- NATURALIDADE

10º Etapa: Match com os seguintes campos (o ano da data de nascimento varia entre +/- 10 anos)

- FREGUESIA RESIDENCIA CAOP13
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ANO DA DATA NASCIMENTO (+/- 10)
- NATURALIDADE

No final destas 10 etapas obtiveram-se os resultados que se apresentam no quadro 4.

Quadro 4 – Síntese de resultados da comparação entre os Censos 2011 e a BDIC 2014

	Nº pares ligados	Duplicados BDIC	Duplicados Censos	NDuplicados BDIC	NDuplicados Censos
1ª Etapa	163.711	5.468	1.453	967	9.016
2ª Etapa	148.914	4.924	1.241	706	8.087
3ª Etapa	56.540	1.334	533	491	2.093
4ª Etapa	61.694	1.352	532	485	2.122
5ª Etapa	26.623	708	267	371	1.253
6ª Etapa	12	0	0	0	0
7ª Etapa	8.935	113	319	658	201
8ª Etapa	8.096	104	352	683	148
9ª Etapa	36.241	105.045	53.499	227.073	2.429
10ª Etapa	32.411	102.491	45.603	215.388	2.162
Total:	543.177	221.539	103.799	446.822	27.511

Na comparação entre os Censos 2011 e a BDIC 2014 encontraram-se 543.177 registos (adicionais relativamente à comparação anterior). Ficaram ainda 986.315 registos dos Censos sem *match*.

Foram obtidos resultados relativos a:

- *Match* entre os Censos e a BDIC_2014
- Registos dos Censos sem *match*

4.5. Comparação CENSOS2011 e Segurança Social 2013

Na comparação das bases de dados dos Censos e da Segurança Social 2013, foram usados apenas os registos sem *match* da comparação com a BDIC 2014 (986.315 registos). Optou-se por fazer 10 etapas de comparações, iniciando na mais restritiva e terminando na menos restritiva.

Tal como nas comparações anteriores, aos registos com *match* foi-lhes atribuído um *score* decrescente. M

1º Etapa: Match com o maior número de campos possível:

Nesta etapa efetuou-se o match com os seguintes campos:

- FREGUESIA RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO
- NATURALIDADE

2º Etapa: Match com o maior número de campos possível (CAOP13)

- FREGUESIA RESIDENCIA CAOP13
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO
- NATURALIDADE

3º Etapa: Match com os seguintes campos (retirou-se a naturalidade)

- FREGUESIA RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- ÚLTIMOS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO

4º Etapa: Match com os seguintes campos (retirou-se a naturalidade)

- FREGUESIA RESIDENCIA CAOP13
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- ÚLTIMOS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO

5º Etapa: Match com os seguintes campos (subiu-se o nível para município e acrescentou-se a naturalidade)

- MUNICIPIO RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO
- NATURALIDADE

6º Etapa: Match com os seguintes campos (subiu-se o nível para município e acrescentou-se a naturalidade)

- MUNICIPIO RESIDENCIA CAOP13
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- DATA NASCIMENTO
- NATURALIDADE

7º Etapa: Match com os seguintes campos (retirou-se o dia da data de nascimento e usou-se a freguesia):

- FREGUESIA RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ANO E MÊS DA DATA NASCIMENTO
- NATURALIDADE

8º Etapa: Match com os seguintes campos (retirou-se o dia da data de nascimento e usou-se a freguesia):

- FREGUESIA RESIDENCIA CAOP13
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ANO E MÊS DA DATA NASCIMENTO
- NATURALIDADE

9º Etapa: Match com os seguintes campos (o ano da data de nascimento varia entre +/- 10 anos)

- FREGUESIA RESIDENCIA
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ANO DA DATA NASCIMENTO (+/- 10)
- NATURALIDADE

10º Etapa: Match com os seguintes campos (o ano da data de nascimento varia entre +/- 10 anos)

- FREGUESIA RESIDENCIA CAOP13
- SEXO
- PRIMEIROS 3 CARACTERES DO PRIMEIRO NOME
- PRIMEIROS 3 CARACTERES DO ÚLTIMO NOME
- ANO DA DATA NASCIMENTO (+/- 10)
- NATURALIDADE

No final destas 10 etapas obtiveram-se os seguintes resultados que se apresentam no quadro 5.

Quadro 5 – Síntese de resultados da comparação entre os Censos 2011 e a Segurança Social 2013

	Nº pares ligados	Duplicados SS	Duplicados Censos	NDuplicados SS	NDuplicados Censos
1ª Etapa	1.556	2.157	255	124	3.937
2ª Etapa	1.524	1.935	224	112	3.537
3ª Etapa	24.346	802	157	83	1.378
4ª Etapa	31.369	670	111	86	1.204
5ª Etapa	632	232	46	73	446
6ª Etapa	0	0	0	0	0
7ª Etapa	435	61	54	116	122
8ª Etapa	427	58	104	192	92
9ª Etapa	9.930	42.595	28.978	77.551	4.129
10ª Etapa	8.584	46.587	26.638	82.351	3.149
Total:	78.803	95.097	56.567	160.688	17.994

Na comparação entre os Censos 2011 e a Segurança Social 2013 fizeram *matching* 78.803 registos (adicionais relativamente às comparações anteriores). No final deste

processo, ficam 907.512 registos dos Censos sem *match* (ou por ligar). Neste âmbito, foram obtidos os seguintes resultados:

- Match entre os Censos e a Segurança Social 2013
- Registos dos Censos sem *match*

4.6. Comparação usando o Quality Stage

Durante os trabalhos de análise da interligação entre ficheiros, após a ligação dos Censos com a BDIC 2010, BDIC 2014 e Segurança Social 2013, foi usado o *software Quality Stage*. Foi possível efetuar o *match* probabilístico usando a morada, combinada com outros atributos, entre os registos dos Censos 2011 que ainda não foram ligados, com os registos que restam sem par, isto é, que não fizeram *matching*, da BDIC 2010.

