M 2015

U.PORTO

FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

Geração Sintética de Microdados utilizando algoritmos de *data mining*

Daniel Fernando Alves da Silva

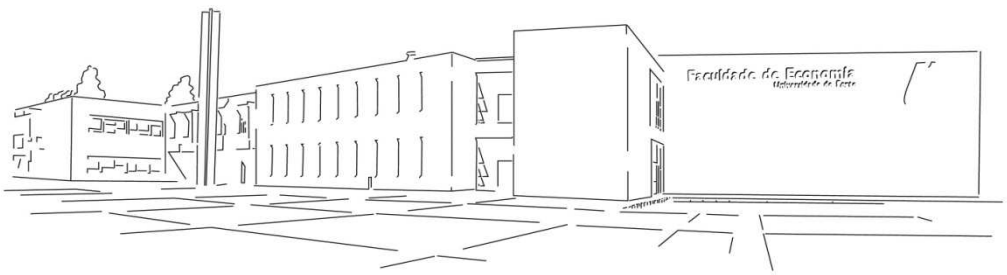
DISSERTAÇÃO DE MESTRADO EM

MODELAÇÃO, ANÁLISE DE DADOS E SISTEMAS DE APOIO À DECISÃO

Orientadores

Prof. Doutor Pedro Campos
Prof. Doutor Pavel Brazdil

Agosto de 2015



NOTA BIBLIOGRÁFICA

Daniel Fernando Alves da Silva, nascido em 11 de dezembro de 1980, natural de Massarelos, Portugal.

Em 2002 terminou a licenciatura em Gestão na Faculdade de Economia da Universidade do Porto com média de 15 valores. No ano de 2010 concluiu o Mestrado em Finanças na mesma faculdade tendo obtido a classificação final de 'Muito Bom'. Em fevereiro de 2015 completou a parte escolar do Mestrado em Análise de Dados e Sistemas de Apoio à Decisão com média final de 18 valores.

Iniciou a sua atividade profissional em outubro de 2002, exercendo a função de analista de risco de crédito na Unidade de Decisões de Operações do grupo Santander Totta. Desde outubro de 2003 trabalha no Banco de Portugal, tendo desempenhado funções técnicas no Núcleo do Fundo de Garantia do Crédito Agrícola Mútuo até maio de 2010. Desde junho do mesmo ano trabalha no Departamento de Estatística, mais precisamente no Núcleo de Análises de Balanços.

AGRADECIMENTOS

Este trabalho não teria sido possível sem o apoio e incentivo de algumas pessoas que me acompanharam durante a elaboração da dissertação.

Aos professores Pedro Campos e Pavel Brazdil, meus orientadores, pelos inúmeros ensinamentos recebidos e colaboração prestada ao longo da realização do presente trabalho.

Quero também expressar os meus agradecimentos a todas aquelas pessoas que comigo convivem e me apoiam constantemente, à minha família, aos meus amigos, colegas de trabalho e, em especial, à minha esposa Tatiana, pela elevada paciência, compreensão e palavras de incentivo.

Por fim, não posso deixar de agradecer a duas instituições que tiveram um papel fundamental. Ao Banco de Portugal por todo o tipo de recursos disponibilizados para que fosse realizado este projeto. À Faculdade de Economia do Porto pelas condições de trabalho e por todos os conteúdos científicos.

A todos eles, um obrigado se impõe!

RESUMO

A questão da confidencialidade e da proteção de dados está intimamente relacionada com os avanços tecnológicos e a proliferação de meios de divulgação da informação. Se por um lado os investigadores estão ávidos por informação cada vez mais detalhada, por outro os detentores da informação têm que lidar com as questões legais relacionadas com a confidencialidade e com os eventuais prejuízos decorrentes da identificação dos respondentes. A geração sintética, enquanto técnica de controlo da divulgação estatística, consiste em obter de forma aleatória bases de dados que preservem algumas propriedades estatísticas e a relação entre as variáveis do ficheiro original. Uma vez que a qualidade dos ficheiros sintéticos depende dos modelos utilizados e as técnicas paramétricas implicam um trabalho intensivo na identificação das relações entre as variáveis, surgiram um conjunto de trabalhos que aplicam algoritmos de *data mining* na geração deste tipo ficheiros (métodos não paramétricos). Este estudo utiliza dois algoritmos de *data mining* (Árvores de Decisão e *Random Forests*) para obter ficheiros totalmente sintéticos considerando uma base de dados de empresas. Para a prossecução deste objetivo foram utilizadas duas operacionalizações distintas, ascendendo a quatro o número de modelos testados. Os resultados obtidos permitem concluir que os modelos que utilizam algoritmos de Árvores de Decisão reproduzem melhor as estatísticas univariadas do ficheiro original e o que utiliza o algoritmo *Random Forests*, onde são consideradas todas as variáveis no modelo de imputação, retrata de forma mais consistente a relação entre as variáveis. A utilização de uma base de dados empresas, a geração de ficheiros totalmente sintéticos e a identificação e aplicação de quatro metodologias diferentes são os elementos inovadores deste trabalho.

ABSTRACT

The issue of confidentiality and data protection is closely related to technological advances and the widespread of ways of disseminating information. On the one hand the researchers are eager for detailed information, on the other hand the data holders have to deal with the legal issues related to confidentiality and damages resulting from the identification of respondents. Synthetic generation, while Statistical Disclosure Control technique, consists of achieving randomly databases that preserve some statistics properties and the relationship between the variables of the original file. The quality of the generated data depends strongly on the quality of underlying models and parametric techniques involve labor-intensive tasks identifying relationships between variables. Recently, a number of works apply data mining algorithms to generate synthetic data (non-parametric methods). This study considers two data mining algorithms (Decision Trees and Random Forests) to create fully synthetic files for a business database. To achieve this goal we used two different approaches, making four the number of models tested. Models that use Decision Trees algorithms reproduce better univariate statistics of the original file and Random Forests, with all variables considered in the imputation model, portrays more consistently the relationship between the variables. The use of a business database, the generation of fully synthetic files and the identification and implementation of four different methods are the innovative elements of this work.

ÍNDICE

ÍNDICE DE TABELAS

ÍNDICE DE FIGURAS

ÍNDICE DE ABREVIATURAS

SDC / CDE – Statistical Disclosure Control / Controlo da Divulgação Estatística
SUF – Scientific Use Files / Ficheiros de Uso Ciêntifico
PUF – Public Use Files / Ficheiros de Uso Público
SRMI – Sequential Regression Multivariate Imputation
(Imputação de Regressões Multivariadas Sequenciais)

1 – INTRODUÇÃO

O aumento da capacidade de processamento aliada à proliferação de novas técnicas de análise e de extração de conhecimento de dados impulsionou a procura de informação cada vez mais detalhada nos mais variados domínios. Se por um lado, o acesso à informação reveste características de bem público contribuindo para o desenvolvimento económico e social, por outro as questões relacionadas com a confidencialidade nunca foram tão pertinentes como atualmente.

Os produtores de informação estatística ao pretenderem facultar a melhor e a maior quantidade de informação possível para os investigadores e decisores têm que lidar com o compromisso e a obrigatoriedade legal de promoverem a confidencialidade. Grande parte da informação obtida pelos detentores da informação é através da garantia da confidencialidade dos respondentes, que se encontram legalmente protegidos de modo a que a sua identidade não seja revelada. No entanto, a questão legal não é a única importante quando se fala em privacidade. Os respondentes que sentirem que a sua privacidade está em risco vão ser relutantes em fornecer informação sensível, podem fornecer informações incorretas ou até mesmo negarem-se a responder, com consequências devastadoras para a qualidade da informação.

Para fazer face à crescente procura de informação e com o objetivo de não prejudicar a produção científica e o desenvolvimento económico e social que depende da utilização da informação estatística, os produtores de estatísticas optaram por estratégias de divulgação da informação em ambiente protegido, tendo-se multiplicado os locais e as formas onde se podem aceder aos ficheiros de uso científico e seguro: i) os laboratórios de dados; ii) os centros de pesquisa e ii) a concessão de acessos restritos (*safe centers*), remotos ou não. Estas opções para além de burocráticas, apenas se encontram acessíveis a investigadores autorizados e legitimados por centros de investigação ou universidades.

Em alternativa aos designados ficheiros de uso ciêntifico (*SUF – Scientific Use File*) e ficheiros de utilização segura (*Secure Use Files*), os ficheiros de uso público (*PUF – Public Use File*) são mais democráticos na medida em que não existem restrições de acesso. Estes ficheiros de dados são anonimizados com o objetivo de serem de acesso livre e não terem problemas de confidencialidade, caracterizando-se por um risco de identificação dos respondentes nulo ou quase nulo.

Na literatura sobre o tratamento do segredo estatístico têm sido propostas diferentes técnicas de proteção e divulgação da informação. Estas técnicas assentam na ideia que a possibilidade de identificar os respondentes pode ser neutralizada através da redução da quantidade de informação facultada, mascarando os dados (por exemplo não facultar a informação na sua totalidade ou perturbando valores), ou pela divulgação de bases de dados sintéticas em que os valores da base de dados original são substituídos. O trabalho de Winkler (2007) demostra as consequências devastadoras na qualidade da informação de alguns métodos que produzem bases de dados mascaradas, chegando a falhar, em alguns casos, o objetivo primordial de proteção de informação dos respondentes. Em alternativa aos métodos tradicionais, a geração sintética consiste em obter de forma aleatória bases de dados que preservem algumas estatísticas e a relação existente entre as variáveis do ficheiro original.

A *Imputação Múltipla* é um dos métodos mais aplicados na geração sintética de informação devido à forte fundamentação teórica e aos bons resultados obtidos em diferentes trabalhos (Reiter, 2005a; Reiter, 2005b; Caiola e Reiter, 2010; Drechsler e Reiter, 2010; Lee *et al.*, 2013; Loong *et al.*, 2013 e Raab *et al.*, 2015). Os modelos de imputação são divididos em duas abordagens: i) paramétricas e ii) não paramétricas. Apesar de alguns autores aplicarem com sucesso abordagens paramétricas, estes métodos implicam um conhecimento profundo da relação entre as variáveis e a sua implementação decorre de métodos sofisticados e tempos de execução elevados (Reiter, 2005a e Loong *et al.*, 2013). Dadas as desvantagens deste tipo de abordagens surgiram outros trabalhos que aplicam métodos não paramétricos (Reiter, 2005b; Caiola e Reiter, 2010; Drechsler e Reiter, 2010; Lee *et al.*, 2013 e Raab *et al.*, 2015). Estes métodos conseguem lidar com diferentes tipos de informação de elevada dimensionalidade, obter relações não lineares e iterações que muito dificilmente são captadas pelas abordagens paramétricas.

Este estudo utiliza dois algoritmos de *data mining* (Árvores de Decisão[1] e *Random Forests*) para gerar ficheiros totalmente sintéticos considerando uma base de dados de empresas composta por atributos[2] qualitativos e quantitativos. Para a prossecução deste

[1] Seguindo Gama *et al.* (2012) ao logo deste trabalho utiliza-se a designação Árvores de Decisão para referir Árvores de Classificação (variáveis qualitativas ou discretas) e Árvores de Regressão (variáveis quantitativas ou contínuas).

[2] No âmbito do presente trabalho o termo 'atributo' e 'variável' são utilizados como sinónimos. Habitualmente no contexto da geração sintética utiliza-se o termo 'variável' e em *data mining* a opção recaí em 'atributo'.

objetivo foram utilizadas duas operacionalizações distintas. Embora em ambas a geração das variáveis seja efetuada de forma sequencial, na primeira operacionalização são utilizadas todas as variáveis como previsores no modelo independentemente de já terem sido sintetizadas, na segunda apenas são considerados os atributos sintetizados em passos anteriores. Assim, no âmbito deste trabalho, foram testados quatro modelos diferentes (dois algoritmos e duas operacionalizações).

O trabalho de Lee *et al.* (2013) é o que mais se aproxima do presente estudo em termos de base de dados, ao utilizar o algoritmo de Árvores de Decisão em variáveis quantitativas (área, custo, colheita e receitas)[3]. Em termos de tipo de ficheiros sintéticos, os trabalhos que utilizam algoritmos de *data mining* optam por gerar ficheiros parcialmente sintéticos de modo a diminuírem a perda de informação (Reiter, 2005b; Caiola e Reiter, 2010; Drechsler e Reiter, 2010 e Lee *et al.*, 2013), sendo o trabalho recente de Raab *et al.* (2015) o único que aplica o algoritmo de Árvores de Decisão para criar ficheiros totalmente sintéticos. No que concerne a algoritmos, o trabalho de Caiola e Reiter (2010) compara os algoritmos de Árvores de Classificação e *Random Forests* para criar ficheiros parcialmente sintéticos sintetizando apenas variáveis qualitativas.

A originalidade deste trabalho pode ser avaliada em quatro parâmetros: i) utilização de uma base de dados empresas; ii) geração de ficheiros totalmente sintéticos; iii) comparação dos algoritmos de Árvores de Decisão e *Random Forests* e iv) comparação de duas operacionalizações. Apesar dos trabalhos de Drechsler (2009) e Hundepool *et al*. (2010) referirem as dificuldades de gerar sinteticamente bases de dados de empresas, não se conhecem estudos que considerem este tipo de informação, que se caracteriza por variáveis maioritariamente quantitativas, fortemente enviesadas e com elevado risco de identificação. Ao gerar ficheiros totalmente sintéticos, este estudo afasta-se dos trabalhos existentes que produzem ficheiros parcialmente sintéticos de modo a preservarem a utilidade de informação. Em termos de comparação dos algoritmos de Árvores de Decisão e *Random Forests* e utilização de duas operacionalizações distintas, os estudos que geram ficheiros sintéticos consideram, maioritariamente, um algoritmo e uma operacionalização. Apenas o trabalho de Caiola e Reiter (2010) utiliza os algoritmos de Árvores de Classificação e *Random Forests*. No entanto, os autores utilizam exclusivamente a operacionalização identificada neste trabalho com o número

[3] Para além de sintetizar apenas variáveis quantitativas, a base de dados utilizada contém apenas 338 registos.

2 e sintetizam apenas variáveis qualitativas, produzindo ficheiros parcialmente sintéticos.

Os resultados permitem concluir que a utilização de algoritmos de *data mining* para a produção de ficheiros sintéticos é uma alternativa aos métodos paramétricos, uma vez que os resultados obtidos considerando o ficheiro original e os ficheiros sintéticos são semelhantes. Em termos de algoritmos, constatou-se que o algoritmo de Árvores de Decisão consegue reproduzir melhor as estatísticas univariadas e o algoritmo *Random Forests*, considerando sempre todos os atributos no modelo de imputação, a relação entre as variáveis (estatísticas multivariadas).

Para além do presente capítulo introdutório, este trabalho está estruturado como se segue. No segundo capítulo faz-se referência à pertinência do tema realçando a problemática da divulgação de microdados e o enquadramento legal. No capítulo seguinte é analisada a questão da confidencialidade, o seu tratamento, o *trade-off* existente entre o risco de divulgação e a utilidade (ou perda de informação) e as características das bases de dados de empresas. Em seguida, o capítulo quatro aborda a questão da geração sintética, começando com uma reflexão sobre a produção de ficheiros sintéticos, a opção entre ficheiros totalmente ou parcialmente sintéticos, terminando com os principais métodos de geração sintética de microdados que utilizam a *Imputação Múltipla*. A análise da base de dados, a descrição detalhada das metodologias e a avaliação dos resultados são expostos no capítulo quinto. E, por fim, encerra-se o trabalho com as principais conclusões e apontam-se algumas linhas futuras de investigação (capítulo sexto).

2 – DIVULGAÇÃO DE INFORMAÇÃO

"Open access to official statistics provides the citizen with more than a picture of society. It offers a window on the work and performance of government itself, showing the scale of government activity in every area of public policy and allowing the impact of public policies and action to be assessed " , UK White Paper on Open Government (1993)[4].

2.1 – DIVULGAÇÃO DE MICRODADOS

A questão da confidencialidade e da proteção de dados está intimamente relacionada com os avanços tecnológicos e a proliferação de meios de divulgação da informação. A Internet e o desenvolvimento tecnológico permitiram um melhor acesso à informação e um aumento da capacidade de processamento e armazenagem dos servidores com uma quebra abrupta dos custos associados. Estes desenvolvimentos implicaram por um lado um aumento da procura de informação cada vez mais detalhada e por outro um incremento do risco de identificação dos titulares dos registos.

A informação pode ser classificada em duas categorias: microdados e dados tabulares ou macrodados. Os ***microdados*** são registos que contêm informação de respondentes individuais associados a uma pessoa, família ou empresa (Hundepool *et al*., 2010). As variáveis comumente encontradas em ficheiros de microdados de registos pessoais são o género, a idade, a ocupação ou o lugar de residência. No caso de bases de dados de microdados de empresas encontram-se referências à atividade económica, setor, número de empregados ou volume de negócios.

Por ***macrodados*** entende-se a informação agregada disponibilizada sobre a forma de tabelas que contêm informações sobre um coletivo, cujos membros têm características comuns. Geralmente, as tabelas de macrodados agregam informação por dispersão geográfica, atividade económica ou outras características comuns aos indivíduos. Segundo Hundepool *et al*. (2010), a publicação de macrodados ou dados agregados foi durante muitos anos a estratégia utilizada pelos institutos de estatística para a divulgação da informação, uma vez que se salvaguardava, em princípio, a informação do individuo ao qual os dados se referiam.

[4] "*O livre acesso às estatísticas oficiais fornecerá ao cidadão mais do que uma fotografia da sociedade. Ele oferece uma janela para o trabalho e o desempenho do governo, mostrando a dimensão da sua intervenção em todas as áreas de políticas públicas, permitindo avaliar o seu impacto*" (tradução livre do autor).

O presente trabalho está focado em bases de dados de microdados, o que não significa que não existam problemas de confidencialidade relacionados com macrodados ou dados tabulares. Embora seja mais difícil identificar um indivíduo num agregado, também existe o risco de identificação dos seus elementos. Por exemplo, quando um agregado resulta de um número reduzido de registos ou no caso de bases de dados dinâmicas, que ao permitirem consultas por diferentes classes pode originar a identificação dos indivíduos através do cruzamento da informação.

No âmbito da divulgação da informação, Ducan *et al.* (2001) esclarecem que uma organização que produz informação estatística não pode 'erguer um muro' em torno da mesma, porque detém um mandato de divulgação. Os autores referem que em sociedades democráticas e livres os clientes da informação são variados. As autoridades estatísticas não facultam apenas informação para os decisores políticos, mas também para os indivíduos, empresas, organizações não-governamentais, meios de comunicação social e grupos de interesse, de modo a promoverem o debate de ideias descentralizado. Drechsler e Reiter (2011) simplificam o debate argumentando que os cidadãos pagam para a recolha de informação através dos impostos, logo devem ter o direito de acesso.

Para Winkler (2005) a própria tecnologia e a evolução da sociedade do conhecimento pressionam para a divulgação de microdados. Segundo este autor, devido ao maior poder de computação, sofisticação das soluções informáticas e incremento da capacidade dos utilizadores para desenvolver os seus próprios programas, os investigadores preferem analisar microdados. Estes investigadores (utilizadores de dados) não ficam satisfeitos com o uso de informação agregada produzida pelos órgãos de estatística (fornecedores de dados). Os utilizadores de dados perceberam que o acesso adequado a microdados permite explorar as bases de dados e encontrar novas questões que estão para além da competência e dos recursos dos detentores da informação.

Segundo Lane (2003) os benefícios decorrentes do acesso a bases de dados de microdados são múltiplos: i) permitem aos decisores políticos colocarem questões complexas e obterem respostas; ii) os analistas e investigadores passam a poder calcular efeitos marginais em vez de optarem pelos tradicionais efeitos médios; iii) promovem a salvaguarda científica, uma vez que os resultados obtidos em estudos podem ser replicados; iv) levam à criação de um ciclo virtuoso de conhecimento para a entidade que divulga a informação, porque o uso de dados, inevitavelmente, revela a sua

qualidade, permitindo identificar anomalias de processamento e novas necessidades de informação e v) beneficia o produtor das estatísticas na medida que dá visibilidade ao seu trabalho, permitindo um incremento da sua legitimidade através do retorno do investimento efetuado na produção da informação.

Diferentes autores concordam que a divulgação de bases de dados de microdados contribuem para o desenvolvimento económico e social, uma vez que dota a comunidade de ferramentas importantes não só no âmbito de trabalhos científicos e académicos, mas também para o exercício da cidadania.

No entanto, para além da divulgação, a questão da confidencialidade aparece também como sendo fundamental, nomeadamente para garantir a qualidade da informação. Para Ducan *et al.* (2001) a confidencialidade é indispensável por questões éticas no âmbito da autonomia dos respondentes e por questões práticas relacionadas com a qualidade e quantidade da informação. Por exemplo, as respostas num inquérito sobre a satisfação no local de trabalho só vão ser sinceras e completas se for garantida a confidencialidade dos respondentes.

A confidencialidade numa organização que produz informação está em constante risco, pois pode ser comprometida por um intruso. Um intruso é alguém que tendo um acesso legítimo à informação tem objetivos e métodos que não estão alinhados com os da missão do produtor das estatísticas[5]. Outros termos utilizados na literatura para intruso são bisbilhoteiro de dados, espião ou atacante[6]. O comprometimento da confidencialidade da informação ocorre quando a informação divulgada permite obter informação ilegítima sobre o respondente individual.

Assegurar a confidencialidade da informação não é uma tarefa elementar de simples eliminação dos designados identificadores diretos como o nome, o número do cartão do cidadão ou a morada, mas obviamente que este é o primeiro passo para minimizar o risco de identificação. Segundo Winkler (2005), a principal razão pela qual a remoção dos elementos identificadores não é suficiente para garantir o anonimato dos respondentes é a facilidade de acesso a bases de dados complementares de nomes e de registos, como por exemplo informação de marketing ou registos de voto. Na presença de informação complementar, o intruso pode aplicar métodos de procura e ligação de

[5] Ducan *et al.* (2001) explicam que um *hacker* que tenta entrar no sistema de um computador para obter informações confidenciais não se enquadra no conceito de intruso.
[6] Termos traduzidos do inglês: *data snooper*, *data spy* ou *attacker*. Tal como na literatura, neste trabalho estes termos são considerados sinónimos. Por questões de coerência optou-se por utilizar o termo intruso.

registos e identificar os respondentes de uma base de dados sem elementos identificadores. O autor faz referência ao trabalho de Bethlehem *et al*. (1990) que perante informação dos serviços tributários holandeses conseguiu identificar alguns indivíduos de uma base de dados anonimizada. Sweeney (2002) conseguiu identificar os registos médicos do governador de Massachusetts numa base de dados pública disponibilizada para investigação sem identificadores diretos, utilizando neste processo uma base de dados de registos de voto que lhe custou 20 dólares[7].

Para a divulgação de bases de dados seguras, a salvo de um intruso, os fornecedores da informação têm que aplicar técnicas de controlo da divulgação estatística que ultrapassem a simples ocultação de campos identificadores diretos. As técnicas de controlo da divulgação estatística (CDE)[8] são definidas como o conjunto de métodos para reduzir o risco de divulgação de informação de indivíduos, empresas ou outras organizações. Estas técnicas têm por objetivo minimizar o risco de divulgação para um nível aceitável, com o propósito de divulgar o máximo de informação possível (Hundepool *et al*., 2010).

2.2 – ENQUADRAMENTO LEGAL

Segundo Hundepool *et al*. (2010), até aos finais da década de 80, os microdados raramente eram transmitidos ao Eurostat pelos Institutos Nacionais de Estatísticas devido aos diferentes enquadramentos legais sobre confidencialidade dos países europeus. Dadas as dificuldades, em junho de 1990, surge o regulamento 1588/90 elaborado e aprovado pelo Conselho Europeu sobre a transmissão de dados confidenciais. Este regulamento autorizava os Institutos Nacionais de Estatística a transmitir informação considerada confidencial ao organismo europeu, enquanto este se obrigava a tomar as medidas necessárias para proteger essa informação. Foi assim dado o primeiro passo na transmissão e proteção de informação confidencial para o exterior das autoridades estatísticas nacionais.

Em janeiro de 1994, no âmbito do Comité de Confidencialidade Estatística, as medidas deste regulamento foram formalmente adotadas pelos Estados Membros. Este comité

[7] Segundo Sweeney (2002), o governador Weld vivia em Cambridge Massachusetts. De acordo com os registos de voto do local de residência, seis pessoas tinham a mesma data de nascimento, apenas três eram homens e apenas uma tinha residência em determinado ZIP code.
[8] Com alguma frequência utiliza-se a sigla inglesa SDC – *Statistical Disclosure Control* (*SDC*).

passou a reunir regularmente para discutir o enquadramento legal da divulgação estatística de microdados e dados agregados (macrodados).

Mais recentemente, em fevereiro de 2005, o Código de Conduta para as Estatísticas Europeias (CCEE) foi adotado pelo Comité do Programa Estatístico da União Europeia e consequentemente pelos Estados Membros. O CCEE é baseado em quinze princípios, sendo o princípio cinco respeitante à confidencialidade estatística*:*

"A privacidade dos fornecedores de dados (famílias, empresas, órgãos da administração pública e outros intervenientes), a confidencialidade das informações que prestam e a sua utilização exclusivamente para fins estatísticos devem ser absolutamente garantidas".

Em 2011, o diploma foi alvo de revisão com o objetivo de reforçar os aspetos relacionados com a gestão da qualidade, fortalecer a independência profissional e melhorar os aspetos associados à apropriação de dados administrativos para efeitos estatísticos.

Em Portugal, a Lei do Sistema Estatístico Nacional (SEN) – Lei n.º 22/2008 – atualmente em revisão, estabelece as bases gerais do sistema estatístico nacional. No âmbito deste trabalho destaca-se o artigo sexto referente ao segredo estatístico. Segundo este artigo, os dados estatísticos individuais recolhidos pelas autoridades estatísticas são de natureza confidencial e constituem segredo profissional, respondendo os funcionários criminalmente pela violação do segredo estatístico. Apesar das preocupações claras do legislador com a confidencialidade, existe a possibilidade dos dados estatísticos poderem ser cedidos em três situações distintas:

- Divulgação de informação anonimizada dos dados estatísticos individuais sobre pessoas singulares, com autorização do respetivo titular ou após autorização do Conselho Superior de Estatística e, neste caso, será apenas quando estejam em causa razões de saúde pública (número cinco);
- No que concerne às pessoas coletivas, o procedimento é idêntico, embora a informação a ceder não seja anonimizada e para além de razões de saúde pública o legislador acrescentou o planeamento e coordenação económica, relações económicas externas ou proteção do ambiente (número seis);
- Para fins científicos, exigindo para isso a divulgação sob a forma anonimizada e apenas mediante o estabelecimento de acordo entre a autoridade estatística

cedente e a entidade solicitante, no qual são definidas as medidas técnicas e organizativas necessárias para assegurar a proteção dos dados de modo a evitar qualquer risco de divulgação ilícita ou de utilização para outros fins aquando da divulgação dos resultados (número sete).

Para além da Lei do Sistema Estatístico Nacional, o princípio do segredo estatístico está consagrado no ornamento jurídico nacional nos seguintes documentos: i) Constituição da República Portuguesa na Lei nº 67/98, de 26 de Outubro, Lei de Proteção de Dados Pessoais; ii) Regulamento (CE) 223/2009 do Parlamento Europeu e do Conselho, de 11 de Março de 2009 (Regulamento das Estatísticas Europeias) e iii) Código de Conduta das Estatísticas Europeias, designadamente o princípio cinco que respeita à Confidencialidade Estatística acordada entre Estados-membros da EU.

3 – CONFIDENCIALIDADE EM MICRODADOS

"The challenge of balancing the competing objectives of allowing statistical analysis of confidential data and maintaining confidentiality is of great interest to national statistical agencies and other data custodians seeking to make their data available for research", Lee *et al.* (2013)[9].

Como referido no ponto anterior, está-se na presença de um ficheiro de microdados quando os registos contêm informação de respondentes individuais (pessoas ou empresas) relativos a um conjunto de variáveis. Quando esta informação é disponibilizada para investigação a primeira ação é a remoção de identificadores diretos óbvios como números únicos, nomes ou moradas. No entanto, nem sempre estas medidas são suficientes para proteger a identidade dos respondentes (Matthews e Harel, 2011).

3.1 – CLASSIFICAÇÃO DAS VARIÁVEIS

Na literatura, de uma forma geral, as variáveis são classificadas no que concerne à confidencialidade em quatro categorias distintas: variáveis identificadoras diretas, variáveis chave, variáveis sensíveis e variáveis não confidenciais (Templ *et al.*, 2014; Hundepool *et al.*, 2010; Domingo-Ferrer e Torra, 2005 e Sweeney, 2002).

Os métodos de CDE (Controlo e Divulgação Estatística) são habitualmente aplicados a variáveis identificadoras diretas, variáveis chave e variáveis sensíveis. As ***variáveis identificadoras diretas*** são aquelas que de forma não ambígua identificam respondentes individuais, tais como o número do cartão do cidadão, número de identificação fiscal, nomes e moradas. Estas variáveis devem ser imediatamente removidas ou substituídas num processo CDE, pois habitualmente não têm impacto na utilidade da informação. As ***variáveis chave*** são aquelas que combinadas podem ser ligadas a informações externas permitindo a identificação dos respondentes. As variáveis chave também são designadas na literatura por identificadores implícitos ou variáveis quase-identificadoras (Hundepool *et al.*, 2010 e Sweeney, 2002).

[9] "*O desafio de conciliar os objetivos concorrentes de permitir a análise estatística dos dados confidenciais e a manutenção da confidencialidade é de grande interesse para os institutos nacionais de estatística e outros depositários de dados que procuram tornar os seus dados disponíveis para investigação*" (tradução livre do autor).

As ***variáveis sensíveis*** são aquelas que não devem ser associadas a qualquer respondente individual. A determinação de variáveis sensíveis é normalmente alvo de preocupações legais ou éticas. Por exemplo, variáveis com informação de registo criminal, médico ou relacionadas com o rendimento não devem ser associados a qualquer elemento da base de dados.

