FACULDADE DE ECONOMIA DA UNIVERSIDADE DO PORTO

Mestrado em Métodos Quantitativos em Economia e Gestão

Confidencialidade de Dados: Aplicação e Comparação de Técnicas de Controlo da Divulgação Estatística

Dissertação com vista à obtenção do Grau de Mestre em Métodos Quantitativos em Economia e Gestão pela Faculdade de Economia da Universidade do Porto

Orientador:

Professor Doutor Pedro campos

Porto, 2010

Nota biográfica

Elsa Cristina Pinto Mendes é natural de Marco de Canaveses, Portugal onde nasceu a 8 de Agosto de 1973.

Estudante na Universidade Portucalense, completou a licenciatura em Economia em Julho de 1999, tendo iniciado a frequência do Mestrado em Métodos Quantitativos em Economia e Gestão no ano lectivo de 2008/2009.

Em termos profissionais, a Elsa exerce a profissão de Economista numa empresa de extracção, transformação de granitos.

Agradecimentos

Em primeiro lugar, gostaria de agradecer ao meu orientador, Professor Pedro Campos, pelos seus valiosos conselhos e orientações ao longo de todo o meu percurso académico e, sobretudo pela sua extrema disponibilidade e dedicação. Para mim foi um enorme privilégio poder trabalhar com o Professor Pedro Campos e seria um prazer poder continuar a colaborar com ele. A si, o meu muito obrigado.

Agradeço à minha família, em especial à minha mãe, o tempo dispensado de privar com eles algumas noites e fins-de-semana, bem como a compreensão e a paciência que tiveram comigo durante estes últimos dois anos de estudo.

Não posso deixar de agradecer aos meus amigos o apoio que me deram durante os últimos dois anos.

Por fim, mas não mais importante, gostaria de agradecer a pessoa que mais me incentivou e apoiou a fazer a dissertação, Pedro Queiroz. Muito obrigada pela tua extrema paciência, compreensão, carinho, dedicação, pela força e pelo teu amor.

Resumo

A procura de informação de qualidade, por parte dos investigadores e do público em geral, tem vindo a crescer rapidamente nos últimos anos. A fim de respeitar a legislação sobre a protecção de dados e ao mesmo tempo fornecer informação estatística de qualidade aos utilizadores da estatística, foram criados métodos e programas de software a partir dos quais se podem aplicar métodos de controlo da divulgação estatística (CDE), procurando diminuir o risco de divulgação de dados. Este trabalho aborda as técnicas de controlo de divulgação estatística, que têm por objectivo, por um lado, a protecção da confidencialidade e por outro lado, uma redução de perda de informação. Foram aplicados vários métodos de CDE a dois tipos de bases de dados que são comuns nos institutos de estatística: informação financeira e famílias. O método mais eficaz é o que, criando ficheiros seguros de dados, conduz à uma menor perda de informação, que nos dois ficheiros em estudo corresponde ao método do arredondamento e à microagregação, respectivamente.

Abstrat

The demand for high quality information from the researchers and the public in general has been growing rapidly in recent years. In order to comply with legislation on data protection, while providing statistical information with quality to final users, statistical methods and software programs were created from which one can apply methods of statistical disclosure control (SDC), trying to reduce the risk data dissemination. This paper covers the techniques of statistical disclosure control, which are aimed, firstly, the protection of confidentiality and on the other hand, a reduction of the information loss. We applied various methods of SDC to two data base types that are common in the national statistical offices: business and households. The most effective method is the one who leads to a smaller loss of information in the files under study, corresponding to the method of the rounding and microaggregation, respectively in each of the files.

Índice de Conteúdos

Índice de quadros

Índice de figuras

Capítulo 1. Introdução

A informação estatística é um bem fundamental nas sociedades modernas. Contribui inequivocamente para o desenvolvimento económico e social e para o reforço da cidadania.

Os serviços de estatística divulgam, essencialmente dois tipos de dados: microdados e dados tabulares ou macrodados. Os utilizadores de microdados assumem um papel duplo na teoria do controlo da divulgação estatística. Por um lado, o utilizador é visto como um cliente de estatística e, por outro, pode ser considerado um possível intruso.

A crescente procura de informação estatística conduz à necessidade de clarificação de um conjunto de procedimentos relacionados com a confidencialidade de dados. Se por um lado, se pretende divulgar a melhor e maior quantidade de informação possível, para que os investigadores e decisores possam desenvolver ideias e promover políticas de desenvolvimento, por outro, existe o compromisso de sigilo dos dados dos entrevistados que, de acordo com a legislação em vigor, os produtores de informação estatística têm de respeitar. Desta forma, opõe-se o direito da sociedade à informação e o direito do indivíduo à privacidade dos seus dados. Para a resolução desta ambiguidade há necessidade de encontrar a melhor forma de proteger os dados sem perder a coerência e a estrutura da informação.

O problema da confidencialidade de dados nem sempre foi visto com a mesma preocupação em todos os países. Até há algumas décadas, era dada maior importância à protecção de dados económicos, mas hoje em dia esse problema é visto de forma diferente. Devido às diferentes regras de confidencialidade adoptadas pelos diversos países, houve a necessidade de harmonizar a legislação, o que aconteceu a partir de 1990 altura em que se começaram a definir métodos de Controlo de Divulgação Estatística (CDE).

O segredo estatístico visa essencialmente salvaguardar a privacidade no domínio das estatísticas e é a chave para a confiança necessária que tem que existir entre os serviços de estatística e os respondentes. O maior desafio para os serviços de estatística prende-

se com facto de minimizar os riscos de divulgação sem alterar de forma significativa os dados, isto é, o risco deve ser gerido eficazmente. Existem diversas formas de gerir o risco, através dos métodos de Controlo da Divulgação Estatística. Estes métodos visam alterar os dados estatísticos de tal forma que aquando da divulgação, as informações individuais sejam suficientemente protegidas contra identificação dos indivíduos ou das empresas. Ao mesmo tempo, oferecer à sociedade o máximo de informação possível, ou seja, encontrar o ponto de equilíbrio, o ponto que maximiza a utilidade da informação e minimiza o risco de divulgação.

Os métodos de Controlo da Divulgação Estatística são normalmente conhecidos por métodos de mascaramento ou anonimização, que dependo dos seus princípios operacionais podem ser classificados em duas categorias: métodos perturbativos e os métodos não perturbativos.

Neste contexto, o contributo essencial da presente dissertação é de âmbito aplicacional e instrumental. Aplicam-se diferentes técnicas de Controlo da Divulgação Estatística e faz-se a sua comparação em microdados provenientes de duas bases de dados. Por um lado, estuda-se a aplicação e comparação das técnicas de controlo da divulgação estatística em dados financeiros e económicos de empresas da indústria extractiva em Portugal (recorrendo a uma base de dados - a base SABI). Por outro, utiliza-se no estudo uma base de dados simulados relativa a famílias semelhante às utilizadas nos inquéritos às famílias (como o Inquérito ao Emprego) realizados pelo Instituto Nacional de Estatística.

Foram aplicadas algumas técnicas de Controlo da Divulgação Estatística disponíveis no software Argus, ferramenta que foi utilizada no âmbito desta tese. As técnicas utilizadas dependem das variáveis em estudo. Quer num ficheiro quer no outro foram estudadas variáveis contínuas e variáveis categóricas. Foram utilizadas métricas de qualidade para comparar as técnicas estudadas. Às variáveis categóricas foi aplicada a recodificação global, a qual não foi comparada com nenhuma outra técnica, uma vez que as métricas de qualidade não são comparáveis com as restantes. Relativamente às variáveis contínuas foram aplicadas diversas técnicas disponíveis no software e para

enriquecer este trabalho foi aplicada uma técnica pouco divulgada, a dos dados híbridos, sobre a qual a autora propõe uma nova metodologia (microagregação híbrida).

As diferentes técnicas utilizadas, como a microagregação, a codificação superior, o arredondamento, rank swapping e os dados híbridos não provocaram alterações significativas na estrutura dos dados. A sua aplicação contribui para uma diminuição significativa do risco individual de divulgação e do número de células inseguras nas duas bases de dados. Quanto à perda de informação, ela varia com os métodos utilizados. O que se pretende é um método que crie um ficheiro de dados seguro e que ao mesmo tempo tenha a menor perda de informação possível.

Os resultados obtidos diferem nos dois ficheiros em estudo: se num dos casos é o método do arredondamento que provoca menor perda de informação (base de dados SABI), no caso do ficheiro de dados relacionados as famílias, o método com menor perda de informação é a microagregação. Em ambas as situações, a técnica de microagregação híbrida proposta no âmbito da tese consegue bons resultados.

A tese encontra-se estruturada da seguinte forma: no Capítulo 2 faz-se um enquadramento do tema da confidencialidade dos dados, referindo a importância geral do segredo estatístico, a protecção dos dados, bem como o quadro jurídico e histórico do segredo estatístico e a filosofia do Controlo da Divulgação Estatística. No Capítulo 3 faz-se uma abordagem detalhada dos microdados, dando especial atenção ao controlo da divulgação estatística, as etapas para a divulgação dos microdados, bem como as diversas técnicas de controlo da divulgação estatística. No Capitulo 4 faz-se uma alusão aos dados tabulares e métodos de controlo da divulgação estatística, apresentando-se um pequeno exemplo de aplicação das técnicas de controlo da divulgação estatística. No capítulo 5 são abordados temas como a qualidade dos dados enumerando as diferentes medidas de qualidade dos dados, quer para dados contínuos, quer para dados categóricos. È também referido neste capítulo o risco de divulgação e as medidas de risco. No Capítulo 6 são apresentados dois casos de estudo, com a aplicação e comparação das técnicas de Controlo da Divulgação Estatística .

Capítulo 2. A Confidencialidade dos dados

Este capítulo faz um enquadramento do tema da confidencialidade dos dados, desde a importância geral do segredo estatístico, à protecção dos dados. O quadro jurídico e histórico do segredo estatístico e a filosofia do Controlo da Divulgação Estatística são também referidos. No final, apresentam-se alguns dos métodos de controlo da divulgação estatística que serão abordados com maior detalhe no capítulo seguinte.

2.1. Importância do segredo estatístico e da sua protecção

A informação estatística é um bem fundamental nas sociedades modernas. Contribui, de modo inequívoco, para o desenvolvimento económico e social e para o reforço da cidadania. A necessidade da protecção da confidencialidade dos dados (também conhecida como protecção do segredo estatístico ou protecção da privacidade de dados) advém por razões legais, relacionadas com a protecção da confidencialidade individual, mas também da obrigação moral com a qual muitas entidades[1] que recolhem informação estatística se comprometem. Com esta obrigação da protecção da confidencialidade dos dados, torna-se mais fácil obter a colaboração dos indivíduos que são seleccionados na amostra de um inquérito. De outra forma, caso seja possível identificar um respondente (ou entrevistado) através dos seus dados, este ficaria reticente à participação em novos inquéritos.

Para o presente trabalho é muito importante o conceito de indivíduo, pois é a ele que se refere a protecção da confidencialidade. Neste trabalho quando se emprega a palavra indivíduo, quer-se referir aos registos individuais, que podem corresponder a pessoas singulares ou pessoas colectivas, como empresas, famílias, etc. Em termos gerais, os indivíduos correspondem às unidades estatísticas de amostragem sobre quem vai incidir a informação dos inquéritos.

[1] Em geral, neste trabalho as entidades produtoras de informação denominam-se Serviços de Estatística (SE) ou responsáveis pela informação. Estas entidades podem ser os institutos de estatística, bancos ou outra fonte ou entidade que recolhe e divulga dados estatísticos.

A discussão em torno da privacidade dos dados não surge apenas devido à informação decorrente dos inquéritos (alguns tão importantes como os Recenseamentos da População e da Habitação) ou aos dados individuais em geral, mas também devido a três outras questões:

1) Qualidade da informação. A informação estatística é recolhida desde há muitos anos e a vários níveis. Os utilizadores da informação tornam-se mais exigentes, obrigando ao aumento da qualidade dos dados estatísticos e ao consequente aumento do risco da divulgação estatística.

2) Crescente presença dos computadores e de programas sofisticados. Hoje em dia vários investigadores em diversas universidades possuem condições para analisar grandes arquivos de dados, o que lhes permite criar os seus próprios cruzamentos de dados. Esta permissão para os investigadores terem acesso aos dados aumenta o risco de identificação dos entrevistados.

3) Informatização da sociedade, onde existem grandes bases de dados contendo uma enorme quantidade de informação sobre os indivíduos (pessoas, empresas e famílias), o que pode permitir a identificação de registos individuais através do cruzamento de várias fontes.

2.2. Protecção dos dados

Quando se produz informação estatística é necessário ter em atenção qual a informação que pode ser divulgada. Se, por um lado, se pretende divulgar a melhor e maior quantidade de informação possível, por outro, à medida que aumenta a qualidade e detalhe da informação, maior é o risco de divulgação dos dados individuais. Desta forma, opõem-se o direito da sociedade à informação e o direito do indivíduo à privacidade dos seus dados.

A Figura 1 apresenta uma análise gráfica de como uma agência (Instituto de Estatística ou outra fonte de informação) fornece dados com utilidade para os utilizadores e reduz o risco de divulgação face ao ataque dos intrusos. Existem

condições para que os dados sejam ao mesmo tempo, analiticamente válidos e analiticamente interessantes, com uma pequena perda de informação e com baixo risco de divulgação, isto é, para que sejam ficheiros seguros.

Zaslavsky e Horton (1998, cf Duncan et al., 2001) utilizam a abordagem decisão - teórica com base na estrutura da Figura 1 para obter um limite óptimo de divulgação para o tamanho da célula mínima nos dados tabulares.

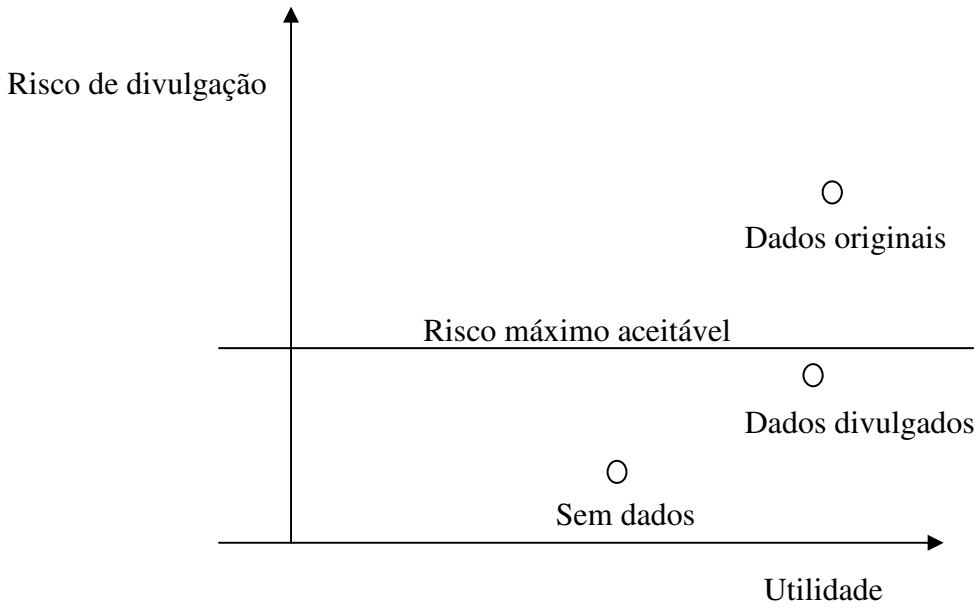

Figura 1 – O problema da limitação da divulgação das estatísticas. Fonte: Duncan et al (2001)

Existem diferentes formas de gerir o risco de divulgação dos dados, nomeadamente os métodos de controlo da divulgação estatística. Os métodos de Controlo da Divulgação Estatística são técnicas que têm por objectivo a protecção da confidencialidade (ou segredo estatístico). A aplicação dessas técnicas conduz a uma perda de informação dos conteúdos estatísticos e afecta a opinião que os utilizadores possam fazer sobre os dados. Para obter um compromisso entre a confidencialidade e a utilidade dos dados, deve-se procurar o ponto óptimo, ou seja, o ponto que maximiza a utilidade dos dados e minimiza o risco de divulgação. É evidente que este ponto é difícil de definir e depende muito das fontes de informação e da base de dados em causa.

O objectivo final do Controlo da Divulgação Estatística (CDE)[2] é a divulgação da informação estatística de tal forma que as informações individuais sejam

[2] Em inglês, a a sigla utilizada habitualmente é SDC – Statistical Disclosure Control

suficientemente protegidas contra a identificação dos indivíduos ou empresas e, ao mesmo tempo, oferecer à sociedade o máximo de informação possível, isto é, encontrar o ponto de equilíbrio, ou seja, o ponto que maximiza a utilidade da informação e minimiza o risco de divulgação dos dados individuais.

Sabendo que a perda de informação aumenta à medida que diminui o risco de identificação, a protecção da confidencialidade deve ser feita de modo a encontrar-se um compromisso entre essas duas quantidades. O risco de identificação é a probabilidade de um intruso identificar pelo menos um entrevistado nos microdados disponibilizados.

Na prática o que se pretende, em primeiro lugar, para um determinado tipo de dados, é saber qual o critério que esses dados têm que cumprir de forma a tornar a sua divulgação segura. Após a identificação do critério, é necessário modificar os dados, que não o satisfazem de tal forma que a perda de informação seja minimizada. Quando a perda de informação é formalmente definida, a modificação dos dados pode ser formulada como um problema de optimização.

Assim numa primeira fase, o protector de dados deve retirar de cada registo os identificadores directos, isto é, as variáveis que permitam identificar directamente o entrevistado, tais como, o nome, a morada, número de identificação (BI, NPC, NIF, etc.), e caso não seja suficiente, deve, numa segunda fase, identificar as combinações raras dos identificadores indirectos que, se não forem suprimidas, possibilitam a identificação de algumas unidades estatísticas.

2.2.1. Breve Benchmarking da protecção de dados

Uma vez que a privacidade dos indivíduos deve ser salvaguardada, vários países têm diferentes quadros jurídicos que a regulam. Nos Estados Unidos a secção nove do Code Protection of Confidential Information, (U. S. Census Bureau, 2006) proíbe nesses inquéritos a divulgação de determinados dados que permitam a identificação de um respondente.

Na Holanda a lei que regula a divulgação dos dados estatísticos económicos responsabiliza o Instituto de Estatística (CBS – Central Bureau for Statistics) pela

confidencialidade dos dados. Em 1990, surgiu uma nova lei que regula os principais aspectos da gestão de dados individuais, contendo várias medidas para salvaguardar a privacidade dos dados individuais. No Reino Unido existe uma lei semelhante: "UK Data Protection ACT 1984".

O problema da confidencialidade dos dados nem sempre é visto com a mesma preocupação em todos os países, pois até há algumas décadas era dada maior importância à protecção dos dados económicos. No entanto, durante as duas ou três últimas décadas a preocupação com as questões da privacidade dos indivíduos tem vindo a crescer rapidamente. Na secção seguinte, apresenta-se um resumo da evolução do segredo estatístico e do correspondente quadro jurídico.

2.3.Quadro jurídico e histórico do segredo estatístico

Durante a comemoração do centenário, em 1985, ISI - International Statistical Institute (www.isi-web.org), organismo internacional com funções de ligação entre os estatísticos, associações e os diversos institutos de estatística foi adoptada a Declaration on Profissional Ethics (ISI - International Statistical Institute, 1985). Esta Declaração resultou de um extenso processo de elaboração e reformulação de consulta com os membros e as secções do Instituto Internacional de Estatística no período de 1979-1985. As cláusulas 4.5 e 4.6 são de extrema importância para o controlo da divulgação estatística, transcrevemo-las de seguida:

4.5 Maintaining confidentiality of records

"Statistical data are unconcerned with individual identities. They are collected to answer questions such as 'how many?' or 'what proportion?' not 'who?'. The identities and records of co-operating (or non- cooperating) subjects should therefore be kept confidential, whether or not confidentiality has been explicitly pledged."

4.6 Inhibiting disclosure of identities

"Statisticians should take appropriate measures to prevent their data from being published or otherwise released in a form that would allow any subject's identity to be disclosed or inferred."

Devem ser tomadas medidas apropriadas para impedir que os dados sejam publicados ou divulgados de forma a permitir a identidade de qualquer indivíduo a ser divulgado.

Até finais de 1980, os microdados[3] raramente eram transmitidos ao Eurostat pelos países membros. A harmonização estatística era, até então, muito difícil de implementar, devido às regras de confidencialidade adoptadas por alguns países. Por esse motivo, em Junho de 1990 surge o Council Regulation (EURATOM) ECC No 1588/90 (Council Regulation (EURATOM, ECC) No 1588/90 of 11 June 1990), um regulamento, elaborado e aprovado pelo Conselho Europeu, sobre a transmissão de dados confidenciais ao Eurostat. Este regulamento autoriza os Institutos de Estatística a transmitir os dados ao Eurostat, enquanto este se obriga a tomar as medidas necessárias à respectiva protecção.

Em 1994, estas medidas também foram definidas e formalmente adoptadas pelos Estados Membros através do Comité de Confidencialidade Estatística. Este comité reúne-se uma vez por ano no Eurostat, para discutir a implementação e a evolução das regulamentações europeias sobre a divulgação de microdados e dados tabulares (macrodados)[4], bem como o quadro jurídico de base estatística.