A análise de qualidade aos resultados das ligações produzidas neste processo do *Quality Stage* ainda está a decorrer, pelo que os pares aqui encontrados não foram adicionados ao resultado total encontrado e referido nos processos descritos nos pontos anteriores. Pretende-se retomar a utilização desta ferramenta após o cruzamento com as fontes adicionais SEF, IEFP, Educação, ...

Dos registos que restam dos Censos 2011 foi necessário retirar aqueles que foram identificados como estrangeiros, 315.498 registos, pois não estão na BDIC 2010; assim dos 1.058.996 registos individuais, ficam 743.498 para comparação. Dos registos da BDIC 2010 foi verificado que 586.924 registos não tinham morada pelo que apenas 3.269.384 registos válidos podem ser comparados.

Nesta fase, após uma análise complementar dos registos considerados inicialmente válidos, verificou-se, não obstante, a existência de moradas com caracteres não padrão. Alguns exemplos de padrões dos caracteres dessas moradas:

****	A	000	.	...	?	???	----	===
------	---	-----	---	-----	---	-----	------	-----

Estas moradas, num total de 121.599 registos, não foram consideradas. Restam assim 3.147.784 de registos válidos e 708.523 de não válidos.

Esquemáticamente, o processo usado nesta etapa, foi o seguinte:



Devido ao volume de registos envolvidos, existiu a necessidade deste processo ser executado ciclicamente para cada distrito. A infraestrutura de suporte ao *software Quality Stage* disponível não permite tratar todos os registos de uma única vez.

Neste processo foram utilizadas duas combinações de variáveis de comparação, além da blocagem por distrito. Embora ao colocar a Freguesia com um dos elementos de comparação esta está na prática a ser usada.

O primeiro grupo a ser usado contém os seguintes campos para comparação:

- NOME
- SEXO
- FREGUESIA
- RUA
- DATA_NASCIMENTO
- NUMERO_PORTA

Com os registos restantes foi novamente excetuado o processo mas agora com a seguinte combinação:

- NOME
- SEXO
- FREGUESIA
- DATA_NASCIMENTO_AM
- NATURALIDADE

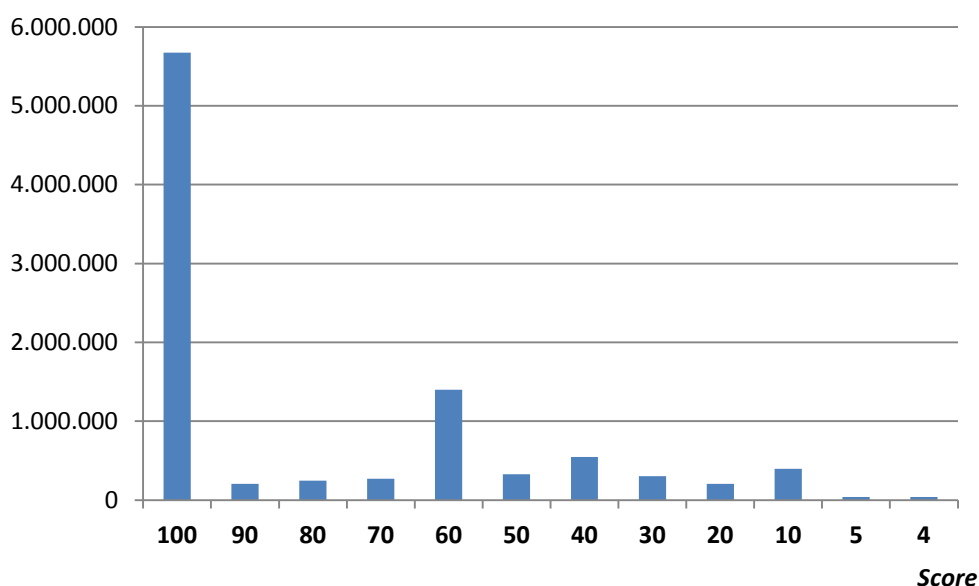
Relativamente à etapa anterior, foi retirado o dia da data de nascimento e número da porta, adicionando-se a naturalidade. Resultaram, desta fase, um total de 199.747 registos ligados. Esses registos não foram considerados nos resultados totais do exercício de *matching*, uma vez que estão ainda em fase de análise.

Existem limitações quanto à utilização deste processo com os mesmos elementos para a BDIC 2014 por não ter o NOME e MORADA completa. Quanto à Segurança Social 2013, os registos podem ser cruzados com os Censos 2011 utilizando os mesmos elementos de caracterização.

4.7. Síntese dos principais resultados

Da ligação entre os Censos 2011 e os três ficheiros utilizados até ao momento para efeitos de *matching*, foi atribuído NIC a 9 654 666 registos censitários, o que corresponde a uma taxa de 91,1%. O gráfico seguinte resume o número de registos ligados por *score* atribuído. Refira-se que o valor de *score* atribuído é classificativo, de ponto de vista da ordenação quanto à precisão, das várias combinações de chaves utilizadas nas diferentes fases do processo, não estando associada a proporcionalidade entre os valores atribuídos.

Figura 3 – Número de registos dos Censos 2011 com match, segundo o score



Os principais resultados colocam em evidência, que dos registos com match cerca de 58,8% (5.672.136) têm *score* 100, o que significa que foram ligados na primeira etapa do processo de *matching*, com a utilização de uma chave de ligação bastante restritiva, em que foi garantida a igualdade de 8 atributos comuns.

O número de registos ligados com *score* 60 representa 14,5%, e resulta da eliminação, na chave de ligação, da variável naturalidade, cujo conceito não é o mesmo nas fontes administrativas e nos censos. Refira-se que com este *score* está acautelada a igualdade de 5 variáveis (freguesia de residência, sexo, data de nascimento, primeiros 3 caracteres do primeiro nome e 3 últimos caracteres do último nome).

O quadro seguinte mostra a contribuição de cada fonte para o total das ligações encontradas, ressaltando no entanto que a procura foi efetuada sequencialmente nas três fontes, o que naturalmente tem impacto no número de ligações encontradas.