Uma variável pode ser ao mesmo tempo identificadora e sensível. Por exemplo, variáveis de rendimento podem ser combinadas com outras variáveis chave para identificar um respondente, mas a variável por si só é sensível, pelo que deve ser confidencial. Variáveis como a ocupação podem não ser sensíveis, mas podem ser combinadas com outras para identificar os respondentes.

Por fim, as ***variáveis não confidenciais*** são aquelas que não se enquadram em nenhuma das três categorias acima identificadas. Em princípio a sua divulgação não envolve qualquer risco.

Quanto à natureza, as variáveis são classificadas em qualitativas (nominais ou ordinais) ou quantitativas. As ***variáveis qualitativas*** são aquelas cuja escala de medida apenas indica a sua presença em categorias de classificação discretas exaustivas e mutuamente exclusivas. Estas variáveis podem ser medidas numa escala nominal (por exemplo o género) quando não é possível estabelecer à partida um qualquer tipo de ordenação, ou ordinal (por exemplo habilitações literárias) caso seja possível estabelecer uma ordem entre as classes. As ***variáveis quantitativas*** são numéricas de tal forma que é possível efetuar operações aritméticas com elas. Uma variável quantitativa não precisa obrigatoriamente de assumir uma infinidade de valores contínuos, como por exemplo a idade (Domingo-Ferrer e Torra, 2005).

3.2 – FORMAS DE DISPONIBILIZAÇÃO DA INFORMAÇÃO

De uma forma geral existe três formas de disponibilização de ficheiros de microdados pelas entidades estatísticas: i) ficheiros de uso público (*Public Use File* ou *PUF*); ii) ficheiros de uso científico (*Scientific Use File* ou *SUF*) e iii) ficheiros de utilização segura (*Secure Use Files*). Os ficheiros de uso público são ficheiros de dados preparados com o objetivo de os tornar de acesso livre e sem problemas de confidencialidade, ou seja, o risco de identificação dos respondentes é nulo ou quase nulo. Os ficheiros de uso ciêntifico são mais detalhados, mas de uso mais restrito, sendo normalmentente enviados ao utilizador. O acesso a este tipo de ficheiros é acompanhado

por uma licença ou uma declaração de confidencialidade e a sua utilização encontra-se regulamentada[10] (Hundepool *et al.*, 2010).

Adicionalmente, algumas entidades fornecem acessos a microdados (ficheiros de utilização segura) em laboratórios de dados, centros de pesquisa ou através de acessos remotos. Os laboratórios de dados ou centros de pesquisa segura (*safe centers*) permitem que utilizadores aprovados tenham acesso em determinado local à informação confidencial. Apesar de terem acesso à informação e a poderem manipular, os utilizadores estão legalmente proibidos de a divulgarem e estão sujeitos a restrições no que concerne à publicação de resultados. Normalmente, o acesso é efetuado em computadores locais com *software* que permite a análise da informação, estando vedada a possibilidade de enviar informação para o exterior. Ao investigador apenas são disponibilizados os resultados obtidos após um rigoroso controlo denominado por *output checking*.

Em termos de acessos remotos, algumas entidades optam pela execução à distância, que consiste em fornecer ao investigador uma descrição detalhada dos microdados, este analisa a informação e envia um programa num formato especificado (*script*) que é executado pela entidade estatística. Os resultados são verificados e enviados para o investigador.

Outro tipo de acesso remoto é a utilização de senhas *(passwords*) ou outros mecanismos de segurança para permitir o acesso a um servidor onde os programas e as bases de dados estão disponíveis. Os investigadores conseguem submeter códigos de análise e em alguns casos ver virtualmente a informação nos seus computadores pessoais. A confidencialidade é garantia através da combinação de um conjunto de procedimentos: i) visualização parcial da informação; ii) modificação da informação; iii) validação automática ou manual dos resultados e iv) através de claúsulas decorrentes de contratos estabelecidos entre o utilizador e a entidade detentora da informação.

Raghunathan *et al.* (2003) esclarecem que normalmente o acesso à informação para investigação obedece a um conjunto de procedimentos administrativos. O investigador tem que submeter uma proposta detalhada do trabalho elencando o conjunto de variáveis, a racionalidade da análise e muitas vezes os moldes em que o projeto de

[10] O Regulamento (UE) n.º 557/2013 da Comissão de 17 de junho, aplica o regulamento (CE) n.º 223/2009 do Parlamento Europeu e do Conselho relativo às Estatísticas Europeias, no que diz respeito aos dados estatísticos confidenciais para fins científicos.

investigação é financiado. As propostas são avaliadas por um *comité* e uma vez aceite é disponibilizado o acesso apenas para as variáveis solicitadas. Pedidos de variáveis adicionais implicam a repetição de todo o processo burocrático.

Embora em qualquer uma das formas de disponibilização da informação (*PUF* ou *SUF*) a questão da confidencialidade deve ser salvaguardada, ela é especialmente pertinente nos ficheiros de uso público devido à impossibilidade de se identificar, pelo menos de forma imediata, o intruso e não existir uma relação contratual e legal que obrigue o utilizador da informação a manter a confidencialidade.

Acerca dos ficheiros de uso público, Winkler (2005) estabelece uma distinção entre ficheiros analiticamente válidos de forma comprovada, analiticamente válidos e ficheios analiticamente interessantes. Os ficheiros analiticamente válidos de forma comprovada são aqueles em que os utilizadores conseguem obter informação de tal forma detalhada que os resultados obtidos são consistentes com os obtidos com a base de dados original. Um ficheiro é analiticamente válido se o utilizador conseguir reproduzir algumas análises estatísticas do ficheiro original de microdados confidencial[11]. Por fim, ficheiros analiticamente interessantes são aqueles que possuem um número suficiente de variáveis que permitam uma análise séria por parte dos investigadores[12].

Segundo o autor, facultar um ficheiro de uso público que satisfaça grandes necessidades analíticas num conjunto alargado de variáveis e ao mesmo tempo seja confidencial é muito difícil ou até mesmo impossível. Tendo em conta a dificuldade em facultar ficheiros analiticamente válidos de forma comprovada respeitando a confidencialidade da informação, de uma forma geral, os métodos propostos na literatura têm por objetivo a criação de ficheiros analiticamente válidos.

3.3 – RISCO E UTILIDADE

Um dos tópicos mais pertinentes quando se fala na disponibilização de ficheiros de dados é o balanceamento entre risco e utilidade. Os fornecedores de informação devem gerir a tensão entre o acesso à informação e a manutenção da confidencialidade. Ducan *et al.* (2001) esclarecem que a resolução deste problema requer a adoção de estratégias

[11] Quando se aborda as medidas de utilidade é explicada a métrica utilizada pelo autor para definir ficheiros analiticamente validos.
[12] Segundo os autores, possuir informação referente a pelo menos cinco variáveis discretas e seis contínuas.

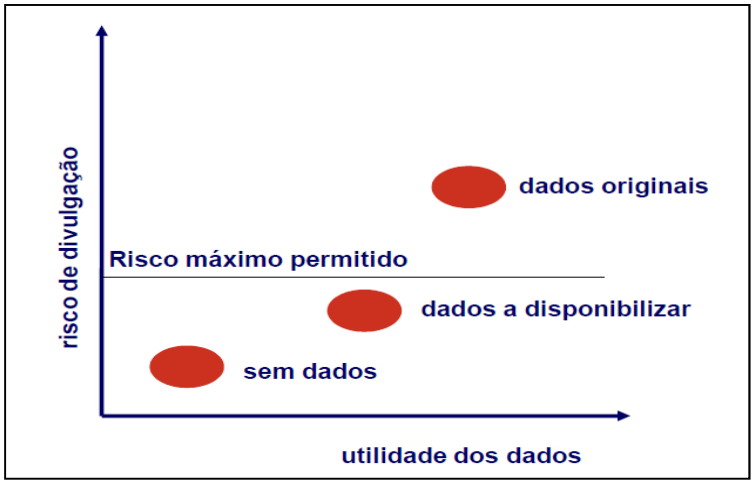
que permitam ter ao mesmo tempo elevada utilidade (informação analiticamente válida) e baixo risco (informação segura).

O risco de divulgação, também definido por alguns autores como risco de identificação, é o risco de um intruso obter informação desconhecida sobre um respondente utilizando dados disponibilizados ao público. Templ *et al.* (2014) com base no trabalho de Lambert (1993) descrevem três tipos de risco de divulgação: i) risco de identidade; ii) risco de atributo e iii) risco de inferência. O ***risco de identidade*** ocorre quando o intruso associa um indivíduo a determinado registo ficando com acesso a informação por ele desconhecida. O ***risco de atributo*** existe quando é possível obter novas características partindo do ficheiro disponibilizado. Por exemplo, se um hospital publica informação em que todos os pacientes masculinos em determinada faixa etária tem determinada doença, fica-se a saber a condição médica dos pacientes sem ter que identificar os indivíduos. Por último, o ***risco de inferência*** ocorre quando se consegue obter de forma mais exata informação que de outra forma não se obteria. Por exemplo, inferir o rendimento com base na informação disponibilizada numa base de dados sobre o consumo.

Como já foi exposto anteriormente, a ocultação ou eliminação de identificadores diretos por si só não garante a confidencialidade da informação, pelo que é necessário aplicar técnicas de CDE. A aplicação destas técnicas resulta, na maior parte das vezes, na perda de informação, que se traduz em perdas de utilidade dos ficheiros. O desafio para os fornecedores de informação consiste em aplicar uma ou várias técnicas aos ficheiros de informação de modo a reduzir o risco de identificação ou obtenção de informação confidencial com o mínimo de perda de utilidade. Para ilustrar o *trade off* entre o risco de divulgação e a utilidade dos dados, na figura 1 é apresentada uma ilustração[13].

[13] Fonte: Adaptado de Hundepool *et al.* (2010).

Figura 1 – Relação entre risco e utilidade



No canto inferior esquerdo o risco de divulgação é reduzido, mas a utilidade é baixa ou nula no caso de não ser disponibilizada qualquer informação. No canto superior direito existe grande utilidade da informação mas com risco elevado. A entidade que divulga a informação deve definir o risco máximo permitido, considerando as suas políticas, a legislação e os códigos de ética. O objetivo é disponibilizar ficheiros com a máxima utilidade possível com um risco abaixo do limiar definido.

Este *trade off* pode ser formulado como um problema de otimização. Sendo *D* a base de dados, *f(D)* o resultado da aplicação de técnicas de CDE, *R*[*f(D)*] a medida do risco de divulgação, *U* [*f(D)*] a medida de utilidade e ε o risco máximo aceitável. O desafio é encontrar a técnica ou técnicas de CDE, nesta caso *f(.)*, dado um *D* e que satisfaça a seguinte condição (Skinner, 2009):

$$Maximização\ U[f(D)],\ sujeito\ a\ \ R[f(D)] < \varepsilon \qquad (1)$$

Dadas as dificuldades em especificar *R(.)*, *U(.)* e o valor especifico de ε de forma exata, o problema de otimização serve apenas conceptualmente. Na prática, diferentes métodos de CDE podem ser comparados e avaliados utilizando diferentes métricas de utilidade e risco.

3.3.1 – Medidas de Risco

A medição do risco de divulgação é feita utilizando cenários de divulgação estatística de modo a tentar identificar como é que o intruso pode obter informação ilegítima sobre os respondentes. Grande parte das medidas utilizadas para medir o risco de divulgação é baseada nas variáveis chave e sensíveis (Matwin *et al.*, 2015).

O risco de divulgação das variáveis qualitativas assenta na ideia de que os registos com combinações únicas de variáveis chave têm maior probabilidade de serem identificados. Sendo f_k a frequência com que determinada combinação de variáveis *k* ocorre na base de dados e F_k o número de unidades da população com o mesmo padrão *k*, um registo único (valor $f_k = 1$) tem uma probabilidade de ser descoberto por um intruso de $1/F_k$, assumindo que ele desconhece os indivíduos que fazem parte da base de dados. Se $F_k = 1$, ou seja, estivermos perante uma população única ou o intruso conhecer os indivíduos presentes na amostra ele consegue identificar o respondente (Templ *et al.*, 2014).

Na prática, apenas se observa a frequência da amostra f_k. Para estimar o risco de divulgação tem-se em consideração estimativas de F_k, assumindo que a população segue uma distribuição binomial negativa ou uma distribuição de Poisson (Hundepool *et al.*, 2010).

3.3.1.1 - *k-anonimato*

Uma forma de proteger a confidencialidade de uma base de dados é garantir que cada padrão de variáveis chave está representado por *k* registos na amostra. Esta abordagem é designada na literatura por *k-anonimato*[14]. Uma prática comum é estabelecer o valor (3) três para *k*, garantido que pelo menos três registos da amostra possuem determinado padrão (utilizando a notação anterior implica que $f_k \geq 3$). O mínimo de três registos com o mesmo padrão deriva do facto de para um valor de dois um intruso conseguindo identificar um registo, por exemplo o dele próprio, conseguiria identificar o do outro respondente por diferença (Hundepool *et al.*, 2010). Mesmo na presença de pelo menos três registos com o mesmo padrão o intruso pode conseguir identificar os respondentes. Os trabalhos de Sweeney (2002)[15] e Templ *et al.* (2014) apresentam diferentes exemplos em que um intruso pode obter informação confidencial com um valor de *k=3*. Normalmente estes cenários são construídos com bases de dados de poucos registos, em que o intruso sabe quem são os respondentes e tem informação específica sobre eles, por exemplo determinado doente tem pouca probabilidade de ter uma doença específica[16], ou as variáveis sensíveis apresentam o mesmo valor.

[14] Traduzido do inglês *k-anonymity*.
[15] Este trabalho é um dos mais exaustivos na análise do k-anonimato.
[16] Exemplo de Sweeny (2002): porque é asiático, tem pouca probabilidade de sofrer de doença cardíaca, logo tem a outra patologia.

3.3.1.2 - *l-diversidade*

Para suprir a limitação apresentada na secção anterior, habitualmente junta-se ao princípio do *k-anonimato* o da *l-diversidade*[17] introduzido por Machanavajjhala *et al.* (2007)[18]. Um grupo de observações que evidenciam o mesmo padrão de variáveis chave tem *l-diversidade* se contém pelo menos *l* valores para as variáveis sensíveis. Segundo os autores, as variáveis sensíveis devem ter pelo menos *l* valores diferentes para cada grupo de variáveis chave. Um exemplo deste risco é uma base de dados de informação clínica em que todos os pacientes do mesmo género, em determinada facha etária e residentes na mesma zona (três variáveis chave comuns) apresentam a mesma patologia (variável sensível com o mesmo valor). Esta violação à privacidade da informação é apelidada pelos autores de ataque de homogeneidade (*homogeneity attack*). Outra forma de violação enunciada neste trabalho é o designado ataque devido ao conhecimento sobre os respondentes, que consiste em utilizar conhecimentos prévios para eliminar valores das variáveis sensíveis, aumentando a certeza acerca do verdadeiro valor da variável. Uma vez que é difícil determinar o conhecimento prévio do intruso sobre os respondentes, também é complicado quantificar este risco. No entanto, pode-se afirmar que quanto maior for o valor *l* menor será a probabilidade de identificar um respondente.

Matwin *et al.* (2015) apresentam outros princípios similares em termos conceptuais existentes na literatura: o *anonimato (α,k)* e a *entropia*. Uma base de dados satisfaz o princípio do *anonimato (α,k)* se cumpre a regra do *k-anonimato* e a probabilidade de inferir uma variável sensível em cada combinação de variáveis é de pelo menos α. Outra variante é garantir que a entropia de cada combinação de variáveis chave deve ser pelo menos $log\,(l)$[19]:

$$\sum_{s \in S} P(EC, s) \log(p(EC, s)) \geq log\ (l) \qquad (2)$$

em que S é o conjunto de valores que as variáveis sensíveis podem assumir e $P(EC, s)$ é a fração de registos de cada combinação de variáveis chave (*EC*) que apresentam valores *s* para a variável sensível. Por detrás da noção de entropia está o facto de à

[17] Traduzido do inglês *l-diversity*.
[18] Segundo Matwin *et al.* (2015), o principio da *l-diversity* é semelhante ao da *p-sensitivity* enunciado por Truta e Vinay (2006).
[19] A sigla EC resulta do termo inglês *equivalence class* e representa uma combinação de variáveis chave.

medida que os valores das variáveis sensíveis se tornam mais uniformes o indicador diminui.

Uma das críticas ao princípio da *l-diversidade* é que não garante a confidencialidade quando as variáveis sensíveis são similares (ataque de similaridade) ou a distribuição das variáveis é assimétrica (ataque de assimetria). O ataque à similaridade acontece quando os valores são diferentes, mas semanticamente similares. Por exemplo, um valor de *l=3* para uma variável que discrimina entre três doenças pulmonares. Um intruso sabe que o paciente com aquela combinação de variáveis tem uma doença pulmonar, embora não consiga identificar a doença específica. Estamos perante um ataque de assimetria, por exemplo, quando numa amostra a probabilidade de ter uma doença é muito diferente da registada na população, levando à identificação da doença de um paciente com uma probabilidade elevada.

3.3.1.3 - *t-proximinade*

Para suprir as limitações expostas da aplicação dos princípios do *k-anonimato* e da *l-diversidade* foi proposto o princípio da *t-proximinade* por Li *et al.* (2007). Uma determinada classe de variáveis chave tem *t-proximinade* se a distância entre a distribuição da variável sensível e a distribuição de toda a base de dados for no máximo *t*. Segundo Matwin *et al.* (2015), ao se garantir uma distribuição semelhante entre as variáveis sensíveis e toda a base de dados assegura-se que não sejam possíveis ataques de assimetria e é pouco provável a existência de similaridade semântica. No entanto, a aplicação deste princípio prejudica a correlação entre as variáveis chave, ou seja, tem consequências sobre a utilidade da informação (Matwin *et al.*, 2015).

3.3.1.4 - Métodos baseados na ligação de registos (*record linkage)*

Assumindo que o intruso tem acesso a uma base de dados que foi alvo de perturbação antes de ser disponibilizada, assim como a informação externa sobre os respondentes, ele pode tentar ligar as bases de dados utilizando as variáveis comuns dos dois ficheiros. Existem diferentes abordagens na literatura: i) modelos baseados nas distâncias (Pagliuca e Seri, 1999 e Mateo-Sanz *et al.*, 2004a); ii) modelos probabilísticos (Domingo-Ferrer e Tora, 2001) e iii) modelos com base em intervalos (Mateo-Sanz *et al.*, 2004a).

Na sua essência, os modelos baseados na ligação dos registos (*record linkage*) tentam entender de que forma bases de dados de registos protegidos podem ser corretamente ligadas a outras complementares revelando a identidade dos respondentes.

3.3.1.5 - Risco em bases de dados hierárquicas e risco global

Algumas bases de dados de microdados apresentam dados hierarquizados ou estruturas de múltiplos níveis. Por exemplo, os indivíduos são classificados em grupos, pelo que um intruso ao identificar um dos elementos pode obter informação de outros membros. Segundo Templ *et al.* (2014) é necessário ter em conta a estrutura hierárquica da base de dados na medição do risco de divulgação. Segundo os autores, o risco de divulgação para um membro de um grupo é maior ou igual do que o risco de identificação de pelo menos um dos seus elementos.

Adicionalmente à medição do risco ao nível do registo, a obtenção de uma medida de risco global é proposta em alguns trabalhos. A forma mais intuitiva é calcular o número esperado de registos em risco na base de dados, utilizando a soma do risco individual de cada combinação de variáveis chave. Outra abordagem é identificar o número de registos com um risco superior a um limiar, permitindo inferir sobre o risco global e identificar valores extremos na base de dados (Templ *et al.*, 2014). Outra forma proposta na literatura é a utilização de modelos log-lineares para estimar o risco global, utilizando as interações das variáveis chave (Skinner e Holmes, 1998).

3.3.2 – Medidas de Utilidade

Após a aplicação de técnicas de CDE a uma base de dados é necessário avaliar os impactos na utilidade ou como comumente é referido na literatura a perda de informação. A ideia é calcular indicadores que permitam inferir o quanto a base de dados protegida difere da original. Embora seja consensual avaliar o resultado de um processo de proteção dos dados, o problema surge na classificação das variações das propriedades estatísticas em termos de utilidade e na forma como se calcula essas diferenças entre a base de dados original e a informação protegida (Yancey *et al.*, 2002).

Segundo Templ *et al.* (2014), existem duas formas complementares de averiguar a perda de informação: i) aplicação de medidas diretas de distâncias entre a base de dados

original e os dados protegidos e ii) uma análise com base em valores de referência, comparando as estatísticas resultantes de cada um dos ficheiros.

3.3.2.1 - Medidas diretas

As medidas diretas de perda de informação são baseadas nas distâncias clássicas entre a base de dados original e a informação protegida. Segundo Domingo-Ferrer e Tora (2001), ocorre uma perda de informação residual se a estrutura analítica da base de dados protegida é similar à estrutura de dados originais.

O trabalho de Winkler (2005) define que uma base de dados analiticamente válida tem que preservar algumas caracteristicas: i) as médias e as covariância num pequeno conjunto de subdomínios; ii) os valores marginais na agregação de alguns dados e iii) pelo menos uma característica da distribuição inalterada.

Para avaliar a eventual perda de informação no ficheiro protegido para as ***variáveis quantitativas*** é necessário determinar alguns elementos estatísticos na comparação das bases de dados. Para as ***variáveis qualitativas*** a análise tem por base, essencialmente, o cálculo de frequências.

Variáveis qualitativas

Em termos de variáveis qualitativas, a perda de informação pode ser medida através da comparação direta das variáveis ou utilizando tabelas de contingência (Domingo-Ferrer e Torra, 2001). A medida de distância para a variável *V* é dada por:

$$dv(c, c') = \begin{cases} 0, se\ c = c' \\ 1, se\ c \neq c' \end{cases} \quad (3)$$

em que *c* corresponde ao valor assumido pela variável no conjunto de dados originais e *c'* ao valor correspondente no ficheiro protegido.

Uma forma alternativa é a utilização de tabelas de contingência. Considerando duas bases de dados *F* e *G* (a original e a protegida, respetivamente) e as suas tabelas de contingência de *t-dimensão* para t ≤ K, pode-se definir a perda de informação para um subconjunto de variáveis *W* da seguinte forma:

$$\sum_{\substack{\{V_{ji}\dots V_{jt},\}\subseteq W \\ |\{V_{ji}\dots V_{jt},\}\leq K|}} \sum_{i_1,\dots,i_t} \left| x^F_{i_1,\dots,i_t} - x^G_{i_1,\dots,i_t} \right| \qquad (4)$$

em que $x^{ficheiro}_{índice}$ é o valor da tabela de contingência do ficheiro na posição dada pelo índice. Esta medida consiste em calcular a diferença das frequências para um conjunto de combinações de variáveis qualitativas entre o ficheiro sintético e o original. Aplicando o estabelecido pelos autores pode-se identificar o número de combinações que foram criadas ou que desapareceram em relação ao ficheiro original.

Variáveis quantitativas

Domingo-Ferrer e Tora (2001) referem um conjunto de métricas baseadas na aplicação de medidas diretas de distâncias entre a base de dados original e os dados protegidos. Seja *X* um ficheiro de microdados, com *n* indivíduos e *p* variáveis quantitativas, para avaliar a eventual perda de informação no ficheiro protegido *X'* é necessário calcular os seguintes elementos (Domingo-Ferrer e Tora, 2001):

- Matrizes de variâncias e covariâncias *V* (em *X*) e *V'* (em *X'*);
- Matrizes de correlações *R* e *R'*;
- Matrizes de correlações *RF* e *RF'* entre *p* variáveis e *p* fatores PC_1, …, PC_p obtidos através da análise de componentes principais;
- Percentagem de cada uma das variáveis que é explicada pelas componentes principais (matriz *C* e *C'*);
- Matriz dos coeficientes dos fatores (*F* e *F'*, onde *F* contem os fatores que devem multiplicar cada variável em *X*, para obter a sua projeção na componente principal.

Uma vez que não é possível sintetizar a informação num único indicador, os autores propõem diferentes formas para avaliar a informação perdida através comparação entre as matrizes obtidas pelos dados originais *X, V, R, RF, C* e *F* e as matrizes obtidas pelos dados protegidos *X'; V'; R'; RF'; C'* e *F'*. A avaliação pode ser efetuada de três formas distintas: i) Erro Quadrático Médio (soma do quadrado das diferenças das componentes entre os pares de matrizes, dividida pelo número de células em cada matriz); ii) Erro Absoluto Médio (soma absoluta das diferenças das componentes entre os pares das

matrizes, dividida pelo número de células em cada matriz) e iii) Variação Média (soma absoluta da variação percentual das componentes da matriz calculada nos dados protegidos no que respeita às componentes da matriz calculada nos dados originais, dividida pelo número de células em cada matriz). No Anexo 1 são apresentadas as equações propostas pelos autores referentes a cada uma das métricas.

3.3.2.2 - Utilização de valores de referência

Esta abordagem consiste em determinar que tipo de análise é que pode ser realizada com a informação disponibilizada e identificar indicadores de referência para as variáveis (Templ *et al.* 2014).

Após a aplicação de técnicas de CDE ao ficheiro original e obter uma base de dados protegida o processo de identificação da perda de informação é o seguinte: i) selecionar um conjunto de indicadores de referência; ii) estimar o valor dos indicadores na base de dados originais e na protegida; iii) comparar as propriedades estatísticas, tais como, estimativas obtidas, variâncias ou intervalos de confiança para cada indicador e iv) averiguar a utilidade da base de dados de microdados protegida.

Segundo Templ *et al.* (2014), esta abordagem é normalmente aplicada a variáveis quantitativas e pode ser utilizada na avaliação de subconjuntos de informação. Se os indicadores calculados para a base de dados original e protegida forem significativamente diferentes, o processo de CDE deve ser reiniciado utilizando outra abordagem.

Em termos empíricos, ao nível da geração sintética de microdados (método a desenvolver mais à frente), encontram-se diferentes abordagens para avaliar a perda de informação:

- Análise exploratória de dados (comparação univariada e multivariada) – cálculo de frequências e médias (Reiter, 2005a; Reiter, 2005b; Drechsler e Reiter, 2010; Lee *et al.*, 2013);

- Análise inferencial[20] – regressão linear (Reiter, 2005a; Reiter, 2005b; Drechsler, 2009 e Drechsler e Reiter, 2010) e regressão logística (Reiter, 2005a; Loong *et al.*, 2013 e Raab *et al.*, 2015).

No âmbito do presente trabalho sugere-se a utilização de um teste não paramétrico para aferir a qualidade do ajustamento para uma avaliação univariada - teste do Qui-quadrado para as variáveis qualitativas e teste de Kolmogorov-Smirnov para as variáveis contínuas. A solução esperada é a não rejeição da hipótese nula, ou seja, não rejeitar a hipótese de que o ficheiro original e o sintético seguem a mesma distribuição.

3.4 – BASES DE DADOS DE EMPRESAS

Neste estudo a geração sintética de microdados é aplicada a uma base de dados de empresas. Este tipo de informação apresenta dificuldades adicionais em relação à informação de indivíduos, quer em termos de confidencialidade, quer em termos de geração variáveis sintéticas.

As bases de dados de empresas contêm geralmente informação sensível, por exemplo informação sobre a produção, os trabalhadores, os clientes e dados financeiros e fiscais. Embora o conceito de privacidade não seja tão valorizado como nos dados de pessoas singulares, as empresas têm interesse em controlar a difusão de informações sobre si mesmas e a entidade detentora da informação geralmente tem a obrigação de manter os dados confidenciais. Embora por razões de sensibilidade comercial, em vez da privacidade dos indivíduos ou das famílias, a entidade detentora da informação tem geralmente a responsabilidade de a proteger.

Segundo Drechsler (2009), as bases de dados de empresas apresentam algumas características que as distinguem das bases de dados de indivíduos:

- As bases de dados de empresas exibem um padrão característico na inclusão das empresas na amostra que aumentam os problemas de confidencialidade para as grandes empresas: enquanto uma grande empresa pela sua relevância estatística e peso no total da economia e setor é sempre incluída nas amostras, uma empresa média é frequentemente incluída e uma pequena empresa pode ser pontualmente selecionada;

[20] Os trabalhos de Drechsler (2009) e Loong *et al.* (2013) replicaram resultados obtidos por estudos que utilizaram a informação original utilizando a informação sintética, para provarem que as conclusões não se alteravam utilizando as bases de dados sintéticas.

- Caracterizam-se por poucas variáveis, a maioria são contínuas em vez de discretas e as distribuições de algumas variáveis são altamente enviesadas;
- Informação corporativa geralmente inclui empresas que são *outliers* em muitas variáveis, habitualmente as grandes empresas do setor.

Ao descrito por Drechsler (2009), pode-se acrescentar outros factos como a população das empresas ser substancialmente inferior à população de indivíduos (pessoas singulares) e que muita da informação acerca das empresas ser do domínio público, aumentando o risco de identificação dos registos nas bases de dados disponibilizadas.

Acerca do facto de grande parte da informação ser quantitativa, Hundepool *et al.* (2010) referem que apesar desta informação, na globalidade, não estar publicamente disponível, as variáveis quantitativas podem ser consideradas identificadoras. Na prática todas as unidades / registos são unicos no que diz respeito a um conjunto de variáveis quantitativas. Os autores afirmam que os cenários de divulgação devem considerar que os registos de empresas são mais vulneráveis do que os dos individuos (pessoas singulares), concluindo que é muito difícil considerar um cenário de divulgação adequado para estas bases de dados, pois considerar o pior cenário pode implicar uma protecção excessiva com muita perda de informação.

A solução para a divulgação de bases de dados de empresas assenta na divulgação de ficheiros totalmente sintéticos, uma vez que, desta forma, nenhum registo original é divulgado e se o modelo for adequado os ficheiros disponibilizados são estatisticamente válidos (permitem obter resultados semelhantes aos do ficheiro original).