Em Fevereiro de 1997 foi elaborado o Council Regulation (EC) No 322/97 (Council Regulation (EC) No 322/97 of 17 February 1997) on Community statistics [5] que define os princípios gerais que regem as comunidades estatísticas, os processos para a produção dessas estatísticas e estabelece as regras de confidencialidade. Este regulamento pode ser considerado como a Lei Geral da Estatística da União Europeia. Em Maio de 2002 surge o Commission Regulation EC No 831/2002 (Commission Regulation (EC) No 831/2002 of 17 May 2002), para a aplicação do Council Regulation (EC) No 322/97 relativo às estatísticas comunitárias em matéria de acesso a dados confidenciais para fins científicos.

[3] Microdados – Conjunto de registos que contem informação de respondentes individuais ou entidades economicas

[4] Macrodados - Informação agregada de entidades e representada em forma de tabelas.

[5] A legislação europeia sobre esta temática encontra-se disponível em: (http://euro-lex.europa.eu),

Em Fevereiro de 2005, o Comité do Programa Estatístico adoptou o European Statistics Code of Practice 2005 (Eurostat, 2005), com 15 princípios. O Princípio 5 respeita à confidencialidade estatística.

Em Portugal, a Lei do Sistema Estatístico Nacional (SEN) - Lei nº 22/2008 de 13 de Maio (Diário da República 1ª série de 13 de Maio de 2008) - estabelece um enquadramento geral da actividade estatística nacional, definindo os princípios fundamentais do SEN, contemplando no seu artigo 6º o princípio do segredo estatístico.

De acordo com esta norma todos os dados estatísticos individuais recolhidos pelas autoridades estatísticas são de natureza confidencial, não podendo ser divulgados de modo a permitirem a identificação directa e indirecta das pessoas singulares ou colectivas a que respeitam.

Com excepção das situações previstas no referido preceito, apenas é permitida a divulgação anonimizada dos dados estatísticos individuais sobre pessoas singulares, com autorização do respectivo titular ou após autorização do Conselho Superior de Estatística e, neste caso, será apenas quando estejam em causa ponderosas razões de saúde pública. Relativamente às pessoas colectivas o procedimento é idêntico, embora a informação a ceder não seja anonimizada e as causas sejam mais diversificadas. Existe ainda a cedência anonimizada de dados estatísticos individuais sobre pessoas singulares e colectivas para fins científicos que se formaliza mediante o estabelecimento de um acordo entre a autoridade estatística cedente e a entidade solicitante, de forma a assegurar a protecção dos dados confidenciais e evitar qualquer risco de divulgação ilícita ou de utilização para outros fins aquando da divulgação dos resultados.

2.4. A filosofia do Controlo da Divulgação Estatística

Para se compreender os princípios do Controlo da Divulgação Estatística torna-se necessário analisar o papel de um intruso, ou seja, alguém com intenções de obter dados confidenciais. O objectivo de um intruso consiste em tentar a combinação de resultados de variáveis de identificação que são raras na população ou na amostra (Willenborg, e Waal, 1996). As combinações de variáveis que permitem identificar registos individuais mais frequentes são menos susceptíveis de provocar a curiosidade do intruso. Se o

intruso tentar encontrar conscientemente registos, vai faze-lo através de valores chave[6] que ocorrem somente algumas vezes. Se essa correspondência é inconsciente e se o utilizador sabe de alguém com esse valor chave raro, então o registo associado a essa particularidade rara pode ser do seu conhecido. Quanto menor for o número de indivíduos com o valor chave correspondente, maior é a probabilidade de uma correcta correspondência.

Se um indivíduo (empresa, família, pessoa) é único na população, sem que esse facto seja perceptível, esse registo não será facilmente identificado. Por outro lado, se o registo não é único na população, mas existe apenas mais um com a mesma chave, o detentor da informação do outro registo é capaz de a identificar. Outro caso pode ocorrer em que o indivíduo não é único na população, mas pertence a um grupo de indivíduos com o mesmo resultado de uma variável sensível[7]. Então, pode ser divulgada informação confidencial sobre esse indivíduo sem que ele seja identificado.

Considere-se um entrevistado que não sendo único, pertence a um pequeno grupo de pessoas. O intruso tem alguma informação sobre essa pessoa, que não é considerada identificativa, mas que está contida no conjunto de microdados divulgados. É possível que o entrevistado seja único na combinação da nova informação com a variável chave, o que torna provável a sua identificação.

A singularidade da população é um problema difícil de verificar, sendo-lhe dada menor importância. Ao invés, fala-se em raridade como sendo um importante factor para a identificação, trazendo uma vantagem adicional. Independentemente de o intruso usar mais ou menos chaves de dimensão superior, na tentativa da divulgação da informação do que as que foram utilizadas pelo responsável da divulgação dos dados, ele (o intruso) em muitos casos não consegue identificar na população as pessoas com essa chave. Se num conjunto de microdados houver vários registos em que alguns dos valores das variáveis são raros, a probabilidade de identificação desses registos é elevada.

[6] Chave – É uma combinação de variáveis identificadoras que identificam inequivocamente o individuo, como por exemplo o nome, o número de identificação fiscal, número do passaporte.

[7] Variáveis sensíveis – São variáveis em que pelo menos um dos seus valores é sensível e para as quais o protector de dados deve ser mais rigoroso na sua protecção, nomeadamente o comportamento sexual, o passado criminal.

2.5.Tipo de variáveis

Para aplicação das técnicas de Controlo da Divulgação Estatística, torna-se necessário definir os tipos de variáveis envolvidas. Uma variável representa algum atributo, característica ou propriedade de um grupo de dados, que assume valores diferentes de indivíduo para indivíduo.

As variáveis podem ser classificadas em qualitativas (nominais e ordinais) e quantitativas (discretas e contínuas). Uma descrição mais detalhada sobre o tipo de variáveis pode ser encontrada no Anexo 1.

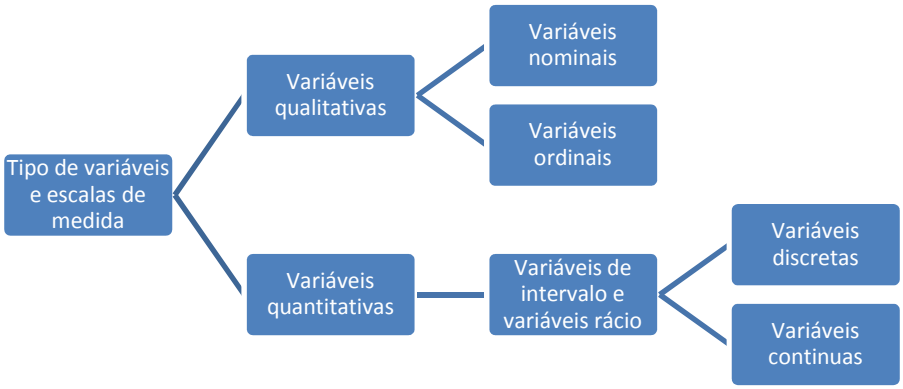

Figura 2 – Tipo de variáveis

Uma outra forma de classificar as variáveis está relacionada com o seu interesse para a detecção do segredo estatístico. Determinar se uma variável é ou não uma variável identificadora[8] não é uma tarefa fácil e não existem regras para o fazer, por esse motivo, elas são seleccionadas por intuição (Willenborg e Waal, 1996). Se se quiser evitar que os dados divulgados sejam combinados com um registo existente, então devem ser consideradas as variáveis que podem potencialmente ser usadas para esse fim como identificadoras. Uma forma de decidir se as variáveis são identificadoras consiste em solicitar a um especialista num determinado tema, para indicar, relativamente a variável de um arquivo de dados, se esta é de identificação ou não.

[8] Variáveis identificadoras - Variáveis que identificam inequivocamente o individuo, como o nome, o NIF.

Existem também as variáveis sensíveis[9]. Por exemplo, o comportamento sexual, o registo criminal são variáveis que podem ser consideradas sensíveis.

Nas tabelas (macrodados[10]) são consideradas variáveis sensíveis as variáveis cujo valor é publicado nas células. Por um lado esta definição é de fácil aplicação, por outro pode ser demasiado protectora, uma vez que em alguns casos existem valores publicados que não são realmente sensíveis.

A chave é um importante conceito na teoria da identificação. A chave é uma combinação de potenciais variáveis identificadoras. Num cenário de divulgação, as combinações chave de variáveis de identificação, são supostamente usadas por um intruso para identificar um entrevistado. A identificação do entrevistado pode ocorrer quando ele é raro na população em relação a uma determinada regra chave, isto é, uma combinação de valores de variáveis de identificação. Por esse motivo, a raridade dos entrevistados na população em relação a alguns valores fundamentais deve ser evitada. Quando um entrevistado parece ser raro na população em relação a um valor chave, devem ser tomadas medidas de controlo da divulgação para proteger esse entrevistado contra a identificação.

Para se definir o que é raro na população deve-se escolher um valor limite para cada chave. O valor é considerado seguro se ele ocorrer mais vezes do que valor limite, caso contrário, a chave é considerada insegura, devendo ser protegida. Os ficheiros para uso público requerem maior protecção dos que os ficheiros usados por investigadores. Os dois tipos de ficheiros serão abordados mais à frente neste trabalho.

2.6.Métodos de controlo da divulgação estatística

Um ficheiro de dados para ser divulgado, tem que ser considerado seguro. Geralmente os ficheiros de dados originais são inseguros, sendo necessário operar um

[9] Variáveis sensíveis – São variáveis em que pelo menos um dos seus valores é sensível e para as quais o protector de dados deve ser mais rigoroso na sua protecção, nomeadamente o comportamento sexual, o passado criminal.

[10] Macrodados - Informação agregada de entidades e representada em forma de tabelas.

conjunto de modificações de modo a que o ficheiro a disponibilizar esteja suficientemente seguro para divulgação. Estas modificações podem ser efectuadas utilizando métodos ou técnicas de controlo de divulgação estatística.

As técnicas utilizadas destinam-se a anonimizar as bases de dados (criar dados anonimizados[11]) e têm como principal objectivo limitar o risco de descobrir informação sensível sobre os respondentes (ou entrevistados) a partir dos dados divulgados a terceiros.

Os métodos para controlo da divulgação estatística podem ser de dois tipos (Hundepool et al., 2009): métodos perturbativos e métodos não perturbativos. Estes métodos são descritos de forma sintética de seguida e de forma mais detalhada no capítulo 3.

A. Métodos perturbativos

Os métodos perturbativos servem para modificar os valores das variáveis de identificação ou das variáveis confidenciais[12] antes da sua publicação. As combinações únicas de variáveis de identificação num conjunto de dados originais podem desaparecer e surgir uma nova combinação única num conjunto de dados alterado, tornando a identificação incerta. Relativamente às variáveis confidenciais, podem ser modificadas e mesmo que ocorra a identificação é o valor errado que está associado e a divulgação do valor original é evitada. Os métodos de perturbação podem ser utilizados quer por dados categóricos quer por dados contínuos, como se pode verificar na seguinte tabela.

Quadro 1 – Métodos perturbativos. Fonte: Hundepool (2009)

Métodos	Dados contínuos	Dados categóricos
Adição de ruído	X	
Microagregação	X	(X)
Hierarquia de troca		X
Arredondamento	X	
Re-amostragem	X	
PRAM		X
MASSC		X

[11] Dados anonimizados – São dados modificados de forma a minimizar o risco de divulgação.

[12] Variáveis Confidenciais - são variáveis que contêm informação sensível sobre o entrevistado, como o salário; religião; filiação política; estado de saúde, etc.

B. Métodos não perturbativos

Os métodos não perturbativos não alteram os valores das variáveis, sejam elas variáveis identificativas ou variáveis confidenciais. Há sim, uma redução de detalhe no conjunto dos dados original e a produção de supressões parciais (Domingo-Ferrer e Torra, 2001). O Quadro 2 indica quais os métodos a utilizar de acordo com o tipo de variáveis.

Quadro 2 – Métodos não perturbativos. Fonte: Hundepool (2009)

Métodos	Dados contínuos	Dados categóricos
Amostragem		X
Recodificação Global	X	X
Codificação Superior e Inferior	X	X
Supressão Local		X

Capítulo 3. Microdados

Microdados são registos que contém informação de respondentes individuais associados a uma pessoa, família, ou empresa. Apesar de haver um elevado interesse no sentido dos microdados serem o mais detalhados possível, os serviços de estatística têm a obrigação de proteger a confidencialidade dos indivíduos envolvidos. (Hundepool et al, 2009).

As variáveis existentes em ficheiro de microdados individuais são variáveis como o sexo, a idade, a ocupação, o lugar de residência, o país de nascimento, etc. No caso de microdados de empresas, são a actividade económica, o número de empregados, etc.

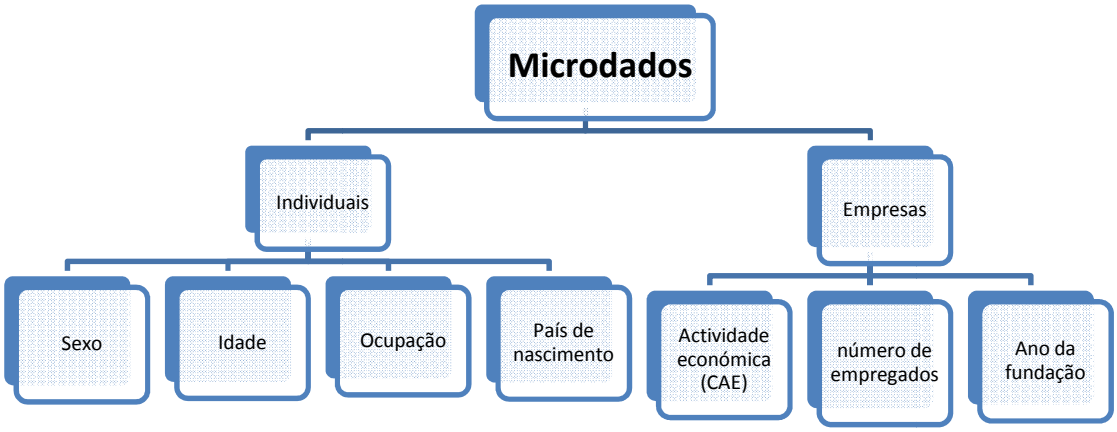

Figura 3 – Tipo de microdados

A aplicação dos métodos de controlo da divulgação estatística para proteger a confidencialidade conduz à perda de informação. A meta para uma estratégia de controlo da divulgação estatística eficaz é a escolha óptima das técnicas que maximizam a utilidade dos dados, minimizando o risco de divulgação.

Existem dois tipos de ficheiros de microdados que podem ser divulgados pelos serviços de estatística: os Ficheiros de Uso Público (FUP) e os Ficheiros de

Investigação (FI)[13]. O risco de divulgação nos FUP e FI são definidos pela aplicação de métodos de CDE e por algumas restrições de acesso e de utilização.

Alguns institutos de estatística permitem o acesso a este e outros ficheiros de microdados em laboratórios de dados, centros de investigação, sendo em alguns casos com acesso/ execução remota. Os utilizadores de dados laboratoriais estão proibidos de revelar informação e estão sujeitos a um controlo bastante rigoroso, por exemplo, verificação de outputs, ajuda no controlo da divulgação. Para os pesquisadores de execução remota são fornecidos microdados com uma descrição mais completa. Depois, os resultados são enviados para os institutos de estatística que executam a análise, fazem a verificação e retorno dos resultados. O acesso remoto a microdados é um recurso on-line seguro, em que os pesquisadores se ligam a um servidor através de uma palavra-chave ou outros dispositivos seguros, onde os dados e os programas estão localizados. O instituto de estatística holandês (CBS) possui um sistema de acesso remoto, o RDC (Research Data Center), que permite o acesso a investigadores autenticados e credenciados. São fornecidos a todos os utilizadores instrumentos estandardizados de metainformação, para cada ficheiro de microdados é produzido um ficheiro de metadados. No caso da Suécia existe o sistema MONA (Microdata on-line acess) que é um sistema de difusão de microdados para a comunidade científica. Este sistema inclui o acesso à meta informação e algumas rotinas para o tratamento de dados no final da pesquisa de informação. O MONA II contempla uma maior diversidade de potencialidades tecnológicas, como o acesso via VPN (acesso virtual aos locais de pesquisa informação), sistemas de base de dados distribuídos, etc. O LISSY é outro tipo de acesso remoto, baseado na base de dados proveniente do LIS (Luxembourg Income Study). A informação contida neste sistema é relativa a inquéritos às famílias de várias zonas do mundo, sendo o seu acesso restringido à investigação nas áreas das ciências sociais.

[13] Os FUP e FI constituem a forma mais comum de divulgar os microdados.

3.1. Os Microdados e o Controlo da Divulgação Estatística

Os utilizadores de microdados assumem um papel duplo na teoria do controlo da divulgação estatística (CDE). Por um lado, o utilizador é visto como um cliente da estatística e, por outro, pode ser considerado um possível intruso (Willenborg, e Waal, 1996).

O utilizador, sendo apenas um cliente de estatística, fica satisfeito com a qualidade dos dados e, normalmente, não está interessado em dados individuais, mas sim em dados estatísticos que resultam da agregação dos dados originais. Num conjunto de microdados divulgados, nem todos os registos têm que ser iguais aos originais. O importante é que o conjunto, como um todo, dê uma correcta ideia da distribuição da população. Por esse motivo, muitas vezes os serviços de estatística alteram os dados por "adição de ruído" ou por troca de registos entre diferentes registos, de forma a reduzir o risco de identificação.

Existem casos em que os utilizadores podem ser encarados como potenciais intrusos. Um intruso tenta combinar registos de conjuntos de microdados com registos de ficheiros de identificação, de indivíduos do seu círculo de conhecimentos ou outros identificadores. Estes últimos podem ser usados para combinar registos de conjuntos de microdados com registos de arquivos identificadores, para facilitar o acesso à identificação dos indivíduos. Neste caso diz-se que houve correspondência entre registos.

A identificação de um registo pode ocorrer se forem satisfeitas as seguintes condições:

1. Os valores da chave são exclusivamente do entrevistado.
2. O entrevistado pertence a um arquivo de identificação ou a um ciclo de conhecimentos do intruso;
3. O entrevistado é um elemento da amostra;
4. O intruso sabe que o registo é único na população sobre a chave;
5. O intruso surge a partir do registo de um conjunto de dados;
6. O intruso reconhece o registo do entrevistado.

Sempre que uma destas condições não se verificar, a identificação pode não ocorrer com toda a certeza. A correspondência pode ocorrer, mas sem certezas para o intruso, se

a primeira ou a quarta condição não se verificarem. A última condição implica que mesmo havendo incompatibilidade de dados, causada por erros de medição ou codificação, entre o conjunto de microdados divulgados e o arquivo de identificação, o intruso consegue reconhecer o registo do entrevistado. Um bom modelo para o risco de identificação deve incorporar aspectos tanto do conjunto de dados como do utilizador.

Para que ocorra a identificação num conjunto de microdados é necessário haver um elevado conhecimento sobre uma população. Em geral, os intrusos têm algum conhecimento do contexto dos dados.

3.2.Guia para a divulgação de ficheiros de microdados

A divulgação do segredo estatístico pode ocorrer de duas formas: por um lado através da divulgação da identidade, que ocorre quando a identidade de um entrevistado corresponde a um registo de dados divulgado (Duncan et al, 2001). Os responsáveis dos serviços de estatística conferem habitualmente maior importância ao risco de divulgação da identidade. Por outro lado, a divulgação pode ocorrer pela divulgação de atributos, que ocorre quando um atributo dos dados divulgados corresponde a um atributo estimado baseado nesses dados. A divulgação de atributos ocorre quando surge algo de novo sobre um entrevistado, ela pode ocorrer com ou sem identificação (Lambert, 1993).

Hundepool et al, (2009), o processo para a divulgação de ficheiros de microdados, ocorre em 5 etapas. Este processo descreve como os dados são processado desde os dados originais até à criação de ficheiros (FUP e FI) para utilizadores externos.

A. Porque é que a protecção da confidencialidade é necessária?
B. Quais são as principais características e utilização dos dados?
C. Riscos de divulgação;
D. Métodos de controlo da divulgação
E. Implementação

Quadro 3 – Processo para a divulgação de ficheiros de microdados. Fonte: Hundepool et al (2009)

Etapas para o processo de divulgação	**Análises a efectuar/Problema resolvido** ↓ **Resultados esperados**
A. Porque é que a protecção da confidencialidade é necessária?	Os dados referem-se a pessoas singulares ou colectivas? ↓ Precisamos proteger a unidade estatística.
B. Quais são as principais características e utilização dos dados?	Análise do tipo/ Estrutura de dados ↓ Visão clara da necessidade de protecção das unidades.
	Análise das metodologias de pesquisa ↓ Tipo de base de amostragem, amostras/enumeração completa dos estratos, análise mais aprofundada da metodologia de pesquisa, calibragem.
	Análise dos objectivos dos Institutos de Estatística ↓ Tipo de ”divulgação” (Ficheiros de Uso Público (FUP), Ficheiros de Investigação (FI)), politicas de divulgação, fenómenos de peculiaridade; coerência entre os várias divulgações (FUP, FI), coerência com as tabelas divulgadas e a base de dados on-line.
	Análise das necessidades dos utilizadores ↓ Variáveis prioritárias, tipo de análises, etc.
	Análise de questionários ↓ Listagem das variáveis a ser removidas, variáveis a ser incluídas, ideias sobre o nível de detalhe das variáveis estruturais.
C. Riscos de divulgação	Cenário de divulgação ↓ Lista de variáveis de identificação
	Definição do risco
	Avaliação do risco ↓ Se o risco é demasiado alto, são necessários métodos de limitação da divulgação.
D. Métodos de controlo da divulgação	Análise do tipo de dados envolvidos, políticas dos Serviços de Estatística e necessidades dos utilizadores ↓ Identificação dos métodos de limitação da divulgação ↓ Análise da perda de informação
E. Implementação	Escolha do software, parâmetros e limites dos diferentes métodos.