Quadro 6 – Número de registos dos Censos 2011 com match, segundo o *score* e a respetiva fonte

<i>Score</i>	BDIC2010	BDIC2014	SS13
Total	9.032.686	543.177	78.803
100	5.506.869	163.711	1.556
90	55.765	148.914	1.524
80	165.908	56.540	24.346
70	177.727	61.694	31.369
60	1.371.175	26.623	632
50	326.207	12	0
40	538.930	8.935	435
30	293.646	8.096	427
20	160.927	36.241	9.930
10	356.464	32.411	8.584
5	40.496	0	0
4	38.572	0	0

4.8. Caraterização dos registos que não foi possível ligar

Conforme já referido em ponto anterior, dos 907 512 registos dos Censos 2011 para os quais não possível obter correspondência nas bases administrativas analisadas, 647 829 são relativos a cidadãos portugueses e 259 683 a cidadãos estrangeiros. Os registos relativos a estes últimos não vão ser caracterizados nesta fase, pois ainda aguardam o cruzamento com a base de dados do SEF.

Relativamente aos indivíduos de nacionalidade portuguesa dos Censos 2011 que, nesta fase, ainda não foi possível ligar são na sua maioria do sexo feminino, e com o estado civil solteiro ou casado.

Figura 4 – Indivíduos de nacionalidade Portuguesa dos Censos 2011, não ligados, segundo o sexo

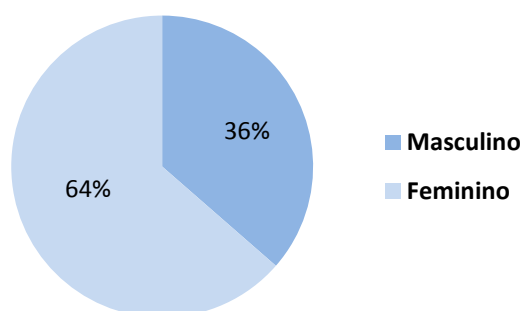
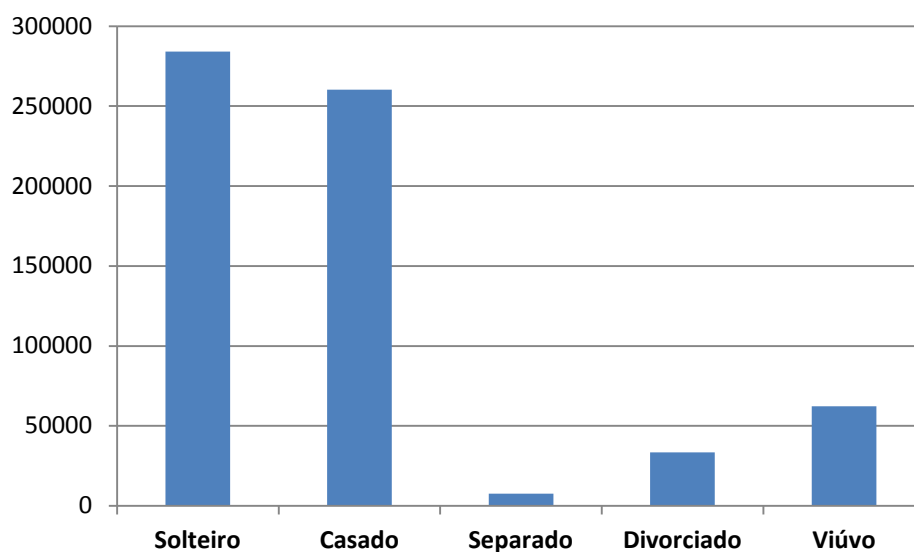


Figura 5 - Indivíduos de nacionalidade Portuguesa dos Censos 2011, não ligados, segundo o estado civil

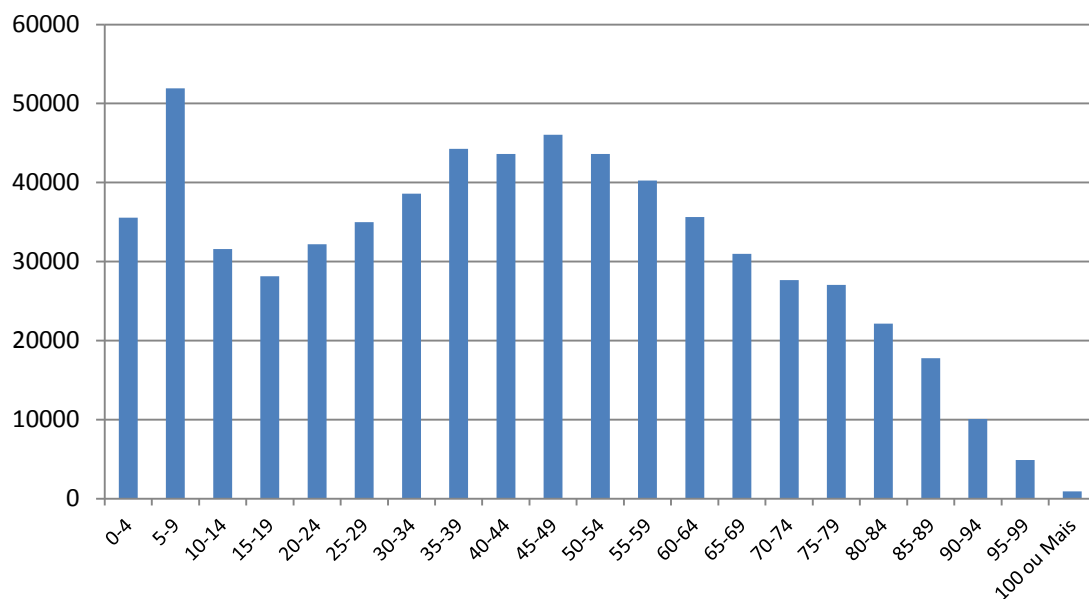


Atendendo ao grupo etário dos indivíduos para os quais não foi possível encontrar ligação, constatou-se a elevada representatividade, por um lado de registos associados aos grupos etários mais novos, com destaque para as idades compreendidas entre os 5 e os 9 anos, e, por outro, dos indivíduos entre os 35 e os 54 anos.