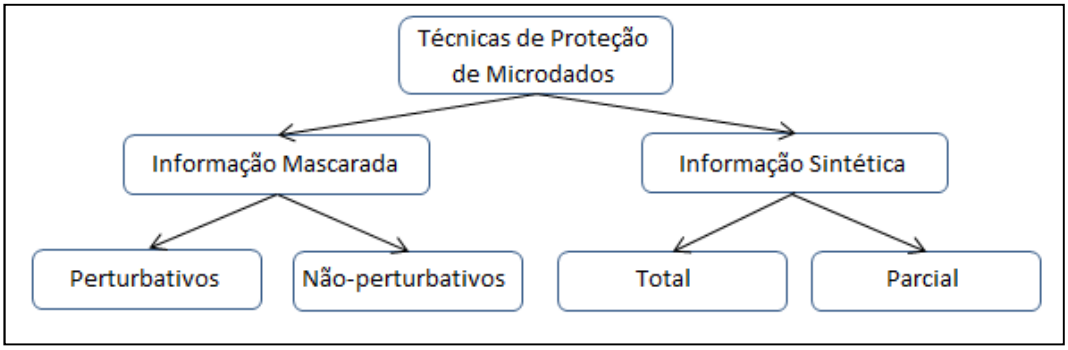
4 – GERAÇÃO SINTÉTICA DE MICRODADOS

"*An alternative to masking the original data is to generate a new data set (a synthetic data set) not from the original data, but from a set of random values that are adjusted in order to fulfill certain statistical requirements*", Mateo-Sanz *et al*. (2004b)[21].

Na literatura têm sido propostas várias técnicas de proteção de divulgação de microdados. Estas técnicas assentam na ideia que a possibilidade de identificar os respondentes pode ser neutralizada através da redução da quantidade de informação facultada, mascarando os dados (por exemplo não facultar a informação na sua totalidade ou perturbando valores), ou pela divulgação de bases de dados sintéticas.

Segundo Ciriani *et al. (*2007), as técnicas de proteção de microdados podem ser utilizadas na produção de dois tipos de informação: informação mascarada e sintética (figura 2)[22].

Figura 2 – Classificação das técnicas de proteção de microdados



No âmbito da informação mascarada existe uma divisão entre métodos perturbativos e não perturbativos. Os ***métodos perturbativos*** distorcem os valores dos dados antes de serem publicados. Nesta abordagem, as combinações únicas de variáveis do ficheiro original podem desaparecer e outras combinações novas podem surgir, beneficiando a preservação da confidencialidade. Os ***métodos não perturbativos*** são aqueles que não alteram os valores dos dados originais, mas por outro lado, os generalizam ou suprimem reduzindo o detalhe da informação (Hundepool *et al*., 2010). Por fim, a ***geração sintética*** consiste em obter de forma aleatória bases de dados que preservem algumas

[21] "*Uma alternativa à informação mascarada é gerar um novo conjunto de dados (dados sintéticos) não a partir dos dados originais, mas a partir de um conjunto de valores aleatórios que são ajustados de forma a cumprir certos requisitos estatísticos*" (tradução livre do autor).
[22] Fonte: Adaptado de Ciriani *et al.* (2007).

propriedades estatísticas e a relação existente entre as variáveis do ficheiro original. Os métodos de geração de informação sintética podem ser aplicados a toda a base de dados (ficheiros totalmente sintéticos) ou apenas a um conjunto de variáveis ou registos específicos (ficheiros parcialmente sintéticos).

Existe uma vasta literatura sobre os métodos perturbativos e não perturbativos. Uma descrição exaustiva destes métodos pode ser encontrada no manual do controlo da divulgação estatística[23] fornecido pelo projeto CENEX-SDC (*Center of Excellence for Statistical Disclosure Control*) fundado pelo Eurostat. No âmbito do presente estudo vão ser abordados métodos que se enquadram no âmbito da geração sintética.

4.1 - CONFIDENCIALIDADE E UTILIDADE NA GERAÇÃO DE DADOS SINTÉTICOS

A publicação de informação sintética ou simulada foi proposta há algum tempo como uma forma de preservar a confidencialidade da informação. Segundo Reiter (2005a), os métodos tradicionais de discretização de variáveis (por exemplo dividir a idade em intervalos), a criação de limiares superiores (por exemplo agregar rendimentos elevados numa única variável), a troca de informação entre determinadas unidades estatísticas ou a adição de ruído aleatório, podem comprometer e distorcer a relação existente entre as variáveis e limitar a análise por parte dos utilizadores, nomeadamente em termos de métodos de análise e perda de utilidade dos ficheiros. Normalmente, o aumento do número de transformações diminui o risco de divulgação, mas também a qualidade das inferências que se podem retirar dos ficheiros. Por exemplo, Caiola e Reiter (2010) argumentam que a troca de registos afeta a correlação entre os registos alterados e os que não foram alvo de intervenção, tendo consequências ao nível das conclusões que se podem retirar dos ficheiros disponibilizados.

Graham e Penny (2007) acrescentam que é difícil identificar e transformar os registos únicos numa base de dados para aplicar os métodos tradicionais (perturbativos e não perturbativos). Segundo os autores, estas tarefas envolvem bastante tempo e juízos de valor, uma vez que se tem de considerar conjuntamente o risco de identificação e a possibilidade de existir ligação com outras bases de dados externas, sem esquecer a perda de informação decorrente do próprio processo de transformação.

[23] Referenciado neste trabalho como Hundepool *et al*. (2010).

O trabalho de Winkler (2007) demostra as consequências devastadoras na qualidade da informação de alguns métodos tradicionais fáceis de implementar, chegando a falhar, em alguns casos, o objetivo primordial de proteção de informação dos respondentes.

Uma abordagem alternativa aos métodos tradicionais é facultar ***bases de dados sintéticas*** que permitem, por um lado preservar a confidencialidade, uma vez que não se facultam valores reais observados (ficheiros totalmente sintéticos) ou se transformam registos críticos (ficheiros parcialmente sintéticos) e por outro permitem aos utilizadores fazer inferências válidas e aplicarem diferentes métodos de análise (Reiter, 2005a e Graham e Penny, 2007).

Em termos de utilidade, Raghunathan *et al.* (2003) afirmam que os ficheiros sintéticos, decorrentes de informação simulada permitem obter inferências válidas. Os autores acrescentam ainda que a informação sintética pode ter vantagens em termos de investigação por resultar de amostras aleatórias e não dos métodos complexos de amostragem utilizados na obtenção da informação original. Segundo os autores, as formas de análise mais complexas podem ser abandonadas e tornar a análise acessível a mais utilizadores[24]. Outra vantagem é a possibilidade de divulgar informação que de outra forma não seria divulgada devido ao risco de identificação dos respondentes (por exemplo informação de áreas geográficas com pouca densidade populacional).

Mathews e Harel (2011) referem que um obstáculo à produção de informação sintética é convencer os pesquisadores de que a sua análise tem méritos. Um ficheiro de dados que garante a confidencialidade dos respondentes e não é aceite pelos investigadores devido à falta de qualidade é de pouco valor, colocando em causa o tempo e o dinheiro investidos na sua produção.

Um aspeto crítico dos ficheiros sintéticos é a validade das inferências que por sua vez depende dos modelos utilizados para gerar a informação. Quando os modelos falham no apuramento de determinadas relações existentes entre as variáveis, os investigadores também não as vão obter. De forma análoga, pressupostos errados nas distribuições vão passar para as análises dos utilizadores. Reiter (2005a) afirma que o *trade-off* entre risco e utilidade continua a existir na produção de informação sintética, tal como nos métodos tradicionais. O autor refere que agrupar rendimentos acima de um limiar numa classe única ou agrupar idades em classes também tem efeitos na utilidade.

[24] Os autores referem este argumento no âmbito da aplicação do método da *Imputação Múltipla* onde são disponibilizados um conjunto de ficheiros sintéticos.

Evidências empíricas demonstram a utilidade da informação sintética, ou seja, se os modelos são adequados muitos dos resultados obtidos são idênticos aos originais. Embora aplicando metodologias diferentes, os trabalhos de Raghunathan *et al.* (2003), Reiter (2005a), Reiter (2005b), Caiola e Reiter (2010), Drechsler e Reiter (2010), Lee *et al.* (2013), Loong *et al.* (2013) e Raab *et al.* (2015) são exemplos de que é possível criar ficheiros sintéticos válidos. Todos estes trabalhos são descritos de forma mais detalhada no Anexo 2.

4.2 – FICHEIROS PARCIALMENTE SINTÉTICOS E HÍBRIDOS

Domingo–Ferrer *et al.* (2009), esclarecem que a criação de dados sintéticos para todas as variáveis pode ser difícil, tendo alguns autores optado por abordagens que congregam a informação existente na base de dados original com informação sintética. Assim, apenas os atributos confidenciais ou os registos que tenham associado um elevado risco de divulgação é que vão ser alvo de transformação, mantendo-se os restantes na forma original (Mathews e Harel, 2011). Por exemplo, Loong *et al.* (2003) optaram por apenas sintetizar variáveis demográficas numa base de dados de informação médica, mantendo inalteradas as variáveis clínicas. Segundo os autores, a produção de ficheiros totalmente sintéticos pode implicar perdas significativas de utilidade, porque a possibilidade do modelo de imputação não reproduzir as relações existentes aumenta com o número de variáveis geradas sinteticamente.

Dadas as dificuldades em produzir ficheiros totalmente sintéticos, muitos dos autores optaram por produzir ficheiros parcialmente sintéticos. Segundo Reiter (2005b), os ficheiros parcialmente sintéticos são atrativos porque permitem manter os benefícios em termos de confidencialidade e superam as dificuldades de obtenção de um ficheiro totalmente sintético plausível, que segundo o autor é difícil de operacionalizar. Em termos empíricos, os trabalhos de Reiter (2005b), Caiola e Reiter (2010), Drechsler e Reiter (2010), Lee *et al.* (2013) e Loong *et al.* (2013) optaram por produzir ficheiros parcialmente sintéticos.

Embora pareça uma decisão acertada apenas intervir nas variáveis chave e sensíveis no sentido de minimizar os impactos em termos de utilidade, Domingo–Ferrer *et al.* (2009) referem que a manutenção de valores originais para algumas variáveis ou registos implica o aumento do risco de divulgação. Neste caso, a probabilidade de existir uma ligação entre o valor divulgado e uma base de dados externa é elevada.

Outra opção referida na literatura é a utilização ficheiros híbridos (Dandekar *et al.*, 2002), que consiste em calcular ficheiros protegidos utilizando uma combinação da base de dados original com a base de dados sintética. Este método tem como vantagem permitir um melhor controlo sobre a manutenção das características do ficheiro original. Segundo os autores, este método levanta duas questões fundamentais: i) como associar os registos sintéticos aos da base de dados original e ii) como combinar a informação dos dois ficheiros. Uma solução para a primeira questão é a aplicação da menor distância entre registos sintéticos e originais, utilizando para isso uma medida de distância (os autores propõem a distância euclidiana).

Em termos de combinação dos dois ficheiros, os autores sugerem a combinação aditiva ou multiplicativa. Na combinação aditiva, o ficheiro de dados híbrido *Z*, é resultado de uma combinação linear dos registos do ficheiro de microdados *X* e do ficheiro de dados sintéticos X_s, assumindo α valores no intervalo [0, 1].

$$Z = \alpha X + (1 - \alpha) X_S \tag{5}$$

Na combinação multiplicativa o ficheiro resulta da seguinte expressão:

$$Z = X^{\alpha} + X_S^{(1-\alpha)} \tag{6}$$

O método híbrido, tal como é definido pelos seus autores, apenas é possível para variáveis quantitativas, constituindo uma limitação.

4.3 – IMPUTAÇÃO MÚLTIPLA

Na literatura existe um conjunto de técnicas para gerar bases de dados sintéticas. Neste ponto são apresentados os métodos baseados na *Imputação Múltipla*, uma vez que é a técnica utilizada no presente trabalho. Como se vai poder avaliar em seguida, a *Imputação Múltipla* por si só não é um método, mas sim um conceito ou linha de orientação que serve de fundamentação teórica para um conjunto de métodos de imputação. No Anexo 3 são descritos outros métodos de geração sintética de microdados identificados na literatura.

O conceito da *Imputação Múltipla* foi inicialmente criado na década de 70 para lidar com valores em falta em bases de dados (*missing values*). Esta técnica, na sua aplicação original, tem por objetivo completar a informação existente numa base de dados, possibilitando assim análises mais completas por parte dos utilizadores. O pressuposto

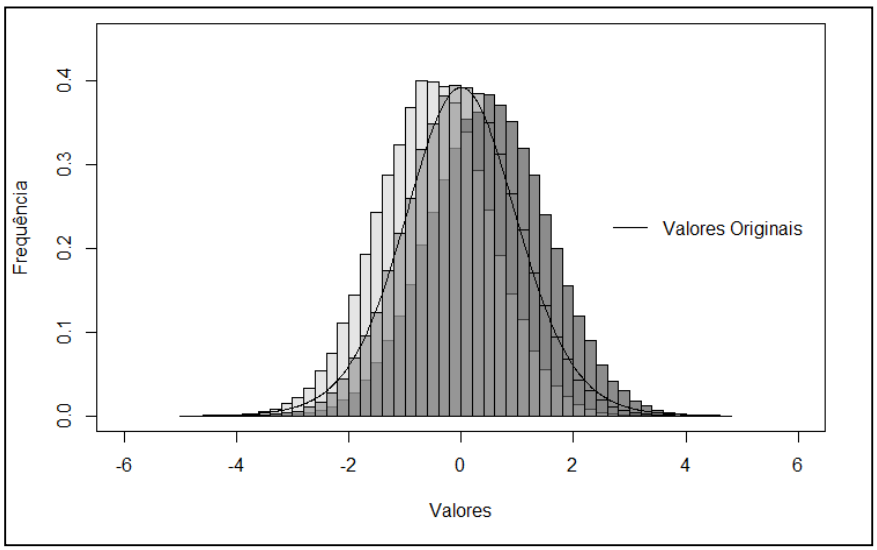
da imputação é assumir que existe uma relação entre as variáveis de um ficheiro de dados, pelo que na presença de valores omissos consegue-se estimar esses valores partindo das restantes variáveis. Se em vez de gerar apenas um valor se associar aleatoriedade aos parâmetros do modelo obtém-se múltiplos valores.

Segundo Refaat (2010), o número de valores imputados não precisa de ser elevado. Demonstra-se que perante um número pequeno de imputações (entre três e dez) conseguem-se obter bons resultados. No âmbito da *Imputação Múltipla* de valores omissos, uma boa prática é substituir o valor em falta pela média ou mediana dos valores obtidos nas múltiplas imputações.

O método de *Imputação Múltipla* aplicado à geração de ficheiros parcialmente sintéticos foi proposto por Rubin (1993). Este método consiste em identificar um conjunto de registos base e tratar todas as observações que se encontram fora desse conjunto como valores em falta (*missing values*), utilizando ferramentas de *Imputação Múltipla* (Domingo-Ferrer *et al.*, 2009). O modelo de imputação é desenhado com o objetivo de preservar as relações existentes no ficheiro original.

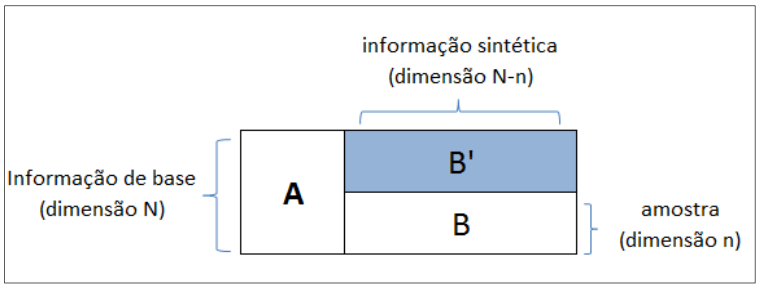
Na figura 3 é apresentado um histograma resultante da simulação da geração de múltiplos valores sintéticos para uma variável (que segue uma distribuição normal de média 0 e desvio padrão 1). A inclusão de aleatoriedade nos parâmetros do modelo de geração sintética permite obter múltiplos valores. Na figura foram geradas três distribuições para a variável original (representada pela linha).

Figura 3 – Histograma resultante da geração de múltiplos valores sintéticos (*m*=3)



O método da *Imputação Múltipla* em termos conceptuais pode ser descrito da forma que se segue (Hundepool *et al*., 2010): considere-se a base de dados de microdados *X* de dimensão *n* retirada de uma população de *N* indivíduos, onde *A* são as variáveis de base e *B* são as variáveis que se pretende gerar sinteticamente (Figura 4). As variáveis de base[25] (*A*) estão disponíveis para os *N* indivíduos e as variáveis *B* apenas estão disponíveis para os *n* registos da amostra. O primeiro passo do método é gerar *m* imputações múltiplas para *N-n* indivíduos, onde *m* é o número de imputações que habitualmente varia entre três e dez. Deste processo resultam *m* populações (compostas pelos *n* registos da base de dados *X* mais *N-n* registos) e *m* matrizes de informação (*B'*) para *N-n* indivíduos que não estão na amostra (Figura 4). No final são disponibilizadas ao público as *m* bases de dados. A variabilidade dos valores imputados garante, pelo menos teoricamente, que inferências validas podem ser obtidas.

Figura 4 – Método da *Imputação Múltipla*



Segundo Hundepool *et al*. (2010), a escolha do modelo de imputação de (*B*) partindo de *A* não é simples, sendo este o principal desafio da *Imputação Múltipla*.

O trabalho de Domingo–Ferrer *et al.* (2009) faz uma avaliação crítica do método proposto por Rubin. Segundo os autores, a definição do modelo de imputação é fundamental, nomeadamente em termos de modelação da relação entre as variáveis consideradas na análise. Se por um lado uma má definição do modelo pode originar uma grande perda de informação, tornando inútil a base de dados gerada, por outro, caso o modelo utilizado seja um descritor perfeito da base de dados, os dados podem ser muito semelhantes ou até mesmo iguais aos dados originais, o que implica um risco de divulgação elevado. Em termos de complexidade, esta cresce com o número de variáveis, pelo que muitas vezes apenas se aplica o método às variáveis chave e sensíveis, gerando-se ficheiros parcialmente sintéticos.

[25] Alguns autores optam por terminologias diferentes. Por exemplo, Reiter (2005b) utiliza o termo '*design variables*', dando como exemplo o estrato, cluster ou indicador de dimensão.

Noutro trabalho, Mateo-Sanz *et al.* (2004b) reconhecem que apesar do método de *Imputação Múltipla* apresentar resultados promissores, ele requer a utilização de modelos e programas informáticos complexos, reduzindo o seu interesse.

Na literatura encontram-se diferentes metodologias de *Imputação Múltipla*. Estes métodos têm como linha principal a proposta de Rubin (1993), diferindo em termos de método de imputação e pelo facto de proporem a criação de ficheiros totalmente ou parcialmente sintéticos. Os métodos são divididos em abordagens paramétricas e não paramétricas, quer assumam ou não a hipótese de normalidade, sendo as inferências condicionadas por esse pressuposto. Em seguida são apresentados os métodos de *Imputação Múltipla* existentes na literatura.

4.3.1 - Distribuição de Probabilidades

Raghunathan *et al.* (2003) propõem a utilização da distribuição de probabilidades no processo de imputação. Utilizando a notação no ponto anterior, este método consiste em gerar valores aleatórios independentes para os registos *B'* utilizando os valores reais existentes para a população (*A* composta por *N* registos) e os valores de *B* para os *n* registos. Os autores utilizaram a distribuição preditiva posterior $P(B'|A,B)$ onde *B'* são os valores gerados. A função probabilidade é condicionada aos valores observados *A* e *B* e aos pressupostos do modelo de probabilidade.

No seu trabalho, os autores utilizaram dados simulados numa abordagem paramétrica e não paramétrica (*bootstraping* bayesiano). Na abordagem paramétrica é utilizada a função de distribuição normal multivariada. Na opção não paramétrica é atribuída a cada observação um peso proporcional à distribuição das probabilidades geradas a partir de uma distribuição *a posteriori* (distribuição *bootstrap* bayesiana)[26].

O passo final do processo consiste em se retirar amostras aleatórias de cada uma das populações de modo a criar bases de dados sintéticas. Os autores sugerem que se efetue testes de confidencialidade antes da disponibilização dos ficheiros de modo garantir que nenhum registo confidencial seja incorporado.

[26] Na abordagem paramétrica, os pressupostos assumidos para a geração das variáveis sintéticas (função de distribuição normal multivariada) coincidiu com o modelo de criação dos dados simulados, o que corresponde a um cenário perfeito em que a distribuição das variáveis pode ser modelada por uma função de distribuição teórica.

Apesar dos bons resultados utilizando bases de dados simuladas, a aplicação deste método a base de dados reais é limitada, pois muito dificilmente se consegue associar uma função de distribuição teórica à informação e a utilização de probabilidades condicionadas apenas é aplicável a variáveis discretas.

4.3.2 – Regressões Multivariadas Sequenciais

Os trabalhos de Reiter (2005a) e Loong *et al.* (2013) fazem uma aplicação do método de *Imputação Múltipla* utilizando regressões. Os modelos são especificados utilizando uma sequência de regressões, ou seja, cada regressão inclui as variáveis correspondentes à informação de base (variáveis que não vão ser sintetizadas) e as variáveis sintéticas entretanto geradas pela regressão anterior[27]. Segundo Reiter (2005a), este método implica duas fases: i) obter valores para os parâmetros da regressão partindo da distribuição posterior, ou aproximações dessas distribuições, dada a informação observada e ii) gerar dados sintéticos com base no valor dos parâmetros estimados, variáveis observadas e, se aplicável, variáveis sintéticas geradas nos processos precedentes.

Este tipo de abordagem é conhecido por imputação de regressões multivariadas sequenciais - *Sequential Regression Multivariate Imputation* (SRMI) sendo considerada uma abordagem paramétrica devido aos pressupostos subjacentes aos modelos de regressão (Drechsler, 2009 e Loong *et al.*, 2013).

O trabalho de Reiter (2005a) utiliza diversos modelos - regressão logística, regressão logística multinominal, *bootstrap* bayesiano e regressão linear - para gerar sete variáveis qualitativas e quantitativas. O autor fez um estudo prévio das relações entre as variáveis com o objetivo de as modelar e sequenciar em cada um dos modelos. Segundo o autor, o responsável pela geração sintética deve tirar vantagem do seu conhecimento e experiência estatística de modo a construir modelos corretos, acrescentando que se deve experimentar diferentes alternativas na construção dos modelos de imputação até se encontrar resultados satisfatórios.

Noutro trabalho, Loong *et al.* (2013) optaram por criar dois modelos de imputação distintos para a mesma base de dados, porque a relação entre as variáveis diferia em função dos valores de uma variável (tipo de doença). Ao contrário do trabalho de Reiter (2005a), os autores apenas sintetizaram variáveis qualitativas (cinco variáveis

[27] Neste trabalho este tipo de abordagem é designada de operacionalização 2 (ponto 5.4.2).

demográficas), utilizando em média cerca de 50 variáveis (num total de 350 do ficheiro original) em cada modelo de imputação.

Apesar dos bons resultados obtidos pelos autores, a aplicação de modelos de regressão no âmbito da geração sintética de informação implica conhecer de forma detalhada a relação entre as variáveis e ter que lidar ou corrigir a violação das hipóteses subjacentes às especificações dos modelos. De entre as eventuais dificuldades encontram-se as seguintes (Drechsler e Reiter, 2011): i) as bases de dados incluem variáveis qualitativas e quantitativas, que algumas vezes não são fáceis de modelar com os modelos de regressão tradicionais[28]; ii) as relações entre as variáveis podem não ser lineares ou podem ser endógenas; iii) a hipótese de normalidade na maior parte das vezes não se verifica e iv) podem existir problemas de multicolinearidade[29].

A principal limitação do SRMI é que a especificação dos modelos é uma tarefa que consome muitos recursos, podendo os divulgadores de informação simplesmente não ter o conhecimento e o tempo necessário para a efetuar. Reiter (2005a) classifica esta tarefa como um processo de 'tentativa e erro', onde se tenta manter a confidencialidade da informação e reproduzir ao máximo a estrutura da base de dados.

Dadas as dificuldades na aplicação de métodos de regressão na geração de informação sintética, surgiram um conjunto de trabalhos que aplicam abordagem não paramétricas. Estes métodos conseguem lidar com diferentes tipos de informação de elevada dimensionalidade, captar relações não lineares e iterações que muito dificilmente são operacionalizadas nos modelos tradicionais. Adicionalmente, a sua implementação é mais simples (não necessita de calibração) e mais rápida, permitindo aos divulgadores de informação cumprirem atempadamente os prazos de divulgação estatística (Drechsler e Reiter, 2011). Em seguida são apresentados os métodos que utilizam os algoritmos de Árvores de Decisão e *Random Forests* para a produção de ficheiros sintéticos.

[28] Segundo Drechsler (2009) para variáveis quantitativas habitualmente utiliza-se a regressão linear e para variáveis quantitativas a regressão logística.
[29] Loong *et al.* (2013) referem que este foi um dos problemas encontrados devido à quantidade de variáveis utilizadas no modelo de imputação.

4.3.3 - Algoritmo de Árvores de Decisão

Os modelos em árvores são designados Árvores de Classificação, no caso de problemas de classificação (variáveis qualitativas ou categóricas), e Árvores de Regressão, nos problemas de regressão (variáveis quantitativas ou contínuas). Quer em Árvores de Classificação, quer em Árvores de Regressão, a interpretação dos modelos assim como os algoritmos de indução das árvores são muito semelhantes. Uma Árvore de Regressão é semelhante a uma Árvore de Decisão pois ambas são formadas por um conjunto de nós de decisão, mas o resultado, em vez de uma categoria (designada de classificação) é um escalar (designado por previsão). Seguindo Gama *et al.* (2012) neste trabalho utiliza-se o termo Árvores de Decisão de uma forma genérica para Árvores de Classificação ou Regressão, referindo-se, sempre que se entender pertinente, as suas diferenças.

Uma Árvore de Decisão usa a estratégia 'dividir para conquistar' para resolver um problema de decisão. Um problema complexo é dividido em problemas mais simples, aos quais recursivamente é aplicada a mesma estratégia. A vantagem de uma abordagem deste género prende-se com a divisão do espaço de instâncias em subespaços e ajustar cada subespaço recorrendo a diferentes formulações (Gama *et al.*, 2012). Esta é a ideia comum aos algoritmos baseados em Árvores de Decisão tais como: *CART* (Breiman *et al.*,1984), *ID3* (Quinlan, 1986), *ASSISTANT* (Cestnik *et al.*, 1987) e *C4.5/J4.8* (Quinlan, 1993)[30].

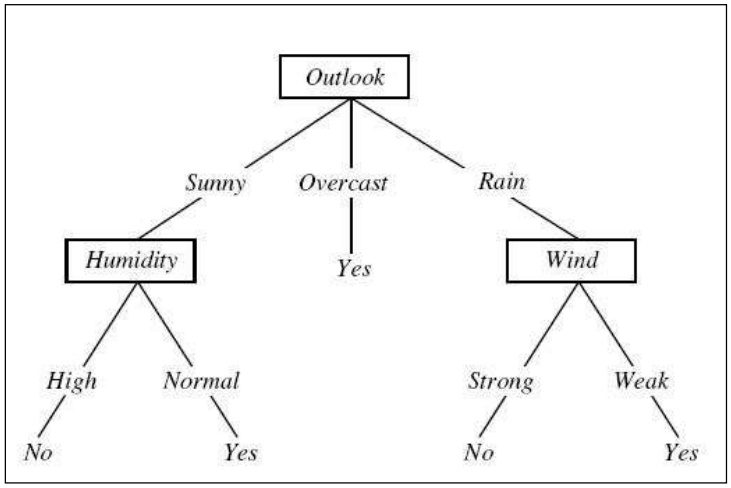
Um algoritmo de Árvores de Decisão traduz o resultado de uma partição recursiva dos dados base da modelação. Uma vez que não estabelece qualquer tipo de restrição em termos de pressupostos e distribuição das variáveis, as Árvores de Decisão são classificadas como um método não-paramétrico. O algoritmo divide o espaço dos previsores para que subconjuntos de unidades formadas pelas partições tenham resultados relativamente similares. As partições podem ser eficazmente representada por uma estrutura em árvore, onde cada folha[31] corresponde a subconjuntos de unidades.

Na figura 5[32] é apresentado um exemplo de Árvore de Decisão (Quinlan,1986), cujo objetivo é obter uma resposta – jogar ou não ténis (atributo classe ou alvo) – partindo de

[30] O classificador J4.8 é uma implementação escrita em Java do algoritmo C4.5, presente no programa Weka. O C4.5 é uma extensão do algoritmo ID3 e foi proposto por Quinlan (1993) no livro '*C4.5: Programs for Machine Learning*'.
[31] No contexto de Árvores de Decisão denomina-se o nó final por folha.
[32] Fonte: Quinlan, (1996)

quatro atributos – *Outlook* (*sunny*, *overcast*, *rain*), *Humidity* (*high*, *normal*), *Wind* (*Strong*, *Weak*) e *Temperature* (*hot*, *mild*, *cool*), utilizando uma base de dados de 14 registos na aprendizagem da árvore.

Figura 5 – Exemplo de Árvore de Decisão



Associada à construção de Árvores de Decisão surge a questão da regra de divisão, ou seja, como se identifica o atributo que vai ser utilizado em cada divisão (por exemplo, na figura 5 o primeiro atributo selecionado foi o *Outlook*). Uma regra de divisão visa avaliar em que medida um dado atributo discrimina as classes. As regras de divisão mais conhecidas e utilizadas são o *Ganho de Informação*, que é usado no algoritmo *C4.5*, e o índice de *Gini*, usado no *CART*. Independentemente do critério utilizado, eles concordam no facto que uma divisão que mantém a proporção de classes em todo o subconjunto não tem utilidade e uma divisão na qual cada subconjunto contém somente exemplos de uma classe tem utilidade máxima (Gama *et al.*, 2012).

Em problemas de regressão (Árvores de Regressão) a função a minimizar é tipicamente o erro quadrático[33], sendo a média a constante que a minimiza (Atkinson e Therneau, 2015). Uma Árvore de Regressão é semelhante à construção de uma Árvore de Classificação tendo em conta a função de custo referida (Gama *et al.*, 2012).