Pretende-se identificar, para cada etapa do processo, as escolhas, o tipo de análise, os problemas a serem abordados e os métodos a seleccionar.

A. Necessidade de protecção da confidencialidade

Existem dados que podem ser divulgados sem necessidade de serem protegidos e outros que necessitam dessa protecção. Por exemplo, a quantidade de chuva que cai numa determinada região, pode ser divulgada. No caso de pessoas singulares ou colectivas há, normalmente, necessidade de maior cuidado com a protecção de dados individuais.

B. Características e utilização dos microdados

Existem diferentes tipos de protecção, dependendo dos utilizadores. O que se pretende neste ponto é analisar as características dos dados e dos utilizadores.

A questão que se coloca é se os microdados a divulgar são para o público em geral (FUP) ou para investigação (FI). Quando os microdados divulgados são para fins de investigação, a protecção tem que ser feita de acordo com procedimentos predefinidos, legais e obrigatórios. As diferenças no tipo de utilizadores, implicam diferentes necessidades, diferentes cenários de divulgação, diferentes tipos de análise que se espera ser cumprida com os dados divulgados, diferentes estatísticas que se pretendem manter e qual a protecção que se quer dar.

A análise das necessidades dos utilizadores envolve um estudo sobre o conteúdo informativo do inquérito subjacente aos dados, que deve ser feito por alguém com elevado conhecimento de pesquisa. Devem ser colocadas questões do tipo:

- Quais as unidades estatísticas envolvidas, pessoas ou empresas?
- Os dados apresentam uma estrutura específica, por exemplo, alunos nas escolas, universitários nas universidades?
- Que tipo de amostragem foi utilizada – Existem estratos que foram recenseados? A enumeração completa de diferentes estratos implica maior risco na amostra.

Nesta fase, as variáveis com um grande poder de identificação são, normalmente, agregadas numa única categoria. Por exemplo: num inquérito às despesas das famílias, para uso público, pode-se evitar a informação muito detalhada sobre as despesas da casa, os anos que a casa tem ou o número de quartos.

C. Riscos de divulgação

No caso da divulgação por diversas formas e com diversas bases de dados do mesmo inquérito, deve ser mantida a coerência entre arquivos diferentes e ao mesmo tempo não deve ser permitida a obtenção de mais informação do que a que se teria para uma base de dados.

A etapa final da avaliação do risco é a definição de um limite para definir quando é que uma unidade ou um arquivo apresentam um risco aceitável e quando, pelo contrário, é considerado um risco inaceitável. Este limite tem a ver com o tipo de medidas adoptadas. Se na avaliação do risco se determina que o risco de divulgação é alto, é necessário tomar medidas de protecção dos dados. A escolha dos cenários e do nível de risco aceitável é fortemente dependente da diferente cultura dos países, das políticas aplicadas nos serviços de estatística e das abordagens da análise estatística. De referir que os diferentes países podem ter situações e fenómenos completamente diferentes, logo diferentes cenários e métodos.

Actualmente não há nenhum acordo sobre a melhor metodologia de risco, apesar de métodos diferentes poderem fornecer respostas semelhantes.

D. Métodos de Controlo da Divulgação Estatística (CDE)

Para Domingo-Ferrer e Torra (2001), o objectivo do Controlo da Divulgação Estatística é fornecer aos utilizadores um conjunto de microdados mascarado (ou alterado) ***V'***, semelhante ao conjunto de microdados original ***V***, para que o risco de divulgação seja baixo e a análise do utilizador em V e em V' tenha o mesmo resultado ou um resultado similar.

Os métodos de protecção de microdados podem gerar um conjunto de microdados protegido ***V'***, quer por mascarar os dados originais, ou seja, gerar uma versão modificada ***V'*** dos dados originais ***V***, quer por gerar dados sintéticos ***S*** que preservam algumas características estatísticas dos dados originais ***V*** (Hundepool et al., 2009).

Vários métodos existem para este fim.

E. Implementação

Nesta etapa procede-se à escolha do software e do risco aceitável, bem como à implementação das medidas CDE.

3.3. Métodos Perturbativos

O método de perturbação utilizado deve ser tal que as estatísticas calculadas sobre o conjunto de dados perturbado (ou alterado) não diferem significativamente das estatísticas que seriam obtidos no conjunto de dados original (Domigo-Ferrer e Torra, 2001).

3.3.1. Adição de ruído

Flossmann e Lechner., (2006) referem que uma forma simples de proteger os dados é através da adição de ruído nas covariâncias. Supondo que a variável explicativa $\boldsymbol{X_i}$ contém informação sensível que tem que ser protegida para a divulgação, essa variável vai ser mascarada e o que se observa é a variável explicativa mascarada $\boldsymbol{X_i^m}$ e não a variável explicativa $\boldsymbol{X_i}$.

$$X_i^m = X_i + u_i \qquad (3.1)$$

u_i – Variável aleatória independente que é adicionada à variável original de forma a ser mascarada.

A variável tem: $E[u_i / X_i]=0$; $V[u_i / X_i]= \sigma_u^2$

Para Domingo-Ferrer e Torra (2001), a adição de ruído consiste na adição de ruído aleatório com a mesma estrutura de correlação dos dados originais. É único método que actualmente pode preservar a correlação.

Existem vários algoritmos de adição de ruído, nomeadamente adição de ruído não correlacionado e adição de ruído correlacionado (Hundepool et al, 2009).

3.3.1.1. Adição de ruído não correlacionado

Neste método, o vector de observações $\boldsymbol{x_j}$ para a j$^{\underline{\text{esima}}}$ variável do conjunto de dados $\boldsymbol{X_j}$ é substituído pelo vector $\boldsymbol{Z_j}$:

$$Z_j=X_j+\varepsilon_j \tag{3.2}$$

ε_j – vector de erros, habitualmente com distribuição normal.

$\varepsilon_j \sim N(0, \sigma^2_{\varepsilon j})$, em que Cov $(\varepsilon_j, \varepsilon_l)$, para t≠1 é um ruído branco

Pressuposto: V(ε_j) é proporcional à variância das variáveis originais:

$$V(X_j) = \sigma^2_j \tag{3.3}$$

$$\sigma^2_{\varepsilon j} = \alpha\sigma^2_j \tag{3.4}$$

O método da adição de ruído não correlacionado mantém a média e a covariância, mas não mantém a variância nem o coeficiente de correlação.

$$E(Z) = E(X) + E(\varepsilon)= E(X)= \mu \tag{3.5}$$

$$Cov(Z_j, Z_l)=Cov((X_j, X_l)\ \forall\ j\neq l \tag{3.6}$$

$$V(Z_j)=V(X_j) + \alpha V(X_j)=(1+\alpha)V(X_j),\quad \forall\ j\neq l \tag{3.7}$$

O coeficiente de correlação é dado pela seguinte função:

$$\rho z_j, z_l = \frac{Cov(Z_j, Z_t)}{\sqrt{V(X_j)V(X_l)}} = \frac{1}{1+\propto}\rho x_j, x_l \tag{3.8}$$

3.3.1.2. Adição de ruído correlacionado

A adição de ruído correlacionado mantém a média e permite a preservação do coeficiente de correlação. Neste método a matriz das covariâncias dos erros é proporcional à matriz das covariâncias dos dados originais, isto é, $\varepsilon \sim (0,\Sigma)$, onde $\Sigma_\varepsilon = \alpha\Sigma$ (ruído correlacionado).

A matriz das covariâncias dos dados mascarados é dada pela seguinte expressão:

$$\sum_z = (1+\alpha)\sum \qquad (3.9)$$

O coeficiente de correlação mantém-se desde que:

$$\rho z_j, z_l = \frac{1+\alpha}{1+\alpha}\frac{Cov(Z_j, Z_t)}{\sqrt{V(X_j)V(X_l)}} = \rho x_j, x_l \qquad (3.10)$$

A adição de ruído correlacionado oferece dados mascarados de melhor qualidade do que a adição de ruído não correlacionado. A adição de ruído simples não é utilizada com muita frequência, uma vez que não oferece um nível de segurança muito grande.

3.3.2. Dados distorcidos pela probabilidade de distribuição

O método da distorção de dados pela probabilidade de distribuição é um método que pode ser utilizado tanto por dados categóricos, como por dados contínuos. Este método processa-se da seguinte forma (Domigo-Ferrer e Torra, 2001):

1. Identifica a função densidade de cada uma das variáveis do conjunto de dados confidenciais e estima os parâmetros associados à função densidade. Selecciona a série original das variáveis confidenciais para determinar qual de um conjunto de funções densidade predeterminada melhor se ajusta aos dados, o que pode ser testado pelo Test Kolmorov-Smirnov;
2. Gera uma série de distorção para cada variável confidencial da função densidade estimada;
3. Substitui a série confidencial pela série distorcida.

3.3.3. Microagregação

Para Domingo-Ferrer e Torra (2001), na microagregação os registos são agrupados em pequenos grupos, de pelo menos k. Para um dado registo em vez de se publicar a variável original V_i, é publica-se a média dos valores de $V_i$do grupo ao qual pertence.

Este método diz o seguinte: as regras de confidencialidade permitem a divulgação do conjunto de microdados se, nos registos correspondentes ao grupo de k ou mais

indivíduos não houver nenhum indivíduo dominante (isto é, que contribua muito) e se k é o valor limite. Os grupos devem ser o mais homogéneos possível para que a perda de informação seja mínima.

Os grupos podem ser de tamanho fixo ou de tamanho variável, os últimos resultam em grupos mais homogéneos, logo com menor perda de informação. Mateo-Sanz e Domingo-Ferrer (1999) optam por investigar métodos de microagregação de conjuntos de dados homogéneos em vez de grupos de tamanho fixo (Hansen e Mukherjee, 2003).

Como mencionado em Hundepool et al., (2009), dado um conjunto de microdados com $\boldsymbol{p}$ variáveis contínuas e n registos (indivíduos), em que um registo particular pode ser visto como, $\boldsymbol{X'} = (X_1, \ldots, X_p)$, $\boldsymbol{X_i}$ são variáveis. São formados **g** grupos com $\boldsymbol{n_i}$ indivíduos no i^{ésimo} grupo ni > k e n = $\sum_{i=1}^{g} n^i$.

x_{ij} – Representa o j^{ésimo} registo do i^{ésimo} grupo;

$\overline{x_i}$ - Média do registo do i^{ésimo} grupo;

$\bar{x}$ - Média do conjunto de "n" indivíduos.

Do ponto de vista da perda de informação, a partição óptima $\boldsymbol{k}$ é a que maximiza a homogeneidade dentro do grupo. Maior homogeneidade no grupo, menor perda de informação. Quanto maior for a homogeneidade do grupo, menor é a soma dos quadrados, ou seja, a partição óptima é a que minimiza a soma dos quadrados. Assim tem-se:

$$\text{SSE} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \overline{x_i})(x_{ij} - \overline{x_i}) \tag{3.11}$$

$$\text{SST} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})(x_{ij} - \bar{x}) \tag{3.12}$$

A microagregação é utilizada para mascarar registos individuais de forma a protege-los da identificação. A microagregação pode ser modelada matematicamente como um problema de clusters, onde o objectivo é agrupar dados em grupos de dimensão igual a $\boldsymbol{k}$ e o mais homogéneos possível (Domingo-Ferrer e Sebé, 2006).

Dado o parâmetro k:

1. Partir os registos X em grupos de registos de pelo menos n registos cada.

2. Substituir cada registo pelo centroide do grupo para obter o conjunto de dados mascarado X'.

Num conjunto de dados microagregados, a identificação não é possível uma vez que todos os registos do grupo são idênticos. O melhor que o intruso pode conseguir é identificar o grupo onde o indivíduo alvo foi mascarado.

Os métodos univariados lidam com conjuntos de dados multivariado por microagregação de uma variável de cada vez, ou seja, as variáveis são sequencialmente e independentemente microagregados.

3.3.4. Re-Amostragem

Para Domingo-Ferrer e Mateo-Sanz (1999 e Heer, 1993 cf Domingo-Ferrer e Torra, 2001) este método foi originalmente proposto para proteger dados tabulares, mas também pode ser utilizado na protecção de microdados.

Seja ***V*** uma variável original num conjunto de dados com ***n*** registos e ***t*** amostras independentes $X_1, \ldots, X_t$. Todas as amostras são ordenadas usando o mesmo critério de classificação, depois é criada uma variável mascarada como $\overline{x_1}, \ldots, \overline{x_n}$, em que:

n – é o número de registos;

$\overline{x_j}$ - é a média do j^{ésimo} valor classificado em $X_1, \ldots, X_t$.

Assuma que os microdados $z_1, \ldots, z_n$ são agregados para criar macrodados numa tabela de contingência ***X***, com ***I*** linhas e ***J*** colunas, e com determinadas especificações, x_{ij} é a frequência original da i^{ésima} linha e da j^{esima} coluna. Com vista a criar uma tabela anonimizada ***X'***, a amostra $z_1', \ldots, z_n'$ é obtida dos dados originais $z_1, \ldots, z_n$ ***n*** vezes e com substituição. Assim, a tabela ***X'*** é uma estimativa da tabela original ***X***, não permitindo obter nenhuma informação precisa de ***X***.

3.3.5. Rank Swapping

A troca de dados foi inicialmente introduzida como sendo um método de controlo da divulgação estatística para variáveis categóricas. A ideia base é transformar um ficheiro de dados através da troca de valores das variáveis confidenciais entre os registos individuais. Para Reiss, Post and Dalenius (1982, cf Hundepool, et al, 2009), a troca de dados foi introduzida para proteger microdados contínuos, por outro lado, Reiss (1984, cf Hundepool, et al, 2009) refere que é utilizada para a protecção de microdados categóricos.

A hierarquia de troca é uma variante da troca de dados, utilizada originalmente por variáveis ordinais (Greenberg, 1987, cf Hundepool et al., 2009), que também pode ser utilizada por variáveis numéricas (Moore, 1996, cf Hundepool et al., 2009). Os valores das variáveis são, classificados por ordem crescente. Depois cada valor ordenado é trocado aleatoriamente, dentro de um intervalo restrito, por outro valor ordenado. As estatísticas calculadas a partir deste algoritmo são menos distorcidas do que as calculadas após uma troca livre. Este método apresenta bons resultados comparando o risco de divulgação com a perda de informação.

3.3.6. Arredondamento

O arredondamento consiste na substituição do valor das variáveis originais por valores arredondados. Os valores arredondados são escolhidos de entre um conjunto de pontos arredondados que definem um conjunto de arredondamento (Domingo-Ferrer e Torra, 2001). Num conjunto de dados originais multivariado, o arredondamento é feito, normalmente, variável a variável - arredondamento univariado (Domingo-Ferrer e Torra, 2001). Para Willenborg e Waal (2001), também é possível haver um arredondamento multivariado.

Por exemplo, considerando a variável contínua X, é determinado um conjunto de pontos de arredondamento $\{p_1,...,p_r\}$, através do arredondamento dos pontos em múltiplos do valor base "b": p_i=b*i, para i=1, …, r. O conjunto de atracção para cada ponto arredondado p_1, é definido como o intervalo $[p_i - b/2, p_i + b/2]$, para i=1, …, (r-

1). Para p_1, o conjunto de atracção é $[0, p_1 + b/2]$; para p_r, o conjunto é dado pelo intervalo $[p_r - b/2, X_{max}]$.

X_{max} – é o maior valor possível para a variável X.

O valor original x de X é substituído pelo valor arredondado do conjunto de atracção.

3.3.7. PRAM (Post Randomization method)

Segundo Hundepool et al., (2009), o PRAM é um método de controlo de divulgação estatística que pode ser aplicado em dados categóricos. É um método perturbativo e probabilístico para protecção de ficheiros de microdados. Alguns métodos, como o método da recodificação global, supressão local e codificação superior e inferior podem conduzir a uma elevada perda de informação para tornar os ficheiros de dados seguros. O método PRAM é uma alternativa, uma vez que é mantida a quantidade de detalhe, enquanto o nível de controlo da divulgação é feito através da introdução da incerteza nos resultados sobre as variáveis de identificação.

PRAM pode ser usado para produzir ficheiros de microdados com a mesma estrutura do ficheiro de microdados original, mas com algum tipo de dados sintéticos. Também pode produzir ficheiros de dados seguros e deixar algumas características do ficheiro mais ou menos inalteradas.

PRAM é um método que surgiu em 1997 e é definido em termos de probabilidades de transição, resumidas numa matriz PRAM (Wolf et al, 1998). Produz ficheiros de microdados em que os valores de algumas variáveis categóricas para determinados registos são alterados em relação aos valores do ficheiro de microdados original. É aplicado normalmente a variáveis de identificação, ou seja variáveis que são usadas para identificar o entrevistado. O resultado é obter ficheiros de microdados com valores incorrectos nas variáveis de identificação, o que torna o risco de identificação pequeno. O método PRAM pode ser considerado como uma forma de erro de classificação.

ξ - Variável categórica no ficheiro de dados original para que PRAM seja aplicado

X – Mesma variável no ficheiro de dados alterado.

ξ e X têm K categorias rotuladas por 1, …, K.

A probabilidade de transição que define o PRAM é dada pela seguinte expressão:

$$p_{kl} = P(X=l/\xi=K) \quad (3.13)$$

A probabilidade de um valor original ξ=K é transformada numa contagem X=l – Probabilidade de transição para todos os K, l=1, …, K.

A matriz Markov ou matriz PRAM, é obtida usando as probabilidades de transição, isto é, a probabilidade um valor ser alterado numa matriz KxK.

P – Matriz Markov ou matriz PRAM

A perda de informação e o risco de divulgação dependem essencialmente da escolha da matrix Markov. A perda de informação e a limitação da divulgação dependem da escolha das probabilidades de transição, por isso, é importante fazer uma escolha adequada dessas probabilidades (Hundepool et al, 2009).

Aplicando PRAM, significa que dado o valor ξ=K para o registo r, o valor X para esse registo é dada pela probabilidade p_{kl}, …, p_{kk}. Este procedimento é feito para cada registo do ficheiro de dados original independentemente dos outros registos (Wolf, 2006).

Considere o seguinte exemplo:

ξ - Variável sexo;

ξ=1 – Se é do sexo masculino;

ξ=2 – Se é do sexo feminino

Em que 110 registos são do sexo masculino e 90 do sexo feminino, a aplicação do método PRAM, com $p_{11}=p_{22}=0,9$produziria um ficheiro de microdados perturbado (ou mascarado, alterado) com um valor esperado de 108 do sexo masculino e 92 feminino. No entanto, 9 desses homens eram originalmente do sexo feminino e 11 mulheres, eram originalmente do sexo masculino.

O efeito de PRAM numa tabela de frequências unidimensional é dado pela seguinte expressão:

$$E(T_X/\xi)= P^tT_{\xi} \quad (3.14)$$

$T_\xi = (T_\xi(1), \ldots, T_\xi(K))^t$ – Frequência da tabela de acordo com o ficheiro de microdados original;

T_X – Frequência da tabela de acordo com o ficheiro de microdados perturbado.

Estimador enviesado da tabela de frequências do ficheiro original:

$$\hat{T}_\xi = (P^{-1})^t T_X \tag{3.15}$$

A aplicação pode ser estendida às tabelas de frequência bi-dimensionais. Pode ser usada a tabela de frequência bi-dimensional $T_{\xi\eta}$ para dados originais e T_{XY} para dados perturbados. Assim tem-se:

$$\hat{T}_{\xi\eta} = (P_X^{-1})^t T_{XY} P_Y^{-1} \tag{3.16}$$

$\mathbf{P}_X$ – Matriz PRAM correspondente à variável categórica X;

$\mathbf{P}_Y$ – Matriz PRAM correspondente à variável categórica Y;

Como referido em Wolf, et al (1998), a regra de limite é utilizada, frequentemente, para determinar se um dado ficheiro de microdados é ou não seguro. Esta regra consiste no seguinte: sempre que uma combinação de resultados de variáveis de identificação for inferior a um determinado valor, essa combinação é considerada insegura.

Para se entender melhor esta regra, considere-se um exemplo em que, a combinação do sexo, ocupação e idade devem ser verificada pela regra de limite. Assumindo que o valor limite[14] é de 50, se apenas existem 43 cirurgiãs do sexo feminino com 57 anos, cada registo que corresponda a uma cirurgiã de 57 anos é considerado inseguro.

Se for utilizado o método PRAM para controlo da divulgação estatística, esta regra não faz muito sentido: desde que um ficheiro perturbado resulte de uma experiência probabilística, os registos inseguros variam em cada realização. Para resolver este problema deve-se considerar o risco de divulgação, isto é, a probabilidade de um dado valor k ser do ficheiro perturbado e e do ficheiro original.

A Regra de Bayes é dada pela seguinte função:

[14] Valor limite – Valor abaixo do qual um registo é considerado inseguro para divulgação

$$R_{PRAM}^{(k)}=IP(\xi=k/X=k)=\frac{IP(X=k/\xi=k)IP(\xi=k)}{\sum_{l=1}^{K}IP(X=k/\xi=l)IP(\xi=l)} \quad (3.17)$$

$$\hat{R}_{PRAM}^{(k)}=\frac{p_{kk}T_{\xi}(k)}{\sum_{l=1}^{K}p_{lk}T_{\xi}(l)} \quad (3.18)$$

$T_{\xi}(k)/n$ - Estimador de $IP(\xi=k)$, em que "n" é o tamanho do ficheiro de microdados original.