A análise efetuada até ao momento não permitiu encontrar motivos que justifiquem esta distribuição. Apesar da BDIC 2010 ter uma baixa representatividade dos grupos etários mais jovens, na BDIC 2014, com a atribuição do NIC à nascença a partir de 2007, esta população em princípio estaria coberta.

De qualquer forma, e no que respeita às crianças, temos a expectativa que este padrão possa ser atenuado ou mesmo desaparecer aquando do cruzamento com o ficheiro da Educação.

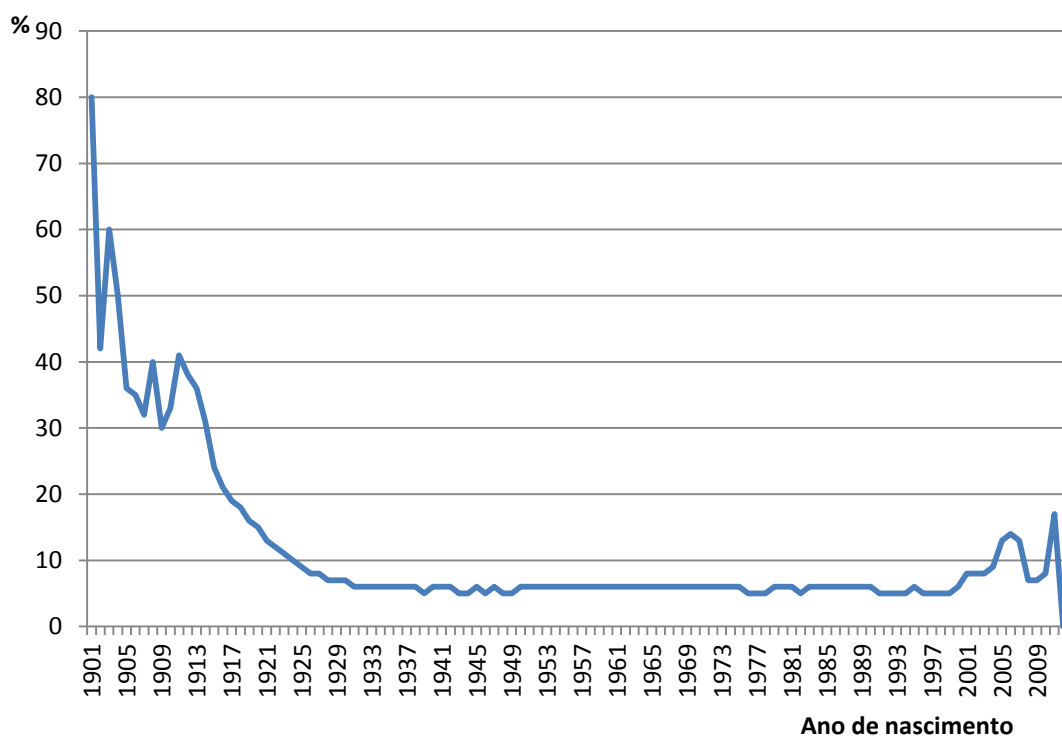
Figura 6 – Indivíduos de nacionalidade Portuguesa dos Censos 2011, não ligados, segundo o grupo etário



Tendo como foco de análise a percentagem de registos que não foi possível ligar, verifica-se que, para a variável ano de nascimento, a taxa com o valor mais baixo é cerca de 5%, sendo este o valor verificado na maior parte dos anos de nascimento.

Destacam-se com maiores percentagens de registos não ligados os indivíduos com anos de nascimento mais recuados, anteriores a 1917 (embora em termos absolutos sejam valores reduzidos) e corroborando os dados anteriores, os indivíduos com ano de nascimento posterior a 2005.

Figura 7 – Percentagem de registos dos Censos 2011, não ligados, segundo o ano de nascimento



Atendendo ao município de residência, a taxa média de registos não ligados foi de 6,4%. A percentagem mais baixa de registos não ligados foi cerca de 4% e, por oposição a taxa mais elevada foi de 12%.

Nos quadros seguintes apresentam-se os municípios com as menores e as maiores percentagens de registos não ligados.

Quadro 7 – Municípios com a menor % de registos dos Censos 2011 sem ligação

Município	Total de pares não ligados	Total Censos	% registos não ligados	Total BDIC 2010
Abrantes	1726	38801	4	41518
Horta	541	14577	4	18005
Arraiolos	321	7248	4	7433
Redondo	277	6876	4	7107
Lajes do pico	191	4587	4	5751
Corvo	16	391	4	409
Braga	9597	176691	5	188617
Guimarães	8415	156577	5	170593
Vila Franca de Xira	6945	129128	5	130838
Leiria	6103	122482	5	134991

Quadro 8 – Municípios com a maior % de registos dos Censos 2011 sem ligação

Município	Total de pares não ligados	Total Censos	% registos não ligados	Total BDIC 2010
Celorico de Basto	1867	19980	9	23098
Vila Franca do campo	973	11126	9	13308
Trancoso	885	9746	9	12880
Idanha-a-Nova	826	9524	9	11571
Vila Nova de Foz Côa	610	7145	9	10288
Aguiar da Beira	464	5408	9	7926
Mêda	442	5140	9	7402
Porto Moniz	232	2694	9	4383
Pampilhosa da Serra	447	4451	10	4895
Freixo de Espada à Cinta	455	3689	12	4458

Da análise das percentagens de registos não ligados por município de residência, não foi possível verificar um padrão geográfico muito definido. Nos municípios de maior dimensão em termos de população residente, a taxa de registos não ligados foi de 7% em Lisboa, Sintra e Porto e de 6% em Vila Nova de Gaia.

5. Avaliação da qualidade do processo de interligação das bases de dados

Neste ponto apresentam-se os primeiros indicadores sobre a qualidade do processo de interligação das bases de dados. Por se tratar de um processo novo, e ainda em construção, e porque avaliar a qualidade, é por si, uma tarefa complexa, esta análise será complementada com elementos adicionais em fases seguintes do projeto.