Num processo de tomada de decisão apenas é necessário seguir a árvore da raiz até á folha para se obter o resultado (classificação ou previsão). No exemplo da figura 5 facilmente se obtém a resposta sobre a realização ou não do jogo de ténis conhecendo o valor dos restantes atributos. Para além da decisão final (no exemplo jogar ou não), em

[33] Um trabalho importante nesta área é a tese de doutoramento de Luís Torgo (Torgo, 2000) que aplica o erro absoluto médio como alternativa ao erro quadrático médio.

cada folha é possível obter a frequência relativa de cada uma das classes (no exemplo seria o número de casos classificados como positivos e negativos), uma vez que os valores das folhas resultam da distribuição condicional das unidades dos dados que satisfazem os critérios que as definem (função de probabilidade condicionada de Y dado X=x_i fixo):

$$f(y_j|\,\mathrm{X} = x_i) = f(y_j|\,x_i) = P(Y = y_j|X = x_i), \forall\, y_j \qquad (7)$$

em que y_j é o valor do atributo classe ou alvo e x_i é o vetor dos valores dos restantes atributos percorrendo a árvore desde á raiz até à respetiva folha para o registo *i*. Nas Árvores de Regressão, uma vez que cada elemento da classe é provavelmente único, a função de probabilidade condicionada é uniforme para os valores que definem a folha.

Na geração sintética para cada registo é aplicada a técnica de amostragem aleatória com reposição em que a probabilidade de seleção de cada y_j é igual à respetiva função de probabilidade condicionada ($f(y_j|\,\mathrm{X} = x_i)$). Para as variáveis quantitativas (ou contínuas) de modo a não se facultar o valor original da variável, os valores resultantes da amostragem são substituídos por uma estimativa de densidade Kernel[34] nesse ponto.

Construindo uma Árvore de Decisão para cada uma das variáveis do ficheiro e gerando valores sintéticos obtém-se um novo ficheiro de dados que preserva a relação entre as variáveis. A aleatoriedade decorrente do processo de amostragem permite obter múltiplos valores para o mesmo registo e consequentemente múltiplos ficheiros de dados. De modo a manter a relação entre as variáveis, as Árvores de Decisão devem ser induzidas utilizando sempre o ficheiro de dados original[35].

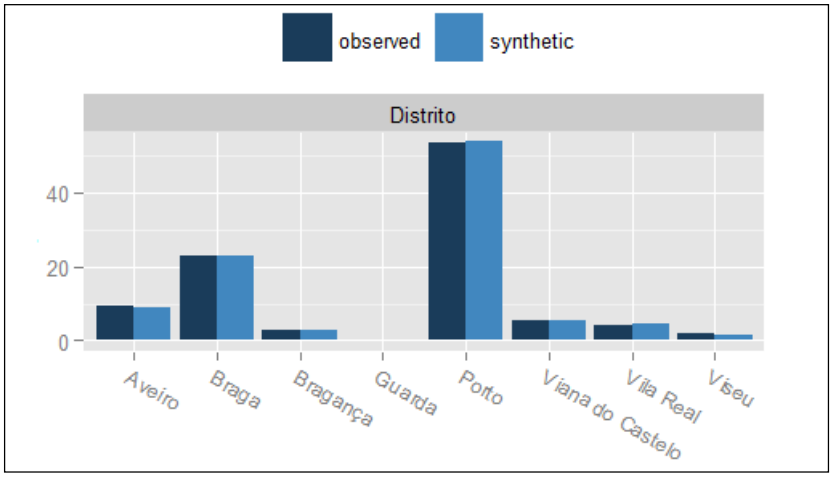
Na figura 6[36] é apresentada uma ilustração do resultado da aplicação de Árvores de Decisão na geração de variáveis sintéticas para a variável *Distrito* na base de dados utlizada neste estudo. Constata-se que as frequências dos valores sintéticos são semelhantes aos valores observados, indo de encontro ao resultado esperado.

[34] Em estatística, estimativa de densidade Kernel (EDK) é uma forma não-paramétrica para estimar a função de densidade de probabilidade de uma variável aleatória. A estimação da densidade Kernel é um problema fundamental de suavização de dados onde inferências sobre a população são feitas com base numa amostra de dados finita (fonte: Guia de programação do software R, *Kernel Density Estimation*).

[35] Nos pontos 5.4.1 e 5.4.2 são descritos e exemplificados com maior detalhe a operacionalização do método de geração sintética.

[36] Este gráfico foi obtido utilizando a função 'compare.synds' do *package synthpop* do programa R (Nowok *et al.*, 2014).

Figura 6 – Distribuição de frequências da variável *Distrito* (valores observados e sintéticos)



Segundo Drechsler e Reiter (2010), a aplicação do algoritmo Árvores de Decisão à geração sintética de microdados tem como vantagens: i) a possibilidade de ser utilizado em diferentes tipos de informação (variáveis quantitativas e qualitativas); ii) a possibilidade de captar relações não lineares e interações complexas e iii) o facto de ser computacionalmente eficiente em bases de dados de elevada dimensão. O trabalho de Reiter (2005b) acrescenta ainda como vantagens o facto da modelação de Árvores de Decisão ser de fácil implementação quando comparado com os métodos paramétricos e o algoritmo possuir uma forma semiautomática de ajustar as relações mais importantes da base de dados, o que pode ser uma vantagem substancial quando se está na presença de muitos previsores.

No entanto, Caiola e Reiter (2010) referem como desvantagens deste algoritmo, relativamente à utilização de modelos paraméricos, a dificuldade de interpretação, a descontinuidade dos limites das partições e a diminuição da eficácia quando as relações são descritas por modelos paramétricos.

Em termos empíricos, o trabalho de Drechsler e Reiter (2010) propõe a criação de ficheiros parcialmente sintéticos utilizando Árvores de Decisão. A ideia básica é substituir as variáveis chave e sensíveis utilizando a *Imputação Múltipla* num conjunto de registos, gerando ficheiros parcialmente sintéticos. Esta técnica compreende várias fases: i) seleção do conjunto de registos a substituir, ou seja, os registos em risco de serem identificados; ii) determinar os modelos de geração sintética utilizando a base de dados disponível, tirando vantagem de toda a informação (fase de modelação) e iii) simular repetidamente os valores para os dados selecionados para criar múltiplas populações protegidas. Para além da abordagem de Drechsler e Reiter (2010) que aplica

a *Imputação Múltipla* a um conjunto de registos classificados como em risco, Reiter (2005b) e Lee *et al.* (2013) utiliza a *Imputação Múltipla* com aplicação de Árvores de Decisão a todos os registos das variáveis sensíveis da base de dados. Todos estes trabalhos enquadram-se no conceito de ficheiros parcialmente sintéticos, enquanto Drechsler e Reiter (2010) apenas substitui unidades específicas e depois divulga dados sintéticos e originais, Reiter (2005b) e Lee *et al.* (2013) aplicam a técnica a todos os registos das variáveis consideradas sensíveis, divulgando ficheiros onde coexistem variáveis ´totalmente sintéticas' e originais. O trabalho de Raab *et al.* (2015) é o único que aplica o algoritmo para criar ficheiros totalmente sintéticos, considerando apenas cinco variáveis em que a idade é a única variável quantitativa.

Para finalizar, importa referir que em algumas situações as Árvores Decisão podem ser descritores perfeitos da base de dados, implicando que os ficheiros gerados sinteticamente sejam iguais ao original, mantendo-se os problemas de confidencialidade. A solução poderá passar pela especificação da árvore, por exemplo obrigar a um número mínimo de registos por folha (pré-poda)[37], ou proceder à poda posterior da árvore de modo a evitar os problemas de sobre ajustamento do modelo aos dados (*overfitting*).

4.3.4 - Algoritmo *Random Forests*

O algoritmo *Random Forests* foi desenvolvido por Breiman (2001) e pode ser entendido como uma extensão da técnica de Árvores de Decisão, uma vez que faz uso de métodos de reamostragem a fim de melhorar a precisão dos modelos construídos. O método *Random Forests* é um classificador do tipo comité ou 'ensemble' constituído por várias Árvores de Decisão.

Esta técnica, segundo Breiman (2001), caracteriza-se por: i) *Bagging* – gerar um conjunto de subamostras selecionadas aleatoriamente com reposição provenientes do ficheiro original; ii) *Boosting* – construir, para cada subamostra, um modelo de Árvores de Decisão, aumentando a ponderação das observações incorretamente classificadas com base nos modelos criados para as outras subamostras e iii) *Randomizing* – em cada árvore é utilizada um subconjunto diferente de atributos.

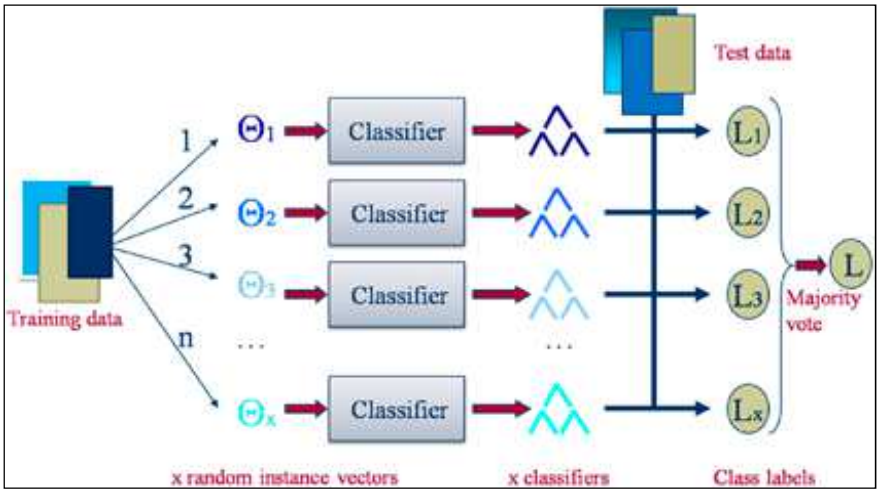
Segundo Liaw *et al.* (2009), o algoritmo *Random Forests* é um conjunto de Árvores de Decisão aplicadas a diferentes combinações de subconjuntos de informação e variáveis

[37] Opção utilizada neste trabalho (ver ponto 5.4.3).

independentes. Este algoritmo aprende um conjunto de Árvores de Decisão aplicando técnicas de amostragem simples à base de dados. A racionalidade subjacente à utilização do algoritmo é que a combinação de diferentes classificadores aumenta a sua precisão. Um elemento será classificado considerando os contributos de cada árvore existente na 'floresta'. Para Árvores de Regressão é aplicada a média dos previsores e para Árvores de Classificação o critério de escolha é o voto maioritário.

Na figura 7 é apresentada uma ilustração do funcionamento do algoritmo *Random Forests* aplicado a um conjunto de Árvores de Classificação[38].

Figura 7 – Funcionamento do algoritmo *Random Forests*



No âmbito da geração sintética e utilizando a notação do ponto anterior, o contributo de cada árvore individual, na figura 7 representados por $L_1, L_2, \dots, L_x$, é considerado para calcular a função de probabilidade condicionada ($f(y_j | X = x_i)$) para cada registo. Uma vez obtida a função, o processo é semelhante ao anteriormente descrito para as Árvores de Decisão. Para as variáveis quantitativas, à semelhança do algoritmo de Árvores de Regressão, o valor é substituído por estimativas Kernel no ponto.

Segundo Caiola e Reiter (2010) normalmente o algoritmo aprende cerca de 500 ou mais árvores utilizando a mesma base de dados. Cada árvore é baseada em diferentes subconjuntos que incluem cerca de 2/3 do total de registos. Os autores referem que o algoritmo *Random Forests* mantém a maioria dos benefícios das Árvores de Decisão mas com algumas vantagens: i) ao aplicar técnicas de amostragem aleatória na construção de cada árvore, o algoritmo permite que diferentes árvores contribuam para a

[38] Fonte: Maragoudakis e Loukis (2012).

previsão final e ii) o contributo agregado de diferentes árvores mais pequenas para a decisão permite reduzir o enviesamento e a variância comuns em árvores maiores[39].

Em termos empíricos, Caiola e Reiter (2010) aplicaram o algoritmo para obter ficheiros parcialmente sintéticos, gerando apenas variáveis qualitativas. Não se conhecem aplicações do algoritmo a variáveis quantitativas ou contínuas, facto que se pretende colmatar com o presente trabalho.

[39] Este é o principal argumento utilizado por Breiman (2001) quando propôs o algoritmo *Random Forests.*

5 – ESTUDO EMPÍRICO

"*Toda a teoria deve ser feita para poder ser posta em prática, e toda a prática deve obedecer a uma teoria. Só os espíritos superficiais desligam a teoria da prática, não olhando a que a teoria não é senão uma teoria da prática e a prática não é senão a prática de uma teoria*", Fernando Pessoa (Palavras iniciais, Revista de Comercio e Contabilidade, 1926).

Neste capítulo apresentam-se de forma detalhada os métodos utilizados e os resultados obtidos com vista à obtenção de ficheiros protegidos utilizando uma base de dados de empresas. Neste estudo são propostos quadro modelos que se dividem em dois algoritmos (Árvores de Decisão e *Random Forests*) e duas operacionalizações (operacionalização 1 e 2) utilizando uma lógica de amostragem com reposição. Com vista a comparar resultados e demonstrar as vantagens da abordagem de amostragem com reposição foram construídos e testados mais quatro modelo que seguem uma abordagem sem amostragem.

As três primeiras secções são dedicadas à análise da base de dados: na secção um (ponto 5.1) é efetuada uma breve descrição dos dados, na secção dois (ponto 5.2) são apresentadas as estatísticas que caracterizam as variáveis (análise exploratória) e na secção três (ponto 5.3) é analisada a questão da confidencialidade. No ponto 5.4 são descritos e discutidos os métodos, os programas utilizados e a parametrização dos algoritmos. Finalmente, na secção cinco (ponto 5.5) são avaliados os ficheiros gerados sinteticamente e analisados os impactos da variação do número de ficheiros gerados sinteticamente.

5.1 – DESCRIÇÃO DA BASE DE DADOS

A amostra utilizada no presente estudo é composta por 10 mil empresas portuguesas em atividade no ano de 2013 (informação mais recente disponível), com sede na região norte do país. A informação foi obtida a partir da base de dados SABI (Sistema de Análise de Balanços Ibéricos[40]) subscrita pela Faculdade de Economia do Porto. Para cada empresa foi obtida a informação referente a um total de nove variáveis, sete se não se considerar os identificadores diretos número de identificação fiscal (*Contribuinte*) e

[40] A SABI é uma base de dados de análise financeira de empresas ibéricas com um histórico de contas anuais - https://sabi.bvdinfo.com/version-2015310/home.serv?product=sabineo.

Nome. Na tabela seguinte são classificadas e descritas cada uma das variáveis e no Anexo 4 é facultada uma descrição mais detalhada.

Tabela 1 – Descrição das variáveis da base de dados

Designação	**Classificação 1** (Natureza)	**Classificação 2** (Confidencialidade)	**Descrição**
Contribuinte	Qualitativa	Identificadora	Número de identificação fiscal de pessoa coletiva
Nome	Qualitativa	Identificadora	Nome da empresa
Distrito	Qualitativa	Chave	Distrito da sede da empresa (8 classes)
Forma_Juridica	Qualitativa	Chave	Forma legal (4 classes)
Setor	Qualitativa	Chave	Setor de atividade de acordo com a secção da CAE Rev. 3 (19 classes)
Total_Ativo	Quantitativa	Sensível	Total do Ativo em 31 de dezembro de 2013
Volume_Negocios	Quantitativa	Sensível	Volume de Negócios em 31 de dezembro de 2013
Capital_Proprio	Quantitativa	Sensível	Capital Próprio em 31 de dezembro de 2013
NPS	Quantitativa	Sensível	Número de Pessoas ao Serviço em 31 de dezembro de 2013

A base de dados deste estudo é composta por cinco variáveis qualitativas e quatro variáveis quantitativas. A variável *Forma Jurídica* resulta da classificação efetuada pelo Código das Sociedades Comerciais e pelo tipo de propriedade (nacional ou estrangeira). O *Setor* corresponde ao primeiro nível de classificação (secção) da Classificação das Atividades Económicas (CAE Rev. 3).

Em termos de confidencialidade, as variáveis *Contribuinte* e *Nome* são ***identificadores diretos***, uma vez que são únicos para cada empresa. Por sua vez, os atributos *Distrito*, *Forma Jurídica* e *Setor* transmitem informação que apesar de por si só não permitirem a identificação imediata das empresas, combinadas podem levar à sua identificação, pelo que são consideradas ***variáveis chave*** ou ***quase identificadoras***. Os atributos quantitativos devido à sua natureza confidencial são consideradas ***variáveis sensíveis***, ou seja, não devem ser associados a qualquer empresa.

5.2 – ANÁLISE EXPLORATÓRIA

5.2.1 – Variáveis Qualitativas

A caracterização das ***variáveis qualitativas*** passa pela análise das frequências. As tabelas de contingência bidimensionais, para além de permitirem analisar frequências, são usadas para avaliar o relacionamento das categorias de duas variáveis.

Na tabela 2, tabela 3 e tabela 4 são apresentadas as tabelas de contingência que relacionam o *Distrito* com a *Forma Jurídica*, o *Setor* com a *Forma Jurídica* e o *Setor* com o *Distrito*, respetivamente.

Tabela 2 – Tabela de contingência: *Distrito* X *Forma Jurídica*

unidades

		Forma_Juridica				
		Cooperativa	Entidade Estrangeira	Sociedade Anónima	Sociedade por Quotas	Total
Distrito	Aveiro	0	1	66	861	**928**
	Braga	1	2	167	2.113	**2.283**
	Bragança	0	0	7	266	**273**
	Guarda	0	0	0	15	**15**
	Porto	3	15	424	4.899	**5.341**
	Viana do Castelo	0	4	16	530	**550**
	Vila Real	0	2	16	399	**417**
	Viseu	0	0	7	186	**193**
Total		**4**	**24**	**703**	**9.269**	**10.000**

Em termos de *Forma Jurídica*, constata-se que maioritariamente as empresas são *Sociedades por Quotas* (9.269 empresas), existindo apenas quatro que são *Cooperativas*. O *Distrito* do *Porto* é o que está mais representado (5.341 empresas) e o da *Guarda* o menos com apenas quinze empresas. Em termos de confidencialidade, verifica-se que as empresas *Cooperativas* e as *Entidades Estrangeiras* nos distritos de *Aveiro*, *Braga* e *Vila Real* se encontram em risco por apresentarem uma frequência inferior a três (considerando o critério do *k*-anonimato para *k*=3).

Tabela 3 – Tabela de contingência: *Seto*r X *Forma Jurídica*

unidades

		Forma_Juridica				
		Cooperativa	Entidade Estrangeira	Sociedade Anónima	Sociedade por Quotas	Total
Setor	A	0	1	13	201	**215**
	B	0	0	5	21	**26**
	C	0	5	195	1.646	**1.846**
	D	0	0	5	7	**12**
	E	0	0	10	19	**29**
	F	2	3	63	917	**985**
	G	2	10	123	2.767	**2.902**
	H	0	1	23	415	**439**
	I	0	0	17	763	**780**
	J	0	1	30	146	**177**
	K	0	1	37	137	**175**
	L	0	1	96	216	**313**
	M	0	0	35	822	**857**
	N	0	1	16	220	**237**
	P	0	0	6	148	**154**
	Q	0	0	19	584	**603**
	R	0	0	9	72	**81**
	S	0	0	1	168	**169**
Total		**4**	**24**	**703**	**9.269**	**10.000**

No que concerne ao setor de atividade, regista-se uma concentração de empresas na secção *G - Comércio por grosso e a retalho; reparação de veículos automóveis e motociclos* com cerca de 29% (2.902 empresas) do total da amostra. O *Setor* menos representado é o *D - Eletricidade, gás, vapor, água quente e fria e ar frio* com apenas doze empresas. Mais uma vez, o cruzamento da informação com a *Forma Jurídica* pode traduzir-se em risco de identificação das empresas para algumas combinações que considerem a *Forma Jurídica* igual a *Cooperativa*, *Entidade Estrangeira* e *Sociedade Anónima*.

Tabela 4 – Tabela de contingência: *Setor* X *Forma Jurídica*

unidades

		Distrito								
		Aveiro	Braga	Bragança	Guarda	Porto	Viana do Castelo	Vila Real	Viseu	Total
Setor	A	16	43	21	2	64	18	28	23	**215**
	B	1	6	2	0	7	4	6	0	**26**
	C	271	558	28	2	827	95	45	20	**1.846**
	D	0	4	1	0	6	0	1	0	**12**
	E	2	10	1	0	14	2	0	0	**29**
	F	84	273	25	1	473	68	32	29	**985**
	G	277	594	92	6	1.558	171	148	56	**2.902**
	H	31	70	17	2	252	31	25	11	**439**
	I	57	138	30	1	455	63	22	14	**780**
	J	13	36	1	0	123	0	3	1	**177**
	K	13	32	3	0	101	7	16	3	**175**
	L	22	80	6	0	184	11	7	3	**313**
	M	57	177	19	0	538	25	28	13	**857**
	N	19	56	2	0	137	8	11	4	**237**
	P	10	32	7	0	89	7	7	2	**154**
	Q	38	120	12	0	374	26	25	8	**603**
	R	7	17	2	1	41	5	5	3	**81**
	S	10	37	4	0	98	9	8	3	**169**
Total		**928**	**2.283**	**273**	**15**	**5.341**	**550**	**417**	**193**	**10.000**

Da análise da tabela de contingência, constata-se que em alguns *Distritos* e *Setores* menos representados existe um conjunto de combinações que podem originar a identificação das empresas da base de dados. Por exemplo, empresas do setor de atividade *E - Captação, tratamento e distribuição de água; saneamento, gestão de resíduos e despoluição* dos distritos de *Aveiro*, *Bragança* e *Viana do Castelo* pela pouca representatividade são facilmente identificadas.

5.2.2 – Variáveis Quantitativas

A caracterização das ***variáveis quantitativas*** passa pelo cálculo de medidas de localização, dispersão, assimetria e achatamento.

Tabela 5 – Medidas de localização e dispersão

	Média	1.º Quartil	Mediana	3.º Quartil	Desvio Padrão	Máximo	Mínimo
Total_Ativo (10^3)	2.105	58	165	502	46.708	3.786.039	0
Volume_Negocios (10^3)	1.334	53	138	419	34.292	3.333.909	0
Capital_Proprio (10^3)	916	4	41	163	26.713	2.306.983	-29.848
NPS (unidades)	12	2	3	8	222	21.383	1

Tabela 6 – Medidas de assimetria e achatamento

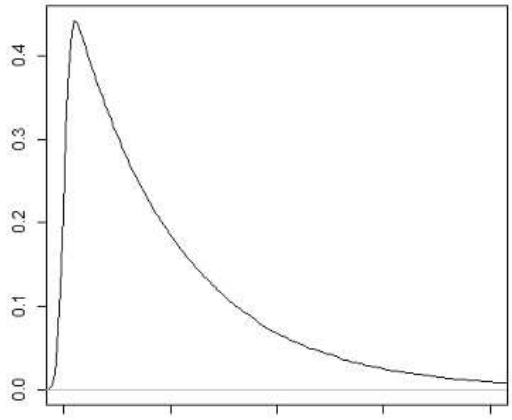
	Total_Ativo	Volume_Negocios	Capital_Proprio	NPS
Skewness	65	92	74	91
Kurtosis	4.757	8.928	5.958	8.638

Quer o coeficiente de assimetria (*skewness*), quer o coeficiente de achatamento (*kurtosis*) são utilizados para comparar a forma da distribuição tendo como referência a distribuição normal. Sempre que os valores absolutos destes coeficientes sejam superiores a um, pode assumir-se que a distribuição dos dados em causa se afasta da distribuição de referência.

A assimetria da distribuição pode ser caracterizada pelo enviesamento (*skewness*) que a distribuição apresenta em relação à média. Da análise dos coeficientes constata-se que para todas as variáveis quantitativas a distribuição se concentra no lado esquerdo com uma longa cauda para a direita (distribuições com enviesamento positivo ou assimetria à direita). O coeficiente de achatamento ou *kurtosis* apresentam valores positivos e muito diferentes de zero, as funções são leptocúrticas, ou seja, apresentam uma forma pontiaguda em relação à distribuição normal[41].

[41] Em relação à distribuição normal: i) existe maior probabilidade de ter valores próximos à média e ii) maior probabilidade de ter valores extremos.

Figura 8 – Função de distribuição assimétrica positiva e leptocúrtica



5.3 – RISCO DE DIVULGAÇÃO

A análise do risco de divulgação tem em consideração diferentes cenários de divulgação estatística.

Neste trabalho assume-se um cenário em que o intruso:

- Sabe quais são as empresas que se encontram na base de dados. Embora seja pouco realista assumir que conhece todos os elementos de uma base de dados de 10 mil empresas, pode assumir-se que ele sabe que um conjunto de empresas consta do ficheiro;
- Conhece os valores para o conjunto de variáveis chave: *Distrito*, *Forma Jurídica* e *Setor*. Esta informação encontra-se disponível no registo comercial e em várias bases de dados disponíveis na Internet, pelo que é plausível conhecer estes atributos.

Figura 9 – *Output* do *package sdcMicro*[42] do R (*k-anonimato*)

```
Number of observations violating
- 2-anonymity:  50 obs
- 3-anonymity:  98 obs
--------------------------
Percentage of observations violating
 - 2-anonymity:  0.5 %
- 3-anonymity:  0.98 %
```

[42] O *package sdcMicro* tem um conjunto de funções que permitem criar ficheiros de dados mascarados utilizando métodos perturbativos e não perturbativos. Por conveniência, no âmbito do presente trabalho, optou-se por utilizar a função 'freqCalc' para o cálculo do risco de divulgação.

No que concerne ao risco de divulgação conclui-se o seguinte:

- *k-anonimato* - 148 registos em que $k \leq 3$ considerando o conjunto as variáveis chave;
- *l-diversidade* - Não se aplica porque as variáveis sensíveis são quantitativas;
- *t-proximidade* - Para prevenir ataques de homogeneidade para cada combinação de variáveis chave calculou-se o coeficiente de variação (desvio-padrão / média) associado às variáveis quantitativas. Neste caso pretende-se identificar os registos que para a mesma combinação das variáveis *Distrito*, *Forma Jurídica* e *Setor*, as variáveis sensíveis (variáveis quantitativas – *Total Ativo*, *Volume Negócios*, *Capital Próprio* ou *NPS*) apresentam valores homogéneos. Se o coeficiente de variação for baixo, apesar do intruso não obter o valor exato da variável, pode estimar um valor aproximado uma vez que todas as empresas apresentam valores semelhantes (risco de inferência)[43]. Na base de dados não se encontraram combinações de variáveis chave que agreguem mais de três empresas (excluir as situações identificadas pelo *k-anonimato*) em que o coeficiente de variação é inferior a 30% para as variáveis sensíveis[44].

Utilizando os métodos de avaliação do risco foram identificados um total de 148 registos em risco em 10 mil empresas (1,48%).

- Total de registos em risco: 148 (*k-anonimato*) + 0 (*l-diversidade*) + 0 (*t-proximidade*).

Se for considerado o estabelecido por Hundepool *et al.* (2010) de que todos os registos são únicos para as variáveis quantitativas e o cenário de divulgação que esta informação está disponível para o intruso, o número de registos em risco é o número de observações da base de dados, ou seja 10 mil registos. Face ao elevado risco de identificação dos respondentes, a solução para a divulgação de bases de dados de empresas é a produção de ficheiros totalmente sintéticos, uma vez que nenhum registo original é divulgado.

No ponto seguinte são apresentados os métodos utilizados para a produção de ficheiros totalmente sintéticos, tal como as opções tomadas.

[43] No ponto 3.3 são explicados de forma detalhada os diferentes tipos de risco.
[44] Para valores superiores a 30% o coeficiente de variação, que capta a dispersão de uma distribuição, é considerado elevado.

5.4 – GERAÇÃO SINTÉTICA

Com o objetivo de produzir ficheiros protegidos utilizando a base de dados descrita nos pontos anteriores, neste estudo são propostos quadro modelos que se dividem em dois algoritmos (Árvores de Decisão e *Random Forests*) e duas operacionalizações (operacionalização 1 e 2) utilizando uma lógica de amostragem com reposição.

Em seguida são apresentados de forma detalhada a aplicação de cada um dos algoritmos e das operacionalizações à geração sintética de microdados.

5.4.1 – Geração Aleatória de Valores com Árvores de Decisão e *Random Forests*

O método de *Imputação Múltipla* é um dos mais aplicados na geração de informação sintética devido à forte fundamentação teórica e aos bons resultados obtidos por diferentes trabalhos. O pressuposto da imputação é assumir que existe uma relação entre as variáveis de um ficheiro de dados, conseguindo-se estimar os valores das variáveis partindo das restantes.

Para um melhor entendimento optou-se por comparar a solução obtida na resolução de um problema de classificação e previsão[45], onde os resultados são determinísticos - abordagem sem amostragem - com a produção de ficheiros sintéticos utilizando técnicas de amostragem - abordagem de amostragem com reposição.

Geração aleatória de valores com Árvores de Decisão

Quando se utiliza o algoritmo de Árvores de Classificação em variáveis qualitativas pode-se obter uma classificação para determinado elemento (classe maioritária) ou as probabilidades associadas para cada classe para esse elemento. Na geração de valores sintéticos utilizando amostragem considera-se a informação de probabilidades da folha em vez de se optar pela classe maioritária. Por exemplo, em determinada folha existem 10 registos em que 6 correspondem ao distrito do *Porto*, 3 correspondem a *Aveiro* e 1 corresponde a *Braga*. Num problema de classificação todos os 10 registos seriam classificados no *Porto*. Na produção de ficheiros sintéticos utilizando amostragem, a geração de valores é aleatória respeitando a distribuição de probabilidade da folha (neste caso 60% no *Porto*, 30% em *Aveiro* e 10% em *Braga*).

[45] As árvores de regressão são em tudo idênticas às árvores de decisão, a diferença principal reside no facto de as folhas das primeiras conterem previsões numéricas e não decisões.

Para as variáveis quantitativas o método é ligeiramente diferente. Por exemplo, se em determinada folha existirem 10 valores numéricos, na resolução de um problema de previsão é atribuído o valor médio. Na geração sintética utilizando amostragem para cada registo é selecionado um valor de entre os 10 e é aplicada uma estimativa de Kernel[46] com base na informação da variável na base de dados de modo a prevenir a divulgação de um valor original.