Um registo é considerado seguro se:

$$\hat{R}_{PRAM}^{(k)} \leq \frac{T_{\xi}(k)}{\tau} \quad (3.19)$$

τ - Limite usado na regra de limite para o ficheiro de microdados original.

Se um registo é seguro de acordo com a regra de limite aplicada ao ficheiro original, então este também é seguro de acordo com a mesma regra.

3.3.8. Microdados Sintéticos

A publicação de dados sintéticos é uma forma de protecção contra a divulgação estatística de dados. Os dados são gerados de forma aleatória, preservando algumas estatísticas ou relações internas do conjunto de dados original (Hundepool et al., 2009).

3.3.8.1. Um precursor: distorção de dados por uma distribuição de probabilidade

A distorção dos dados através da distribuição de probabilidades foi proposta em 1985 por Liew, Choi e Liew (1985). Este método pode ser utilizado tanto em variáveis categorias, como em variáveis contínuas. Pretende-se obter um conjunto de dados protegido aleatoriamente a partir do conjunto de dados original.

A distorção de dados é realizada em três etapas, conforme se segue:

1. *Identificar a função densidade subjacente para cada variável confidencial*[15] *no conjunto de dados e estimar os parâmetros associados com a função densidade*

   As séries originais das variáveis confidenciais (por exemplo, salários) são analisadas de forma a determinar se um dado conjunto de uma função densidade se ajusta melhor aos dados. O que pode ser verificado pelo teste de Kolmogorov- Smirnov. Se forem aceites várias funções de densidade para um dado nível de significância, deve-se escolher a que apresenta o menor valor do teste Kolmogorov-Smirnov. No caso de nenhum conjunto predeterminado pela função densidade se ajustar aos dados, deve ser utilizada a frequência imposta pelo método da distorção.

   No método da distorção, a série original é dividida em vários intervalos, dentro do qual são contadas as frequências, para a série original, tornando-se numa orientação para gerar as séries de distorções. A série distorcida é gerada até a sua frequência se tornar a mesma frequência da série original. Se, em alguns intervalos, houver frequências em excesso elas são descartadas.

2. *Gerar uma série obtida aleatoriamente a partir da função densidade, para cada variável confidencial*

   Depois de escolhida a melhor função densidade, são estimados os parâmetros para gerar um valor aleatório e produzir a serie distorcida.

3. *Mapeamento*

   O mapeamento consiste em classificar a série alterada e a série original na mesma ordem e substitui cada elemento da série original com o correspondente elemento da serie alterada.

   O mapeamento e fase de substituição são necessários apenas se as variáveis alteradas estiverem a ser utilizadas em conjuntos com outras variáveis não alteradas.

[15] Variáveis confidenciais – São variáveis que contêm informação sensível sobre o entrevistado, como o salário, a religião, estado de saúde, filiação política, etc.

3.3.8.2. Abordagem dos microdados híbridos

A abordagem de microdados híbridos consiste no cálculo de dados mascarados como uma combinação de dados originais e de dados sintéticos. Esta combinação permite um melhor controlo dos dados totalmente sintéticos sobre as características individuais dos dados mascarados (Dandekar et al, 2002).

O "mascaramento híbrido" envolve a combinação de dados originais com dados sintéticos.

Exemplo:

Considere um conjunto de dados original com n registos e um conjunto de dados sintéticos com m registos e que os dois conjuntos de dados têm o mesmo número de variáveis numéricas d. A distância euclidiana pode ser usada para combinar os dados originais com os sintéticos da seguinte forma:

1. As variáveis nos dois conjuntos de dados são estandardizadas (subtrair os valores da cada variável pelo seu valor médio e dividir pelo seu desvio-padrão).
2. Cada par de registos do ficheiro de dados estandardizado com o registo mais próximo do conjunto de dados sintéticos estandardizado, em que o mais próximo significa a distância euclidiana mais pequena.

Depois de determinar os pares é necessário um modelo para misturar as varáveis em pares de registos de forma a obter um conjunto de dados híbridos mascarado. Estes autores sugerem, para as variáveis numéricas, a combinação aditiva e a combinação multiplicativa para combinar um registo original com um registo sintético X_S.

No entanto estes dois modelos têm como inconveniente o facto de os pares misturados utilizados no modelo híbrido resultarem num conjunto de dados mascarado com o mesmo número de registos dos dados originais, o que origina uma perda de flexibilidade comparado com os dados sintéticos puros. Uma forma de superar esta situação é utilizar dados re-amostrados originais em vez dos dados originais.

Supondo que o conjunto de dados originais é composto por n registos e que o conjunto de dados mascarado tem n' registos, os n' registos mascarados podem ser obtidos através do seguinte algoritmo:

1. Re- amostragem com a substituição de ***n*** registos do conjunto de dados original para obter ***n'*** de um conjunto de dados re-amostrados.
2. Cada registo re-amostrado ***n'*** deve ser emparelhado com o registo mais próximo do conjunto de dados sintéticos, onde o mais próximo significa a menor distância euclidiana.
3. Dentro de cada par de registos, a mistura de variáveis pode ser feita através da combinação aditiva ou da combinação multiplicativa.

O produto da combinação aditiva é dado pela seguinte expressão:

$$Z=X\,\alpha + (1-\alpha)X_S \qquad (3.20)$$

O produto da combinação multiplicativa é dado por:

$$Z= X^{\alpha}X_S^{(1-\alpha)} \qquad (3.21)$$

α – Parâmetro de entrada entre $[0,1]$

Z – Registo híbrido

3.3.8.3. Microagregação híbrida

No âmbito deste trabalho, introduz-se aqui um método novo para mascaramento da informação através de dados híbridos. Este método pode ser visto como uma combinação da migroagregação numérica e do mascaramento híbrido. Consiste em combinar os resultados provenientes de um processo de microagregação com os dados originais. O processo ocorre da seguinte forma:

1. Criação de k classes para cada variável (trata-se de um tratamento univariado)
2. Calcular a média para cada uma das classes e substituir todos os valos da classe pela sua média
4. Para cada registo, efectuar uma combinação aditiva Z da seguinte forma:

$$Z=X\,\alpha+ (1-\alpha)X_S \qquad (3.22)$$

Sendo:

α – Parâmetro de entrada entre $[0,1]$

Z – Registo híbrido

3.4.Métodos não perturbativos

Os métodos não perturbativos não dependem da distorção dos dados originais, mas sim da supressão ou da redução do detalhe. Existem diferentes tipos de métodos perturbativos, são eles (Domingo-Ferrer e Torra, 2001):

3.4.1. Amostragem

No método da amostragem, os dados a serem publicados não são os microdados originais, mas sim, uma amostra $\boldsymbol{S}$ desse ficheiro. Este método é utilizado em microdados categóricos e em microdados contínuos.

Num cenário de divulgação geral a amostragem não é tão adequada, uma vez os valores de uma variável contínua $\boldsymbol{V_i}\ (\bullet)$ não perturbada (ou mascarada ou alterada) persistem em todos os registos da amostra S. Se uma variável contínua $\boldsymbol{V_i}$ pertencer a um ficheiro público externo, no caso de haver dois entrevistados é improvável que o valor de $\boldsymbol{V_i}$ seja igual para os dois registos, ou seja, $V_i\,(O_1) = V_i\,(O_2)$ se $O_1 \neq O_2$.

Para Willenberg e Wall (1996, cf Domingo-Ferrer e Torra, 2001), se a variável identificativa é contínua e se o registo de um entrevistado é aproximadamente conhecido de um intruso, então a variável deve ser protegida pelo método da amostragem.

3.4.2. Recodificação global

A recodificação global é uma técnica mais apropriada para microdados categóricos, que ajuda a disfarçar os registos com combinações "estranhas" de variáveis categóricas (Domingo-Ferrer, Torra, 2001).

Dada uma variável categórica V_i, são combinadas várias categorias para criar novas categorias, que resultam numa nova variável ***V'i***, com $\left|D(V'_i) < D(V_i)\right|$, em que $\left|\bullet\right|$ é o operador de cardinalidade. A recodificação global em variáveis contínuas significa a substituição da variável V_i pela variável V'_i.

- Exemplos: Supondo que há um registo com o estado civil de viúva/o e com idade de 47 anos, a recodificação global pode ser aplicada ao estado civil, através da criação de uma categoria mais ampla que contempla o estado de viúva/o e divorciada/o. Esta nova categoria reduziria a probabilidade de o registo ser único.
- A recodificação global também pode ser aplicada à variável ocupação, combinando as categorias, de estatístico e matemático em apenas uma categoria: estatístico ou matemático. Se o número de mulheres estatísticas e o número de mulheres matemáticas na cidade de Urk é suficientemente alto, pode-se considerar que a combinação das variáveis: local de residência = Urk; Sexo = feminino; ocupação = Estatístico ou matemático é uma combinação segura para ser divulgada.

A recodificação global é aplicada a todos os dados de um conjunto e não apenas à parte insegura do conjunto, o que é feito para obter uma classificação uniforme de cada variável, (Hundepool et al, 2009).

3.4.3. Codificação superior e inferior

A codificação superior e inferior é um caso específico do método de recodificação global, que pode ser usado em variáveis ordinais contínuas ou categóricas. Este método baseia-se no seguinte: os valores superiores são agrupados para formar uma nova categoria, o mesmo é feito com os valores inferiores (Domingo-Ferrer e Torra, 2001).

3.4.4. Supressão local

A supressão local quando é utilizada para protecção dos dados, os valores inseguros da combinação são suprimidos, ou seja, passam a ter um valor em falta. Este método é fundamentalmente orientado para variáveis categóricas, no entanto também pode ser utilizado em variáveis contínuas (Hundepool et al, 2009).

Considerando o exemplo anterior, a combinação: local de residência = Urk; sexo = feminino; ocupação = estatístico, insegura, podemos protege-la através da supressão do valor da ocupação. A combinação segura obtida com esta supressão seria: local de residência = Urk; sexo = feminino; ocupação = missing.

A supressão local é apenas aplicada a um valor particular, se o valor de uma variável for suprimido num dado registo, isso não implica que o valor dessa variável seja suprimido noutros registos.

O facto de se ter liberdade para seleccionar os valores que devem ser suprimidos, permite minimizar o número de supressões locais. A forma mais fácil para determinar os valores das variáveis que devem ser suprimidos localmente é fazendo-o para cada combinação a ser verificada e para cada registo separadamente, o que pode ser feito de duas formas (Willenborg e Waal, 1996):

1. Definir de imediato o valor suprimido como missing, resultando num conjunto de microdados, que é utilizado para determinar se as combinações são ou não seguras.
2. Através da utilização do conjunto de microdados original para determinar se a combinação é ou não segura.

No entanto estas duas formas têm alguns inconvenientes, quando é aplicado o primeiro método, pode parecer, incorrectamente, que algumas combinações não ocorrem com frequência suficiente, considerando a combinação insegura. Na realidade a combinação pode ocorrer com frequência suficiente para a considerar segura.

Se o que se pretender é reduzir o número de supressões locais, não se pode decidir quais os valores que devem ser suprimidos para cada combinação insegura e registo separadamente, mas sim em simultâneo. Quando há um grande número de combinações inseguras é recomendável utilizar a supressão local para suprimir algumas dessas combinações.

Capítulo 4. Dados tabulares (macrodados)

As tabelas estatísticas mostram a soma das observações de uma variável quantitativa através de todas as observações e / ou dentro de grupos de observações de um dado conjunto. Cada observação refere-se apenas a um indivíduo. O conjunto das observações é definido pelas variáveis categóricas observadas por cada um dos entrevistados, que podem ser indivíduos, famílias, empresas, etc. Normalmente uma tabela contém informações sobre um colectivo, cujos membros têm características em comum (Hundepool et al, 2009).

Geralmente, o conjunto de variáveis fornece informação geográfica, económica, etc. sobre os entrevistados. As células da tabela são definidas por combinações cruzadas de agrupamentos de variáveis. Assim, cada tabela refere-se a um grupo de entrevistados.

A dimensão duma tabela é dada pelo número de conjuntos de variáveis usados para a especificar. Uma tabela contém células marginais, se nem todas as células de uma tabela são especificadas pelo mesmo número de conjuntos de variáveis. Quanto menor for o número de conjuntos de variáveis, maior é o nível de células marginais.

Cada célula duma tabela apresenta a soma de uma variável quantitativa, como a renda, o volume de negócios, despesas, etc., estes montantes são os valores das células de uma tabela de magnitude. As observações individuais da variável são os contributos para o valor da célula.

Existem muitas semelhanças entre os microdados e os dados tabulares no controlo da divulgação estatística, como é o caso da segurança dos dados, a recodificação global e a supressão local. Nas tabelas tem-se por exemplo, o colapso das linhas e/ou colunas, o redesenho da tabela (recodificação global) e a supressão de células (supressão local).

Outra semelhança entre os dois tipos de dados no controlo da divulgação estatística é a adição de "ruído" nas variáveis sensíveis. Normalmente a adição de ruído nas tabelas é feito de uma forma muito ordenada: todos os valores na tabela original são arredondados para um dos dois mais próximos múltiplos de um valor base escolhido. As tabelas marginais dificultam este arredondamento.

É importante referir que também existem diferenças entre as tabelas e os microdados no controlo da divulgação estatística. Os critérios utilizados nos microdados para

considerar as combinações raras são diferentes dos utilizados nas tabelas. Nas tabelas com dados contínuos é utilizado o critério da regra de dominância. Ao aplicar a regra de dominância numa tabela de frequência de dados implica declarar que as células são sensíveis se o seu valor for inferior a um determinado limite. Situação semelhante à que acontece com o conjunto de microdados quando as combinações são raras num conjunto de microdados inseguro. Outra diferença tem a ver com o facto de nas tabelas se assumir que a informação apresentada é de uma população e não apenas de uma amostra.

4.1.Tabela com dados de magnitude

O controlo da divulgação estatística dos dados tabulares tem como objectivo impedir o utilizador dos dados de inferir com precisão pequenos valores de dados categóricos ou de contribuições de um entrevistado para um valor total das células em dados de magnitude. Durante muito tempo a supressão de dados era o único método adequado ao controlo da divulgação estatística nos dados tabulares. Mais recentemente, em dados categóricos, esse controlo pode ser alcançado através de outros métodos, tais como o arredondamento e a perturbação (Willenborg e Waal, 1996).

4.2.Procedimentos para o Controlo de Divulgação Estatística (CDE)

Os procedimentos de limitação da divulgação estatística são analisados de acordo com as seguintes etapas:

1º Determinação das células sensíveis

Uma célula tabular é considerada sensível se o verdadeiro valor da célula divulgado for susceptível de identificar o contribuinte. Existem variadas formas de determinação de células sensíveis, a mais utilizada é a regra de dominância – (***n, k***), que diz o seguinte: "Uma célula é considerada sensível se a soma das ***n*** maiores contribuições representam mais do que ***K%*** do valor total das células."

Esta regra significa que se o valor de uma célula é dominado pelo valor de um entrevistado, a contribuição desse entrevistado pode ser estimada com elevada precisão pelo valor total da célula.

Geralmente é utilizado um valor baixo para o parâmetro n, (n <5) e ***K*** assume um valor alto (k>100). Deve-se dar especial atenção quando n=1 e n=2. Nestes casos, facilmente se consegue fazer uma boa estimativa do valor da célula, uma vez que ela é dominada predominantemente pela contribuição de um ou dois inquiridos respectivamente. Por este motivo, o número mínimo de inquiridos de uma célula deve ser três.

Além da regra da dominância para a determinação das células sensíveis existe também a regra "priori-posterior", que utiliza os parâmetros ***p*** e ***q***, em que ***p <q***. Esta regra diz o seguinte: todos os inquiridos podem estimar a contribuição de cada um dos outros dentro de ***q%*** do seu respectivo valor.

2º Reformulação da tabela

Nesta etapa procede-se à recolha de algumas linhas e/ou colunas que contêm muitas células sensíveis e constrói-se uma tabela com uma nova classificação. Após a construção da nova tabela, deve-se verificar se ainda existem células sensíveis ou não, no caso de ainda existirem muitas células sensíveis, é recomendável agregar mais variáveis. Caso contrário, podem ser tomadas as medidas locais de controlo da divulgação estatística.

3º Supressão

Deve-se proceder à supressão das restantes células sensíveis e caso seja necessário, deve-se adaptar células adicionais na tabela de forma a proteger as células sensíveis.

A supressão de uma célula é chamada de supressão primária, que normalmente não é suficiente para obter uma tabela de dados segura para divulgação. O valor de uma célula que foi suprimida pode ser recalculado através dos totais marginais, o que se designa por supressão secundária, que visa optimizar a função objectivo, expressa em perda de informação. Por exemplo, pode-se tentar minimizar o número de entrevistados cujos dados são suprimidos na tabela ou tentar minimizar o valor total de dados que sejam suprimidos.

A supressão secundária acarreta alguns problemas, nomeadamente a selecção da função objectivo, que não é simples de determinar, muitas vezes baseada em considerações subjectivas. Outra questão prende-se com a possibilidade de calcular o intervalo para os valores das células "mentira". Para o intruso é mais fácil obter boas estimativas se o intervalo dos valores viáveis nas células suprimidas for pequeno.

4.3.Métodos de controlo da divulgação

À semelhança do que acontece com os microdados existem métodos de controlo da divulgação estatística para macrodados. De seguida faz-se uma pequena abordagem desses métodos.

4.3.1. Reformulação da tabela

Uma tabela que contenha várias células sensíveis, torna-se necessário proceder à sua reformulação, isto é, alterar o seu esquema de classificação, reduzindo o detalhe da informação estatística. Pretende-se que, com a redução do detalhe da informação na tabela, o número de células sensíveis diminua.

Para compreender melhor esta medida, considera-se o seguinte exemplo relativo a investimentos (x 1 milhão de florins) das empresas, de acordo com a região e a actividade.

Quadro 4 – Investimentos das empresas. Fonte: Willenborg e Waal (1996)

Actividades	Região A	Região B	Região C	Total
Actividade 1	20	50	10	80
Actividade 2	8	19	22	49
Actividade 3	17	32	12	61
Total	**45**	**101**	**44**	**190**

Supondo que a maior parte das células da actividade 2 e da actividade 3 são células sensíveis. Como medida de protecção dos dados, as linhas correspondentes a estas actividades foram agregadas, conforme Quadro 5.

Quadro 5 – Investimentos após reformulação. Fonte: Willenborg e Waal (1996)

Actividades	Região A	Região B	Região C	Total
Actividade 1	20	50	10	80
Actividade 2 e 3	25	51	34	110
Total	**45**	**101**	**44**	**190**

Se a agregação das células mencionada no Quadro 5, referente às actividades 1 e 2, não for suficiente para tornar as variáveis não sensíveis, devem ser tomadas medidas adicionais de controlo da divulgação estatística. Por exemplo, é possível uma redução adicional do detalhe, sobretudo se o número de células sensíveis ainda for grande. Quando o número de células sensíveis é reduzido, as medidas de controlo da divulgação estatística que devem ser aplicadas são o arredondamento e a supressão local.

A reformulação da tabela é recomendada como um método simples que minimiza o número de células de risco e preserva a contagem original. Este método pode ser aplicado com os métodos de controlo da divulgação pós tabular ou pré tabular ou ser aplicado por conta própria (Hundepool et al, 2009).

4.3.2. Supressão de células

O método da supressão de células consiste na eliminação dos valores das células sensíveis e a colocação de ***X*** no seu lugar (Willenborg e Waal, 1996).

Pegando no exemplo anterior e considerando que a célula correspondente à actividade 2 e região C é uma célula sensível de acordo com a regra da dominância o valor da célula deve ser suprimido, conforme Quadro 6:

Quadro 6 – Investimentos após supressão primária. Fonte: Willenborg e Waal (1996)

Actividades	Região A	Região B	Região C	Total
Actividade 1	20	50	10	80
Actividade 2	8	19	X	49
Actividade 3	17	32	12	61
Total	**45**	**101**	**44**	**190**

A supressão das células sensíveis, normalmente não é suficiente porque facilmente se consegue obter o valor da mesma através dos totais marginais.

Uma solução poderia ser através da construção de uma tabela sem os totais marginais. No entanto, esta solução pode não ser aceitável por parte dos utilizadores, uma vez que há uma perda de informação. A partir do momento em que se verifica que os totais marginais não são sensíveis, eles podem ser publicados.

A outra solução é a supressão adicional dos valores internos das células não sensíveis, a chamada supressão secundária. Por exemplo, pode-se suprimir os valores correspondentes às actividades 2 e 3 da região A e o valor da actividade 3 da região C, conforme Quadro 7. Neste caso seria difícil obter os valores das células sensíveis.

Quadro 7 – Investimentos após supressão primária e secundária. Fonte: Willenborg e Waal, (1996)

Actividades	Região A	Região B	Região C	Total
Actividade 1	20	50	10	80
Actividade 2	X	19	X	49
Actividade 3	X	32	X	61
Total	**45**	**101**	**44**	**190**

Nesta tabela é fácil escolher as supressões secundárias, o mesmo não acontece quando as tabelas são maiores nestas, quando o objectivo é minimizar a perda de informação, a escolha das supressões secundárias é complexa.