O objetivo da aplicação da metodologia de *matching* é localizar o mesmo indivíduo em diferentes bases de dados, e assim criar uma base de registos com informação proveniente de diferentes ficheiros, registos esses a utilizar na construção de uma base de população residente em Portugal para fins censitários. Trata-se de um processo exaustivo de ligar registos relativos a pessoas, pelo que avaliar a qualidade torna-se essencial.

Esta análise afigura-se também pertinente, na fase corrente dos trabalhos, pois qualquer limitação identificada na qualidade das ligações pode ser tida em conta na análise dos dados ligados, havendo assim hipótese de melhorar os resultados. Adicionalmente, da avaliação da qualidade pode resultar um *feedback* que permita ajudar na definição de estratégias de melhoria da recolha e processamento dos dados junto das fontes.

Neste ponto, recorda-se que se considerou como hipótese inicial de trabalho para aplicação da metodologia de ligação de registos através de atributos comuns que *match* é uma ligação verdadeira entre um par de registos em que os dois registos pertencem efetivamente à mesma pessoa (na prática, os registos são comparados aos pares com base na hipótese de que fazem um *match* exato).

5.1. Como determinar a qualidade das ligações?

Os métodos para avaliar um processo de *matching* dependem, à *priori*, da qualidade e dimensão dos dados envolvidos e dos próprios métodos de ligação, designadamente da qualidade e das estratégias de definição das chaves utilizadas bem como do *software* envolvido.

A qualidade do emparelhamento entre registos, concretamente determinar as ligações que correspondem verdadeiramente ao mesmo indivíduo, ou que fazem *match*, é um fator essencial e decisivo neste processo.

Relativamente à qualidade dos dados, as bases foram previamente limpas, normalizadas e identificados os valores duplicados. Este processo, já descrito, tornou os dados mais consistentes e reduziu a redundância nas ligações.

Por seu turno, para maximizar o número de ligações entre as bases de dados, foram aplicadas chaves de ligação em várias etapas sucessivas, alterando as chaves de acordo com os atributos disponíveis (do mais restritivo para o mais abrangente), de forma a captar a variabilidade dos dados. Neste sentido, há um conjunto de regras de decisão que determinam se um par está ou não está ligado. A cada chave de ligação estão associados ponderadores, que se baseiam no volume e qualidade de informação com que essas características contribuem para a probabilidade dos registos pertencerem ao mesmo indivíduo ou a indivíduos diferentes.

Em termos de implementação, para além de ser usado um processo automatizado para ligação através de chaves determinísticas (uma vez que estão disponíveis nas diferentes fontes atributos como a naturalidade, o sexo ou a data de nascimento), foi também utilizado o *software Quality Stage* para identificar ligações a partir dos registos que não fizeram *matching* pela via anterior e com recurso a atributos que se verificou que embora possam pertencer ao mesmo indivíduo não apresentam um formato fechado. Neste caso, quando há variações dos atributos nas diversas fontes, como nos de natureza geográfica e relativos a endereços, por exemplo, em vez de verificarem igualdades exatas entre chaves de ligação para localizar registos comuns, verificam-se semelhanças entre os atributos.

Para avaliar a qualidade dos dados ligados podem ser usados vários métodos. Nesta fase identificam-se, de entre os pares ligados, os que correspondem a ligações verdadeiras, ou *matches*, e os que são ligações falsas (usualmente designados, em qualidade, como falsos positivos). As várias possibilidades de ligações, entre cada par de

registos, podem ser resumidas no quadro seguinte (os pares não ligados serão objeto de estudo em fase posterior).

Quadro 9 – Relação entre ligações e *matches*

		Pares verdadeiros		
		<i>Match</i>	<i>No Match</i>	
Pares ligados	Ligação	Ligações verdadeiras	Ligações falsas	Total de ligações
	Sem ligação	Não ligações falsas	Não ligações verdadeiras	
		Total de <i>matches</i>		

5.2. Exatidão das ligações

Apresenta-se, de seguida, o processo de estimação de uma medida de exatidão das ligações, que se baseia na estimação do número de ligações corretas, de entre os registos da base de dados resultante do processo de interligação entre as diferentes bases de dados.

$$\text{Exatidão das ligações} = \frac{\text{Ligações verdadeiras}}{\text{Total de ligações}} \times 100$$

Da estimação desta medida pode deduzir-se, por sua vez, a própria estimação do número de ligações falsas, ou falsos positivos (pares de registo ligados com possibilidade de pertencerem ao mesmo indivíduo mas que afinal pertencem a indivíduos diferentes), isto é:

$$\text{Taxa de falsos positivos} = 100 - \text{Exatidão das ligações}$$

Assim, para estimar as ligações verdadeiras, foi aplicada uma rotina de verificação de pares de nomes com o objetivo de selecionar de entre os pares que foram ligados, os

que estão corretos, isto é, que correspondem a ligações de registos relativos a um mesmo indivíduo ou que fizeram *matching*. Para o efeito utilizaram-se os campos relativos aos nomes completos. Esta reflexão surgiu da observação aleatória de alguns registos da tabela de pares ligados, associados sobretudo aos *scores* mais baixos, em que há fortes evidências de que os pares de nomes não correspondem ao mesmo indivíduo. De seguida apresenta-se o método desenvolvido para o efeito.

Criação de uma função de verificação de coincidência de nomes, que, para cada registo:

1. Mediu a dimensão total, em termos de caracteres da *string* do campo nome completo, em cada uma das bases de dados em comparação (ex.: nome NOME_BDIC vs NOME_C2011)
2. Selecionou, de entre os dois campos dos nomes completos, o relativo à *string* com menor dimensão e partiu em *substrings* a partir dos espaços em branco (ex.: comparando o registo “ANA GERTRUDES VIEIRA ALEXANDRE FEIJAO”, correspondente campo NOME_BDIC, com o registo “ANA FEIJAO”, correspondente campo NOME_BC2011, escolheu-se para partir o registo do campo NOME_BDIC que deu origem às *substrings* “ANA” e “FEIJAO”)
3. Verificou se todas as *substrings* originadas pertencem à *string* maior do campo do nome completo que não foi partido (no exemplo do ponto anterior verifica se “ANA” e “FEIJAO” estão em “ANA GERTRUDES VIEIRA ALEXANDRE FEIJAO”) e nesse caso valida como correto aquele par.