Em termos práticos, na geração sintética com amostragem para cada variável é construída uma Árvore de Decisão em função das restantes variáveis (que poderá variar de acordo com a operacionalização 1 ou 2). No passo seguinte cada registo é classificado numa folha e identificadas as respetivas probabilidades[47]. O passo final consiste em aplicar uma função que gera valores aleatórios com base nessas mesmas probabilidades. Repetindo o passo final obtém-se múltiplos valores para cada registo. Numa abordagem sem amostragem ou de classificação e previsão 'clássica' os resultados são determinísticos, logo apenas é possível obter um valor para cada registo.

Geração aleatória de valores com *Random Forests*

A geração de variáveis através do algoritmo *Random Forests* tem por base os resultados de classificação de cada uma das árvores. Para as variáveis qualitativas, a geração sintética utilizando amostragem respeita a probabilidade associada às previsões do total das árvores para aplicar posteriormente a técnica de amostragem com reposição. Por exemplo, se para 500 árvores geradas, 300 classificam o registo no *Porto*, 100 em *Braga*, 50 em *Aveiro* e 50 em *Viana do Castelo*, obtém-se uma probabilidade associada ao registo de 60%, 20%, 10% e 10%, respetivamente. Na geração sintética utilizando amostragem, estas probabilidades são utilizadas para gerar valores através de amostragem com reposição. Num problema de classificação, o registo seria classificado no *Porto* (voto maioritário).

Tal como para as variáveis qualitativas, para as variáveis quantitativas a aplicação do algoritmo vai gerar para cada registo um conjunto de resultados igual ao número de árvores definidas pelo utilizador. Na resolução de um problema de previsão o resultado seria a média destes valores. Na geração sintética utilizando amostragem, o valor vai ser

[46] Em estatística, estimativa de densidade Kernel (EDK) é uma forma não-paramétrica para estimar a função de densidade de probabilidade de uma variável aleatória. A estimação da densidade Kernel é um problema fundamental de suavização de dados onde inferências sobre a população são feitas com base numa amostra de dados finita (fonte: Guia de programação do software R, *Kernel Density Estimation*).
[47] Para as variáveis quantitativas cada valor da folha tem igual probabilidade de ser selecionado.

obtido através de amostragem utilizando os valores das árvores para o registo em causa. Com o objetivo de não facultar os valores originais, os valores são substituídos por uma estimativa de densidade Kernel.

A aplicação do algoritmo *Random Forests* na geração sintética utilizando amostragem é mais simples do que no algoritmo de Árvores de Decisão, pois a função utilizada permite parametrizar a obtenção da probabilidade associada a cada registo. Após a construção do modelo e da obtenção das probabilidades associadas foi aplicada a função que gera valores aleatórios para obter os valores sintéticos. Replicando esta função é possível obter múltiplos valores. Numa abordagem sem amostragem, tal como no algoritmo anterior, os resultados são determinísticos, sendo apenas possível gerar um valor para cada registo.

5.4.2 – Operacionalização dos Métodos

A aplicação da *Imputação Múltipla* compreende duas fases: i) a construção do modelo (fase da modelação[48]) e ii) a geração de valores sintéticos.

Neste estudo são utilizadas duas operacionalizações. Em ambas, a sintetização das variáveis é feita de forma sequencial e na fase de modelação, com o objetivo de manter a relação entre as variáveis, os modelos são sempre construídos utilizando a base de dados original. No entanto, enquanto na operacionalização 1 todos os atributos são sempre utilizados na definição do modelo, na operacionalização 2 apenas são utilizadas as variáveis já sintetizadas.

Operacionalização 1

Esta operacionalização segue o trabalho de Reiter (2005b) que foi o precursor na utilização de algoritmos de *data mining* para a geração sintética de microdados. Os trabalhos posteriores de Drechsler e Reiter (2010) e Lee *et al.* (2013) seguiram o especificado pelo autor.

Sendo r as variáveis a sintetizar, $X_{(0)}$ o vetor das variáveis com valores ainda não sintetizados e $X_{(i)}$ a variável *i-ésima* na síntese, a operacionalização da imputação processa-se da seguinte forma:

[48] Optou-se pelo termo modelação por ser mais genérico e ser aplicado a diferentes métodos. No âmbito dos algoritmos de Árvores de Decisão e *Random Forests* a designação mais adequada seria indução ou aprendizagem.

1) Executar o algoritmo para modelar $X_{(1)}$ em função de $X_{(0)}$. Obter $X'_{(1)}$ usando o sintetizador correspondente para $X_{(1)}$. Considere-se $X'_{(1)}$ o valor sintético de $X_{(1)}$.
2) Executar o algoritmo para modelar $X_{(2)}$ em função de $(X_{(0)}, X_{(1)})$. Utilizar os valores de $X'_{(1)}$e $X_{(0)}$ para prever os novos valores para $X_{(2)}$. Considere-se $X'_{(2)}$ o valor sintético de $X_{(2)}$.
3) Para cada *i*, onde *i* = 3, …, *r*, executar o algoritmo para modelar $X_{(i)}$ em função de $(X_{(0)}, X_{(1)},\ldots, X_{(i-1)})$. Substituir cada $X_{(i)}$ utilizando o sintetizador com base nos valores de $(X_{(0)}, X'_{(1)}, X'_{(2)},\ldots, X'_{(i-1)})$.

O resultado é um conjunto de dados sintéticos. Para obter *m* conjuntos, estas três etapas devem ser repetidas. A vantagem de obter mais do que um ficheiro de dados decorre da fundamentação do método da *Imputação Múltipla*: a qualidade da informação é superior se em vez de se gerar apenas um valor (imputação simples), for acrescentada aleatoriedade aos parâmetros e forem gerados múltiplos valores. Ao se gerar os valores para as variáveis através de amostragem a aleatoriedade é assegurada, permitindo gerar múltiplos ficheiros.

Operacionalização 2

Outra abordagem é a considerada nos trabalhos que utilizam regressões multivariadas sequenciais (Reiter, 2005a e Loong *et al.*, 2013) e que foi implementada recentemente em estudos que utilizam algoritmos de *data mining* em Caiola e Reiter (2010) e Rabb *et al*. (2015)[49]. Tal como foi referido anteriormente, a diferença desta operacionalização em relação à anterior é o facto de apenas se considerar na fase de modelação e de geração de valores sintéticos como previsores as variáveis já geradas sinteticamente em processos anteriores em vez de todas as variáveis da base de dados. Como no início do processo de sintetização ainda não existe qualquer variável gerada sinteticamente recorre-se à amostragem com reposição utilizando o peso dos valores observados no total da base de dados para se gerar a primeira variável.

Considerando a notação exposta anteriormente, esta abordagem consiste no seguinte:

[49] Esta abordagem é a utilizada no *package synthpop* do *software R*, que permite gerar sinteticamente bases de dados. O trabalho Rabb *et al*. (2015) é dos mesmos autores do *package* e tem por objetivo demonstrar as suas potencialidades na geração de valores sintéticos.

1) Gerar $X_{(1)}$ através de amostragem com reposição utilizando a função de probabilidades associadas aos valores observados. Considere-se $X'_{(1)}$ o valor sintético de $X_{(1)}$.
2) Executar o algoritmo para modelar $X_{(2)}$ em função $X_{(1)}$. Utilizar os valores de $X'_{(1)}$ para prever os novos valores para $X_{(2)}$. Considere-se $X'_{(2)}$ o valor sintético de $X_{(2)}$.
3) Para cada *i*, onde *i* = 3, …, *r*, executar o algoritmo para modelar $X_{(i)}$ em função $(X_{(1)}, X_{(2)},\ldots, X_{(i-1)})$. Substituir cada $X_{(i)}$ utilizando o sintetizador com base nos valores em $(X'_{(1)}, X'_{(2)},\ldots, X'_{(i-1)})$.

O resultado é um conjunto de dados sintéticos. Para obter *m* conjuntos, estas três etapas devem ser repetidas.

Para ficheiros parcialmente sintéticos, em alternativa ao passo 1 pode-se modelar $X_{(1)}$ utilizando as variáveis que não vão ser geradas sinteticamente como previsores e utilizar os seus valores para prever $X'_{(1)}$. Em ficheiros totalmente sintéticos, tal como referido anteriormente, no passo 1 como ainda não existem variáveis geradas sinteticamente, os valores são obtidos através de amostragem considerando as probabilidades associadas à primeira variável alvo de síntese.

5.4.3 – Resumo dos modelos propostos

Neste trabalho são propostos quadro modelos - dois algoritmos (Árvores de Decisão e *Random Forests*) e duas operacionalizações (operacionalização 1 e 2) - utilizando uma lógica de amostragem com reposição. Adicionalmente foram construídos mais quatro moledos baseados numa abordagem de classificação ou previsão ‘clássica’, ou seja, sem utilização de técnicas de amostragem. O objetivo destes quatro modelos adicionais é demonstrar as vantagens de utilizar amostragem com reposição na produção de ficheiros sintéticos.

Na figura 10 são enquadrados os oito modelos utilizados – quatro utilizando técnicas de amostragem e outros quatro que não utilizam amostragem. No sentido de tornar mais intuitivo o acompanhamento de cada um dos modelos optou-se por incluir na sua nomenclatura os fatores que os caracterizam: i) o tipo de abordagem (A+ e A- se utiliza ou não amostragem); ii) o algoritmo (AD e RF para Árvores de Decisão e *Random Forests*, respetivamente) e iii) o número correspondente à operacionalização (1 ou 2).

Figura 10 – Modelos de geração de microdados utilizados

Abordagem	Algoritmo	Operacionalização 1	Operacionalização 2
Sem Amostragem	*Árvores de Decisão (AD)*	A- AD_1	A- AD_2
	Random Forests (RF)	A- RF_1	A- RF_2
Amostragem com Reposição	*Árvores de Decisão (AD)*	A+ AD_1	A+ AD_2
	Random Forests (RF)	A+ RF_1	A+ RF_2

5.4.4 – Implementação dos Modelos

5.4.4.1 – *Software* utilizado

O *software* utilizado no presente estudo foi o R (versão 3.1.3) e o SAS Enterprise Guide (versão 6.1).

O R é uma linguagem de programação utilizada no processamento e tratamento de dados. Este *software* permite um eficiente tratamento e armazenamento de dados e tem uma linguagem de programação que se pode considerar simples e eficaz, podendo ser implementada em vários sistemas operativos. Mesmo não sendo uma ferramenta comercial, disponibiliza um conjunto de *packages*, permitindo a sua utilização em diversos problemas de diferentes áreas de conhecimento. Para a implementação dos diferentes modelos foram utilizados os *packages rpart* (Therneau *et al.*, 2010)[50] *e randomForest* (Liaw *et al.*, 2009). O primeiro foi utilizado para a construção de Árvores de Decisão e o último na aplicação do algoritmo com o mesmo nome, permitindo ambos abordar problemas de classificação e de regressão. Para a prossecução deste estudo foram ainda utilizadas funções do *package* s*dcMicro* (Templ, 2008) para a análise da questão da confidencialidade e do *package synthpop* (Nowok *et al.*, 2014) para a visualização gráfica dos ficheiros gerados sinteticamente.

O nome SAS resulta do acrônimo *Sistema de Análises Estatísticas* (*Statistical Analysis System*). O SAS é definido como um sistema integrado de aplicações para a análise de dados. No âmbito do presente trabalho, este *software* foi utilizado na avaliação dos ficheiros gerados sinteticamente, nomeadamente no apuramento de subconjuntos de

[50] O *rpart* é uma implementação do código aberto (*open-source)* do algoritmo *CART*.

dados mais frequentes e cálculo das medidas de sobreposição dos intervalos de confiança.

5.4.4.2 – Parametrização dos modelos

No que concerne à parametrização dos modelos, no algoritmo de ***Árvores de Decisão*** considerou-se um número mínimo de registos por nó final igual a cinco. Este valor é igual ao utilizado na literatura que aplica este algoritmo, funcionando como um garante da confidencialidade da informação. O valor cinco evita situações de sobre ajustamento (*overfitting*) aos dados e ao mesmo tempo permite uma diversidade de soluções, não sendo demasiado restritivo.

Em problemas 'clássicos' de Árvores de Decisão, o estabelecimento de regras de paragem (designada por pré-poda) previne a construção de árvores com muitos ramos e com pouco poder de generalização. No âmbito da geração de dados sintéticos, o objetivo é conseguir descrever de forma correta a relação das variáveis, garantindo ao mesmo tempo a confidencialidade, pelo que a dimensão da árvore é uma questão secundária. Para além da pré-poda, outra alternativa seria gerar uma árvore completa sobre ajustada aos dados e podar posteriormente. Esta solução tem como desvantagem ser uma parametrização mais lenta e como vantagem ser mais confiável, uma vez que se analisa e se decide podar ou não cada uma das subárvores.

Em termos do algoritmo ***Random Forests***, as parametrizações utilizadas foram a aprendizagem de 500 árvores, a utilização de todas as variáveis como previsores em cada árvore e a consideração de 2/3 da base de dados por árvore.

Em termos de ordenação das variáveis a sintetizar, optou-se por utilizar a ordem dos atributos no ficheiro de dados: i) *Distrito*; ii) *Forma Jurídica*; iii) *Setor*; iv) *Total do Ativo*; v) *Volume de Negócios*; vi) *Capital Próprio* e vii) *NPS*. Nesta ordenação as variáveis qualitativas são as primeiras a ser sintetizadas e posteriormente as quantitativas.

Dos trabalhos que utilizam algoritmos de *data mining*, apenas o trabalho de Caiola e Reiter (2010) refere a questão da ordenação das variáveis. Segundo os autores não existe nenhuma teoria a sugerir a sua ordenação, podendo diferentes formulações originar diferentes perfis de utilidade e risco. No entanto, os autores referem que quando duas ou mais variáveis apresentam o mesmo número de valores a substituir, uma solução é selecionar de forma aleatória a ordenação. Outra solução é imputar primeiro

as variáveis qualitativas com um número menor de valores por uma questão de eficiência computacional. Uma terceira alternativa é experimentar diferentes combinações e optar pela solução mais desejada (combinação utilidade / risco). De referir que em Caiola e Reiter (2010) apenas foram sintetizadas três variáveis qualitativas, tendo os autores aplicado três ordenações em seis combinações possíveis.

Finalmente, em termos de número de ficheiros, optou-se por gerar dez ficheiros totalmente sintéticos (*m*=10) para cada modelo que utiliza amostragem. Este número decorre, por um lado, do estabelecido na literatura acerca da *Imputação Múltipla* (a geração de múltiplos valores apresenta vantagens em relação à imputação simples) e por outro por se acreditar que um número mais elevado de imputações permite obter melhores resultados. Na literatura o número de ficheiros gerados sinteticamente varia entre os cinco[51] e dez[52]. No ponto 5.5.3 faz-se uma avaliação do impacto da variação deste número na qualidade da informação sintética.

No Anexo 5 são apresentados dois exemplos das linhas de código desenvolvidas para este trabalho. Por parcimónia optou-se por facultar um exemplo para cada algoritmo e operacionalização, utilizando uma abordagem com reposição – modelos A+AD_2 e A+RF_1.

5.4.5 – Análise de Múltiplos Ficheiros Gerados Sinteticamente

Uma vez facultado um conjunto de ficheiros totalmente sintéticos, o utilizador precisa de conhecer algumas regras no sentido de combinar a informação dos *m* ficheiros.

Segundo Drechsler (2009), de modo a obter um valor *Q* referente à população, que pode ser por exemplo a média de uma variável, o coeficiente de correlação entre duas variáveis ou o coeficiente de regressão numa regressão linear, para cada base de dados sintética d_i, o utilizador da informação deve calcular uma estimativa q e a sua variância v.

Sendo $i = i, \dots, m$ (onde *m* é o número de ficheiros sintéticos gerados) considerando q_i e v_i como sendo os valores respetivos de q e v em d_i, os seguintes valores devem ser calculados para obter inferências para *Q*:

[51] Trabalhos que consideram m=5: Reiter (2005b), Caiola e Reiter (2010), Drechsler e Reiter (2010) e Loong *et al*. (2010)
[52] Trabalhos que utilizam *m*=10: Reiter (2005a) e Lee *et al.* (2013)

$$q_m = \sum_{i=1}^{m} \frac{q_i}{m} \tag{8}$$

$$b_m = \sum_{i=1}^{m} \frac{(q_i - q_m)^2}{(m-1)} \tag{9}$$

$$v_m = \sum_{i=1}^{m} \frac{v_i}{m} \tag{10}$$

O valor q_m deve ser utilizado como um estimador de Q e a variância (v_{syn}) de q_m resulta da seguinte aplicação:

$$v_{syn} = \left(1 + \frac{1}{m}\right) b_m - v_m \tag{11}$$

Inferências para Q podem ser obtidas utilizando a distribuição *t-student* com os seguintes graus de liberdade (Reiter, 2002):

$$GL = (m-1)\left(1 - \frac{1}{r_m}\right)^2, \text{ em que } r_m = \frac{\left(1+\frac{1}{m}\right)b_m}{v_m} \tag{12}$$

A estimativa com um intervalo de confiança (*IC*) de $100\ (1-\alpha)\%$ para Q é dada pela seguinte formula:

$$IC = q_m \pm t_{student\ \alpha/2}\sqrt{v_{syn}} \tag{13}$$

Quando *n* é elevado, as inferências para o escalar *Q* podem ser baseadas na distribuição normal, logo um intervalo de confiança de 95% para *Q* é dado seguindo a formulação (Reiter, 2005a):

$$IC_{95\%} = q_m \pm 1{,}96\sqrt{v_{syn}} \tag{14}$$

À semelhança do que acontece na literatura, neste trabalho considera-se que *n* é elevado, pelo que se utiliza a aproximação à distribuição normal (equação 14).

Estas equações aplicam-se a ficheiros totalmente sintéticos. Para ficheiros parcialmente sintéticos os cálculos da variância e dos graus de liberdade da estatística *t* diferem. A demonstração destas equações pode ser consultada em Raghunathan *et al.* (2003) para ficheiros totalmente sintéticos e em Reiter (2003) para ficheiros parcialmente sintéticos.

A sobreposição dos intervalos de confiança (J_q) é uma medida citada e utilizada na literatura para medir a perda de informação. Esta medida foi definida por Karr *et al.* (2006) e é calculada da seguinte forma para uma estimativa q:

$$J_q = \frac{1}{2}\left(\frac{U_{over,q} - L_{over,q}}{U_{0,q} - L_{0,q}} + \frac{U_{over,q} - L_{over,q}}{U_{syn,q} - L_{syn,q}}\right) \quad (15)$$

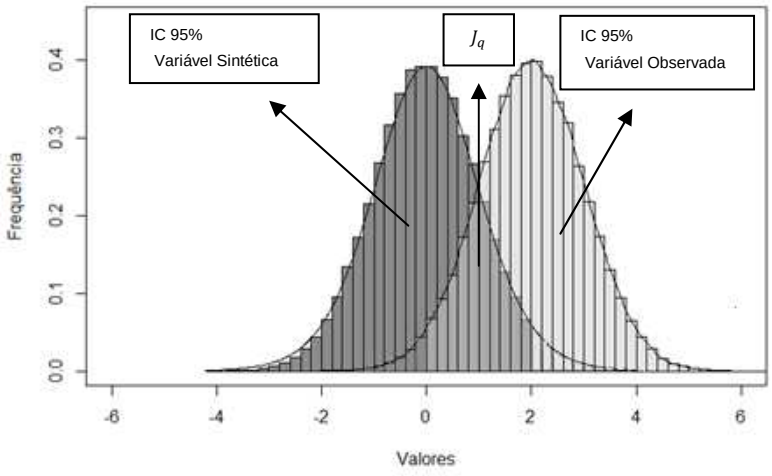
em que $(L_{0,q}, U_{0,q})$ correspondem aos limites inferiores e superiores do intervalo de confiança para a estimativa q utilizando a base de dados original ($q_{obs} \pm 1{,}96\sqrt{v_{obs}}$), $(L_{syn,q}, U_{syn,q})$ são os valores correspondentes da informação sintética e $(L_{over,q}, U_{over,q})$ são os pontos inferiores e superiores de intersecção de ambos os intervalos de confiança.

Valores elevados de sobreposição correspondem aos seguintes valores:

$$0{,}9 \leq J_q \leq 1 \quad (16)$$

Quando os intervalos de confiança do ficheiro original e do sintético são idênticos, corresponde a uma utilidade elevada, J_q é próximo de um. Quando os intervalos não se sobrepõem está-se perante uma baixa utilidade, J_q é zero. O segundo termo da equação 15 diferencia entre intervalos de igual utilidade mas de diferentes amplitudes, beneficiando obviamente o menor.

Figura 11 – Representação da sobreposição de Intervalos de Confiança (J_q)



5.5 – AVALIAÇÃO DOS FICHEIROS GERADOS SINTETICAMENTE

A avaliação de ficheiros gerados sinteticamente assenta na comparação da base de dados original com a base de dados sintética. Esta avaliação foi efetuada utilizando duas abordagens complementares: comparação univariada e multivariada.

A ***comparação univariada*** é a mais simples e comtemplou os oito modelos de geração sintética de microdados. Nesta avaliação comparou-se frequências (variáveis qualitativas) e estatísticas descritivas (variáveis quantitativas) entre os ficheiros gerados sinteticamente e a base de dados original. No final foi realizado um teste estatístico para validar se o ficheiro sintético e o original seguem a mesma distribuição.

A ***comparação multivariada*** foi efetuada apenas para os modelos que demonstraram ter qualidade na comparação univariada. Nesta avaliação calculou-se o número de combinações novas e que desapareceram nos ficheiros sintéticos considerando as variáveis qualitativas, comparou-se as matrizes de variâncias e covariâncias e de correlação do ficheiro original e do ficheiro sintético e procedeu-se ao cálculo da sobreposição dos intervalos de confiança para as doze combinações de variáveis qualitativas mais representadas na base de dados.

Neste trabalho foram calculados um total de 44 ficheiros sintéticos, um por cada modelo que não utiliza amostragem[53] e dez (valor de *m*) para cada modelo que utiliza a técnica de amostragem.

A análise dos múltiplos ficheiros gerados sinteticamente é efetuada utilizando as formulações do ponto 5.4.4, que correspondem aos procedimentos que um utilizador tem de efetuar para analisar um conjunto de ficheiros sintéticos.

Na figura 12, a título de exemplo, são apresentados os cinco primeiros registos do ficheiro original e de um ficheiro sintético gerado através de amostragem com reposição, para o algoritmo de Árvores de Decisão considerando a operacionalização 1 (A+AD_1). A introdução de aleatoriedade permite gerar um novo ficheiro de dados diferente do original.

[53] A lógica que não utiliza amostragem é determinística permitindo apenas gerar um ficheiro.

Figura 12 – Ficheiro original vs ficheiro sintético – 5 primeiros registos

N	Distrito	Forma_Juridica	Setor	Total_Ativo	Volume_Negocios	Capital_Proprio	NPS
Ficheiro original							
1	Porto	Sociedade Anónima	K	597.565.834	16.153.326	4.171.561.742	35
2	Vila Real	Sociedade por Quotas	Q	15.576.612	64.01	-4.517.073	3
3	Aveiro	Sociedade Anónima	C	950.475.807	1.081.779.057	259.111.316	89
4	Viana do Castelo	Sociedade por Quotas	G	1.513.846	66.921.787	1.433.845	9
5	Porto	Sociedade por Quotas	F	53.599.762	26.041.141	26.981.753	4
Ficheiro sintético - modelo A+AD_1							
1	Porto	Sociedade Anónima	C	9.328.260.516	1.266.319.131	3.041.022.298	40
2	Braga	Sociedade por Quotas	F	18.471.697	69.704.961	59.804.748	2
3	Braga	Sociedade Anónima	C	806.649.078	740.671.545	-3.567.825	51
4	Porto	Sociedade por Quotas	G	11.288.005	7.441.279	39.428.899	1
5	Porto	Sociedade por Quotas	N	51.293.766	12.544.306	-2.051.397	21

5.5.1 – Comparação Univariada

5.5.1.1 – Estatísticas descritivas

Uma forma simples de averiguar a qualidade dos ficheiros gerados sinteticamente é calcular o valor de algumas estatísticas. Para as variáveis qualitativas a avaliação passa pelo cálculo de frequências e para as variáveis quantitativas optou-se por calcular medidas de localização (média, mediana e quartis) e dispersão (variância ou desvio-padrão). Em cada uma das tabelas é apresentado o valor do ficheiro original e o obtido para cada um dos modelos.

De modo a tornar a apresentação dos resultados mais fluída optou-se por separar a abordagem sem amostragem (A-AD_1, A-AD_2, A-RF_1, A-RF_2) da amostragem com reposição (A+AD_1, A+AD_2, A+RF_1, A+RF_2).

No corpo do texto principal são apresentadas apenas as tabelas de frequências da variável *Distrito* e as estatísticas descritivas das variáveis quantitativas. Os resultados das restantes variáveis qualitativas são idênticos, pelo que as respetivas tabelas de frequência e alguns exemplos de histogramas obtidos com o *package synthpop* são apresentados em anexo (Anexo 6).

Neste ponto a análise resulta da comparação das estatísticas univariadas dos ficheiros gerados sinteticamente com as do ficheiro original. Esta análise é completada e validada utilizando o teste não paramétrico de qualidade do ajustamento.

Estatísticas descritivas dos modelos sem amostragem

Tabela 7 – Tabela de frequências da variável *Distrito* (sem amostragem)

unidades

		Original	A- AD_1	A- AD_2	A- RF_1	A- RF_2
Distrito	Aveiro	928	0	898	342	949
	Braga	2.283	0	2.307	2.029	2.221
	Bragança	273	0	280	61	228
	Guarda	15	0	15	0	23
	Porto	5.341	10.000	5.287	7.250	5.419
	Viana do Castelo	550	0	596	181	533
	Vila Real	417	0	418	99	418
	Viseu	193	0	199	38	209

Da análise da tabela 7 constata-se que uma abordagem que não utiliza amostragem implica perdas de informação elevadas. A razão para estes resultados reside no facto da informação do ficheiro original se encontrar desbalanceada, pelo que os algoritmos tendem a classificar os registos na classe maioritária.

Os modelos A-AD_1 e A-AD_2 e A-RF_2 classificam pelo menos uma variável qualitativa em apenas uma classe, alterando a estrutura da informação. Nestas situações os algoritmos não encontraram formas alternativas de classificação, concluindo que classificar todos os elementos na classe dominante implica uma probabilidade de erro menor. Por exemplo, classificar todos os elementos no distrito do *Porto* implica uma taxa de acerto de 53% e classificar todos os registos na forma jurídica *Sociedade por Quotas* corresponde a uma taxa de 93%. Apesar do modelo A-RF_1 não ter classificado todos os registos na mesma classe, constata-se, à semelhança dos restantes, uma preponderância das classes maioritárias em relação ao ficheiro original.

Tabela 8 – Estatísticas descritivas das variáveis quantitativas (sem amostragem)

	Modelo	Média	1.º Quartil	Mediana	3.º Quartil	Desvio Padrão	Máximo	Mínimo
Total_Ativo (10^3)	**Original**	2.105	58	165	502	46.708	3.786.039	0
	A- AD_1	2.110	681	681	681	42.509	2.369.554	681
	A- AD_2	499	499	499	499	0	499	499
	A- RF_1	1.916	83	192	534	31.339	2.333.644	12
	A- RF_2	727	545	900	900	189	900	401
Volume_Negocios (10^3)	**Original**	1.334	53	138	419	34.292	3.333.909	0
	A- AD_1	1.091	466	466	466	24.670	1.684.440	466
	A- AD_2	551	551	551	551	0	551	551
	A- RF_1	1.293	93	186	541	18.291	1.539.084	17
	A- RF_2	812	695	946	946	152	946	508
Capital_Proprio (10^3)	**Original**	916	4	41	163	26.713	2.306.983	-29.848
	A- AD_1	1.027	251	251	251	22.494	1.257.121	251
	A- AD_2	251	251	251	251	0	251	251
	A- RF_1	1.041	20	57	186	13.880	816.600	-1.600
	A- RF_2	222	156	276	276	59	276	132
NPS (unidades)	**Original**	12	2	3	8	222	21.383	1
	A- AD_1	9	6	6	6	165	11.622	6
	A- AD_2	6	6	6	6	0	6	6
	A- RF_1	12	3	5	9	76	5.483	1
	A- RF_2	9	8	10	10	1	10	6

Em termos de variáveis quantitativas (Tabela 8), os resultados confirmam a disparidade entre os ficheiros sintéticos e o ficheiro original: i) para o modelo A-AD_2 todos os valores sintetizados são iguais (desvio-padrão = 0); ii) apenas o modelo A-RF_1 gerou valores negativos para a variável *Capital Próprio* (como no ficheiro original) e iii) a amplitude e o desvio padrão dos valores gerados para as variáveis quantitativas em todos os modelos é inferior nos ficheiros sintéticos.