Devem ser tidos em atenção alguns aspectos nomeadamente:

- *As células sensíveis devem ser bem protegidas pela escolha da supressão secundária; os intervalos dos valores das células suprimidas não devem ser demasiado pequenos.*

É necessário ter em atenção que o cálculo dos intervalos é possível quando o valor das células é de alguma forma restringido. Quando os intervalos são demasiado pequenos é fácil para o intruso fazer uma boa estimativa do valor da célula.

- *A perda de informação, devido à supressão secundária, deve ser minimizada*

A perda de informação depende das diferentes escolhas da supressão secundária. Essa perda será quantificada através da atribuição de um peso $\mathbf{w_{ij}}$ para cada célula $(\boldsymbol{i, j})$.

- *Nenhuma célula de valor zero ou célula vazia deve ser suprimida*

Uma célula vazia ou com valor zero não deve ser suprimida, uma vez que essa supressão pode levar à divulgação do valor de outra célula suprimida.

4.3.3. Intervalos viáveis

Uma tabela sendo protegida pelo método da supressão de células é sempre possível obter limites superiores e inferiores para o verdadeiro valor da célula suprimida da tabela. O intervalo dado por esses limites é chamado de intervalo viável (Hundepool et al, 2009).

Os intervalos viáveis variam de acordo com a supressão secundária adoptada (Willenborg e Waal, 1996).

Considerando o exemplo anterior, da supressão secundária, obtêm-se os seguintes intervalos viáveis da supressão das células:

Quadro 8 – Investimentos com intervalos viáveis da supressão de células.
Fonte: Willenborg, L. e Waal, T. (1996)

Actividades	Região A	Região B	Região C	Total
Actividade 1	20	50	10	80
Actividade 2	0-25	19	5-30	49
Actividade 3	0-25	32	4-29	61
Total	**45**	**101**	**44**	**190**

4.3.4. Arredondamento

O método do arredondamento é outro método que pode ser utilizado para tornar a tabela de dados segura. Este método consiste no arredondamento dos valores das células para um número inteiro através da multiplicação do valor por uma base fixa (Willenborg e Waal, 1996).

Como referido por Hundepool et al, (2009), o arredondamento envolve o ajustamento dos valores de todas as células de uma tabela para uma dada base, de forma a criar incerteza sobre os valores reais de qualquer célula ao adicionar um determinado valor aceitável para a distorção dos dados.

Existem variadas formas de arredondamento dos dados, tais como, o arredondamento convencional, arredondamento aleatório, pequeno ajustamento das células, arredondamento controlado e o arredondamento semi-controlado.

4.4.Dados tabulares baseados em amostras

Existem alguns inconvenientes com os dados, por um lado, na maior parte dos casos, os dados referem-se a amostras, em que apenas uma parte da população é entrevistada. Por outro lado, existem casos em que não se obtêm respostas, provocando também a exclusão de alguns membros da população alvo do inquérito. Estes aspectos implicam que na estimativa de parâmetros populacionais de cada observação, os dados tenham que ser ponderados com um factor adequadamente escolhido. Estes factores são baseados num sistema de amostragem na natureza, na extensão da não resposta e no procedimento para os estimar (Willenborg e Waal, 1996).

Capítulo 5. Qualidade da informação e risco de divulgação

Como referido em Karr et al (2005), a qualidade dos dados avalia-se pela capacidade dos mesmos serem utilizados de forma eficaz, económica e rápida para informar e avaliar no suporte às decisões. A qualidade dos dados é uma medida multidimensional, indo para além do nível do registo, incluindo factores como a acessibilidade, pertinência, actualidade, metainformação, documentação, capacidade e expectativa dos utilizadores.

Acrescenta-se a todos estes factores a utilidade da informação. Os mesmos autores definem a utilidade dos dados como sendo a capacidade de preservar as mesmas inferências a partir de microdados divulgados para microdados protegidos. Vários autores abordam a questão da utilidade da informação: para (Haworth et al, 2001, cf Kennickell e Lane, 2006, a utilidade dos dados é a totalidade dos recursos ou características de um produto ou de um serviço que afectam a sua capacidade de satisfazer as necessidades explícitas ou implícitas dos utilizadores. Para (Duncan et al, 2001), a utilidade dos dados é uma medida do valor da informação estatística que a fonte fornece a um utilizador

Ao abordar o problema da limitação da divulgação estatística, há que ponderar duas situações: satisfazer os utilizadores dos dados e tranquilizar os entrevistados.

5.1. Medidas de qualidade da informação

A utilidade dos dados é uma expressão positiva da perda de informação. Foram propostas várias medidas para determinar a utilidade dos dados. Por exemplo, Özsoyoğlu and Chung (1986, cf Duncan et al. 2001), sugeriram como medida de utilidade dos dados tabulares, em que se utiliza o método da supressão de células para limitação da divulgaçãoa percentagem de células suprimidas. Da mesma forma, Waal e Willenborg (1996) consideram várias opções para escolher as supressões locais (ou seja,

os valores de variáveis específicas nos registos específicos), incidindo sobre o número total das supressões, ou no número de categorias efectuadas pelas supressões locais.

Para medir a qualidade dos dados, Domingo-Ferrer e Torra (2001), adoptam outra abordagem baseada em estatísticas de informação de um conjunto de dados divulgado e um conjunto de dados original, como se pode ver mais à frente na secção 5.1.1.

A perda de informação depende da utilização dos dados. A utilização dos dados potenciais é muito diversa, tornando a sua identificação mais difícil no momento da divulgação. Quando a estrutura analítica do conjunto de dados mascarados é semelhante à estrutura do conjunto de dados original, pode-se dizer que há uma pequena perda de informação. De facto, é importante manter a estrutura do conjunto de dados para garantir que os dados mascarados sejam analiticamente válidos e interessantes.

De acordo com Winkler, (2005), um conjunto de dados é analiticamente válido se:

- As médias e as covariância num pequeno conjunto de subdomínios forem mais ou menos preservadas;
- Os valores marginais de pequenas tabulações de dados forem mais ou menos preservados;
- Pelo menos uma característica da distribuição for mais ou menos preservada.

Existem algumas formas complementares de avaliar a preservação da estrutura do conjunto de dados original, nomeadamente:

- Comparação entre os dados originais e os dados mascarados. Quanto mais similar for o método de controlo da divulgação estatística para a função identidade menor, é o impacto mas maior é o risco de divulgação; Este será o tipo de medidas a utilizar no âmbito deste trabalho.
- Comparação de algumas estatísticas do conjunto de dados original e do conjunto de dados mascarado. Uma pequena perda de informação significa pequenas diferenças entre estatísticas;
- Analisar o comportamento do método de controlo de divulgação estatística usado para medir o impacto sobre a estrutura do conjunto de dados original.

5.1.1. Medidas de qualidade para dados contínuos

Domingo-Ferrer e Torra, (2001), adoptam uma abordagem baseada em estatísticas de informação de um conjunto de dados divulgado e um conjunto de dados original. Considerando um conjunto de microdados com ***n*** indivíduos (registos), I_1, I_2, …, I_n e ***p*** variáveis contínuas $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$, …, $\boldsymbol{Z_p}$) em que:

X - Representa a matriz dos microdados originais, em que as linhas representam os registos e as colunas representam as variáveis;

X' - Representa a matriz dos microdados mascarados

Existem diversas ferramentas para caracterizar a informação contida no conjunto de dados, tais como:

- Matrizes de 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*** factores $\boldsymbol{PC_1}$, …, $\boldsymbol{PC_p}$ obtidos através da análise das componentes principais;
- A comunalidade entre cada uma das p variáveis ***p*** e a primeira componente principal $\boldsymbol{PC_1}$ (ou outras PC's). A comunalidade é a percentagem de cada variável que é explicada por $\boldsymbol{PC_1}$ (ou $\boldsymbol{PC_i}$), sendo ***C*** o vector das semelhanças para ***X*** e ***C'*** o vector correspondente a ***X'***.
- Matriz dos coeficientes do factor de pontuação ***F*** e ***F'***. A matriz ***F*** contem os factores que devem multiplicar cada variável em ***X*** para obter a sua projecção na componente principal. ***F'*** é a correspondente matriz para ***X'***.

Estas ferramentas não são simples medidas quantitativas que reflectem por completo as diferenças estruturais. Assim, são propostas outras formas para medir a informação perdida, através da discrepância entre as matrizes obtidas pelos dados originais, ***X***; ***V***; ***R***; ***RF***; ***C*** e ***F*** e as matrizes obtidas pelos dados mascarados ***X'***; ***V'***; ***R'***; ***RF'***; ***C'*** e ***F'***. A discrepância entre as correlações está relacionada com a informação perdida para os utilizadores de dados.

A matriz das discrepâncias pode ser medida de três formas:

- Erro quadrático médio – Soma do quadrado das diferenças das componentes entre os pares de matrizes, dividida pelo número de células em cada matriz;
- Erro absoluto médio – Soma absoluta das diferenças das componentes entre os pares das matrizes, dividida pelo número de células em cada matriz;
- Variação média - Soma absoluta da variação percentual das componentes da matriz calculada nos dados mascarados no que respeita às componentes da matriz calculada nos dados originais, dividida pelo número de células em cada matriz.

Quadro 9 – Medidas de utilidade de microdados contínuos. Fonte: Kennickel e Lane (2006

	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'	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}(f_{ij}-f'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\lvert f_{ij}-f'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\frac{\lvert f_{ij}-f'_{ij}\rvert}{\lvert f_{ij}\rvert}}{p^2}$

As componentes das matrizes são representadas pelas correspondentes letras minúsculas, por exemplo, x_{ij} é a componente da matriz ***X***. Relativamente às medidas ***X-X'***, devem ser calculadas a partir das médias das variáveis em vez de todos os dados, o que se chamaria de $\overline{X}-\overline{\overline{X}}'$. É também preferível utilizar a medida ***V-V'*** para comparar apenas as variâncias das variáveis, ou seja, para comparar a diagonal da matriz de covariâncias em vez da matriz inteira, o que se chamaria de ***S-S'***.

p- número de variáveis;

n – número de registos;

5.1.2. Medidas de qualidade para dados categóricos

Uma vez que as medidas de utilidade mencionadas no Quadro 9 não podem ser utilizadas para dados categóricos, foram sugeridas alternativas a essas medidas, nomeadamente:

1) *Comparação directa dos valores categóricos*

Na comparação directa dos valores categóricos, a comparação entre as matrizes ***X*** e ***X'*** para dados categóricos requer a definição da distância para as variáveis categóricas. A definição apenas considera a distância entre os pares de categorias que podem aparecer quando comparado um registo original com um registo mascarado.

Numa variável nominal ***V*** (uma variável categórica que assume valores sobre um conjunto desordenado) só é permitida a comparação da igualdade, o que nos leva à seguinte definição de distância:

$$d_y(c,c') \begin{cases} 0, c = c' \\ 1, c \neq c' \end{cases}$$

c - variável categórica do conjunto de dados original

c' - variável categórica do conjunto de dados mascarado

Para uma variável ordinal V_i, a distância entre a categoria ***a*** e ***b***, com ***a*** $\geq$***b***, ***a*** representa o registo original e ***b*** o registo mascarado, pode ser calculada através da seguinte expressão (Domingo-Ferrer e Torra, 2005):

$$D_{ORD}(a, b) = \frac{\left|\{i \setminus a \leq i < b\}\right|}{\left|D(Vi)\right|} \qquad (5.1)$$

Esta distância calcula o número de categorias que separa ***a*** de ***b*** a dividir pelo número de categorias do intervalo da variável (a divisão é usada para estandardizar a distância entre 0 e 1).

Os operadores da média usados nas variáveis ordinais são a média e a mediana convexa.

Se a função de frequências ***f*** das categorias é transformada numa função convexa ***f'***, a mediana sobre ***f'*** é chamada de mediana convexa.

$$f'(ci) = \min(\max_{c_j \le c_i}(f(c_j)), \max_{c_j \ge c_i}(f(c_j))) \tag{5.2}$$

2) ***Comparação de tabelas de contingência***

Uma alternativa à comparação directa dos valores das variáveis, é a comparação das tabelas de contingência. Dado o conjunto de dados original F e o conjunto de dados mascarados G e as correspondentes tabelas de contingência para $t \le K$, pode-se definir uma medida de perda de informação baseada na tabela de contingência (PIBTC) para um sub-conjunto W de variáveis, como:

$$\text{PIBTC}(F, G; W, K) = \sum \sum_{i_1 \ldots i_t} \left| x^F_{i_1 \ldots i_t} - x^G_{i_1 \ldots i_t} \right| \tag{5.3}$$

$$\{V_{j1} \ldots V_{jt}\} \subseteq W$$
$$\left|\{V_{j_1} \ldots V_{jt}\}\right| \le K$$

$x^{file}_{subscripts}$ - Entrada da tabela de contingência do ficheiro na posição dada por subscritos.

3) ***Medidas baseadas em entropia***

Em Willenborg e Waal (1999, cf Domingo-Ferrer e Torra, 2001) e Kooiman et al (1998, cf Domingo-Ferrer e Torra, 2001) o uso da entropia de Shannon para medir a perda de informação, pode ser utilizado na supressão local, na recodificação global, e no PRAM. A entropia é uma medida de informação teórica que pode ser usada no controlo da divulgação estatística se, o processo de mascaramento é modelado como o ruído que seria adicionado ao conjunto de dados original no caso de ter sido transmitido por um canal ruidoso.

Uma vez que o método PRAM é um método que generaliza outros como o ruído, a supressão e a recodificação, a entropia será limitada ao PRAM.

Considerando ***V*** uma variável do conjunto de dados original e ***V'*** a correspondente variável no conjunto de dados alterado pelo método PRAM e que $P_{V,\ V'} = \{p(V' = j \mid V = i)\}$é a matriz PRAM Markov. A incerteza condicional de ***V*** dado ***V'= j*** é dada pela seguinte expressão:

$$H(V\backslash V'j) = -\sum_{i=1}^{n} p(V = i \mid V' = j) \log p(V = i \mid V' = j) \qquad (5.4)$$

As probabilidades na equação (5.5) podem ser derivadas utilizando a fórmula de Bayes.

A medida de perda de informação baseada na entropia (PIBE) é obtida pela acumulação da equação (5.5) para todos os indivíduos r no conjunto de dados mascarados ***G***, assim PIBE é dada pela seguinte expressão:

$$PIBE\ (P_{V,\ V'}, G) = \sum_{r \in G} H(V \mid V' = j_r) \qquad (5.5)$$

j_r – é o valor do registo r

5.2.O risco de divulgação

A identificação ocorre quando um registo no ficheiro divulgado e um registo no arquivo externo pertencem ao mesmo indivíduo na população. A hipótese subjacente é que o intruso irá sempre tentar igualar um registo da amostra s a ser divulgada e um registo no arquivo externo através das variáveis de identificação. É provável que o intruso pretenda identificar as unidades da amostra que são únicas sobre as variáveis de identificação. A identificação ocorre quando, com base numa comparação de resultados sobre as variáveis de identificação, um registo ***i**** no arquivo externo é seleccionado correctamente como correspondente a um registo ***i*** da amostra, assim a informação confidencial sobre o indivíduo é divulgada usando os identificadores directos (Hundepool et al, 2009).

Os microdados apresentam muitas vantagens sobre os dados agregados, mas também colocam questões de divulgação mais graves devido à quantidade de variáveis divulgada. Nos microdados, a divulgação ocorre quando um indivíduo é identificado por um intruso que usa a informação de um ficheiro de dados (ou quando consegue obter informação confidencial). Para serem divulgados, os ficheiros de microdados não podem conter variáveis identificadoras, tais como o nome, a direcção, o número de identificação. No entanto, existem outras variáveis nos microdados que podem ser usadas como variáveis de identificação indirectamente.

Dado que a fonte responsável pelos dados é capaz de manter um risco de divulgação suficientemente baixo, então deve-se procurar maximizar a utilidade dos dados. À medida que aumenta a perda de informação devido à limitação da divulgação, a utilidade dos dados torna-se mais baixa. Simultaneamente, o risco de divulgação também diminui. Este quadro conceptual pode ser usado para comparar métodos alternativos de limitação da divulgação.

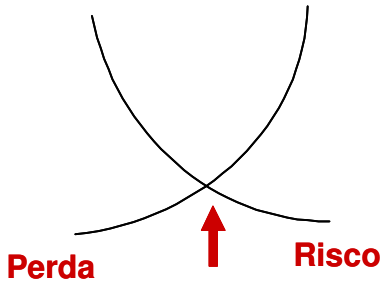

Figura 4 – Evolução comparativa do risco de divulgação e da perda de informação

O risco de divulgação de uma amostra sozinha ou de ambas, amostra e população, é uma função da variável de identificação/chave. Para avaliar o risco de divulgação é necessário considerar algumas hipóteses realistas sobre o que o intruso pode saber sobre os inquiridos e, qual a informação disponível que, para ele, pode levar a uma identificação e divulgação. Com base num cenário de divulgação, podem-se determinar as variáveis de identificação. As outras variáveis do ficheiro são variáveis confidenciais ou sensíveis, e representam dados que não devem ser divulgadas.

Para microdados obtidos através dos censos, o risco de divulgação é saber que se têm todas as variáveis de identificação disponíveis da população. No caso dos microdados

de amostras da população os riscos de divulgação são desconhecidos ou são parcialmente conhecidos através de uma distribuição marginal. Baseados na informação disponível na amostra, são utilizados modelos probabilísticos ou heurísticas para estimar as medidas de risco de divulgação da população.

Quando as variáveis de identificação são categóricas, como é normal em inquéritos sociais, o risco é convertido em termos de células da tabela de contingência, construída pela tabulação cruzada das variáveis de identificação: a chave. Consequentemente todos os registos na mesma célula têm o mesmo valor de risco.

5.2.1. Medidas de risco

O risco de identificação pode ser calculado de diversas formas. De seguida são apresentadas algumas medidas de risco.

5.2.1.1. Medidas de risco baseadas em chaves da amostra

Uma unidade está em risco se a combinação de resultados sobre as variáveis de identificação está abaixo de um determinado limite Hundepool et al., (2009).

5.2.1.2. Medidas de risco baseadas em chaves da população efectuadas por modelos estatísticos ou heurísticas para estimar as quantidades

A preocupação neste caso tem a ver com o risco individual determinado pela combinação de resultados das variáveis de identificação da população e a probabilidade de identificação. Assim, o indivíduo está em risco se o valor do seu risco estiver acima de um determinado limite.

Uma vez que a frequência da população é normalmente desconhecida, ela pode ser estimada através de um processo de modelização. Por exemplo, o risco baseado numa distribuição binomial negativa ou numa distribuição de Poisson.

1) ***Modelo de Poisson***

Considere F_j independente que segue uma distribuição de Poisson, com média λ_k e uma amostragem de Bernouli. $F_k - f_k$ são independentes e seguem uma distribuição de Poisson, como se segue:

$f_k \mid \lambda_k \sim \text{Poisson}(\pi\lambda_k)$

$F_k - f_k \mid \lambda_k \sim \text{Poisson}((1-\pi)\lambda_k)$

O risco individual da amostra é dado pela seguinte expressão:

$$r_k = E_{\lambda_k}\left(\frac{1}{F} \mid f_k = 1\right) = \frac{1}{\lambda_k(1-\pi)}\left[1 - e^{-\lambda k(1-\pi)}\right] \quad (5.6)$$

Nesta abordagem, o parâmetro λ_k é estimado pelo modelo logaritmo linear tendo em conta a estrutura e a dependência dos dados. Considerando que a frequência da amostra f_k é uma distribuição de Poisson independente com a média $u_k = \pi\lambda_k$, o modelo do logaritmo linear para u_k é dado pela expressão $\log(u_k) = x'_k\,\beta$, em que x_k representa o vector dos principais efeitos e interacções do modelo para as variáveis chave. Através da utilização de procedimentos estandardizados obtêm-se as estimativas de máxima probabilidade de Poisson para o vetor β e calculam-se os valores ajustados $\hat{u}_k = \exp(x'_k\,\hat{\beta})$. A estimativa $\hat{\lambda}_k$ é dada pela expressão: $\hat{\lambda}_k = \frac{\hat{u}_k}{\pi}$

A medida global do risco de divulgação é dado por:

$$\hat{\tau}_2 = \sum_{k \in SU} \hat{r}_k = \sum_{k \in SU} \frac{1}{\hat{\lambda}_k(1-\pi)}\left[1 - e^{-\hat{\lambda}_k(1-\pi)}\right] \quad (5.7)$$

SU – Conjunto de todas as amostras únicas

*2) **Modelo Binomial Negativo***

Outro método que pode ser utilizado para a avaliação do risco é baseado na distribuição Binomial Negativa, em que $f_k \sim NB\left(\alpha_k, p_k = \frac{1}{1+N\pi_k\beta_k}\right)$ e $F_k \mid f_k \sim NB\left(\alpha_k + f_k, \rho_{k} = \frac{1+N\pi_k\beta_k}{1+N\beta_k}\right)$, em que π_k é uma fracção da amostra.

As medidas de risco de divulgação são estimadas com base num modelo de Distribuição Binomial Negativo.

Assim, o risco global é dado pela expressão (Rinnot e Sholmo, (2005, 2006), cf Hundepool et al, 2009):

$$\hat{\tau}_2 = \sum_{k \in SU} \hat{r}_k = \sum_{k \in SU} \frac{\hat{\rho}_k \left(1 - \hat{\rho}\hat{k}\right)^{\alpha_k}}{\hat{\alpha}k\left(1 - \hat{\rho}_k\right)} \qquad (5.8)$$

5.2.1.3. Modelos baseados na teoria "record linkage"

Quando uma variável de identificação é contínua não se pode explorar o conceito de raridade da chave, pode-se transforma-lo num conceito de raridade numa vizinhança de registos. Uma forma de medir a raridade na vizinhança é através de técnicas de linkage (ligação).