5.3. Resultados – Indicadores de Qualidade

Para o efeito da implementação da metodologia acima descrita, foi criada a variável NOME_CONTIDO. Foi atribuído o valor 1 aos pares identificados no *matching* e que consideramos corresponderem a ligações verdadeiras (9148759 dos pares ligados) e o valor 0 (atribuído a 505907 dos pares ligados) às ligações que consideramos falsas e não correspondem a *matching*, isto é, os falsos positivos.

Estes valores apontam para os seguintes resultados:

- *Exatidão das ligações* ~ 95%
- *Taxa de falsos positivos* ~ 5%

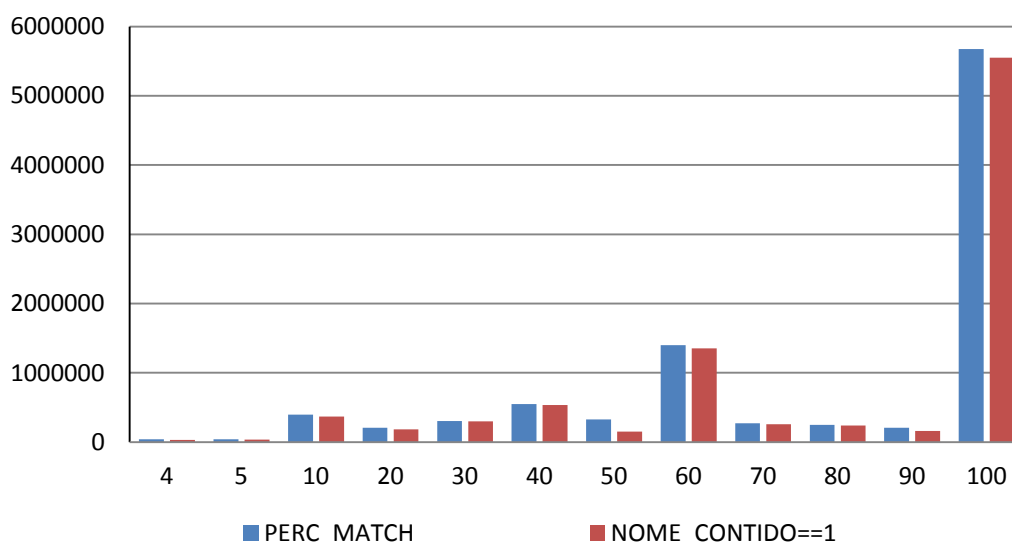
Nesta fase do processo, consideram-se estes valores bastantes satisfatórios e adequados para prosseguir com a aplicação da metodologia. Ainda assim, terá que haver um trabalho complementar, através de mais indicadores, de modo a uma confirmação exaustiva destes falsos positivos, uma vez que foram ligados pares de registos com pequenas diferenças nos nomes e que possivelmente dizem respeito ao mesmo indivíduo (Exemplo: Isaltino Soares Mendes e Izaltino Soares Mendes, a que se atribui NOME_CONTIDO=0).

No quadro 10 e figura 8 apresentam-se os resultados da aplicação deste método, por *score* (correspondente ao peso ou fiabilidade atribuída a cada uma das etapas de ligação com a utilização de diferentes chaves). Como espectável, os valores obtidos para o NOME_CONTIDO vão ao encontro das percentagens de *match*, com valores de exatidão das ligações, que na maior parte das etapas rondam, ou superam, os 90%.

Quadro 10 - Número de ligações, taxas de falsos positivos e exatidão da ligação, por *score*

Scores	Nº pares ligados	Nº ligações falsas	Taxa de falsos positivos	Exatidão da ligação
100	5672136	123929	2,2%	97,8%
90	206203	47044	22,8%	77,2%
80	246794	9311	3,8%	96,2%
70	270790	12270	4,5%	95,5%
60	1398430	47342	3,4%	96,6%
50	326219	177385	54,4%	45,6%
40	548300	14151	2,6%	97,4%
30	302169	5628	1,9%	98,1%
20	207098	24809	12,0%	88,0%
10	397459	27604	6,9%	93,1%
5	40496	6675	16,5%	83,5%
4	38572	9759	25,3%	74,7%
Total	9654666	505907	5,2%	94,8%

Figura 8 – Percentagem de *match* e de nome contido, por *score*



5.4. Outros indicadores da qualidade das ligações

Para aferir a qualidade das ligações identificadas, foram criadas, para além da variável NOME_CONTIDO, outras variáveis, que ainda estão em estudo e que poderão ser também indicadores de qualidade. Nesse sentido, foi construída uma tabela de resultados contendo, para as ligações estabelecidas com os registos da base dos Censos 2011, a seguinte informação adicional:

NOME_C2011_CONV – Nome dos censos sem DE, DA, DOS, DAS, DO e plicas

NOME_OUTRO_CONV – Nome doutras fontes sem DE, DA, DOS, DAS, DO e plicas

TAM_C2011 – Tamanho do campo NOME_C2011_CONV

TAM_OF – Tamanho do campo NOME_OUTRO_CONV

NOME1 – Primeiro nome do nome mais pequeno entre censos e outras fontes

NOME2 – Segundo nome do nome mais pequeno entre censos e outras fontes

...