Estatísticas descritivas dos modelos de amostragem com reposição

Tabela 9 – Tabela de frequências da variável *Distrito* (amostragem com reposição)

unidades

		Original	A+ AD_1	A+ AD_2	A+ RF_1	A+ RF_2
Distrito	Aveiro	928	**928**	**920**	949	920
	Braga	2.283	**2.277**	**2.281**	2.294	2.281
	Bragança	273	**267**	**276**	287	276
	Guarda	15	**17**	**15**	17	15
	Porto	5.341	**5.344**	**5.327**	5.269	5.327
	Viana do Castelo	550	**551**	**563**	561	563
	Vila Real	417	**417**	**417**	427	417
	Viseu	193	**199**	**200**	196	200

Nota: valores médios resultantes de 10 ficheiros sintéticos (m =10) - equação 8

Tabela 10 – Estatísticas descritivas das variáveis quantitativas (amostragem com reposição)

	Modelo	Média	1.º Quartil	Mediana	3.º Quartil	Desvio Padrão	Máximo	Mínimo
Total_Ativo (10³)	**Original**	2.105	58	165	502	46.708	3.786.039	0
	A+ AD_1	**2.069**	**57**	**164**	**500**	**42.259**	**2.929.990**	**0**
	A+ AD_2	**2.171**	**58**	**164**	**498**	**43.607**	**2.847.464**	**0**
	A+ RF_1	1.813	58	160	479	36.526	2.966.490	0
	A+ RF_2	728	544	910	910	196	910	395
Volume_Negocios (10³)	**Original**	1.334	53	138	419	34.292	3.333.909	0
	A+ AD_1	**1.417**	**53**	**137**	**418**	**37.142**	**3.045.164**	**0**
	A+ AD_2	**1.308**	**52**	**138**	**422**	**30.923**	**2.756.418**	**0**
	A+ RF_1	1.070	54	136	405	20.854	1.872.844	0
	A+ RF_2	1.054	721	1.391	1.391	366	1.391	499
Capital_Proprio (10³)	**Original**	916	4	41	163	26.713	2.306.983	-29.848
	A+ AD_1	**918**	**4**	**41**	**164**	**25.977**	**1.880.922**	**-24.334**
	A+ AD_2	**1.008**	**4**	**41**	**163**	**29.048**	**1.973.676**	**-28.898**
	A+ RF_1	771	6	44	165	19.404	1.626.966	-11.501
	A+ RF_2	262	161	349	349	93	349	129
NPS (unidades)	**Original**	12	2	3	8	222	21.383	1
	A+ AD_1	**12**	**2**	**3**	**8**	**174**	**13.861**	**1**
	A+ AD_2	**13**	**2**	**3**	**8**	**237**	**19.718**	**1**
	A+ RF_1	10	2	3	7	87	7.670	1
	A+ RF_2	9	8	10	10	2	10	6

Nota: valores médios resultantes de 10 ficheiros sintéticos (m =10) - equação 8

A geração sintética utilizando amostragem com reposição obtém resultados próximos do ficheiro original para os modelos A+AD_1 e A+AD_2 que utilizam o algoritmo de Árvores de Decisão e A+RF_1 que utiliza o algoritmo *Random Forests*.

Os resultados obtidos com a aplicação do modelo A+RF_2 são substancialmente diferentes dos obtidos para o ficheiro original. Estes resultados derivam do facto de na operacionalização 2 as variáveis utilizadas como previsores serem apenas as previamente sintetizadas. Considerando a ordem de imputação, constatou-se que a variável *Distrito* tem pouca capacidade de prever a *Forma Jurídica*, todas as 500 árvores classificaram os elementos na forma jurídica *Sociedade por Quotas* (Anexo 6). Uma vez que na operacionalização 2 utiliza-se na sintetização da variável seguinte apenas as variáveis já sintetizadas como previsores, os problemas foram propagados para toda a base de dados.

5.5.1.2 – Teste de qualidade do ajustamento

Para avaliar a qualidade da informação obtida por cada um dos oito modelos e validar as conclusões do ponto anterior é pertinente efetuar um teste não paramétrico à qualidade do ajustamento. Para as variáveis discretas (qualitativas) utilizou-se o teste do Qui-

quadrado (χ^2) e para as variáveis contínuas (quantitativas) o teste de Kolmogorov-Smirnov (K-S). Em ambas as situações o objetivo é averiguar se o ficheiro sintético segue a mesma distribuição do ficheiro original.

O princípio básico do teste χ^2 é comparar as possíveis divergências entre as frequências observadas e esperadas para um certo evento. Pode-se afirmar que dois grupos se comportam de forma semelhante se as diferenças entre as frequências observadas e as esperadas em cada categoria forem próximas a zero.

O teste K-S é aplicado a variáveis continuas e baseia-se na comparação da função de distribuição. Este teste determina o ponto onde se verifica a maior distância vertical entre as duas funções (a empírica e a de teste). A distribuição da estatística do teste K-S é definida de forma rigorosa, sendo considerado mais potente do que o χ^2 quando aplicado a variáveis contínuas.

A hipótese nula (H0) e a hipótese alternativa (H1) são as seguintes para ambos os testes:

H0: O ficheiro sintético e o original seguem a mesma distribuição

H1: O ficheiro sintético e o original não seguem a mesma distribuição

Os *p-values* dos testes de qualidade de ajustamento são apresentados na tabela 11. Para um *p-value* superior a 0,05 não se rejeita a hipótese nula para um nível de significância de 5% (valores a negrito na tabela).

Tabela 11 – Teste de qualidade de ajustamento (*p-values*)

Abordagem		**Sem Amostragem**				**Amostragem com Reposição ***			
Modelo		**A- AD_1**	**A- AD_2**	**A- RF_1**	**A- RF_2**	**A+ AD_1**	**A+ AD_2**	**A+ RF_1**	**A+ RF_2**
χ^2	Distrito	0,000	**0,542**	0,000	**0,520**	**0,537**	**0,554**	**0,592**	**0,554**
	Forma_Juridica	0,000	0,000	0,000	0,000	**0,399**	**0,715**	**0,607**	0,000
	Setor	0,000	0,000	0,000	0,000	**0,634**	**0,492**	**0,699**	0,000
K-S	Total_Ativo	0,000	0,000	0,000	0,000	**0,737**	**0,828**	**0,348**	0,000
	Volume_Negocios	0,000	0,000	0,000	0,000	**0,782**	**0,874**	**0,321**	0,000
	Capital_Proprio	0,000	0,000	0,000	0,000	**0,852**	**0,825**	0,001	0,000
	NPS	0,000	0,000	0,000	0,000	**0,865**	**0,911**	**0,596**	0,000

* Corresponde à média dos *p-values* obtidos para 10 ficheiros sintéticos (m=10) - equação 8.
Nota: No Anexo 7 são apresentados os *p-values* obtidos para cada um dos ficheiros individuais.

De acordo com o esperado e o testemunhado no ponto anterior (comparação univariada), os modelos que utilizam uma abordagem sem amostragem produzem resultados inferiores aos seus equivalentes que aplicam uma lógica de amostragem com reposição. No modelo A+RF_2 a hipótese nula também é rejeitada para a quase totalidade das variáveis e o modelo A+RF_1 não consegue replicar a distribuição do atributo *Capital Próprio*, levando à rejeição da hipótese nula para esta variável para um nível de significância de 5%.

A não rejeição da hipótese nula para a variável *Distrito* nos algoritmos que seguem a operacionalização 2 deve-se ao facto deste atributo ser gerado utilizando amostragem, pois segundo esta operacionalização apenas são consideradas nos modelos de imputação as variáveis previamente sintetizadas. Como a variável *Distrito* é a primeira a ser sintetizada, a opção foi gerar a variável através de amostragem respeitando as frequências associadas a cada valor, para depois poder ser considerada na estimação da variável seguinte.

Face aos resultados superiores obtidos pelos modelos A+AD_1, A+AD_2 e A+RF_1 em relação aos restantes, por parcimónia apenas tem sentido continuar a avaliação considerando estes modelos. Conclui-se que uma abordagem sem amostragem e com amostragem com reposição utilizando o algoritmo *Random Forests* numa operacionalização tipo 2 (modelo A+RF_2) tem impactos significativos na utilidade da informação, pelo que estes não são bons métodos de geração sintética de bases de dados de empresas.

5.5.2 - Comparação Multivariada

5.5.2.1 – Variáveis qualitativas

A avaliação multivariada dos atributos qualitativos passa por avaliar o número de combinações novas e que desaparecem nos ficheiros sintéticos em relação ao ficheiro original. Quanto maior o número de combinações e o número de registos que lhe estão associados, maior a perda de informação resultante do processo de sintetização.

Na tabela seguinte são apresentados os valores médios, o máximo e o mínimo para o total dos dez ficheiros gerados para cada modelo (equação 8).

Tabela 12 – Avaliação das variáveis qualitativas (*m*=10)

unidades

	N.º de combinações			N.º de registos		
Modelo	**Média**	**Máximo**	**Mínimo**	**Média**	**Máximo**	**Mínimo**
Combinações novas						
A+ AD_1	44	63	33	177	414	257
A+ AD_2	59	106	41	608	1.323	583
A+ RF_1	**37**	**54**	**26**	**65**	**70**	**35**
Combinações que desapareceram						
A+ AD_1	39	57	28	**159**	615	36
A+ AD_2	47	94	31	383	1.264	42
A+ RF_1	**32**	**46**	**21**	247	**576**	**25**

Num total de 216 combinações presentes no ficheiro original em 608 possíveis[54], o melhor resultado (modelo A+RF_1) corresponde a uma taxa média de combinações novas e que desapareceram de 17% (37 em 216) e 15% (32 em 216), respetivamente. O impacto em termos de registos é bastante menor, cifrando-se nos 0,65% (65 em 10 mil) e 2,47% (247 em 10 mil), respetivamente. Este modelo é apenas superado em termos de número de registos associados às combinações que desaparecem pelo modelo A+AD_1. Os melhores valores obtidos para cada métrica estão representados a negrito na tabela 12.

Os modelos que utilizam os algoritmos de Árvores de Decisão (modelo A+AD_1 e A+AD_2) tiveram mais dificuldades em reproduzir o ficheiro original, apresentando taxas médias de combinações novas de 20% (44 em 216) e 27% (59 em 216) e taxas de combinações que desapareceram de 18% (39 em 216) e 21% (47 em 216), respetivamente. Em termos de registos, os valores diminuem para 2% (177 em 10 mil) e 6% (608 em 10 mil) para as combinações novas e 2% (159 em 10 mil) e 4% (383 em 10 mil) para as combinações que desaparecem.

Os resultados permitem concluir que existe uma maior dificuldade dos modelos em reproduzirem as combinações menos representadas. Estes resultados já eram esperados e decorrem da própria formulação dos modelos que consideram um número mínimo de registos por folha. De destacar o facto do modelo com melhores resultados (modelo A+RF_1) o impacto em termos de número de registos ser bastante reduzido e da superioridade da operacionalização 1 nos modelos que utilizam Árvores de Decisão.

[54] 8 * 4 * 19 (8 distritos, 4 formas jurídicas e 19 setores).

5.5.2.2 – Variáveis quantitativas

De acordo com o estabelecido por Domingo-Ferrer e Tora (2001), para medir a utilidade dos ficheiros sintéticos para as variáveis quantitativas foram calculadas medidas de distâncias entre a base de dados original e os dados sintetizados utilizando as matrizes de variâncias e covariâncias (*V* e *V'* no Anexo 1) e correlação (*R* e *R'* no Anexo 1).

Este processo compreendeu diferentes fases: i) calcular a matriz de variâncias e covariâncias e correlação para cada um dos ficheiros gerados sinteticamente (total de dez para cada modelo) e para o ficheiro original; ii) estimar uma matriz por modelo de geração sintética calculando a média dos dez valores para cada elemento da matriz (equação 8); iii) calcular a distância entre a matriz do ficheiro original e a matriz de cada modelo estimada no ponto anterior utilizando diferentes métricas. No Anexo 8 são apresentadas as matrizes de variâncias e covariâncias e de correlação para o ficheiro original e para cada um dos modelos.

Tabela 13 – Medidas de utilidade – variáveis quantitativas (*m*=10)

Modelo	Erro quadrático médio	Erro absoluto médio	Variação Média
matriz de variâncias e covariâncias			
A+ AD_1	5,6 2E+22	217.002.014	0,9950
A+ AD_2	**3,14 E+22**	**153.128.917**	0,9953
A+ RF_1	20,27 E+22	366.828.101	**0,8915**
matriz de correlação			
A+ AD_1	0,0523	0,2378	0,9439
A+ AD_2	0,0504	0,2132	0,9480
A+ RF_1	**0,0364**	**0,1792**	**0,6641**

A hierarquização dos modelos em termos de utilidade difere de acordo com a métrica utilizada (a negrito os melhores resultados – quanto menor a distância entre o ficheiro original e o sintético mais fielmente o ficheiro sintético reproduz as relações do ficheiro original). Quando se utiliza a matriz de variâncias e covariâncias e como métrica o erro quadrático médio e o erro absoluto médio, os ficheiros gerados utilizando o modelo A+AD_2 são os que apresentam uma menor distância em relação ao ficheiro original.

Os ficheiros que utilizam o modelo A+RF_1 são os que obtém melhores resultados quando se considera a matriz de correlação e para a métrica variação média considerando a matriz de variâncias e covariâncias.

5.5.2.3 – Análise baseada em intervalos de confiança

No sentido de avaliar a relação entre as variáveis qualitativas e quantitativas foram calculados os intervalos de confiança para o conjunto de combinações de variáveis mais representadas. Calculou-se a sobreposição dos intervalos de confiança em 95% (equação 15) entre os ficheiros sintéticos e o original para as doze combinações de variáveis qualitativas mais representadas. Quanto maior a sobreposição dos intervalos de confiança (valores próximos de um) maior é a capacidade dos ficheiros sintéticos reproduzirem as estatísticas do ficheiro original. Esta métrica é uma das mais utilizadas na literatura para avaliar ficheiros gerados sinteticamente utilizando a *Imputação Múltipla*.

Estes doze intervalos representam cerca de 60% (6.017) dos registos, onde a combinação mais representada agrega 15% dos registos (1.470) e a menos apenas 2% (234). As médias dos valores obtidos para as doze combinações em dez ficheiros sintéticos (*m*) são apresentadas na tabela seguinte. No Anexo 9 encontram-se os resultados para cada uma das doze combinações por modelo.

Tabela 14 – Média da sobreposição dos intervalos de confiança a 95% (*m*=10)

		Média da sobreposição I.C. a 95%			
Modelo	**Frequência**	**Ativo**	**Volume de Negócios**	**Capital Próprio**	**NPS**
A+ AD_1	**12**	0,794	0,755	0,675	0,682
A+ AD_2	**12**	0,760	0,759	0,704	0,708
A+ RF_1	**12**	**0,856**	**0,903**	**0,740**	**0,916**

Da análise da tabela verifica-se que o modelo A+RF_1 é o que obtém melhores resultados em todas as variáveis (valores a negrito), seguido do modelo A+AD_2. O modelo A+RF_1 obtém uma sobreposição dos intervalos de confiança superior a 90% para as variáveis *Volume de Negócios* e *NPS* e superior a 85% para o *Ativo*. O *Capital Próprio* em comparação às restantes variáveis é o que apresenta uma sobreposição inferior em todos os modelos.

5.5.3 – Número de Ficheiros Sintéticos

Importa também avaliar em que medida o número de ficheiros sintéticos tem impacto nos resultados obtidos. Na literatura o número de ficheiros sintéticos varia entre os cinco e os dez: Reiter (2005a) e Lee *et al.* (2013) consideraram dez ficheiros sintéticos e Reiter (2005b), Caiola e Reiter (2010), Drechsler e Reiter (2010) e Loong *et al.* (2010) consideraram apenas cinco. Neste trabalho optou-se por gerar dez ficheiros sintéticos por se acreditar que um número maior de imputações iria aumentar a qualidade da informação. Para medir o impacto do número de ficheiros sintéticos considerou-se quatro alternativas: três, cinco, sete e dez.

5.5.3.1 – Comparação univariada

Na avaliação do impacto da variação de *m* é utilizada a média das variáveis quantitativas como referência e a sobreposição dos intervalos de confiança para um nível de significância de 95% (equação 15) entre o ficheiro original e os ficheiros sintéticos.

Tabela 15 – Avaliação univariada da variação do número de ficheiros sintéticos

		m=3		m=5		m=7		m=10	
	Modelo	**Média**	**Sobreposição I.C.**	**Média**	**Sobreposição I.C.**	**Média**	**Sobreposição I.C.**	**Média**	**Sobreposição I.C.**
	Original	2.105							
Total_Ativo	**A+ AD_1**	**2.256**	**0,946**	**2.116**	**0,993**	**2.087**	**0,967**	**2.069**	**0,962**
(10^3)	**A+ AD_2**	1.721	0,836	1.832	0,840	**2.167**	**0,985**	**2.171**	**0,995**
	A+ RF_1	**1.917**	**0,940**	**1.837**	**0,913**	**1.821**	**0,908**	1.813	0,898
	Original	1.334							
Volume_Negocios	**A+ AD_1**	1.548	0,892	**1.449**	**0,927**	**1.414**	**0,945**	**1.417**	**0,942**
(10^3)	**A+ AD_2**	1.166	0,813	**1.251**	**0,901**	**1.328**	**0,970**	**1.308**	**0,978**
	A+ RF_1	989	0,795	1.081	0,888	**1.821**	**0,908**	1.813	0,898
	Original	916							
Capital_Proprio	**A+ AD_1**	794	0,877	773	0,866	**825**	**0,929**	**918**	**0,978**
(10^3)	**A+ AD_2**	**877**	**0,964**	**960**	**0,942**	**1.077**	**0,901**	**1.088**	**0,936**
	A+ RF_1	**870**	**0,968**	**822**	**0,937**	**806**	**0,909**	771	0,880
	Original	12							
NPS	**A+ AD_1**	10	0,608	11	0,866	11	0,883	**12**	**0,973**
(unidades)	**A+ AD_2**	**12**	**0,995**	**13**	**0,906**	**13**	**0,927**	**13**	**0,946**
	A+ RF_1	9	0,560	10	0,811	10	0,824	10	0,774

Para a generalidade dos modelos, os valores de sobreposição são superiores a 90% (a negrito na tabela), indo de encontro ao estabelecido anteriormente sobre a qualidade dos modelos. Em termos de número de ficheiros sintéticos, os resultados confirmam o estabelecido na literatura: perante um número pequeno de imputações (entre três e dez) consegue-se obter bons resultados.

No que concerne aos modelos, o modelo A+AD_1, para 10 imputações obtém os melhores resultados, as quatro variáveis quantitativas apresentam valores para a sobreposição dos intervalos de confiança superiores a 90%, das quais três apresentam valores acima dos 95%.

O modelo A+AD_2 para sete imputações já apresenta as quatro variáveis com valores de sobreposição superiores a 90%. No entanto, para valores de *m*=10 regista-se uma melhoria para todas as variáveis.

O modelo A+RF_1 é o que apresenta piores taxas de sobreposição. O melhor resultado é obtido com sete imputações, onde três das quatro variáveis apresentam valores maiores que 90%.

Apesar destas conclusões, em algumas situações pontuais para valores de *m* iguais a três e cinco encontram-se taxas de sobreposição superiores a 90%.

5.5.3.2 – Comparação multivariada

Replicando a análise efetuada anteriormente para as doze combinações de variáveis qualitativas mais representadas os resultados são os seguintes (a negrito as sobreposições superiores a 90%).

Tabela 16 – Avaliação multivariada da variação do número de ficheiros sintéticos

		Média da sobreposição I.C. a 95%			
m	**Modelos**	**Ativo**	**Volume de Negócios**	**Capital Próprio**	**NPS**
3	**A+ AD_1**	0,763	0,743	0,695	0,703
	A+ AD_2	0,747	0,734	0,718	0,720
	A+ RF_1	0,846	**0,904**	0,757	**0,926**
5	**A+ AD_1**	0,765	0,751	0,693	0,741
	A+ AD_2	0,736	0,738	0,720	0,716
	A+ RF_1	0,850	0,879	0,775	**0,913**
7	**A+ AD_1**	0,764	0,753	0,702	0,717
	A+ AD_2	0,750	0,742	0,726	0,712
	A+ RF_1	0,813	**0,902**	0,745	**0,933**
10	**A+ AD_1**	0,794	0,755	0,675	0,682
	A+ AD_2	0,760	0,759	0,704	0,708
	A+ RF_1	0,856	**0,903**	0,740	**0,916**

Para as doze combinações de variáveis qualitativas mais representadas confirma-se os resultados obtidos anteriormente, ou seja, o número de imputações com melhores resultados é entre sete e dez para todos os modelos. No entanto, nesta análise vislumbram-se diferenças entre as variáveis analisadas: para a variável *Ativo* e *Volume*

de Negócios, o número de imputações com melhores resultados são dez, mas para a variável *Capital Próprio* e *NPS* as sete imputações são as que apresentam uma sobreposição dos intervalos de confiança superiores para a generalidade dos modelos.

5.5.4 – Conclusões da Avaliação dos Ficheiros Sintéticos

Nos pontos precedentes efetuou-se uma avaliação dos ficheiros gerados sinteticamente resultantes de diferentes modelos, pretende-se agora apresentar um resumo do que foi descrito anteriormente e retirar algumas conclusões.

Os métodos que aplicam uma ***abordagem sem amostragem*** (A-AD_1, A-AD_2, A-RF_1, A-RF_2) implicam perdas significativas de utilidade, gerando ficheiros onde os valores das variáveis são concentrados nos valores maioritários, alterando de forma significativa a estrutura dos dados (teste de qualidade do ajustamento). Estes resultados vão de encontro o esperado e visam mostrar as potencialidades e a necessidade de recorrer à amostragem para a produção de ficheiros sintéticos.

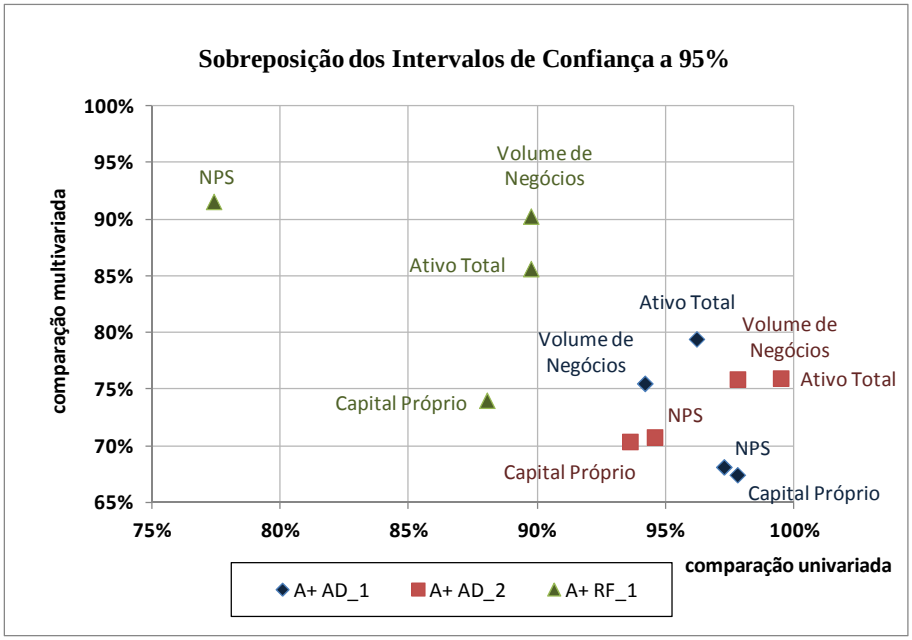
Com exceção do modelo A+RF_2, a utilização de algoritmos de Árvores de Decisão e *Random Forests* aplicando uma ***abordagem de amostragem com reposição*** permite criar ficheiros sintéticos que reproduzem as estatísticas univariadas das variáveis sem perdas significativas de utilidade. O cálculo de frequências, estatísticas descritivas e a aplicação de um teste estatístico para medir a qualidade do ajustamento permitiram concluir que as variáveis geradas sinteticamente seguem a mesma distribuição do ficheiro original para a generalidade das variáveis.

Os modelos que utilizam ***Árvores de Decisão*** (A+AD_1 e A+AD_2) obtêm os melhores resultados univariados, apresentando para as variáveis quantitativas valores de sobreposição dos intervalos de confiança superiores a 90%. No modelo A+AD_1 três das quatro variáveis quantitativas apresentam valores superiores a 95% e no modelo A+AD_2 apenas duas.

O modelo que utiliza o algoritmo ***Random Forests*** (A+RF_1), apesar de obter os piores resultados univariados de entre os três modelos em que não se rejeitou a hipótese dos ficheiros sintéticos e dos ficheiro original seguirem a mesma distribuição (A+AD_1, A+AD_2 e A+RF_1) é o que de forma mais consistente consegue reproduzir a relação entre as variáveis, quer ao nível das combinações das variáveis qualitativas, quer ao nível de variáveis quantitativas.

Na figura 13 é apresentado um gráfico que resume os resultados da comparação univariada para as variáveis quantitativas (eixo das abcissas) e multivariada das doze combinações mais frequentes (eixo das ordenadas) para a sobreposição dos intervalos confiança a 95% para a geração de dez ficheiros sintéticos. Todos os valores do gráfico já foram apresentado anteriormente, pretende-se agora efetuar uma representação bidimensional dos resultados obtidos.

Figura 13 – Representação gráfica da sobreposição dos intervalos de confiança a 95%



O canto superior direito do gráfico representa uma sobreposição total dos intervalos de confiança dos resultados (univariados e multivariados) e o canto inferior esquerdo o cenário oposto. O objetivo da geração sintética é que as variáveis se concentrem no canto superior direito. O gráfico demonstra o que foi referido anteriormente, ou seja, uma melhor capacidade do algoritmo *Random Forests* reproduzir a relação entre as variáveis (eixo das ordenadas) e do algoritmo de Árvores de Decisão capta as estatísticas univariadas (eixo das abcissas).

6 - CONCLUSÕES E COMENTÁRIOS FINAIS

Perante as limitações dos métodos tradicionais de perturbação dos dados, quer devido às consequências nefastas na qualidade da informação, quer devido ao falhanço do objetivo primordial de proteção da identidades dos respondentes, os métodos de controlo de divulgação estatística evoluíram para a aplicação de metodologias mais sofisticadas. A produção de ficheiros sintéticos apareceu como uma solução credível, uma vez que se garante a confidencialidade da informação, já que nenhum registo original é divulgado, com impactos pouco significativos na utilidade dos dados.

A geração de ficheiros sintéticos utilizando o método da *Imputação Múltipla* apresenta uma forte fundamentação teórica e é uma das metodologias mais utilizadas[55]. Neste trabalho utilizaram-se algoritmos de *data mining*, mais precisamente os algoritmos de Árvores de Decisão e *Random Forests*, para gerar ficheiros sintéticos aplicando a metodologia de *Imputação Múltipla* desenvolvida por Rubin (1993). Os métodos não paramétricos conseguem lidar com diferentes tipos de informação de elevada dimensionalidade, obter relações não lineares e iterações que muito dificilmente são captadas pelas abordagens paramétricas. Uma vez que a modelação se processa de forma semiautomática, estes métodos implicam um menor investimento na determinação da relação entre as variáveis e na construção dos modelos face aos métodos paramétricos.

A criação de ficheiros totalmente sintéticos para uma base de dados de empresas constituiu um desafio neste estudo. Os trabalhos que aplicam a *Imputação Múltipla* habitualmente comtemplam apenas um conjunto limitado de registos ou um número restrito de variáveis, na sua maioria qualitativas, no sentido de manterem a utilidade da informação. A informação das empresas comporta riscos adicionais de identificação dos respondentes, devido à existência de uma diversidade de bases de dados disponíveis ao público, pelo que a geração de ficheiros parcialmente sintéticos comportaria um risco elevado de identificação dos respondentes.

[55] Exemplos de trabalhos que utilizam a *Imputação Múltipla* na produção de ficheiros sintéticos: Reiter, (2005a), Reiter (2005b), Drechsler (2009), Drechsler e Reiter (2010), Loong *et al.* (2013) e Raab *et al.*(2015).

Os resultados obtidos permitem concluir que a utilização dos algoritmos de Árvores de Decisão e *Random Forests* são eficazes na obtenção de ficheiros sintéticos que reproduzem na generalidade as propriedades estatísticas da base de dados original.

Da avaliação dos resultados, constata-se que os três modelos (A+AD_1, A+AD_2 e A+RF_1) reproduzem de forma competente a distribuição do ficheiro original (teste de qualidade do ajustamento). Apesar dos modelos que utilizam Árvores de Decisão obterem os melhores resultados univariados, o modelo *Random Forests* é o que de forma mais consistente consegue reproduzir a relação entre as variáveis, quer ao nível das combinações das variáveis qualitativas, quer ao nível de variáveis quantitativas. Em termos de número de ficheiros sintéticos, conclui-se que o número ideal se situa entre as sete e as dez imputações, variando de acordo com o modelo e as variáveis consideradas.

Em termos de investigação ainda existe algum ceticismo na utilização deste tipo bases de dados, mas a questão da confidencialidade é cada vez mais premente, pelo que a opção será muitas vezes entre a utilização da informação sintética ou nenhuma informação. Obviamente que é impossível garantir que os ficheiros sintéticos consigam reproduzir de forma exata toda e qualquer análise do ficheiro original e ao mesmo tempo garantir a confidencialidade dos respondentes. Para desmistificar a falta de qualidade dos ficheiros sintéticos, os centros de dados numa atitude pedagógica podem facultar aos utilizadores os resultados obtidos utilizando o ficheiro original no sentido de se comparar os valores e certificar que os resultados são semelhantes e as conclusões permanecem inalteradas.

No que concerne a potenciais contributos para a literatura, este estudo concluiu que vale a pena considerar algoritmos de *data mining* para a produção de ficheiros totalmente sintéticos de bases de dados de empresas que congregam variáveis quantitativas e qualitativas, resultados semelhantes aos existentes para ficheiros parcialmente sintéticos utilizando bases de dados de indivíduos (pessoas singulares). Ao se aplicar dois algoritmos e duas operacionalizações distintas na mesma base de dados, este trabalho comparou quatro alternativas metodológicas, permitindo retirar conclusões importantes sobre os impactos da sua utilização na produção de ficheiros sintéticos. Os trabalhos existentes utilizam apenas algumas destas alternativas que são aplicadas a bases de dados distintas, tornando os resultados muito difíceis de comparar. Ao se identificar e

agrupar as diferentes abordagens, este trabalho embora não propondo uma abordagem inovadora contribui para a literatura também numa perspetiva metodológica.

Por fim importa apontar algumas linhas de futura investigação. Embora seja difícil lidar ao mesmo tempo com variáveis qualitativas e contínuas no mesmo algoritmo, a aplicação de algoritmos de *data mining* à produção de ficheiros sintéticos não se esgota neste trabalho, pelo que se sugere a utilização de outras especificações. Visando suprir a limitação descrita sugere-se a possibilidade de considerar mais do que um algoritmo na produção dos mesmos ficheiros sintéticos, por exemplo considerar um algoritmo para as variáveis quantitativas e outro para as variáveis qualitativas. A ordem de imputação das variáveis e trabalhos que utilizem bases de dados de empresas são áreas que permanecem pouco exploradas e carecem de mais trabalhos. Ainda como linha futura de investigação e no âmbito do algoritmo de Árvores de Decisão sugere-se na fase de modelação a aprendizagem de árvores completas, ou seja, sem utilização de estratégias de paragem, e posteriormente aplicar estratégias de poda de modo a evitar o sobre ajustamento (*overfitting*) do modelo aos dados.