A técnica record linkage (ligação de registos) consiste em ligar cada registo ***a*** no ficheiro protegido ***A*** para cada registo ***b*** no ficheiro de dados original ***B***. O par (***a, b***) é uma correspondência se ***b*** se tornar o registo original correspondente a ***a***.

Para a aplicação deste método, assume-se que o intruso tem um conjunto de dados externo que partilha algumas variáveis com o conjunto de dados protegido e divulgado e adicionalmente contém algumas variáveis identificadoras, por exemplo, número de passaporte; nome completo, etc. Através das variáveis partilhadas, o intruso tenta ligar o conjunto de dados protegidos com o conjunto de dados externo. O número de correspondências fornece uma estimativa do número de registos protegidos cujo

entrevistado pode ser identificado pelo intruso. Assim, o risco de divulgação é definido como a percentagem de correspondências entre o número total de registos em ***A***.

1) ***Record linkage baseada na distância***

O método "record linkage" baseado na distância foi proposto inicialmente por Pagliuca e Seri (1999, cf Hundepool, 2009) para avaliar o risco de divulgação após a microagregação. O método consiste em ligar cada registo **a** do ficheiro ***A*** com o registo ***b*** do ficheiro ***B*** mais próximo.

A aplicação deste método obriga à definição de uma função de distância que expressa a proximidade entre os registos. A distância dos registos pode ser definida a partir da função da distância das variáveis, no entanto exige uma estandardização das variáveis para evitar problemas de escala e atribuir a cada variável um peso na distância do registo. Pagliuca e Seri (1999, cf Hundepool, 2009) utilizam a Distância Euclidiana e pesos iguais para todas as variáveis. Domingo-Ferrer e Torra, 2001 utilizam outros "record linkage" baseado na distância, como o "record linkage" probabilístico.

2) ***Record linkage probabilístico***

Este método tal como os anteriores tem como objectivo ligar pares de registos de conjuntos de dados. Para cada par de registos é determinado um índice **R**, onde são utilizados dois limites para classificar os pares: LT e NLT. Se o índice estiver acima de LT, o par é ligado, se estiver abaixo de NLT, o par não é ligado. Quando o índice está no intervalo entre LT e NLT diz-se que o par é um "par de escritório". Um "par de escritório, é uma par que não pode ser classificado à partida classificado como ligado ou não ligado, é necessária uma verificação para o classificar (Domingo-Ferrer e Torra 2003).

O índice R (a, b) é calculado pela seguinte expressão:

$$\text{R (a, b)} = \log\left(\frac{P(a=b \mid (a,b)\in M}{P(a=b) \mid (a,b)\in U}\right) \qquad (5.9)$$

M – Conjunto de pares correspondentes

U – Conjunto de pares não correspondentes

É um método mais complicado do que o método baseado na distância, mas tem a vantagem de não ser necessário ponderar as variáveis (Hundepool, 2009).

Quando as variáveis são independentes, o índice pode ser calculado a partir de probabilidades condicionais para cada uma das variáveis, conforme se segue:

a) P (1|M) - Probabilidade dos valores das variáveis de dois registos ***a*** e ***b*** coincidirem, dado que esses registos são uma correspondência real;
b) P (O|U) - Probabilidade dos valores das variáveis de dois registos ***a*** e ***b*** não coincidirem, dado que não existe uma correspondência real.

Os limites LT e NLT são calculados a partir de:

a) P (LP|U) – Probabilidade de ligar um par que não é correspondente (ligação falsa);b)
b) P (NP|M) – Probabilidade de não ligar um par correspondente (não ligação falsa).

Em microdados, as medidas do risco de divulgação quantificam o risco de identificação. As medidas de risco de divulgação individual são úteis para identificar os registos de alto risco e orientar os métodos de CDE. Estas medidas de risco individual podem ser agregados para obter o arquivo global dos riscos de divulgação. As medidas de risco global são particularmente úteis para os serviços de estatística, para o seu processo de decisão nos microdados (se são ou não seguros para serem divulgados) e permitem comparações entre diferentes arquivos.

5.2.1.4. O risco individual no Argus

O Argus calcula o risco individual $r_i^{ind} = r_{k(i)}^{ind}$ que representa o risco individual ***i*** da combinação das variáveis chave k(i)=k, baseado na seguinte expressão (Hundepool et al, 2008):

$$r_{k(i)}^{ind} = r_k^{ind} = \left(\frac{\hat{p}_k}{1-\hat{p}_k}\right)^{f_k} \left\{ A_0 \left(1 + \sum_{j=0}^{f_k-3} (-1)^{j+1} \prod_{l=0}^{j} B_l \right) + (-1)^{f_k} \log(\hat{p}_k) \right\} \qquad (5.10)$$

Em que,

$$\hat{p}_k = \frac{f_k}{\hat{F}_k} = \frac{f_k}{\sum_{i:k(i)=k} w_i} \tag{5.11}$$

w_i – peso individual

f_k - Frequência das combinações da variável chave na amostra.

$\hat{F}_k$ - Estimativa da frequência das combinações da variável chave na população.

$$B_l = \frac{(f_k - 1 - l)^2}{(l+1)(f_k - 2 - l)} \frac{\hat{p}_k^{l+2-f_k} - 1}{\hat{p}_k^{l+1-f_k} - 1} \tag{5.12}$$

$$A_0 = \frac{\hat{p}_k^{1-f_k} - 1}{(f_k - 1)} \tag{5.13}$$

Uma vez que o risco individual indicado na expressão (3.49) é de difícil execução, foi introduzida uma aproximação a esta expressão para frequências superiores a 40, conforme se segue:

$$r_k = \frac{\hat{p}_k}{fk - (1 - \hat{p}_k)} \tag{5.14}$$

Dado que existem outros factores que influenciam o risco, nomeadamente a qualidade das variáveis, é usado um factor de multiplicação. Assim a fórmula do risco é dada pela seguinte expressão:

$$\rho i = \pi * r_{k(i)}^{ind} \tag{5.15}$$

Capítulo 6. Estudo de caso

6.1.Metodologia de Investigação

O objectivo deste estudo é a aplicação e comparação de alguns métodos de controlo da divulgação estatística, nomeadamente a microagregação, codificação superior, rank swapping, arredondamento (disponíveis no software Argus) e a aplicação da microagregação híbrida. Neste capítulo começa-se por apresentar o software Argus, utilizado neste trabalho e prossegue-se com as etapas que se utilizam como guião para a divulgação dos dados. Em, seguida, para cada uma das base de dados em estudo, depois da respectiva descrição, faz-se a aplicação dos métodos de Controlo de Divulgação Estatística.

6.2.Software "Argus"

O software Argus surgiu no projecto CASC – Computational Aspects of Statistical Confidentiality[16], onde se exploram novas formas de controlo da divulgação estatística e onde se alargaram os métodos e ferramentas já existentes. O objectivo é dar maior ênfase a ferramentas práticas e a trabalhos de investigação para as desenvolver. Este software pode ser utilizado em microdados e macrodados. No primeiro caso é utilizado o μ-Argus; quando se trabalha com dados tabulares utiliza-se o t-Argus.

A crescente procura de dados estatísticos por parte de investigadores e o mais detalhados possível, leva a uma grande preocupação: a da violação da privacidade dos entrevistados. Os entrevistados devem ser protegidos sem que essa protecção leve a uma grande perda de informação. Algumas questões se colocam:

[16] O projecto CASC (http://neon.vb.cbs.nl/casc/index.htm) foi o embrião de um conjunto de projectos onde, a nível do Eurostat, se promoveu o desenvolvimento e partilha de conhecimento sobre métodos de Controlo de Divulgação Estatística.

1) Como alterar um conjunto de microdados para que sua divulgação tenha um risco aceitável e ao mesmo tempo o mínimo de informação perdida?
2) Como se pode definir exactamente o risco de divulgação?
3) Como se pode quantificar a informação perdida?

Por esta razão desenvolveu-se um software que responde a todas estas questões. Uma das respostas é o Argus.

Para além do software Argus, existem outros que podem ser utilizados para a protecção de dados, nomeadamente SUDA, R e SAS.

O ponto de partida do µ-Argus é a aplicação de limites para a identificação de registos inseguros e os procedimentos de recodificação global e da supressão local.

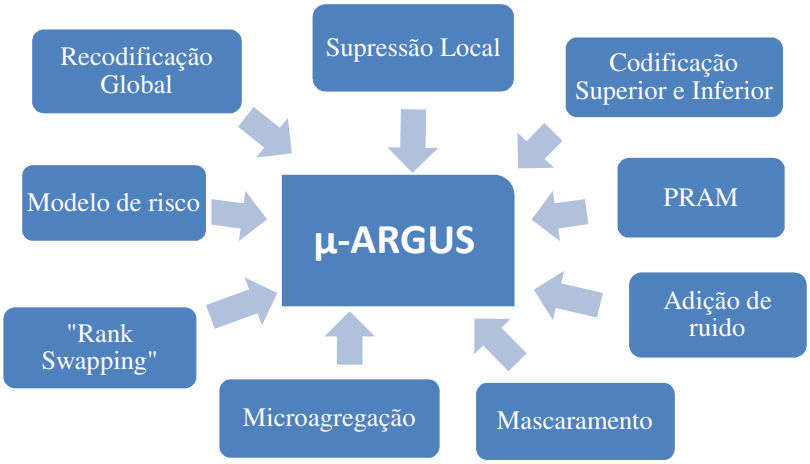

Figura 5 – Software µ-Argus. Fonte: Hundepool et al (2009)

O trabalho no µ-Argus decorre de acordo com as seguintes etapas:

1) *Metadados*

   É necessário definir a estrutura dos dados, não apenas os aspectos gerais, mas também informações adicionais do controlo da divulgação estatística;

2) *Regra limite/Modelo de risco*

3) Seleccionar e calcular uma tabela de frequências onde os métodos de controlo da divulgação estatística (como os modelos de risco, regra do limite) se podem basear.
4) *Recodificação Global*

   Seleccionar as variáveis para recodificar e verificar os resultados.

5) *Seleccionar e aplicar outro método de protecção de dados*

   Métodos como a microagregação; o PRAM; o arredondamento; a codificação superior e inferior; "rank swapping"e a adição de ruído.

6) *Modelo de risco*

   Seleccionar o nível de risco.

7) *Gerar um micro ficheiro de dados seguro*

   Todas as transformações de dados durante este processo são especificadas. Nesta fase todas as combinações inseguras são protegidas através da supressão local. É gerado um relatório.

A figura seguinte faz uma descrição do funcionamento do software Argus em dados tabulares e microdados para gerar ficheiros de dados seguros.

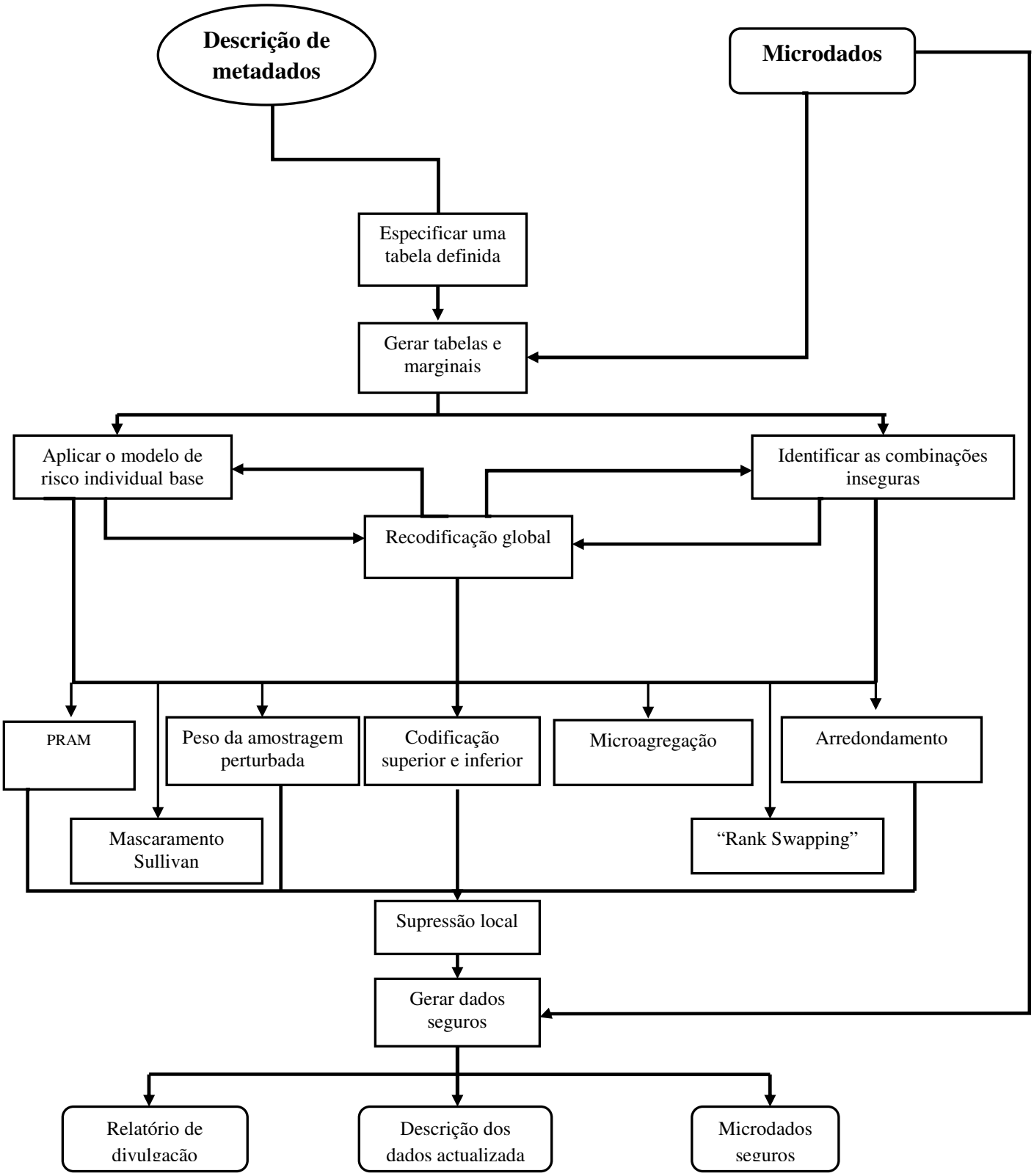

Figura 6 – Funcionamento do µ-Argus. Fonte: Hundepol et al (2008)

6.3.Estudo da base de dados SABI

Foram utilizadas duas bases de dados neste trabalho, representativas dos tipos mais comuns de dados existentes nos institutos de estatística: informação sobre empresas e informação sobre famílias. Os dados utilizados para a elaboração deste capítulo foram recolhidos através da base de dados SABI, correspondentes ao primeiro caso (dados sobre empresas) e dados gerados aleatoriamente relativos ao inquérito às famílias realizados pelo INE (dados sobre famílias). Como se referiu anteriormente, serão aplicadas técnicas de controlo de divulgação para microdados.

A base de dados SABI - Sistema de Análise de balanços Ibéricos, é a mais completa base de dados de análise financeira sobre empresas portuguesas e espanholas com um histórico de contas anuais até 10 anos. É a única base de dados ibérica com informação económica - financeira sobre mais de 1 milhão de empresas de Portugal e Espanha (Coface Serviços Portugal, SA).

A informação contida na base SABI é actualizada periodicamente. Esta informação é obtida junto de algumas fontes oficiais, em Portugal o Registo Comercial, o Diário da República, em Espanha o Borme e a imprensa, entre outros. Com esta base de dados pode-se obter informação geral e informação sobre as contas anuais de mais de 50.000 empresas portuguesas e 530.000 espanholas.

SABI é resultado da colaboração entre três empresas:

- Coface Serviços Portugal, SA, responsável pela base de dados de empresas portuguesas;
- Informa, responsável pela base de dados de empresas espanhola;
- Bureau Van Dijk, responsável pelo software de pesquisa, tratamento e análise de dados.

A base de dados SABI fornece diversa informação, como a morada; localidade/Concelho; nº contribuinte; descrição da actividade, rácios financeiros; número de empregados; etc. Contém também variadas possibilidades de pesquisa, nomeadamente a localização por comunidades autónomas, por províncias, por

localidade, a actividade por códigos CAE, etc., os dados financeiros por qualquer rubrica das contas anuais, dos rácios Informa, dos rácios Coface, dos rácios europeus, etc., a estrutura do capital por accionistas, participações ou nacionalidade de ambos; a consolidação, etc.

6.3.1. Etapas para a divulgação dos dados

Como referido anteriormente no capítulo 3, o processo para a divulgação de ficheiros de microdados, ocorre em 5 etapas, nas quais são descritos os passos que os dados devem seguir até serem divulgados.

Quadro 10 – Guia para a divulgação do ficheiro da base de dados SABI

Etapas para o processo de divulgação	**Análises a efectuar/Problema resolvido** ↓ **Resultados esperados**
1. Porque é que a protecção da confidencialidade é necessária?	**Os dados referem-se a pessoas singulares ou colectivas?** ↓ Os dados referem-se a pessoas colectivas, mais precisamente empresas da indústria extractiva em Portugal, pelo que se torna necessária a sua protecção.
2. Quais são as principais características e utilização dos dados?	**Análise dos dados/Estrutura dos dados** **Os dados apresentam uma estrutura específica?** ↓ Os dados referem-se a variáveis financeiras e económicas e informações gerais das empresas da indústria extractiva. O capítulo 6.3.3 faz uma análise preliminar dos dados
	Análise das metodologias de pesquisa ↓ Os dados referem-se às 5% maiores empresas da indústria extractiva em Portugal.
	Análise dos objectivos dos Institutos de Estatística **Que tipo de divulgação?** ↓ Na realidade não vai haver divulgação dos dados. Os dados são para efectuar um estudo da aplicação de alguns métodos CDE e a comparação dos mesmos.
	Análise das necessidades dos utilizadores ↓ (Não se aplica)
	Análise de questionários **Listagem das variáveis a ser removidas, variáveis a ser incluídas** ↓ As variáveis incluídas na base de dados estão identificadas no Quadro 11. As variáveis utilizadas neste estudo são as mencionadas no Quadro 13.

Etapas para o processo de divulgação (Continuação)	Análises a efectuar/Problema resolvido ↓ Resultados esperados
3. Riscos de divulgação	**Cenário de divulgação - Lista de variáveis identificadoras** ↓ Nome da empresa **Lista das variáveis identificadoras indirectas**[17] ↓ Região, Antiguidade e Empregados 07
	Definição do risco ↓ O Eurostat recomenda um risco individual máximo de 4%. No entanto este nível de rico justifica-se apenas em grandes bases de dados (com milhões de registos). Neste caso assume-se um risco individual máximo de 20% dada a pequena dimensão do ficheiro de dados.
	Avaliação do risco ↓ A avaliação do risco é feita no capítulo 6.3.4 e no qual se pode verificar que os dados são inseguros para divulgação dado o risco definido no ponto anterior.
4. Métodos de controlo da divulgação	**Análise do tipo de dados envolvidos, políticas dos Serviços de Estatística e necessidades dos utilizadores.** **Identificação dos métodos de limitação da divulgação** ↓ Recodificação global, microagregação, codificação superior, rank swapping, arredondamento e microagregação híbrida (capítulo 6.3.6 e 6.3.7).
	Análise da perda de informação ↓ Análise da perda de informação é realizada no capítulo 6.3.9 através do SSE; EAM e VA.
5. Implementação	Escolha do software, parâmetros e limites dos diferentes métodos ↓ µ-Argus e SPSS

6.3.2. Amostra (ver a necessidade de explicar o porquê de retirar algumas empresas)

O investigador exercendo a sua actividade na área da indústria extractiva, achou por bem fazer um estudo de alguns dados financeiros e económicas das maiores empresas da indústria extractiva do nosso país retiradas da base de dados SABI. Para além disso, a indústria extractiva é uma área que não tendo muitas empresas tem maior risco de divulgação de dados, por isso tornou-se importante realizar o estudo nesta área.

[17] Variáveis identificadoras indirectas - variáveis que possibilitam deduzir as unidades estatísticas a partir de informação que não conste das variáveis identificadoras directas.

Inicialmente foram retiradas da base SABI 215 empresas, das quais apenas 147 foram alvo deste estudo. Algumas empresas foram eliminadas da base de dados inicial, uma vez quer não possuíam dados actualizados ou indicavam o seu ramo de actividade.

O que se pretende com este estudo é verificar o impacto da aplicação de alguns métodos de Controlo da Divulgação Estatística, nomeadamente a microagregação, codificação superior, arredondamento, rank swapping e a microagregação híbrida, na qualidade dos dados a serem divulgados. Vão ser analisadas algumas variáveis do foro financeiro e económico e algumas variáveis identificadoras indirectas, como a região, antiguidade e o número de empregados, conforme se pode verificar no Quadro 11.

Mais à frente também é realizada uma análise a um ficheiro aleatório de dados familiares semelhante aos do Instituto de Estatística e provenientes de inquéritos familiares. A análise contempla a aplicação e comparação de métodos de controlo da divulgação estatística para microdados.