NOME11 – Décimo primeiro nome do nome mais pequeno entre censos e outras fontes

NOME_CONTIDO – 1 se todos os nomes de uma fonte estão contidos na outra; 0 caso contrário

DIST_NOME – Nº de caracteres diferentes entre os nomes dos censos e das outras fontes

DIST_PRIM_ULT_NOME – Nº de caracteres diferentes entre o primeiro + último nomes dos censos e das outras fontes

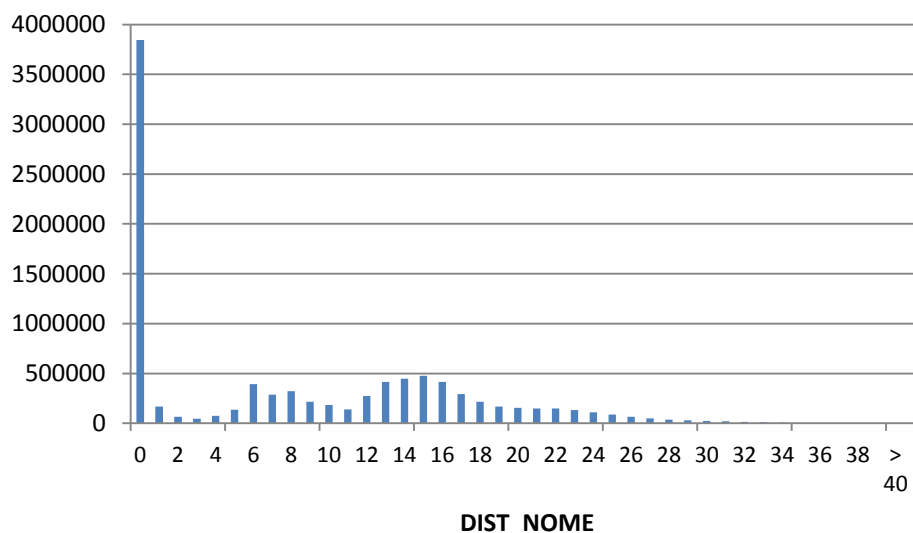
DIST_DATA_NASC – Nº de caracteres diferentes entre as datas de nascimento
DIST_NAT – Nº de caracteres diferentes entre as naturalidades
DIST_FR – Nº de caracteres diferentes entre as freguesias de residência
DIST_RUA – Nº de caracteres diferentes entre as ruas (não normalizadas)
SEMELHANCA_NOME – Semelhança entre os nomes completos – Nº de 0 a 100 completa semelhança; 0 nenhuma semelhança
SEMELHANÇA_PRIM_ULT_NOME – Semelhança entre os primeiros e últimos nomes (100 semelhança completa; 0 nenhuma semelhança)
PERC_DIST_NOME – proporção entre o nº de caracteres no nome (dos Censos 2011) e o nº de caracteres diferentes entre o nome dos censos e o nome de outras fontes

De entre este conjunto de variáveis destacam-se duas: DIST_NOME e SEMELHANCA_NOME, que se analisam sumariamente, a seguir.

- **DIST_NOME:** Traduz a diferença entre o número de caracteres de dois nomes ligados.

Considerando o quadro 10, que ilustra os resultados a variável NOME_CONTIDO, os valores associados aos *scores* 90 e 50 são significativamente mais elevados que os restantes. Estes podem ser reduzidos se as variáveis NOME_CONTIDO e DIST_NOME forem trabalhadas em conjunto. A título de exemplo refira-se que se, para além do NOME_CONTIDO, forem consideradas em simultâneo a distância entre os nomes em que há diferença apenas em 1 carácter o valor estimado para a segunda etapa de ligação de registos diminui de 22,8% para 17,6% e o da quinta etapa de 54,4 para 49,4%. Na figura 9, ilustra-se a distribuição das diferenças entre pares de nomes ligados (note-se que em cerca de um terço dos registos, 3 843 299, os nomes ligados têm exatamente o mesmo número de caracteres).

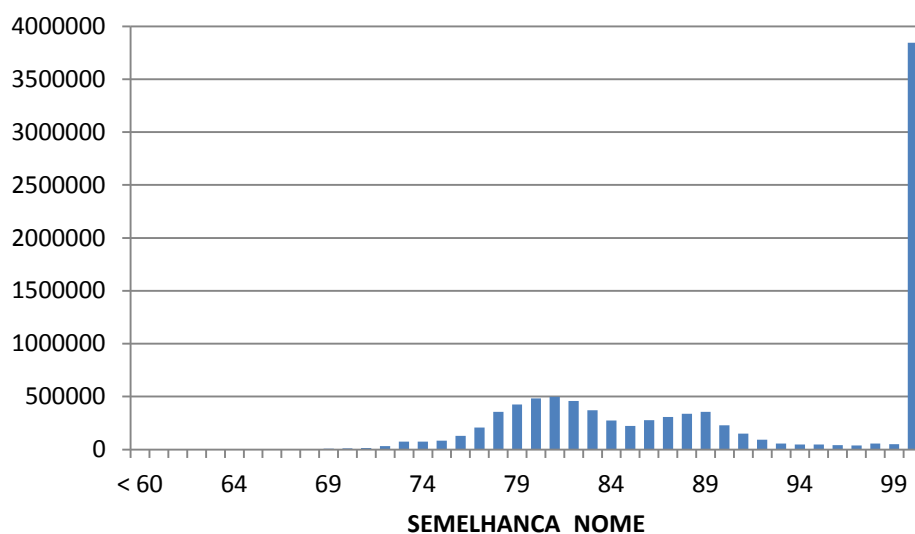
Figura 9 – Número de registos por diferença no número de caracteres entre nomes



Globalmente, se para além dos pares de registos com NOME_CONTIDO = 1, se considerar aqueles que tendo NOME_CONTIDO = 0 tiverem DIST_NOME = 1, o valor estimado da taxa de falsos positivos pode reduzir-se para cerca de 4% (ou mesmo 3,5% se consideramos $\text{DIST_NOME} \leq 2$). A este propósito, considerem-se, por exemplo, os nomes Maria Lurdes Ferreira e Maria Lourdes Ferreira, a que se atribui $\text{NOME_CONTIDO}=0$ e $\text{DIST_NOME}=1$).