REFERÊNCIAS BIBLIOGRÁFICAS

Atkinson, E. J., & Therneau, T. M. (2015). An introduction to recursive partitioning using the RPART routines. Rochester: Mayo Foundation.

Bethlehem, J. G., Keller, W. J., & Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85(409), 38-45.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Caiola, G., & Reiter, J. P. (2010). Random Forests for Generating Partially Synthetic, Categorical Data. *Transactions on Data Privacy*, 3(1), 27-42.

Cestnik, B., Kononenko, I., & Bratko, I. (1987, May). ASSISTANT 86: A Knowledge-Elicitation Tool for Sophisticated Users. *In EWSL* (pp. 31-45).

Ciriani, V., Di Vimercati, S. D. C., Foresti, S., & Samarati, P. (2007). Microdata protection. In *Secure data management in decentralized systems* (pp. 291-321). Springer US.

Código de Conduta para as Estatísticas Europeias - Adotado pelo Comité do Sistema Estatístico Europeu em 28 de setembro de 2011.

Council Regulation (EURATOM, EEC) No 1588/90 of 11 June 1990 on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities, Official Journal of the European Union, OJ No L151, 15.6.1990.

Dandekar, R., Cohen, M., & Kirkendall, N. (2001). Applicability of Latin Hypercube Sampling to create multi variate synthetic micro data. Proceedings of ETK-NTTS, Eurostat, Luxemburg, 839-847.

Dandekar, R. A., Domingo-Ferrer, J., & Sebé, F. (2002). LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In *Inference Control in Statistical Databases* (pp. 153-162). Springer Berlin Heidelberg.

Domingo-Ferrer, J., & Torra, V. (2001). Disclosure control methods and information loss for microdata. Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies, 91-110.

Domingo-Ferrer, J., & Torra, V. (2005). Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2), 195-212.

Domingo-Ferrer, J., Drechsler, J., & Polettini, S. (2009). ESSNET-SDC Deliverable Report on Synthetic Data Files.

Drechsler, Jörg. Generating Multiply Imputed Synthetic Datasets: *Theory and Implementation*. Doctoral dissertation, 2009.

Drechsler, J., & Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105(492), 1347-1357.

Drechsler, J., & Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, 55(12), 3232-3243.

Duncan, G. T., Keller-McNulty, S. A., & Stokes, S. L. (2001). Disclosure risk vs. data utility: The RU confidentiality map. In *Chance*.

Gama, C. F., Carvalho, A., Oliveira, M., Faceli, K., & Lorena, A. (2012). *Extração de Conhecimento de Dados*, Data Mining.

Graham, P., & Penny, R. (2007). Multiply imputed synthetic data files. *Official Statistics Research* series, 1.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., ... & De Wolf, P. P. (2010). *Handbook on statistical disclosure control*. *ESSnet on Statistical Disclosure Control*.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & De Wolf, P. P. (2011). *Statistical disclosure control*. John Wiley & Sons..

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., & Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60(3), 224-232.

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics-Stockholm*, 9, 313-313.

Lane, Julia (2003). Uses of Microdata: Keynote Speech. Statistical Confidentiality and Access to Microdata. Proceedings of the Seminar Session of the 2003 Conference of European Statisticians. United Nations.

Lee, J. H., Kim, I. Y., & O'Keefe, C. M. (2013). On regression-tree-based synthetic data methods for business data. *Journal of Privacy and Confidentiality*, 5(1), 5.

Lei do Sistema Estatístico Nacional (SEN) Lei 22/2008. Diário da República 2008 (1ª série de 13 de Maio.

Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007*. ICDE 2007. IEEE 23rd International Conference on (pp. 106-115). IEEE.

Liaw, A., Wiener, M., Breiman, L., & Cutler, A. (2009). Package "randomForest". Retrieved December, 12, 2009.

Loong, B. (2012). Topics and Applications in Synthetic Data. Doctoral dissertation, Harvard University

Loong, B., Zaslavsky, A. M., He, Y., & Harrington, D. P. (2013). Disclosure control using partially synthetic data for large-scale health surveys, with applications to CanCORS. *Statistics in medicine*, 32(24), 4139-4161.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 3.

Maragoudakis, M., & Loukis, E. (2012). Using Ensemble Random Forests for the extraction and exploitation of knowledge on gas turbine blading faults identification. *OR insight*, *25*(2), 80-104.

Mateo-Sanz, J. M., Sebé, F., & Domingo-Ferrer, J. (2004a, January). Outlier protection in continuous microdata masking. In *Privacy in Statistical Databases* (pp. 201-215). Springer Berlin Heidelberg.

Mateo-Sanz, J. M., Martínez-Ballesté, A., & Domingo-Ferrer, J. (2004b, January). Fast generation of accurate synthetic microdata. In *Privacy in Statistical Databases* (pp. 298-306). Springer Berlin Heidelberg.

Matwin, S., Nin, J., Sehatkar, M., & Szapiro, T. (2015). A Review of Attribute Disclosure Control. In *Advanced Research in Data Privacy* (pp. 41-61). Springer International Publishing.

Matthews, G. J., & Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, 5, 1-29.

Nowok, Beata, Gillian M. Raab, and Chris Dibben (2014). "synthpop: Bespoke creation of synthetic data in R."

Pagliuca, D., and G. Seri. "Some results of individual ranking method on the system of enterprise accounts annual survey." Esprit SDC Project, Deliverable MI-3 D 2 (1999): 1999.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1), 81-106.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers Inc. San Francisco, USA.

Raab, Gillian, Beata Nowok, and Chris Dibben. "A simplified approach to generating synthetic data for disclosure control." arXiv preprint arXiv:1409.0217 (2015).

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*-Stockholm, 19(1), 1-16.

Refaat, M. (2010). *Data preparation for data mining using SAS*. Morgan Kaufmann.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of official Statistics* -, 18(4), 531-544.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2), 181-188.

Reiter, J. P. (2005a). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society*: Series A (Statistics in Society), 168(1), 185-205.

Reiter, J. P. (2005b). Using CART to generate partially synthetic public use microdata. *Journal of official Statistics* - Stockholm -, 21(3), 441.

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of official Statistics*, *9*(2), 461-468.

SABI (2015), Empresas com sede na região norte do país. Informação retirada em 10/03/2015, https://sabi.bvdinfo.com/version-2015310/home.serv?product=sabineo.

Skinner, C. J., & Holmes, D. J. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14(4), 361-372.

Skinner, C. (2009). Statistical disclosure control for survey data. *Handbook of statistics*, 29, 381-396.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty*, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

Templ, M. (2008). Statistical disclosure control for microdata using the R-package sdcMicro. na.

Templ, M., Meindl, B., & Kowarik, A. (2014). Introduction to Statistical Disclosure Control (SDC).

Therneau, T. M., Atkinson, B., & Ripley, B. (2010). rpart: Recursive Partitioning. R package version 3.1-46. *Computer software program retrieved from http://CRAN. R-project. org/package= rpart.*

Torgo, L. Aprendizagem Indutiva de Modelos de Regressão baseados em Árvores. Tese de Doutoramento. LIACC. Universidade do Porto.

Truta, T. M., & Vinay, B. (2006, April). Privacy Protection: p-Sensitive k-Anonymity Property. In ICDE workshops (p. 94).

UK White Paper on Open Government (1993), HM Government.

Viana, I. (2014). Método de Geração de Dados Sintéticos para a Criação de Microdados de Uso Público. Tese de Mestrado. FEP.

Winkler, W. E. (2005). Re-identification methods for evaluating the confidentiality of analytically valid microdata. *Statistics*, 09.

Winkler, W. E. (2007). Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified. Washington, DC: Statistical Research Division, US Bureau of the Census. Tech. Rep.

Yancey, W. E., Winkler, W. E., & Creecy, R. H. (2002). Disclosure risk assessment in perturbative microdata protection. In *Inference control in statistical databases* (pp. 135-152). Springer Berlin Heidelberg.

ANEXOS

A.1 MEDIDAS DE UTILIDADE DE MICRODADOS – VARIÁVEIS QUANTITATIVAS

	Erro quadrático médio	Erro absoluto médio	Variação média
X-X'	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\lvert x_{ij}-x'_{ij}\rvert}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{np}$
V-V'	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(v_{ij}-v'_{ij})^2}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert v_{ij}-v'_{ij}\rvert}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{p(p+1)}{2}}$
R-R'	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(r_{ij}-r'_{ij})^2}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert r_{ij}-r'_{ij}\rvert}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert r_{ij}-r'_{ij}\rvert}{\lvert r_{ij}\rvert}}{\frac{p(p-1)}{2}}$
RF-RF'	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}(rf_{ij}-rf'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\lvert rf_{ij}-rf'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\frac{\lvert rf_{ij}-rf'_{ij}\rvert}{\lvert rf_{ij}\rvert}}{p^2}$
C-C'	$\frac{\sum_{i=1}^{p}(c_i-c'_i)^2}{p}$	$\frac{\sum_{i=1}^{p}\lvert c_i-c'_i\rvert}{p}$	$\frac{\sum_{i=1}^{p}\frac{\lvert c_i-c'_i\rvert}{c_i}}{p}$
F-F'	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}(f_{ij}-f'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\lvert f_{ij}-f'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\frac{\lvert f_{ij}-f'_{ij}\rvert}{\lvert f_{ij}\rvert}}{p^2}$

Fonte: Domingo-Ferrer e Tora, 2001

A.2 ESTUDOS QUE GERARAM FICHEIROS DE MICRODADOS SINTÉTICOS

Artigo	Informação sintética total ou parcial ?	Tipo de variáveis	Base de dados	Técnica utilizada	Resultados e principais evidências
Mateo-Sanz *et al.* (2004b)	Totalmente sintético	Quantitativas	US Census Bureau - 13 variáveis quantitativas.	Decomposição de Cholesky (método que permite preservar a média, a variância e as covariâncias das variáveis)	Este método é apenas aplicável às variáveis quantitativas. Segundo os autores as principais características deste método são: i) permite um diferente número de registos no ficheiro original e no ficheiro sintético; ii) a complexidade computacional é linear ao n.º de registos; iii) é um método não iterativo, pelo que o tempo pode ser previsto antes da execução (ao contrário de outros - Hipercubo Latino de Amostragem); iv) o estudo empírico demonstrou que o risco de divulgação é menor do que os métodos tradicionais (produção de ficheiros mascarados).
Reiter (2005a)	Apesar de mencionarem que o ficheiro é totalmente sintético o autor não gerou 3 variáveis demográficas que funcionaram como input's aos modelos.	Qualitativas e Quantitativas	US Current Population Survey (março de 2000) - 10 variáveis (sexo, etnia, estado civil, habilitações literárias, idade, pensão de alimentos, recebimentos da segurança social, impostos sobre o património, rendimento da família e pensão de alimentos para o cônjuge).	Imputação de regressões multivariadas sequenciais - *Sequential Regression Multivariate Imputation* : o autor apenas sintetizou 7 das 10 variáveis (todas as regressões incluem as variáveis demográficas sexo, etnia e idade). O processo de gerar variáveis é sequencial, ou seja, em cada regressão são utilizadas as 3 variáveis demográficas e as variáveis sintetizadas anteriormente se existirem.	O estudo coloca ênfase na necessidade de definir de forma correta o modelo de imputação. O autor demonstrou um grande conhecimento da base de dados, nomeadamente da relação causal entre as variáveis. No processo foram testadas e implementadas diferentes regressões para subconjuntos de elementos da população (por exemplo: separou os membros da família com idades inferiores e superiores a 15 anos e criou modelos com especificações diferentes para cada subconjunto).
Reiter (2005b)	Parcialmente sintético (apenas sintetizam um conjunto de registos)	Qualitativas e Quantitativas	US Current Population Survey (março de 2000) - 10 variáveis (sexo, etnia, estado civil, habilitações literárias, idade, pensão de alimentos, recebimentos da segurança social, impostos sobre o património, rendimento da família e pensão de alimentos para o cônjuge).	Algoritmo de Árvores de Decisão para as variáveis quantitivas e qualitativas para um conjunto de registos (variáveis com valores superiores a determinado limite mantendo os restantes registos).	Os ficheiros parcialmente sintéticos são atrativos porque permitem manter os benefícios dos ficheiros totalmente sintéticos em termos de confidencialidade e superam as dificuldade de obtenção de um ficheiro totalmente sintéticos plausível (segundo o autor difícil de operacionalizar).
Caiola e Reiter (2010)	Parcialmente sintético (todos os registos com informação em determinada variável)	Qualitativas	US Current Population Survey (março de 2000) - 12 variáveis (sexo, etnia, estado civil, habilitações literárias, idade, pensão de alimentos, recebimentos da segurança social, impostos sobre o património, rendimento da família, pensão de alimentos para o cônjuge, n.º de pessoas na família e n.º de dependentes).	Algoritmo Random Forests para variáveis qualitativas (etnia, estado civil e sexo). Apesar de considerarem a idade uma variável identificadora e qualitativa, esta não foi sintetizada devido ao facto de existir uma multiplicidade de valores (varia de 0 a 90).	O algoritmo teve uma performance razoável quando aplicado às variáveis qualitativas. A relação entre variáveis nos grupos mais pequenos é mais difícil de manter do que nos grandes grupos. A ordem de imputação das variáveis determinou alteração dos resultados. Os autores concluem que é difícil saber qual a melhor ordem de imputação, mesmo em trabalhos que utilizam a técnica da imputação múltipla no tratamento de valores omissos.
Drechsler e Reiter (2010)	Parcialmente sintético (apenas sintetizam um conjunto de registos em risco)	Qualitativas	US Current Population Survey (março de 2000) - 10 variáveis (sexo, etnia, estado civil, habilitações literárias, idade, recebimentos da segurança social, impostos sobre o património, rendimento da família, n.º de pessoas na família e n.º de dependentes).	Algoritmo de Árvores de Decisão para as variáveis qualitativas (idade, etnia e estado cívil) e para um conjunto de registos em risco. Utilizaram na definição do modelo todas as variáveis que não são alvo de intervenção mais as variáveis geradas em processos anteriores. De modo a manter a coerência da informação construiram o modelo de imputação utilizando apenas os registos em risco.	Os autores apenas sintetizaram um conjunto de registos em risco (1089/51016=2%) de modo a manterem a utilidade do ficheiro a disponibilizar.

Artigo	Informação sintética total ou parcial ?	Tipo de variáveis	Base de dados	Técnica utilizada	Resultados e principais evidências
Lee *et al.* (2013)	Ficheiro parcilamente sintético (todos os registos com informação em determinada variável)	Quantitativas	Sugar Farms Data (1982) survey of the sugar cane industry in Queensland, Australia - 5 variáveis (região, area, custo, colheita e receitas).	Algoritmo de Árvores de Decisão aplicado apenas às variáveis quantitativas da base de dados e algorimo de Árvores de Decisão para cada uma das 4 regiões separadamente.	Obtiveram resultados similares numa análise exploratória univariada considerando as bases de dados sintéticas e a base de dados original. Os algoritmos de Árvores de Decisão subestimaram as correlação entre as variáveis. Os melhores resultados foram obtidos para um valor de k=2 (número mínimo de registos em cada nó folha final), o que pode comprometer a confidencialidade dos dados. Os autores concluem que as bases de dados de empresas possuem uma estrutura de correlações muito forte entre as variáveis, exacerbado pela presença de outliers, e que a aplicação imediata de algoritmos baseados em árvores de regressão no contexto de base de dados de empresas pode não ser possível.
Loong *et al.* (2013)	Ficheiro parcilamente sintético (todos os registos com informação em determinada variável)	Qualitativas	CanCORS (Cancer Care Outcomes and Surveillance) - 350 variáveis (variáveis demográficas: Idade (dividida em 5 grupos), estado cívil, etnia, sexo, habilitações literátias e variáveis clínicas: restantes).	Imputação de regressões multivariadas sequenciais - *Sequential Regression Multivariate Imputation* : cada modelo continha em média 50 variáveis. Os autores criaram 2 modelos diferentes para os diferentes tipos de doença (cancro nos pulmões e cancro coloretal) porque a relação entre as variáveis difere entre as diferentes doenças.	Este estudo replicou outros 2 estudos que utilizaram a base de dados original, concluindo que os resultados obtidos foram semelhantes. Os autores referem que a informação sintética deve ser vista pelos utilizadores externos como uma forma de disponibilizar a informação balanceando a coerência com o risco de divulgação. Mesmo na presença de 'síntese imperfeita', os utilizadores podem usar a informação para ganhar familiaridade com a estrutura dos dados e formularem e testarem hipóteses preliminares sem os custos envolvidos no acesso à informação original.
Raab *et al.* (2015)	Totalmente sintético	Qualitativas e Quantitaivas	UK Longitudinal Studies - 5 variáveis referentes aos anos de 1991 e 2001 - idade, sexo, estado cívil, etnia e doença de longa duração). Apenas a idade é variável quantitativa.	Algoritmo de Árvores de Decisão e métodos paramétricos (modelos de regressão logística e modelos de regressão linear).	Trabalho que utiliza o package synthpop do R. O algoritmo de Árvores de Decisão obteve melhores resultados do que os métodos paramétricos, não tendo sido necessário customizar a aplicação do algoritmo à base de dados. Os métodos paramétricos tiveram resultados menos satisfatórios na reprodução a distribuição marginal de algumas variáveis.

A.3 OUTROS MÉTODOS DE GERAÇÃO SINTÉTICA

1. Distorção dos dados através da distribuição de probabilidade

Segundo Hundepool *et al*. (2010), a distorção de informação através da distribuição de probabilidade foi o percursor na obtenção de uma base de dados sintética. Esta técnica consiste em construir uma base de dados utilizando as funções de distribuição das variáveis, podendo ser utilizado em variáveis quantitativas e qualitativas.

A distorção dos dados através da distribuição de probabilidades compreende três passos: i) identificar a função densidade probabilidade teórica subjacente para as variáveis chave e confidenciais no conjunto de dados e estimar os parâmetros associados; ii) gerar uma série aleatória obtida a partir da função densidade teórica para cada variável e iii) mapear a série alterada com a série original e substituir os valores originais pelos valores protegidos.

A dificuldade deste método é encontrar a função de probabilidade teórica aplicável às variáveis. A solução é experimentar diferentes funções e aplicar um teste de qualidade de ajustamento. Na ausência de uma função teórica, a opção é utilizar as frequências relativas das variáveis. Para as variáveis contínuas é necessário proceder à sua descritização, calcular as frequências associadas e depois gerar valores que as respeitem.

Hundepool *et al*. (2010) fazem a ressalva que este método foi proposto para gerar uma variável. Segundo os autores, a utilização de uma função de probabilidade multivariada torna o método mais complexo, quer no momento da definição da função, quer no mapeamento final e pode implicar perdas de informação consideráveis.

2. Bootstrap

Os dados sintéticos podem ser ainda gerados utilizando técnicas de amostragem com reposição. Segundo Hundepool *et al*. (2010), este método tem alguma similaridade com a distorção dos dados através da distribuição de probabilidades e o método da *Imputação Múltipla*. Sendo X a base de dados de microdados com p-variáveis, pode-se calcular a função de probabilidade acumulada (ou simplesmente função de distribuição) F para cada uma das p variáveis. Em vez de se distorcer a informação original para obter um ficheiro protegido, pode-se alterar a função de distribuição F para F'. O passo final é aplicar técnicas de amostragem (*bootstraping*) para obter um ficheiro sintético Z.

Uma vez alterada a função de distribuição, a informação gerada por este método vai ser diferente da original, podendo diferir de forma significativa caso a distorção aplicada na função de distribuição seja elevada.

3. Hipercubo latino de amostragem

A técnica assente no hipercubo latino de amostragem[56] permite gerar dados sintéticos que reproduzem as características univariadas (atributos considerados isoladamente) da base de dados original (médias e covariâncias). Dandekar *et al.* (2001) aplicaram a técnica utilizando a matriz de correlação da ordenação dos valores para reproduzir não só estatísticas univariadas mas também multivariadas (manter a correlação da ordenação). Segundo Hundepool *et al*. (2010) e Mateo-Sanz *et al.* (2004b), esta técnica é bastante demorada mesmo para amostras pequenas, uma vez que o método tem por base o refinamento iterativo, pelo que o processamento depende do número de variáveis a serem processadas e do valor utilizado para iniciar o processo. Esta técnica pode ser utilizada em variáveis quantitativas e qualitativas.

4. Decomposição de *Cholesky*

Mateo-Sanz *et al.* (2004b) propõem um algoritmo para gerar variáveis quantitativas que reproduz a média e a matriz de covariâncias do ficheiro original e consequentemente os coeficientes de correlação entre as variáveis.

O algoritmo proposto consiste em partir de uma matriz de *n* registos e *m* variáveis (matriz *A*), calcular a matriz de covariâncias (matriz *C*) do ficheiro original *X* e aplicar a decomposição de *Cholesky* de modo a gerar uma matriz *U*.

$$C = U^t \times U \qquad (17)$$

(onde *U* é uma matriz triangular e U^t é a sua transposta)

Depois obtém-se o ficheiro sintético *X'* através do produto das matrizes *A* e *U*.

$$X' = \mathrm{A} \times U \qquad (18)$$

[56] No contexto de amostragem estatística, uma grade quadrada contendo as posições da amostra é um quadrado latino se (e somente se) há apenas uma amostra em cada linha e em cada coluna. Um hipercubo latino é a generalização deste conceito para um número arbitrário de dimensões, em que cada amostra é única em cada hiperplano (fonte: Guia de programação do software R, *Random Latin Hypercube*).

Demonstra-se que a matriz de variâncias e covariâncias de *X'* é igual à de *X*.

Segundo os autores, as vantagens deste método são a simplicidade do algoritmo, quando comparado com o método da *Imputação Múltipla*, e um tempo computacional linear em relação ao número de registos. O facto de apenas ser aplicável a variáveis quantitativas constitui uma limitação.

5. Mínimo Erro Absoluto das Combinações das Variáveis (MEAC)

Viana (2014) apresenta uma metodologia de geração de ficheiros totalmente sintéticos que utiliza probabilidades condicionais.

Esta metodologia, apesar dos bons resultados obtidos em relação a um ficheiro de variáveis discretas criado a partir do ficheiro original de variáveis qualitativas e quantitativas, tem como principal limitação o facto de apenas ser aplicável a variáveis qualitativas (ou discretas). Em Viana (2014) as variáveis quantitativas são discretizadas e tratadas como variáveis qualitativas. Neste método cada variável é gerada de forma sequencial utilizando a probabilidade condicionada às variáveis sintetizadas anteriormente e o número de combinações existentes no ficheiro original, minimizando o erro absoluto entre as observações dos dados originais e as observações geradas.

O MEAC obtém bons resultados quando aplicado a variáveis qualitativas, conseguindo preservar a estrutura do ficheiro de dados. No entanto, a sua aplicação a variáveis quantitativas implica perdas significativas de utilidade, uma vez que o processo de transformação destas variáveis origina a agregação e a simplificação da informação do ficheiro original com impactos na qualidade do ficheiro de dados.

A.4 DESCRIÇÃO DAS VARIÁVEIS

Designação	Classificação	Unidade	Descrição
Contribuinte	Qualitativa	N.A.	Número de Identificação Fiscal (nif) de pessoa coletiva
Nome	Qualitativa	N.A.	Nome da empresa
Distrito	Qualitativa	N.A.	Distrito da Sede da Empresa (8 classes): Aveiro; Braga; Bragança; Guarda; Porto; Viana do Castelo; Vila Real e Viseu.
Forma_Juridica	Qualitativa	N.A.	Forma legal da empresas (4 classes): Cooperativa; Entidade Estrangeira; Sociedade Anónima e Sociedade por Quotas.
Setor	Qualitativa	N.A.	Setor de Atividade de acordo com a secção da CAE Rev. 3 (19 classes): A-Agricultura, produção animal, caça, floresta e pesca; B-Indústrias extrativas; C-Indústrias transformadoras; D-Eletricidade, gás, vapor, água quente e fria e ar frio; E-Captação, tratamento e distribuição de água; saneamento, gestão de resíduos e despoluição; F-Construção; G-Comércio por grosso e a retalho; reparação de veículos automóveis e motociclos; H-Transportes e armazenagem; I-Alojamento e restauração; J-Atividades de informação e de comunicação; K-Atividades financeiras e de seguros; L-Atividades imobiliárias; M-Atividades de consultoria, científicas, técnicas e similares; N-Atividades administrativas e dos serviços de apoio; O-Administração Pública, Defesa e Segurança Social; P-Educação; Q-Atividades de saúde humana e apoio social; R-Atividades artísticas, de espetáculos, desportivas e recreativase e S-Outras atividades de serviços.
Total_Ativo	Quantitativa	Milhares de euros	Total do Ativo em 31 de dezembro de 2013
Volume_Negocios	Quantitativa	Milhares de euros	Volume de Negócios em 31 de dezembro de 2013
Capital_Proprio	Quantitativa	Milhares de euros	Capital Próprio em 31 de dezembro de 2013
NPS	Quantitativa	Milhares de euros	Número de Pessoas ao Serviço em 31 de dezembro de 2013

A.5 CÓDIGO R DOS MODELOS IMPLEMENTADOS

Em seguida é apresentado o código R desenvolvido para a criação dos ficheiros sintéticos. Devido ao elevado número de linhas optou-se por apresentar apenas o código referente a dois modelos (modelos A+AD_2 e A+RF_1) que utilizam algoritmos e operacionalizações diferentes e que obtiveram bons resultados.

Modelo A+AD_2 (Árvores de Decisão, operacionalização 2)

```
## Variável Distrito

# Sampling

freq <-summary(mydata$Distrito)/length(mydata3$Distrito)
i <- names(freq)
probs<-unname(freq)

Distrito <- sample(i,length(mydata$Distrito), replace=T, prob=probs)

X1 <- data.frame(Distrito) # Variável X'(1)


## Variável Forma Jurídica

X2<- rpart(Forma_Juridica ~ Distrito, data=mydata,
control=rpart.control(minsplit=5))

prob_2<-predict(X2,X1,type="prob")

Sim<-matrix(0,length(mydata$Forma_Juridica),1)

for(i in 1:length(mydata$Forma_Juridica)){

  Forma_Juridica <- sample(colnames(prob_2),1, replace=T,
prob=prob_2[i,])

  Sim[i,1]<-Forma_Juridica

}

colnames(Sim)<-"Forma_Juridica"

X2 <- data.frame(X1,Sim) #Acrescentar X'(2)


## Variável Setor

X3<- rpart(Setor ~ Distrito + Forma_Juridica, data=mydata,
control=rpart.control(minsplit=5))

prob_3<-predict(X3,X2,type="prob")

Sim<-matrix(0,length(mydata$Setor),1)

for(i in 1:length(mydata$Setor)){

  Setor <- sample(colnames(prob_3),1, replace=T, prob=prob_3[i,])
```

```
  Sim[i,1]<-Setor

}

colnames(Sim)<-"Setor"

X3<-data.frame(X2,Sim) #Acrescentar X'(3)


## Variável Total_Ativo

X4<- rpart(Total_Ativo ~ Setor + Distrito + Forma_Juridica, data=mydata,
method="anova", control=rpart.control(minsplit=5))

leafnr <- floor(as.numeric(row.names(X4$frame[X4$where, ])))

X4$frame$yval <- as.numeric(row.names(X4$frame))

nodes <- predict(object = X4, newdata = X3)

uniquenodes <- unique(nodes)

new <- vector("numeric", nrow(X3))

y <- mydata$Total_Ativo
ys <- 1:length(new)

#bootstrap

for (j in uniquenodes) {
  donors <- y[leafnr == j]
  new[nodes == j] <- sample(donors, size = sum(nodes == j), replace = T)
}

#smoothing == "density"

ys <- 1:length(new)
maxfreq <- which.max(table(new))
maxcat <- as.numeric(names(table(new))[maxfreq])
if (table(new)[maxfreq]/sum(table(new)) > 0.7)
  ys <- which(new != maxcat)
if (10 * table(new)[length(table(new)) - 1] < tail(table(new),
                                                   n = 1) -
table(new)[length(table(new)) - 1]) {
  ys <- ys[-which(new == max(y))]
  maxy <- max(y)
}

densbw <- density(new[ys], width = "SJ")$bw
new[ys] <- rnorm(n = length(new[ys]), mean = new[ys],
                 sd = densbw)
if (!exists("maxy"))
  maxy <- max(y) + densbw
new[ys] <- pmax(pmin(new[ys], maxy), min(y))
new[ys] <- round(new[ys], 5)

Sim<-data.frame(new)

colnames(Sim)<-"Total_Ativo"

X4<-data.frame(X3,Sim) #Acrescentar X'(4)
```

```
## Variável Volume_Negocios

X5<- rpart(Volume_Negocios ~ Total_Ativo + Setor + Distrito +
Forma_Juridica, data=mydata, method="anova",
control=rpart.control(minsplit=5))

leafnr <- floor(as.numeric(row.names(X5$frame[X5$where, ])))

X5$frame$yval <- as.numeric(row.names(X5$frame))

nodes <- predict(object = X5, newdata=X4)

uniquenodes <- unique(nodes)

new <- vector("numeric", nrow(X4))

y <- mydata$Volume_Negocios
ys <- 1:length(new)

#bootstrap

for (j in uniquenodes) {
  donors <- y[leafnr == j]
  new[nodes == j] <- sample(donors, size = sum(nodes == j), replace = T)
}

#smoothing == "density"

ys <- 1:length(new)
maxfreq <- which.max(table(new))
maxcat <- as.numeric(names(table(new))[maxfreq])
if (table(new)[maxfreq]/sum(table(new)) > 0.7)
  ys <- which(new != maxcat)
if (10 * table(new)[length(table(new)) - 1] < tail(table(new),
                                                   n = 1) -
table(new)[length(table(new)) - 1]) {
  ys <- ys[-which(new == max(y))]
  maxy <- max(y)

densbw <- density(new[ys], width = "SJ")$bw
new[ys] <- rnorm(n = length(new[ys]), mean = new[ys],
                 sd = densbw)
if (!exists("maxy"))
  maxy <- max(y) + densbw
new[ys] <- pmax(pmin(new[ys], maxy), min(y))
new[ys] <- round(new[ys], 5)

Sim<-data.frame(new)

colnames(Sim)<-"Volume_Negocios"

X5<-data.frame(X4,Sim) #Acrescentar X'(5)


## Variável Capital_Proprio

X6<- rpart(Capital_Proprio ~ Volume_Negocios + Total_Ativo + Setor +
Distrito + Forma_Juridica, data=mydata, method="anova",
control=rpart.control(minsplit=5))

leafnr <- floor(as.numeric(row.names(X6$frame[Y6$where, ])))
```

```
X6$frame$yval <- as.numeric(row.names(X6$frame))

nodes <- predict(object = X6, newdata=X5)

uniquenodes <- unique(nodes)

new <- vector("numeric", nrow(X5))

y <- mydata3$Capital_Proprio
ys <- 1:length(new)

#bootstrap

for (j in uniquenodes) {
  donors <- y[leafnr == j]
  new[nodes == j] <- sample(donors, size = sum(nodes == j), replace = T)
}

#smoothing == "density"

ys <- 1:length(new)
maxfreq <- which.max(table(new))
maxcat <- as.numeric(names(table(new))[maxfreq])
if (table(new)[maxfreq]/sum(table(new)) > 0.7)
  ys <- which(new != maxcat)
if (10 * table(new)[length(table(new)) - 1] < tail(table(new),
                                                   n = 1) -
table(new)[length(table(new)) - 1]) {
  ys <- ys[-which(new == max(y))]
  maxy <- max(y)
}

densbw <- density(new[ys], width = "SJ")$bw
new[ys] <- rnorm(n = length(new[ys]), mean = new[ys],
                 sd = densbw)
if (!exists("maxy"))
  maxy <- max(y) + densbw
new[ys] <- pmax(pmin(new[ys], maxy), min(y))
new[ys] <- round(new[ys], 5)

Sim<-data.frame(new)

colnames(Sim)<-"Capital_Proprio"

X6<-data.frame(X5,Sim) # Acrescentar X'(6)


## Variável NPS

X7<- rpart(NPS ~., data=mydata3,
method="anova",control=rpart.control(minsplit=5))

plot(X7, uniform=TRUE, main="Regression Tree for NPS ")
text(X7, use.n=TRUE, all=TRUE, cex=.8)

leafnr <- floor(as.numeric(row.names(X7$frame[X7$where, ])))

X7$frame$yval <- as.numeric(row.names(X7$frame))

nodes <- predict(object = X7, newdata =X6)
```

```
uniquenodes <- unique(nodes)

new <- vector("numeric", nrow(X6))

y <- mydata$NPS
ys <- 1:length(new)

#bootstrap

for (j in uniquenodes) {
  donors <- y[leafnr == j]
  new[nodes == j] <- sample(donors, size = sum(nodes == j), replace = T)
}

#smoothing == "density"

ys <- 1:length(new)
maxfreq <- which.max(table(new))
maxcat <- as.numeric(names(table(new))[maxfreq])
if (table(new)[maxfreq]/sum(table(new)) > 0.7)
  ys <- which(new != maxcat)
if (10 * table(new)[length(table(new)) - 1] < tail(table(new),
                                                   n = 1) -
table(new)[length(table(new)) - 1]) {
  ys <- ys[-which(new == max(y))]
  maxy <- max(y)
}

densbw <- density(new[ys], width = "SJ")$bw
new[ys] <- rnorm(n = length(new[ys]), mean = new[ys],
                 sd = densbw)
if (!exists("maxy"))
  maxy <- max(y) + densbw
new[ys] <- pmax(pmin(new[ys], maxy), min(y))
new[ys] <- round(new[ys], 5)

Sim<-round(data.frame(new))

colnames(Sim)<-"NPS"

X7<-data.frame(X6,Sim) # Acrescentar X'(7)
```