Quadro 11 – Variáveis financeiras, económicas e outras das empresas da industria extractiva

Região	Proveitos Operacionais 07 (PO07)	Capital Próprio 07 (CP07)	Depósitos bancários e caixa 07 (DBCX07)	Fundo de Maneio 07 (FM07)	Custos com Pessoal 07 (CP 07_A)
Antiguidade (ANTIG)	Proveitos Operacionais 06 (PO06)	Capital Próprio 06 (CP06)	Depósitos bancários e caixa 06 (DBCX06)	Fundo de Maneio 06 (FM06)	Custos com Pessoal 06 (CP 06_A)
Nº de empregados 07 (EMP07)	Resultados Correntes 07 (RC07)	Imobilizações Corpóreas 07 (IC07)	Total Activo 07 (ACT07)	Custo Mercadorias Vendidas Matérias Consumidas 07 (CMVMC07)	Outros Custos Operacionais 07 (OCO07)
Nº de empregados 06 (EMP06)	Resultados Correntes 06 (RC06)	Imobilizações Corpóreas 06 (IC06)	Total Activo 06 (ACT06)	Custo Mercadorias Vendidas Matérias Consumidas 06 (CMVMC06)	Outros Custos Operacionais 06 (OCO06)
Resultados Operacionais 07 (RO07)	Proveitos Ganhos Financeiros 07 (PGF 07)	Custos e Perdas Financeiras 07 (CPF07)	Imposto s/ Rendimento do Exercício 07 (ISREND07)	Resultado Liquido do Exercício 07 (RLE07)	Margem Bruta 07 (MB07)
Resultados Operacionais 06 (RO06)	Proveitos Ganhos Financeiros 06 (PGF 06)	Custos e Perdas Financeiras 06 (CPF06)	Imposto s/ Rendimento do Exercício 06 (ISREND06)	Resultado Liquido do Exercício 06 (RLE06)	Margem Bruta 06 (MB06)
Amortizações do Exercício 07 (AMTEX07)	Amortizações do Exercício 06 (AMTEX06)	Valor Acrescentado Bruto 06 (VAB06)	Valor Acrescentado Bruto 07 (VAB07)	Volume de Negócios 07 (VN07)	Volume de Negócios 06 (VN06)
Juros Suportados 07 (JS07)	Juros Suportados 06 (JS06)	Nome das empresas			

6.3.3. Análise preliminar dos dados

Com o intuito de se realizar uma análise preliminar da base de dados SABI procedeu-se inicialmente à identificação da matriz de dados e à análise univariada dos dados originais.

6.3.3.1. Matriz (Quadro) de dados

Os dados são representados por uma matriz X, em que n representa os indivíduos (empresas) em linha: w_i, i=1, 2, …., n e p, as variáveis (atributos) em coluna: y_j, j=1, 2, …, p (dados financeiros).

Quadro 12 – Matriz X

X	Y_1	Y_2	…	Y_p
W_1	X_{11}	X_{12}	…	X_{1p}
W_2	X_{21}	X_{22}	…	X_{2p}
…..	…	…	…	…
W_n	X_{n1}	X_{n2}	…	X_{np}

A matriz de dados das variáveis e indivíduos em estudo é apresentada no Quadro 13.

Quadro 13 – Matriz de dados

Nome das empresas	VN 07 Y_1 (€)	CMVMC 07 Y_2 (€)	MB 07 Y_3 (€)	VAB 07 Y_4 (€)
A. BENTO VERMELHO, LDA.	3.111.517	676.406	2.420.855	1.496.595
……	…	…	…	…
SOCIEDADE DAS PEDREIRAS DO MARCO, LDA.	4.732.592	1.189.035	3.536.246	2.404.655
…….	…	…	…	…
VIMIBRITA - SOCIEDADE DE EXPLORAÇAO DE GRANITOS, S.A	1.678.095	275.224	1.577.480	964.078

6.3.3.2. Análise univariada das variáveis

De seguida são apresentadas as medidas de localização, nomeadamente a média, mediana, moda e percentis, os outliers e as medidas de dispersão, variância e desvio padrão das variáveis contínuas. Relativamente às variáveis categóricas é apresentada a tabela de frequências bem como a identificação das classes.

1) Variáveis Contínuas

a) Medidas de Localização

O Quadro 14 apresenta os resultados das medidas de localização para as variáveis contínuas CMVMC 07; MB 07; VAB 07 e VN 07, no qual se podem verificar os valores da média, mediana, moda e percentis para as 147 empresas, bem como os valores válidos e os missing's.

Quadro 14 – Medidas de localização

		CMVMC07	MB07	VAB07	VN07
	Valores válidos	142	147	147	147
	Missing's	5	0	0	0
Média		1.128.227,69	3.553.014,81	1.745.529,01	4.446.503,59
Mediana		418.275,50	1.811.377,00	987.953,00	2.338.340,00
Moda		15[a]	668091[a]	80672[a]	977070[a]
Percentis	25	181.492,50	1.314.073,00	700.607,00	1.654.261,00
	50	418.275,50	1.811.377,00	987.953,00	2.338.340,00
	75	1.413.988,50	4.260.367,00	2.043.121,00	5.176.496,00

b) Boxplot (Caixa de bigodes)

A caixa de bigodes é uma representação gráfica, em que a caixa ou rectângulo situa os quartis de distribuição. Nos extremos de cada bigode estão posicionadas as observações mínima e máxima. Todas as observações que se situam fora dos bigodes são outliers. Os outliers são observações aberrantes que podem representar erros na introdução dos dados, neste caso devem ser eliminados, ausência de valores de dados ou podem fazer parte de um fenómeno em estudo e aqui devem ser mantidos, assinalando-se a sua existência. Dependendo do seu afastamento relativamente às outras observações, os outliers podem ser severos ou moderados.

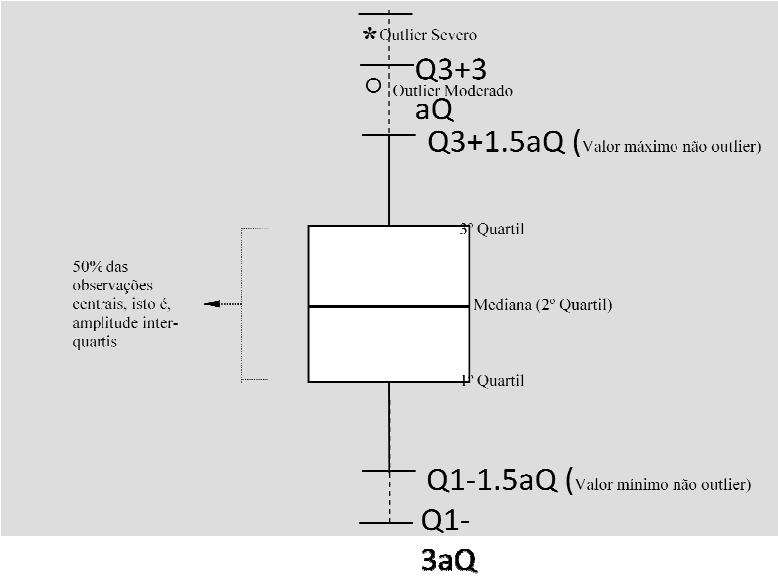

Figura 7 – Caixa de bigodes (Boxplot)

Recorreu-se à representação gráfica das quatro variáveis contínuas CMVMC 07; MB 07; VAB 07 e VN 07 para verificar a existência de outliers, que existindo, significa que há empresas com valores muito acima ou muito abaixo da maioria das outras empresas. Ou seja, são empresas com elevado risco de identificação por parte dos intrusos.

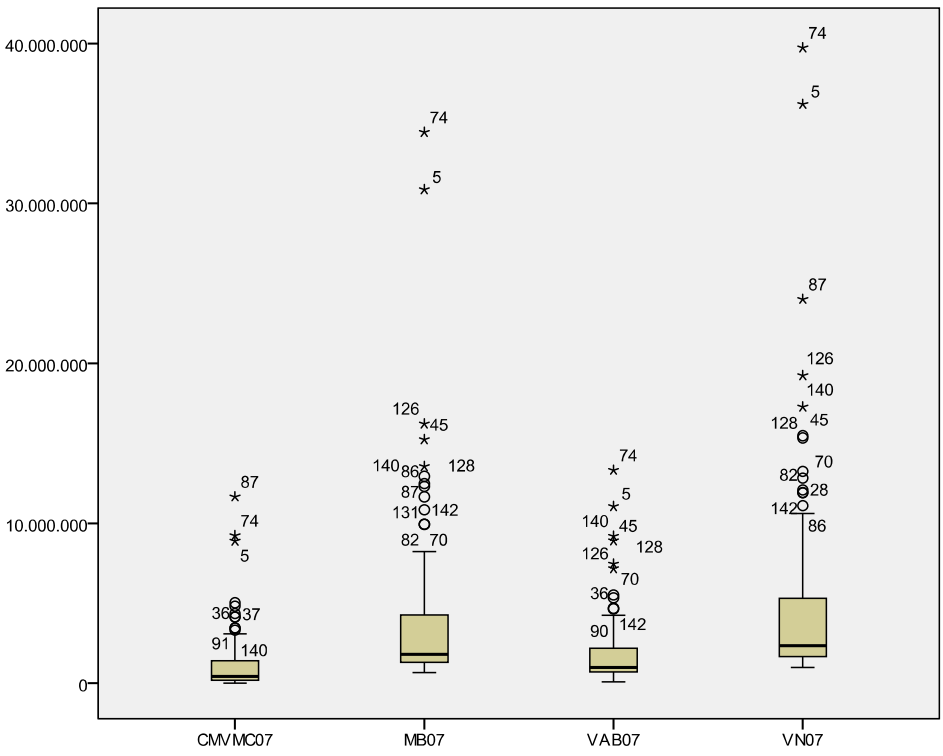

Figura 8 – Empresas outliers

Como se pode verificar na Figura 8, existem empresas que têm um elevado risco de divulgação nas variáveis em estudo, como se pode verificar no Quadro 15.

Quadro 15 – Empresas outliers

Nome das empresas		CMVMC 07	MB 07	VAB 07	VN 07
Agrepor Agregados, SA	5	8 888 120 €	30 861 580 €	11 056 038 €	36 202 695 €
Calbrita, SA	28	4 622 367 €			11 100 093 €
Cunha Duarte, SA	36	3 446 375 €		4 687 104 €	
Domingues & Contente, SA	37	3 314 275 €			
Ferbritas, SA	45		15 232 395 €	8 913 440 €	15 475 340 €
Granbeira, SA	51	5 025 984 €			
Iberobrita, SA	70	4 151 216 €	9 934 251 €	5 518 617 €	12 828 179 €
Irmãos Cavaco, SA	74	9 228 613 €	34 446 659 €	13 320 246 €	39 736 780 €
José Aldeia Lagoa & Filhos, SA	82		9 937 619 €		12 086 678 €
Lafarge Agregados – Unipessoal, Lda	86		12 304 937 €		11 925 128 €
Lena Agregados, SA	87	11 659.375 €	12 934 880 €		24 007 407 €
Lusolevantina Portugal, Lda.	89	4 357 151 €			
M. dos Santos & Ca. SA	90			5 342 544 €	
Madeira inerte, Lda	91	3 456 536 €			
R & G - Rogranit Gralpe, Lda.	120	4 821 398 €			
Secil Britas, SA	126		16 216 989 €	7 163 647 €	19 241 059 €
Sibelco Portuguesa, SA	128		12 481 301 €	7 462 460 €	15 337 305 €
Sifucel, SA	131		11 655 067 €		13 247 209 €
Solubema – Soc. Luso-Belga, SA	140	4 138 529 €	13 553 164 €	9 190 435 €	17 279 559 €
Sorgila, SA	142		10 845 127 €	4 652 493 €	11 905 212 €

c) *Medidas de dispersão*

No Quadro 16 podem ser verificados os valores para o desvio padrão e variâncias das variáveis em estudo, como se segue:

Quadro 16 – Medidas de dispersão

	CMVMC07	MB07	VAB07	VN07
Valores válidos	142	147	147	147
Missing	5	0	0	0
Desvio padrão	1.719.809,508	4.598.413,176	1.975.837,225	5.551.866,211
Variância	2,958E12	2,115E13	3,904E12	3,082E13

2) Variáveis categóricas

Antes de qualquer análise às variáveis identificadoras indirectas, também referidas neste documento como variáveis categóricas como a região, antiguidade e número de empregados 07, foram criadas classes para as mesmas, para isso recorreu-se ao software SPSS Statistical. Assim, o quadro abaixo indica a classes e intervalos de valores para as variáveis categóricas.

Quadro 17 – Tabela de classes e frequências das variáveis região, antiguidade e empregados 07

Classes	Região	Frequência	Antiguidade (anos)	Frequência	Empregados 07 (nºs)	Frequência
1	Norte	45	Até 16	25	Até 16	26
2	Centro	56	17 – 20	25	17 – 24	23
3	Lisboa	18	21 – 23	24	25 – 35	28
4	Alentejo	15	24 – 28	25	36 - 45	20
5	Algarve	4	29 – 36	24	46 - 72	24
6	Madeira	5	+ 36	24	+ 72	24
7	Açores	4	_____	_____	_____	_____

A análise das variáveis categóricas é feita através da tabela de frequências. Aplicando a tabela de frequências às variáveis região, antiguidade e empregados obtêm-se os resultados do Quadro 17. Neste quadro constata-se que o maior número de empresas incide sobre a classe dois, há uma frequência de cinquenta e seis empresas na região Centro. A região do Algarve e a Região Autónoma dos Açores são as que contribuem com o menor número de empresas, apenas quatro em cada uma das regiões. A classe de empregados com maior frequência é a classe com um número de empregados entre 25 e 35. Relativamente à variável antiguidade a frequência varia entre as 24 e 25 empresas nas diversas classes.

6.3.4. Avaliação do risco

Nesta fase há que definir o risco individual máximo aceitável, ou seja, definir um limite a partir do qual o ficheiro de dados apresenta um risco aceitável ou, pelo contrário, é considerado um risco inaceitável. Como já referido anteriormente o risco individual máximo aceitável nestes dados é de 20%. Assim procedeu-se à análise do risco individual como se pode verificar na Figura 9.

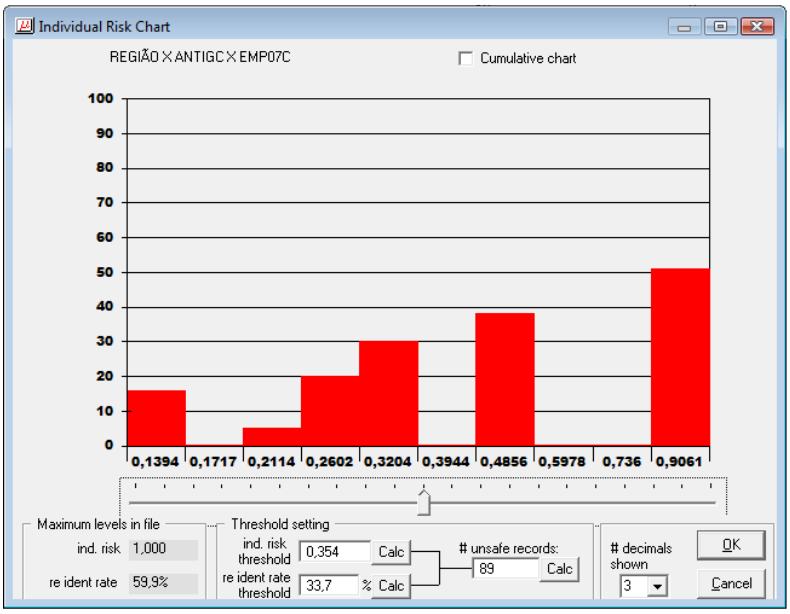

Figura 9 – Risco individual do ficheiro de dados original

A Figura 9 indica que a probabilidade de identificação dos dados originais é de 59,9%, isto é o número esperado de identificações é de 88.

Este ficheiro de dados originais apresenta um risco individual de 35,4%, o que significa que é um risco superior ao risco individual aceitável (20%). O software Argus calcula o risco através da equação 5.10.

Uma vez calculado o risco individual do ficheiro de dados original e antes de se proceder à aplicação de métodos de Controlo da Divulgação Estatística, uma vez que o risco individual é superior ao inicialmente estabelecido, faz-se uma análise no Argus e no SPSS para identificar e determinar quantas células inseguras existem.

6.3.5. Análise das variáveis no Argus

Para se verificar a existência de células inseguras é necessário definir um valor limite de indivíduos em cada célula para esta ser considerada insegura. O valor utilizado por alguns serviços de estatística é 3, tendo-se definido também 3 neste estudo. Este valor significa que se houver um número de empresas igual ou inferior a três numa célula de uma tabela resultante de um cruzamento das variáveis identificadoras o ficheiro é considerado inseguro.

Definido está que o ficheiro não é seguro para divulgação, procedeu-se ao cruzamento das variáveis categóricas região, antiguidade empregados 07.

O µ-Argus apenas indica a existência de combinações inseguras entre as variáveis categóricas, no entanto não indica quais as células em que existe essa insegurança, o que é conseguido recorrendo-se software SPSS, como se pode verificar mais à frente.

Seguidamente é realizada uma análise às variáveis categóricas no que respeita à sua segurança para divulgação.

1) Região

O cruzamento das variáveis região, antiguidade e empregados no software µ-Argus indica o número de combinações inseguras, conforme se pode verificar no Quadro 18.

Quadro 18 – Cruzamento das variáveis Região x Antiguidade x Número de Empregados no µ-Argus

Código	Frequência	Cruzamento de duas variáveis	Cruzamento de três variáveis
1	45	0	21
2	56	0	21
3	18	10	13
4	15	8	12
5	4	7	4
6	5	7	5
7	4	5	4

O Quadro 18 para além de indicar a existência de 37 inseguras no cruzamento de duas variáveis e 80 no cruzamento de três variáveis, também indica a frequência das empresas em cada uma das classes e por cada variável.

Quadro 19 – Combinações inseguras da variável Região

Células inseguras	Variável 1	Variável 2	Variável 3
20	Região	Antiguidade C	
17	Região	Empregados 07 C	
80	Região	Antiguidade C	Empregados 07 C

O Quadro 19 apenas indica a existência de combinações inseguras resultantes do cruzamento da variável região com as variáveis antiguidade e empregados 07 C.

O Quadro 20 identifica as células inseguras resultantes do cruzamento das variáveis região e antiguidade. Todas as células com valores iguais ou inferiores a 3 são células inseguras, identificadas no Quadro 20 com sombreado.

Quadro 20 – Cruzamento das variáveis Região x Antiguidade

		ANTIGUIDADE C						
		1	2	3	4	5	6	Total
REGIÃO	1	11	6	11	4	7	6	45
	2	8	12	4	15	9	8	56
	3	2	2	3	3	2	6	18
	4	4	2	2	2	2	3	15
	5	0	0	2	0	1	1	4
	6	0	1	1	1	2	0	5
	7	0	2	1	0	1	0	4
Total		**25**	**25**	**24**	**25**	**24**	**24**	**147**

O Quadro 21 identifica as 17 células inseguras entre as variáveis região e empregados 07, conforme se segue:

Quadro 21 – Cruzamento das variáveis Região e Número de empregados 07

		EMPREGADOS 07C							
		.	1	2	3	4	5	6	Total
REGIÃO	1	0	6	6	6	10	5	12	45
	2	1	11	11	13	4	10	6	56
	3	1	3	2	3	3	2	4	18
	4	0	2	2	4	3	4	0	15
	5	0	1	1	1	0	0	1	4
	6	0	3	1	1	0	0	0	5
	7	0	0	0	0	0	3	1	4
Total		**2**	**26**	**23**	**28**	**20**	**24**	**24**	**147**

2) <u>Antiguidade</u>

Analisando a variável antiguidade, obtém-se o Quadro 22 onde se pode verificar o número de células inseguras dessa variável com as restantes.

Quadro 22 – Cruzamento das variáveis Antiguidade x Região x Número de Empregados no µ-Argus

Código	Frequência	Cruzamento de duas variáveis	Cruzamento de três variáveis
1	25	4	13
2	25	6	14
3	24	6	16
4	25	6	9
5	24	9	14
6	24	5	14

O Quadro 22 indica a frequência de empresas em cada classe de antiguidade e o número de células inseguras no cruzamento das variáveis.

O cruzamento da variável antiguidade com outra variável, isto é, o cruzamento de duas variáveis sendo uma delas a antiguidade, resulta em 36 células inseguras, enquanto no cruzamento das três variáveis observam-se 80 células inseguras. O que pode ser verificado mais pormenorizadamente no Quadro 23 e Quadro 24.