- SEMELHANCA_NOME** – Avalia, com recurso ao *software Quality Stage* e através de uma função predefinida a semelhança dos nomes entre duas fontes, baseando-se na avaliação da semelhança/ proximidade entre duas *strings*.
 A este propósito refira-se que a percentagem de nomes exatamente iguais, entre os pares ligados, é de cerca de 40%, sendo que com uma aproximação de 90% estão 48,2% dos pares de registos identificados como *matching* (Exemplo: MARIA HELENA COUTINHO MOREIRA MARIA HELENA MOREIRA têm $\text{SEMELHANCA_NOME}=90\%$). De resto, uma grande parte dos pares ligados apresenta uma percentagem de semelhança do nome que varia entre os 72 e os 92%, conforme se ilustra no gráfico seguinte, o que se considera bastante satisfatório nesta fase em que ainda é necessário um trabalho mais exaustivo sobre estes indicadores.

Figura 10 – Número de registos por percentagem de semelhança entre nomes



5.5. Considerações finais sobre a qualidade do processo

Conforme referido anteriormente, na fase corrente dos trabalhos ainda não foi possível definir exaustivamente todos os indicadores de qualidade para avaliar o processo de interligação entre os vários ficheiros administrativos com o objetivo de encontrar registos comuns que correspondam a uma mesma pessoa. Considera-se, no entanto, este primeiro conjunto de resultados bastante promissores, uma vez que este é um trabalho em progresso e que apresenta potencial para ser melhorado.

Pelo exposto, através da utilização simultânea de vários indicadores, dos próprios scores das chaves, e eventualmente de outros indicadores (número de caracteres diferentes no nome, nº de caracteres diferentes na data de nascimento, etc...) será possível continuar a trabalhar na redução da taxa de falsos positivos, ou seja na melhoria da medida de exatidão do processo de ligação, que, numa altura em que não estão ainda esgotadas todas as hipóteses de ligação, se situa nos 95%.

Embora não se afigure necessário, nesta fase, a avaliação da qualidade poderá permitir uma eventual reformulação de alguma etapa do processo de ligação entre as bases de dados.

Outras abordagens para avaliar a qualidade, como técnicas de validação cruzadas através de chaves de ligação ou definição de amostras para verificação manual podem ainda ser consideradas em fases futuras do processo.

Note-se que ainda não foi possível usar alguns dos ficheiros administrativos disponíveis na identificação de ligações entre registos, como o ficheiro do SEF ou da Educação, que entre outros, poderão ajudar na identificação de registos comuns a um mesmo indivíduo (no caso se este for estrangeiro ou frequentar um nível de ensino), o que se traduzirá numa atualização dos resultados e potencialmente no aumento da dimensão da tabela resultante do exercício de interligação entre as diferentes bases de dados.

Refira-se ainda que no final deste processo de ligação entre ficheiros a avaliação incidirá sobre os resultados do *matching* exato (ou determinístico) através de chaves de ligação, do *matching* probabilístico, através do *software Quality Stage* (ou *automatching*) que se baseia nas semelhanças entre as chaves ou campos de ligação e do *matching* manual (que se considera aplicar aos registos remanescentes, não ligados e que não fizeram *matching* através das técnicas anteriores). Esta última abordagem será um passo obrigatório neste processo exaustivo de localizar registos relativos à população residente. A sua implementação ainda está em estudo, mas terá que passar pela verificação manual de cada possível par, identificando a sua situação (*match* ou não *match*) utilizando toda a informação disponível restante (mesmo a que não foi usada nos processos automáticos de ligação), ajustando-se no final os registos que não tiveram correspondência.

Da análise dos pares sem ligação mas que em análise complementar se revelarem verdadeiros *match* (falsos negativos), será ainda possível melhorar os resultados da base resultante do processo de ligação de registos únicos, no que se refere à sua dimensão.

6. Conclusões e trabalho futuro

A não existência de um identificador único comum nas diferentes fontes administrativas, é uma condicionante que torna mais complexa a ligação entre ficheiros. No entanto, as chaves de *matching* com base em atributos, utilizadas neste exercício, produziram resultados muito encorajadores.

Os resultados até agora obtidos na ligação entre os Censos 2011 e os ficheiros administrativos BDIC e Segurança Social, apontam para um elevado potencial no processo de interligação de ficheiros através das chaves de *matching* utilizadas, dado que foi possível ligar 9 654 666 registos dos Censos 2011, o que corresponde a cerca de 91% do universo censitário em análise. De referir que, a maioria dos registos foram ligados nas primeiras etapas de *matching*, com a utilização de chaves de ligação

restritivas, o que oferece mais garantias de que se trata efetivamente de registos do mesmo indivíduo.

A qualidade do processo de *matching* foi avaliada através da exatidão das ligações e da estimação dos falsos positivos. Os resultados obtidos apontam para uma taxa de exatidão das ligações de 95%, ou seja, pares que pertencem ao mesmo indivíduo.

O processo de ligação de registos tem que ser objeto de maior aprofundamento, nomeadamente, na definição da estratégia a utilizar para os cerca de 9% dos registos censitários que falta ligar. As variáveis disponibilizadas nos ficheiros administrativos e a sua qualidade têm aqui um papel determinante como *input* para os métodos a utilizar.

No que respeita à interligação de ficheiros a partir de identificadores numéricos, é essencial assegurar que os identificadores numéricos, nomeadamente o NIC, sejam recolhidos e registados de forma harmonizada/consistente pelas diferentes entidades.

A articulação institucional e a cooperação por parte dos organismos detentores da informação administrativa são determinantes num projeto desta natureza.

A normalização dos dados das diferentes fontes administrativas, embora morosa, é decisiva, no sentido de assegurar bons resultados nos processos de ligação de registos.

Com a manutenção das características dos ficheiros e com a eventual possibilidade de serem colmatadas parte das limitações associadas a alguns atributos, a par de um maior aperfeiçoamento dos processos de *matching*, poderão estar reunidas as condições para assegurar a interligação de ficheiros, a um nível que permita a construção de uma base de população com base em informação administrativa.