Modelo A+RF_1 (*Random Forests*, operacionalização 1)

```
## Variável Distrito

X1<- randomForest(Distrito ~ ., data=mydata, ntree=500,mtry=6)

prob_1<-predict(X1, mydata,type="prob")

Sim<-matrix(0,length(mydata$Distrito),1)

for(i in 1:length(mydata$Distrito)){

  Distrito <- sample(colnames(prob_1),1, replace=T, prob=prob_1[i,])
  Sim[i,1]<-Distrito

}

aux<-mydata[c(-1)]

colnames(Sim)<-"Distrito"

Sim_RF<-data.frame(aux,Sim) #substituir X(1) por X'(1)


## Variável Forma Jurídica

X2<- randomForest(Forma_Juridica ~ ., data=mydata, ntree=500,mtry=6)

prob_2<-predict(Y2,Sim_RF,type="prob")

#calcular o ficheiro sintético

Sim<-matrix(0,length(mydata$Forma_Juridica),1)

for(i in 1:length(mydata$Forma_Juridica)){

  Forma_Juridica <- sample(colnames(prob_2),1, replace=T,
prob=prob_2[i,])

  Sim[i,1]<-Forma_Juridica

}

aux<-Sim_RF[c(-1)]

colnames(Sim)<-"Forma_Juridica"

Sim_RF<-data.frame(aux,Sim) #substituir X(2) por X'(2)


## Variável Setor

X3 <- randomForest(Setor ~ ., data=mydata, ntree=500, mtry=6)

prob_3<-predict(Y3,Sim_RF,type="prob")


Sim<-matrix(0,length(mydata$Setor),1)

for(i in 1:length(mydata$Setor)){
```

```
  Setor <- sample(colnames(prob_3),1, replace=T, prob=prob_3[i,])
  Sim[i,1]<-Setor
}

aux<-Sim_RF[c(-1)]

colnames(Sim)<-"Setor"

Sim_RF<-data.frame(aux,Sim) #substituir X(3) por X'(3)


## Variável Total_Ativo

X4<- randomForest(Total_Ativo ~ ., data=mydata3, ntree=500, mtry=6)

Previsao_X4 <- predict(object = X4, newdata = Sim_RF, predict.all=TRUE)

Previsao_X4<-Previsao_X4$individual

Sim<-matrix(0,length(mydata$Total_Ativo),1)

densbw <- density(mydata$Total_Ativo, width = "SJ")$bw

for(i in 1:length(mydata$Total_Ativo)){

x<-Previsao_X4[i,]

Total_Ativo <- rnorm(n = 1, mean=sample(x[!x %in%
boxplot.stats(x)$out],1), sd = densbw)
  Sim[i,1]<-Total_Ativo
}

Sim <- round(Sim, 5)

aux<-Sim_RF[c(-1)]

Sim<-data.frame(Sim)

colnames(Sim)<-"Total_Ativo"

Sim_RF<-data.frame(aux,Sim) #substituir X(4) por X'(4)


## Variável Volume_Negocios

X5<- randomForest(Volume_Negocios ~ ., data=mydata, ntree=500, mtry=6)

Previsao_X5 <- predict(object = X5, newdata = Sim_RF, predict.all=TRUE)

Previsao_X5<-Previsao_X5$individual

Sim<-matrix(0,length(mydata$Volume_Negocios),1)

densbw <- density(mydata$Volume_Negocios, width = "SJ")$bw

for(i in 1:length(mydata$Volume_Negocios)){

x<-Previsao_X5[i,]

  Volume_Negocios <- rnorm(n = 1, mean=sample(x[!x %in%
boxplot.stats(x)$out],1), sd = densbw)
```

```
  Sim[i,1]<-Volume_Negocios
}

Sim <- round(Sim, 5)

aux<-Sim_RF[c(-1)]

Sim<-data.frame(Sim)

colnames(Sim)<-"Volume_Negocios"

Sim_RF<-data.frame(aux,Sim) #substituir X(5) por X'(5)

## Variável Capital_Proprio

X6<- randomForest(Capital_Proprio ~ ., data=mydata, ntree=500, mtry=6)

Previsao_X6 <- predict(object = X6, newdata = Sim_RF, predict.all=TRUE)

Previsao_X6<-Previsao_Y6$individual

Sim<-matrix(0,length(mydata$Capital_Proprio),1)

densbw <- density(mydata$Capital_Proprio, width = "SJ")$bw

for(i in 1:length(mydata$Capital_Proprio)){

  x<-Previsao_X6[i,]

  Capital_Proprio <- rnorm(n = 1, mean=sample(x[!x %in%
boxplot.stats(x)$out],1), sd = densbw)

  Sim[i,1]<-Capital_Proprio
}

Sim <- round(Sim, 5)

aux<-Sim_RF[c(-1)]

Sim<-data.frame(Sim)

colnames(Sim)<-"Capital_Proprio"

Sim_RF<-data.frame(aux,Sim) #substituir X(6) por X'(6)

## Variável NPS

X7<- randomForest(NPS ~ ., data=mydata, ntree=500, mtry=6)

Previsao_X7 <- predict(object = X7, newdata = Sim_RF, predict.all=TRUE)

Previsao_X7<-Previsao_X7$individual

Sim<-matrix(0,length(mydata$NPS),1)

densbw <- density(mydata$NPS, width = "SJ")$bw
```

```
for(i in 1:length(mydata3$NPS)){

 x<-Previsao_X7[i,]

NPS <- rnorm(n = 1, mean=sample(x[!x %in% boxplot.stats(x)$out],1), sd =
densbw)

  Sim[i,1]<-NPS
}


Sim <- round(Sim, 0)


aux<-Sim_RF[c(-1)]

Sim<-data.frame(Sim)

colnames(Sim)<-"NPS"

Sim_RF<-data.frame(aux,Sim) #substituir X(7) por X'(7)
```

A.6 COMPARAÇÃO UNIVARIADA

TABELAS DE FREQUÊNCIA DOS FICHEIROS GERADOS SINTETICAMENTE

Sem Amostragem

unidades

		Original	A- AD_1	A- AD_2	A- RF_1	A- RF_2
Forma_Juridica	Cooperativa	4	0	0	2	0
	Entidade Estrangeira	24	11	0	1	0
	Sociedade Anónima	703	577	0	592	0
	Sociedade por Quotas	9.269	9.412	10.000	9.405	10.000

unidades

		Original	A- AD_1	A- AD_2	A- RF_1	A- RF_2
Setor	A	215	0	0	87	0
	B	26	0	0	14	0
	C	1.846	2.223	0	1.769	0
	D	12	0	0	2	0
	E	29	0	0	4	0
	F	985	445	0	782	0
	G	2.902	5.123	10.000	4.324	10.000
	H	439	290	0	281	0
	I	780	343	0	696	0
	J	177	0	0	47	0
	K	175	0	0	88	0
	L	313	307	0	229	0
	M	857	1.175	0	1.032	0
	N	237	0	0	75	0
	P	154	0	0	59	0
	Q	603	94	0	441	0
	R	81	0	0	19	0
	S	169	0	0	51	0

Amostragem com Reposição

unidades

		Original	A+ AD_1	A+ AD_2	A+ RF_1	A+ RF_2
Forma_Juridica	Cooperativa	4	6	4	3	0
	Entidade Estrangeira	24	27	24	25	0
	Sociedade Anónima	703	696	702	719	0
	Sociedade por Quotas	9.269	9.271	9.271	9.253	10.000

Nota: valores médios resultantes de 10 ficheiros sintéticos (m =10)

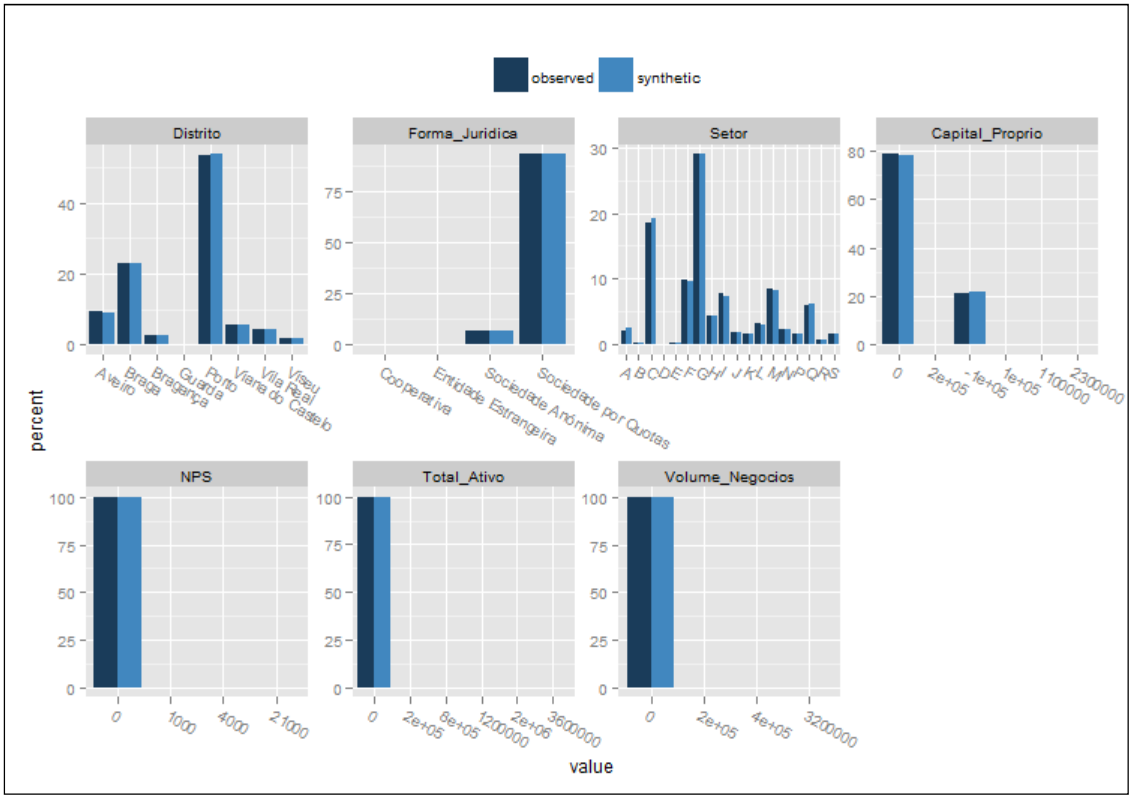
unidades

		Original	A+ AD_1	A+ AD_2	A+ RF_1	A+ RF_2
Setor	A	215	219	217	211	0
	B	26	25	29	26	0
	C	1.846	1.853	1.852	1.845	0
	D	12	12	11	12	0
	E	29	29	28	28	0
	F	985	977	1.003	986	0
	G	2.902	2.899	2.894	2.869	10.000
	H	439	435	438	436	0
	I	780	781	772	780	0
	J	177	177	180	182	0
	K	175	176	177	175	0
	L	313	319	317	312	0
	M	857	850	840	876	0
	N	237	231	228	239	0
	P	154	157	160	157	0
	Q	603	606	599	611	0
	R	81	83	83	85	0
	S	169	172	173	171	0

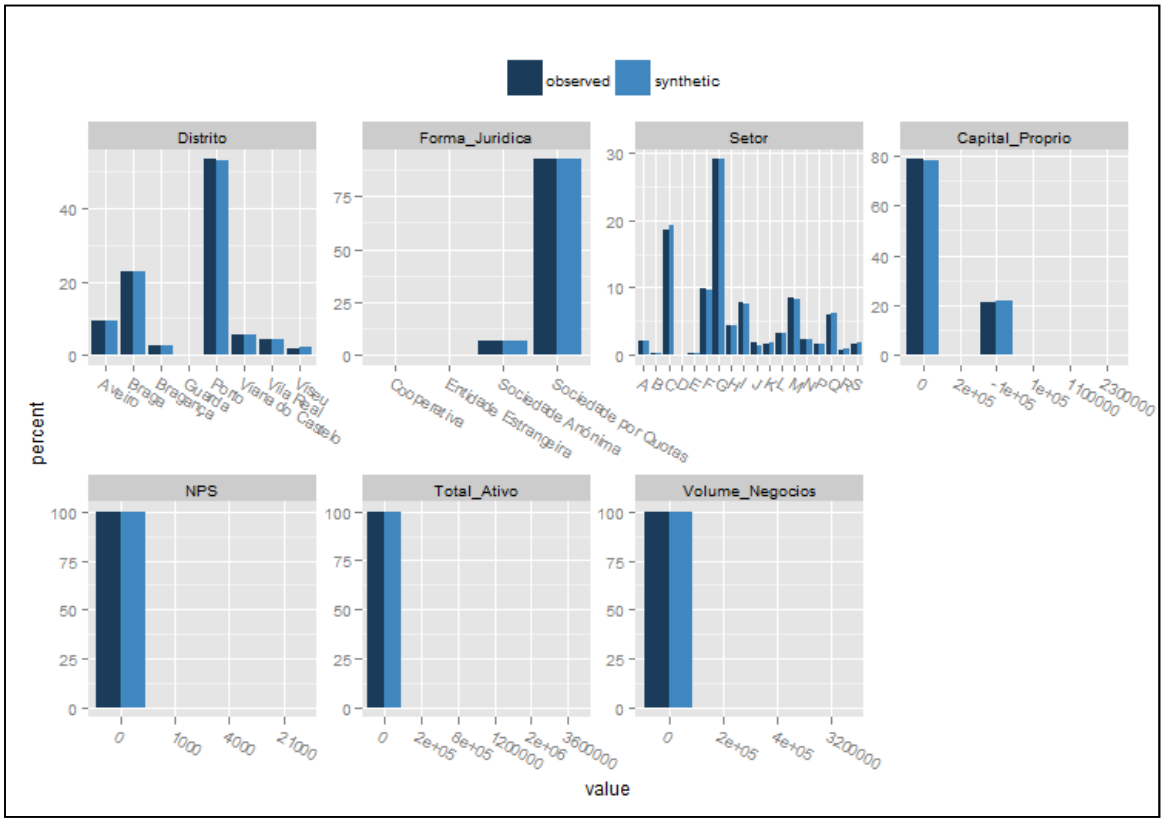
Nota: valores médios resultantes de 10 ficheiros sintéticos (m =10)

REPRESENTAÇÃO GRÁFICA DAS FREQUÊNCIAS DE UM FICHEIRO GERADO SINTETICAMENTE

Modelo A+AD_1



Modelo A+RF_2



A.7 TESTE DE QUALIDADE DE AJUSTAMENTO

Modelo A+AD_1

		Operacionalização 1										
	Algoritmo	Árvores de Decisão										
	Modelo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Média (m=10)**
χ^2	Distrito	0,309	0,841	0,127	0,490	0,141	0,583	0,857	0,771	0,610	0,640	0,5255
	Forma_Juridica	0,782	0,020	0,569	0,699	0,018	0,551	0,078	0,093	0,475	0,703	0,3650
	Setor	0,447	0,585	0,711	0,744	0,269	0,956	0,477	0,665	0,888	0,596	0,6382
K-S	Total_Ativo	0,478	0,963	0,981	0,204	0,251	0,998	0,853	0,994	0,958	0,688	0,7423
	Volume_Negocios	0,999	0,468	0,928	0,941	0,947	0,834	0,664	0,823	0,457	0,758	0,7845
	Capital_Proprio	0,935	0,958	0,987	0,991	0,890	0,769	0,426	0,664	0,953	0,953	0,8412
	NPS	1,000	0,581	0,890	0,881	0,813	0,890	0,999	0,999	0,998	0,604	0,8944

Modelo A+AD_2

		Operacionalização 2										
	Algoritmo	Árvores de Decisão										
	Modelo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Média (m=10)**
χ^2	Distrito	0,031	0,644	0,978	0,738	0,799	0,939	0,476	0,420	0,420	0,094	0,5538
	Forma_Juridica	0,875	0,881	0,821	0,739	0,889	0,706	0,215	0,708	0,979	0,338	0,7152
	Setor	0,014	0,761	0,035	0,705	0,632	0,523	0,419	0,693	0,927	0,214	0,4923
K-S	Total_Ativo	0,992	0,652	0,723	0,881	0,688	0,906	0,975	0,557	0,997	0,914	0,8285
	Volume_Negocios	1,000	0,723	0,758	0,981	0,941	0,489	0,914	0,984	0,963	0,987	0,8740
	Capital_Proprio	0,688	0,664	0,813	0,780	0,963	0,664	0,963	0,914	0,802	0,998	0,8247
	NPS	0,640	0,872	0,998	0,999	0,998	0,999	0,997	0,652	0,987	0,971	0,9112

Modelo A+RF_1

		Operacionalização 1										
	Algoritmo	*Random Forests*										
	Modelo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Média (m=10)**
χ^2	Distrito	0,375	0,913	0,863	0,548	0,438	0,870	0,963	0,070	0,770	0,109	0,5918
	Forma_Juridica	0,739	0,995	0,542	0,299	0,369	0,893	0,373	0,586	0,278	0,995	0,6068
	Setor	0,514	0,575	0,696	0,748	0,935	0,925	0,576	0,841	0,260	0,923	0,6992
K-S	Total_Ativo	0,273	0,331	0,468	0,281	0,244	0,500	0,244	0,187	0,297	0,652	0,3476
	Volume_Negocios	0,478	0,339	0,198	0,192	0,376	0,136	0,357	0,305	0,395	0,436	0,3214
	Capital_Proprio	0,001	0,001	0,002	0,000	0,000	0,000	0,001	0,000	0,000	0,000	0,0007
	NPS	0,266	0,436	0,478	0,557	0,863	0,557	0,928	0,426	0,758	0,688	0,5956

Modelo A+RF_2

		Operacionalização 2										
	Algoritmo	*Random Forests*										
	Modelo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Média (m=10)**
χ^2	Distrito	0,031	0,644	0,978	0,738	0,799	0,939	0,476	0,420	0,420	0,094	0,5538
	Forma_Juridica	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,0000
	Setor	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,0000
K-S	Total_Ativo	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,0000
	Volume_Negocios	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,0000
	Capital_Proprio	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,0000
	NPS	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,0000

A.8 MATRIZES DE VARIÂNCIAS E COVARIÂNCIAS E DE CORRELAÇÃO

Matriz de variâncias e covariâncias

Ficheiro original

	Ativo	Volume de Negócios	Capital Próprio	NPS
Ativo	2.181.594.313	733.916.217	1.129.195.628	4.558.087
Volume de Negócios	733.916.217	1.175.929.898	115.705.307	7.324.130
Capital Próprio	1.129.195.628	115.705.307	713.609.377	629.709
NPS	4.558.087	7.324.130	629.709	49.208

Modelo A+AD_1 (*m*=10) – equação 8

	Ativo	Volume de Negócios	Capital Próprio	NPS
Ativo	1.863.053.165	511.364.803	881.076.146	1.378.890
Volume de Negócios	511.364.803	1.505.470.709	366.574.326	3.437.883
Capital Próprio	881.076.146	366.574.326	777.393.152	1.098.963
NPS	1.378.890	3.437.883	1.098.963	44.032

Modelo A+AD_2 (*m*=10) – equação 8

	Ativo	Volume de Negócios	Capital Próprio	NPS
Ativo	2.227.615.673	911.720.419	1.017.857.134	3.727.218
Volume de Negócios	911.720.419	1.074.684.751	402.079.269	4.732.977
Capital Próprio	1.017.857.134	402.079.269	938.125.326	1.437.974
NPS	3.727.218	4.732.977	1.437.974	62.030

Modelo A+RF_1 (*m*=10) – equação 8

	Ativo	Volume de Negócios	Capital Próprio	NPS
Ativo	1.378.983.348	216.616.676	664.414.730	387.898
Volume de Negócios	216.616.676	605.573.524	110.055.999	2.318.358
Capital Próprio	664.414.730	110.055.999	413.031.838	184.574
NPS	387.898	2.318.358	184.574	14.761

Matriz de correlação

Ficheiro original

	Ativo	Volume de Negócios	Capital Próprio	NPS
Ativo	1	0,458	0,905	0,440
Volume de Negócios	0,458	1	0,126	0,963
Capital Próprio	0,905	0,126	1	0,106
NPS	0,440	0,963	0,106	1

Modelo A+AD_1 (m=10) – equação 8

	Ativo	Volume de Negócios	Capital Próprio	NPS
Ativo	1	0,313	0,787	0,243
Volume de Negócios	0,313	1	0,354	0,641
Capital Próprio	0,787	0,354	1	0,285
NPS	0,243	0,641	0,285	1

Modelo A+AD_2 (m=10) – equação 8

	Ativo	Volume de Negócios	Capital Próprio	NPS
Ativo	1	0,591	0,833	0,355
Volume de Negócios	0,591	1	0,447	0,647
Capital Próprio	0,833	0,447	1	0,245
NPS	0,355	0,647	0,245	1

Modelo A+RF_1 (m=10) – equação 8

	Ativo	Volume de Negócios	Capital Próprio	NPS
Ativo	1	0,340	0,889	0,207
Volume de Negócios	0,340	1	0,334	0,705
Capital Próprio	0,889	0,334	1	0,171
NPS	0,207	0,705	0,171	1

A.9 SOBREPOSIÇÃO DOS INTERVALOS DE CONFIANÇA A 95% - 12 COMBINAÇÕES

					Ativo Total		Volume de Negócios		Capital Próprio		NPS	
Distrito	**Forma_Juridica**	**Setor**	**Modelo**	**N**	**Média**	**Sobreposição I.C.**	**Média**	**Sobreposição I.C.**	**Média**	**Sobreposição I.C.**	**Média**	**Sobreposição I.C.**
Porto	Sociedade por Quotas	G	Original	1.470	528,634		818,603		140,922		5,151	
			A+ AD_1	1.459	725,053	0,954	734,049	0,813	269,015	0,658	9,161	0,596
			A+ AD_2	1.509	501,181	0,863	643,078	0,924	267,673	0,671	8,290	0,716
			A+ RF_1	1.448	494,077	0,859	626,988	0,904	165,011	0,603	4,651	0,882
Porto	Sociedade por Quotas	C	Original	735	1018,614		819,133		554,720		15,482	
			A+ AD_1	866	894,318	0,823	822,217	0,830	339,470	0,683	9,024	0,834
			A+ AD_2	719	501,855	0,736	640,259	0,764	301,961	0,720	8,090	0,764
			A+ RF_1	728	926,455	0,934	776,118	0,891	432,848	0,854	13,205	0,883
Braga	Sociedade por Quotas	G	Original	561	507,362		741,569		153,814		5,103	
			A+ AD_1	622	667,987	0,707	710,438	0,789	267,630	0,622	9,305	0,551
			A+ AD_2	593	503,715	0,637	710,027	0,632	294,035	0,605	8,398	0,607
			A+ RF_1	564	527,544	0,719	720,447	0,815	161,672	0,869	4,917	0,936
Porto	Sociedade por Quotas	M	Original	514	185,401		161,172		66,533		3,465	
			A+ AD_1	433	657,685	0,639	699,523	0,586	255,023	0,580	7,983	0,621
			A+ AD_2	458	439,657	0,687	591,323	0,612	218,055	0,599	8,675	0,613
			A+ RF_1	512	168,228	0,938	134,634	0,901	68,579	0,962	3,238	0,980
Braga	Sociedade por Quotas	C	Original	503	1034,899		1100,909		338,708		17,429	
			A+ AD_1	370	823,288	0,888	801,928	0,961	338,898	0,688	8,776	0,822
			A+ AD_2	520	484,153	0,730	609,643	0,799	278,070	0,868	8,095	0,779
			A+ RF_1	495	952,201	0,963	1010,860	0,878	386,325	0,696	15,597	0,931
Porto	Sociedade por Quotas	I	Original	443	159,890		162,749		8,242		5,546	
			A+ AD_1	397	678,647	0,562	648,449	0,559	256,390	0,570	8,468	0,657
			A+ AD_2	440	463,826	0,615	641,742	0,546	222,194	0,551	10,764	0,532
			A+ RF_1	437	145,624	0,951	159,394	0,990	20,492	0,740	5,246	0,924
Porto	Sociedade por Quotas	F	Original	432	577,817		405,924		182,814		8,558	
			A+ AD_1	482	770,091	0,700	742,236	0,626	281,589	0,709	9,105	0,781
			A+ AD_2	478	438,818	0,936	564,543	0,775	241,246	0,851	8,073	0,803
			A+ RF_1	434	587,044	0,683	377,949	0,983	203,266	0,679	7,872	0,958
Porto	Sociedade por Quotas	Q	Original	365	294,912		186,425		63,493		3,312	
			A+ AD_1	313	666,876	0,772	754,813	0,552	236,803	0,796	10,299	0,524
			A+ AD_2	342	549,513	0,624	656,097	0,570	271,228	0,830	8,896	0,601
			A+ RF_1	355	247,298	0,799	172,587	0,946	143,895	0,593	3,073	0,819
Aveiro	Sociedade por Quotas	G	Original	268	459,204		745,777		143,748		4,623	
			A+ AD_1	252	725,453	0,601	628,148	0,843	284,379	0,600	10,279	0,527
			A+ AD_2	243	427,314	0,837	613,223	0,887	216,462	0,629	8,124	0,670
			A+ RF_1	270	436,509	0,977	629,097	0,936	150,445	0,896	4,393	0,937
Braga	Sociedade por Quotas	F	Original	253	509,703		337,134		127,393		8,304	
			A+ AD_1	201	671,739	0,710	727,095	0,593	265,463	0,620	8,694	0,782
			A+ AD_2	238	472,549	0,837	588,330	0,626	284,135	0,619	8,051	0,835
			A+ RF_1	242	540,273	0,733	381,268	0,706	150,922	0,703	8,094	0,995
Porto	Sociedade por Quotas	H	Original	239	493,881		591,312		182,784		6,766	
			A+ AD_1	220	729,704	0,857	769,276	0,791	265,690	0,929	8,378	0,829
			A+ AD_2	245	683,824	0,644	669,619	0,810	276,201	0,797	8,731	0,789
			A+ RF_1	225	428,508	0,917	501,777	0,929	167,868	0,882	6,312	0,937
Aveiro	Sociedade por Quotas	C	Original	234	1003,099		984,353		370,151		14,910	
			A+ AD_1	154	696,487	0,947	703,608	0,972	252,828	0,884	8,918	0,912
			A+ AD_2	218	504,564	0,814	629,857	0,862	263,141	0,876	8,495	0,929
			A+ RF_1	226	1057,390	0,685	917,103	0,969	430,173	0,604	13,873	0,931