Quadro 23 - Combinações inseguras da variável Antiguidade

Células inseguras	Variável 1	Variável 2	Variável 3
20	Região	Antiguidade C	
16	Antiguidade C	Empregados 07 C	
80	Região	Antiguidade C	Empregados 07 C

Quadro 24 – Cruzamento das variáveis Empregados x Antiguidade

		ANTIGUIDADE C						
		1	2	3	4	5	6	Total
EMPREGADOS 07C	.	1	0	0	1	0	0	2
	1	8	7	5	3	1	2	26
	2	5	6	4	4	3	1	23
	3	3	1	6	13	3	2	28
	4	3	2	4	3	3	5	20
	5	4	5	1	1	9	4	24
	6	1	4	4	0	5	10	24
Total		**25**	**25**	**24**	**25**	**24**	**24**	**147**

Como referido anteriormente, o cruzamento das três variáveis origina 80 células inseguras. O Quadro 25 identifica as células inseguras, como se segue:

Quadro 25 – Região x Antiguidade x Número de empregados

EMPREGADOS 07C			ANTIGUIDADE C						Total
			1	2	3	4	5	6	
1	REGIÃO	1	3	1	1	0	0	1	6
		2	3	4	1	3	0	0	11
		3	2	0	0	0	0	1	3
		4	0	1	1	0	0	0	2
		5	0	0	1	0	0	0	1
		6	0	1	1	0	1	0	3
	Total		**8**	**7**	**5**	**3**	**1**	**2**	**26**
2	REGIÃO	1	1	0	2	3	0	0	6
		2	2	5	0	1	2	1	11
		3	0	1	1	0	0	0	2
		4	2	0	0	0	0	0	2
		5	0	0	1	0	0	0	1
		6	0	0	0	0	1	0	1
	Total		**5**	**6**	**4**	**4**	**3**	**1**	**23**
3	REGIÃO	1	2	0	4	0	0	0	6
		2	0	0	2	8	2	1	13
		3	0	0	0	2	0	1	3
		4	1	1	0	2	0	0	4
		5	0	0	0	0	1	0	1
		6	0	0	0	1	0	0	1
	Total		**3**	**1**	**6**	**13**	**3**	**2**	**28**
4	REGIÃO	1	3	1	2	1	2	1	10
		2	0	1	0	2	0	1	4
		3	0	0	1	0	0	2	3
		4	0	0	1	0	1	1	3
	Total		**3**	**2**	**4**	**3**	**3**	**5**	**20**
5	REGIÃO	1	1	2	0	0	2	0	5
		2	2	1	0	1	4	2	10
		3	0	1	0	0	1	0	2
		4	1	0	0	0	1	2	4
		7	0	1	1	0	1	0	3
	Total		**4**	**5**	**1**	**1**	**9**	**4**	**24**
6	REGIÃO	1	1	2	2		3	4	12
		2	0	1	1		1	3	6
		3	0	0	1		1	2	4
		5	0	0	0		0	1	1
		7	0	1	0		0	0	1
	Total		**1**	**4**	**4**		**5**	**10**	**24**

3) *Número de empregados*

Analisando a variável empregados, obtêm-se os seguintes quadros:

Quadro 26 – Cruzamento das variáveis Antiguidade x Região x Número de Empregados

Código	Frequência	Cruzamento de duas variáveis	Cruzamento de três variáveis
1	26	7	15
2	23	6	12
3	28	7	11
4	20	6	14
5	24	4	15
6	24	3	13

O Quadro 26 e Quadro 27 indicam a existência de 33 combinações inseguras no cruzamento da variável empregados 07 com outra variável e 80 células inseguras no cruzamento de três variáveis.

Quadro 27 – Cruzamento da variável Empregados 07

Células inseguras	Variável 1	Variável 2	Variável 3
17	Região	Empregados 07 C	
16	Antiguidade C	Empregados 07 C	
80	Região	Antiguidade C	Empregados 07 C

6.3.6. Aplicação dos métodos de Controlo da Divulgação Estatística nas variáveis categóricas

A divulgação de dados, como já foi referido anteriormente, apenas deve ocorrer quando a confidencialidade dos entrevistados está protegida. Como se verificou anteriormente, os dados em estudo não são seguros para divulgação, tornando-se necessário a sua protecção.

Seguidamente é aplicado um método de Controlo da Divulgação Estatística, a recodificação global, de forma a criar um ficheiro seguro para divulgação.

6.3.6.1. Recodificação global

Numa primeira fase aplicou-se o método da recodificação global nas variáveis categóricas região, antiguidade e empregados 07. Este método consiste na criação de novas classes, mais amplas, tornando assim a identificação dos entrevistados menos provável.

Para a aplicação deste método foi necessário criar novas classes como se pode verificar no Quadro 28.

Quadro 28 – Novas classes para as variáveis região, antiguidade e empregados 07

Classes	Região	Antiguidade (anos)	Empregados 07 (n.º)
1	Norte e Centro	0 – 20	0 - 24
2	Lisboa	21 – 28	25 – 45
3	Alentejo e Algarve	+ 29	+ 46
4	Madeira e Açores	_____	_____

A aplicação da recodificação global contribui para uma significativa diminuição do número de combinações inseguras para divulgação, como se pode verificar no Quadro 29 e Quadro 30.

Quadro 29 – Cruzamento de variáveis após recodificação global – Variável Região

Código	Frequência	Cruzamento de duas variáveis	Cruzamento de três variáveis
1	101	0	0
2	18	0	7
3	19	0	6
4	9	4	7

Quadro 30 – Cruzamento das variáveis Região e Antiguidade após recodificação global

		ANTIGUIDADE RG			
		1	2	3	Total
REGIÃO RG	1	37	34	30	101
	2	4	6	8	18
	3	6	6	7	19
	4	3	3	3	9
Total		**50**	**49**	**48**	**147**

Após a aplicação do método da recodificação global no cruzamento das variáveis categóricas região com antiguidade existem apenas 3 combinações inseguras, todas elas referentes à região 4, ou seja, Madeira e Açores, com 3 empresas em cada classe de antiguidade, conforme se pode verificar no Quadro 30.

Quadro 31 - Região x Empregados 07 após recodificação global

		EMPREGADOS 07CRG			
		1	2	3	Total
REGIÃO RG	1	34	33	33	100
	2	5	6	6	17
	3	6	8	5	19
	4	4	1	4	9
Total		**49**	**48**	**48**	**145**

No Quadro 31 pode-se verificar a existência de uma combinação insegura. A célula insegura recai sobre uma empresa da Madeira ou dos Açores (classe 4), com um número de empregados igual ou inferior 24 (classe 2).

Quadro 32 – Antiguidade e Número de Empregados após recodificação global

		EMPREGADOS 07CRG			
		1	2	3	Total
ANTIGUIDADE C RG	1	26	9	14	49
	2	16	26	6	48
	3	7	13	28	48
Total		**49**	**48**	**48**	**145**

Analisando o Quadro 32 verifica-se que não existem células inseguras na combinação das variáveis antiguidade e empregados 07.

Quadro 33 – Antiguidade x Número de Empregados x Região após recodificação global

			EMPREGADOS 07CRG			
REGIÃO RG			1	2	3	Total
1	ANTIGRG	1	19	7	10	36
		2	11	19	4	34
		3	4	7	19	30
2	ANTIGRG	1	3	0	1	4
		2	1	3	1	5
		3	1	3	4	8
3	ANTIGRG	1	3	2	1	6
		2	3	3	0	6
		3	0	3	4	7
4	ANTIGRG	1	1	0	2	3
		2	1	1	1	3
		3	2	0	1	3

Após a recodificação global, o cruzamento das variáveis categóricas região, antiguidade e empregados ainda contém combinações inseguras embora em muito menor número (20 combinações inseguras), conforme se pode verificar no Quadro 33.

6.3.7. Aplicação dos métodos de controlo da divulgação estatística nas variáveis contínuas

Seguidamente são aplicados alguns métodos de Controlo da Divulgação Estatística, como a microagregação, codificação superior, rank swapping, arredondamento e microagregação híbrida, também referida como dados híbridos, às variáveis contínuas, CMVMC 07, MB 07, VAB 07 e VN 07.

6.3.7.1. Microagregação numérica

Para obter dados de um conjunto de registos microagregados, os dados são combinados de forma a obter g grupos de tamanho, pelo menos, k. Para cada variável, é calculado o valor médio de cada grupo. Os valores das variáveis originais das variáveis são substituídos pelo valor médio de cada grupo. Os grupos são formados com um critério de similaridade máxima. A aplicação deste método nas variáveis em estudo agrega 5 variáveis em cada grupo.

De seguida é apresentado um pequeno exemplo do que faz este método.

Quadro 34 – Exemplo da aplicação do método da microagregação

Empresas	Variáveis	Valor médio
1; 35; 68; 116; 139	CMVMC 07	284 931 €
1; 35; 68; 116; 139	MB 07	2 008 425 €
1; 35; 68; 116; 139	VAB 07	1 884 186 €
1; 35; 68; 116; 139	VN 07	2 306 959 €

O Quadro 34, exemplifica o método da microagregação nas variáveis em estudo, onde a partir dos valores originais das variáveis CMVMC 07; MB 07; VAB 07 e VN 07, das empresas 1; 35; 68; 116 e 119 se calcularam os valores médios de cada variável desse

grupo e se substituíram os valores originais pelos valores médios mencionados no Quadro 34.

Depois da aplicação do método da microagregação, há que fazer uma nova análise das variáveis no que respeita à média, mediana e desvio padrão, conforme o quadro seguinte:

Quadro 35 – Análise descritiva das variáveis após a microagregação

	CMVMC07		MB07		VAB07		VN07	
	Original	Microagregado	Original	Microagregado	Original	Microagregado	Original	Microagregado
Valores válidos	142	147	147	147	147	147	147	147
Missing	5	0	0	0	0	0	0	0
Média	1 128 227,69	1 089 852,50	3 553 014,81	3 553 014,78	1 745 529,01	1 745 529,00	4 446 503,59	4 446 503,48
Mediana	418 275,50	408 542,00	1 811 377,00	2 008 425,00	987 953,00	1 001 998,00	2338 340, 00	2 306 959,00
Desvio padrão	1 719 809,51	1 550 593,69	4 598 413 18	4 207 586,25	1 975 837,23	1 829 876,82	5 551 866,21	5 232 547,88

O Quadro 35 contém o valor da média, mediana e desvio padrão das variáveis em estudo antes e após a microagregação. Analisando os seus valores verifica-se que não existe diferença (apenas nos cêntimos, fruto dos arredondamentos) nos valores da média, à excepção da variável custo das mercadorias vendidas e matérias consumidas.

O CMVMC 07 apresenta um valor diferente na média devido aos missing's existentes no ficheiro de dados original. Após a microagregação não existem missing's, uma vez que é atribuída a cada unidade inquirida o valor médio do seu grupo. Todas as medidas descritivas desta variável sofrem uma diminuição após a microagregação.

Relativamente à mediana e ao desvio padrão verificam-se algumas alterações dos seus valores relativamente aos dados originais, a MB e o VAB apresentam um valor mais elevado da mediana, enquanto o desvio padrão apresenta um valor menor.

A variável VN 07, após a microagregação tem valores inferiores quer na mediana quer no desvio padrão.

6.3.7.2. Codificação superior

O método de codificação superior também foi utilizado para modificar as variáveis em estudo. Este método consiste em substituir os valores das variáveis acima de um determinado limite por um dado valor.

Neste caso recorreu-se à análise das empresas outliers para identificar as empresas e valores a partir dos quais existe maior risco de identificação. No Quadro 36, podem-se verificar os valores máximos de cada variável em estudo e as respectivas empresas às quais pertencem esses valores.

Quadro 36 – Valores máximos das variáveis CMVMC 07, MB 07, VAB 07 e VN 07

Variáveis contínuas	Identificação das Empresas	Nº da Empresa	Valor
Custo das mercadorias vendidas e matérias consumidas 07 (CMVMC 07)	Domingues & Contente – Britas e Asfaltos, SA	37	3 314 275 €
Margem Bruta 07 (MB 07)	Iberobrita, SA	70	9 934 251 €
Valor Acrescentado Bruto (VAB 07)	Sorgila, SA.	142	4 652 493 €
Volume de Negócios 07 (VN 07)	Calbrita, SA	28	11 100 093 €

Após identificar os valores máximos das variáveis CMVMC 07, MB 07, VAB 07 e VN 07, aplicou-se o método da codificação superior para diminuir o risco de identificação. Todas as unidades inquiridas (empresas) que têm valores das variáveis acima dos valores máximos mencionados no Quadro 36, esses valores são substituídos pelos valores máximos de cada variável, conforme se pode verificar no Quadro 37.

Quadro 37 – Exemplo da aplicação do método da codificação superior

Identificação das Empresas	Variáveis	Novo Valor
5; 28; 36; 37; 51; 70; 74; 87; 89; 91; 120; 140	CMVMC 07	3 314 275 €
5; 45; 70; 74; 82; 86; 87; 126; 128; 131; 140; 142	MB 07	9 934 251 €
5; 36; 45; 70; 74; 90; 126; 128; 140; 142	VAB 07	4 652 493 €
5; 28; 45; 70; 74; 82; 86; 87; 126; 128; 131; 140; 142	VN 07	11 100 093 €

Não é de mais referir que o ficheiro de dados em estudo não tem empresas outliers inferiores e, uma vez que o critério utilizado para a codificação foi baseado na

representação gráfica da caixa de bigodes, o método da codificação inferior não foi aplicado.

O Quadro 38 apresenta a análise descritiva das variáveis contínuas CMVMC07, MB 07, VAB 07 e VN 07 após a aplicação do método da codificação superior.

Quadro 38 – Análise descritiva após a codificação superior

	CMVMC07		MB07		VAB07		VN07	
	Original	Codificação Superior	Original	Codificação Superior	Original	Codificação Superior	Original	Codificação Superior
Valores válidos	142	142	147	147	147	147	147	147
Missing	5	5	0	0	0	0	0	0
Média	1 128 227,69	935 702,06	3 553 014,81	3 068 708,98	1 745 529,01	1 536 127,01	4 446 503,59	3 792 956,41
Mediana	418 275,50	418 275,50	1 811 377,00	1 811 377,00	987 953,00	987 953,00	2 338 340, 00	2 338 340,00
Desvio padrão	1 719 809,51	1 021 562,05	4 598 413 18	2 616 855,83	1 975 837,23	1 211 390,59	5 551 866,21	3 086 323,31

Analisando o Quadro 38 constata-se que não existem variações no valor da mediana, o que seria de esperar, uma vez que esta é uma medida do centro. Uma vez que este método apenas alterou o valor das variáveis acima de um determinado valor e considerou nelas o valor limite, o valor central continua o mesmo. Relativamente ao valor médio e ao desvio padrão verifica-se uma significativa diminuição, uma vez que todos os valores alterados foram substituídos por valores inferiores.

6.3.7.3. Arredondamento

O método do arredondamento, como referido anteriormente, consiste no arredondamento individual das variáveis.

A base de arredondamento utilizada neste trabalho foi de 10 000 €. Os valores das variáveis originais foram arredondados para valores "certos" de acordo com o critério de arredondamento normal.

Tomando como exemplo a empresa 1 – A. Bento Vermelho, Lda., o Quadro 39 exemplifica o método de arredondamento.

Quadro 39 – Exemplo do Método do arredondamento

Variáveis	Valor original	Valor arredondado
CMVMC 07	676 406 €	680 000 €
MB 07	2 420 855 €	2 420 000 €
VAB 07	1 496 595 €	1 500 000 €
VN 07	3 111 517 €	3 110 000 €

Após a aplicação do método do arredondamento fez-se uma análise descritiva das variáveis contínuas, conforme o quadro abaixo:

Quadro 40 – Análise descritiva após o método do arredondamento

	CMVMC07		MB07		VAB07		VN07	
	Original	Arredonda.	Original	Arredonda.	Original	Arredonda.	Original	Arredonda.
Valores válidos	142	142	147	147	147	147	147	147
Missing	5	5	0	0	0	0	0	0
Média	1 128 227,69	1 128 521,13	3 553 014,81	3 553 061,22	1 745 529,01	1 745 510,20	4 446 503,59	4 446 666,67
Mediana	418 275,50	415 000,00	1 811 377,00	1 810 000,00	987 953,00	990 000,00	2 338 340, 00	2 340 000,00
Desvio padrão	1 719 809,51	1 719 799,98	4 598 413 18	4 598 274,04	1 975 837,23	1 975 682,36	5 551 866,21	5 552 357,93

O Quadro 40 indica que não existem diferenças significativas nos valores da média, mediana e desvio padrão das variáveis após a aplicação do método do arredondamento.

6.3.7.4. Rank Swapping

O método de Controlo da Divulgação Estatística rank swapping, consiste em ordenar por ordem crescente os valores das variáveis contínuas originais CMVMC 07; MB 07; VAB 07 e VN 07, cada valor ordenado é trocado aleatoriamente por outro valor. Esse valor é escolhido dentro de um intervalo de 15% do número total de linhas do ficheiro, neste caso o valor tem que estar dentro das 22 linhas mais próximas da posição ordenada da variável original.

Por exemplo, o valor do custo das mercadorias vendidas e matérias consumidas 07 original, da empresa 1 (A. Bento e Vermelho, Lda) é de 676 406 €, posicionada na linha 82 do ficheiro ordenado, após a aplicação do método, esse valor foi substituído por 304 324 € que pertence à empresa 123 (Sanchez, SA), posicionada na linha 60. A troca processa-se de uma forma aleatória. O Quadro 41 oferece um pequeno exemplo do seu funcionamento.

Quadro 41 – Exemplificação do método rank swapping

Empresa	Valores		CMVMC 07	MB 07	VAB 07	VN 07
1	Valor original	Valor	676 406 €	2 420 855 €	1 496 595 €	3 111 517 €
		Linha	82	92	94	93
	Novo valor	Valor	304 324 €	2 411 075 €	2 179 694 €	2 870 692 €
		Linha	60	91	113	90
2	Valor original	Valor	118 295 €	1 249 822 €	813 673 €	1 041 527 €
		Linha	29	35	56	6
	Novo valor	Valor	76 669 €	1 140 098 €	943 395 €	1 106 610 €
		Linha	22	24	68	8

À semelhança das análises anteriores, fez-se também após a aplicação do método rank swapping a análise descritiva das variáveis contínuas, conforme Quadro 42.

Quadro 42 – Análise descritiva após o método rank swapping

	CMVMC07		MB07		VAB07		VN07	
	Original	Arredonda.	Original	Arredonda.	Original	Arredonda.	Original	Arredonda.
Valores válidos	142	142	147	147	147	147	147	147
Missing	5	5	0	0	0	0	0	0
Média	1 128 227,69	1 089 852,60	3 553 014,81	3 553 014,81	1 745 529,01	1 745 529,01	4 446 503,59	4 446 503,59
Mediana	418 275,50	399 714,00	1 811 377,00	1 811 377,00	987 953,00	987 953,00	2 338 340, 00	2 338 340,00
Desvio padrão	1 719 809,51	1 702 516,31	4 598 413, 18	4 598 413,18	1 975 837,23	1 975 837,23	5 551 866,21	5 551 866,21

A aplicação deste método apenas alterou os valores da média, mediana e desvio padrão da variável custo das mercadorias vendidas e matérias consumidas. Esta alteração, tal como no método da microagregação, deve-se ao facto dos dados originais terem missing's e após a aplicação destes dois métodos eles deixam de existir.

Relativamente às restantes variáveis não sofreram alterações nestas medidas, uma vez que os valores são os mesmos apenas "mudaram de lugar".

6.3.7.5. Microagregação Híbrida

A aplicação da microagregação híbrida num ficheiro de dados para divulgação consiste em calcular dados mascarados como uma combinação de dados originais e de dados sintéticos. Tomando por base o ficheiro original, foram calculadas classes para cada uma das variáveis contínuas e calculada a média para cada classe, que foi atribuída às respectivas empresas. O valor final da variável é calculado com base na seguinte

equação 3.22 apresentada anteriormente na secção 3.3.8.3 : Variável final = 0,5*Variável original + 0,5*Variável média.

É apresentado de seguida um pequeno exemplo da aplicação dados híbridos.

Quadro 43 – Exemplificação da aplicação dos dados híbridos

Variáveis	Valor original	Classe	Valor médio da classe	Valor Final
CMVMC 07	676 406 €	4	655 914 €	666 160 €
MB 07	2 420 855 €	4	2 268 387 €	2 344 621 €
VAB 07	1 496 595 €	4	1 443 775 €	1 470 185 €
VN 07	3 111 517 €	4	2 929 240 €	3 020 379 €

À semelhança dos métodos anteriores, depois de aplicar os dados híbridos efectuou-se a análise descritiva do ficheiro alterado, como se segue:

Quadro 44 – Análise descritiva após o método dos dados híbridos

	CMVMC07		MB07		VAB07		VN07	
	Original	Dados híbridos	Original	Dados híbridos	Original	Dados híbridos	Original	Dados híbridos
Valores válidos	142	142	147	147	147	147	147	147
Missing	5	5	0	0	0	0	0	0
Média	1 128 227,69	1 107 100,93	3 553 014,81	3 553 014,81	1 745 529,01	1 745 529,01	4 446 503,59	4 446 503,59
Mediana	418 275,50	405 846,50	1 811 377,00	1 811 377,00	987 953,00	987 953,00	2 338 340, 00	2 338 340,00
Desvio padrão	1 719 809,51	1 524 547,41	4 598 413, 18	4 050 975,88	1 975 837,23	1 812 553,24	5 551 866,21	4 995 922,01

A variável CMVMC 07, tal como acontece noutros métodos, é a que apresenta maiores diferenças na análise descritiva, o que se deve à presença de missing´s no ficheiro de dados. Relativamente às restantes variáveis apenas se verificam alterações no desvio padrão.

6.3.8. Análise global do ficheiro seguro

Depois da aplicação dos métodos de Controlo da Divulgação Estatística devem ser gerados ficheiros seguros para divulgação. A criação de ficheiros seguros leva à supressão de algumas células, como se pode verificar no Quadro 45.

Quadro 45 – Supressão de células

Variável	Supressões
Região	47
Antiguidade	4
Empregados 07 C	0
Total	**51**

6.3.8.1. Conclusão da aplicação dos métodos de CDE nas variáveis contínuas

Após a aplicação de qualquer um dos métodos de Controlo da Divulgação Estatística, como seria de esperar, o número de células inseguras diminuiu substancialmente. O resultado obtido relativamente ao número de combinações inseguras é o mesmo independentemente do método aplicado.

1) Região

A combinação da variável região com as restantes variáveis após a aplicação dos métodos de Controlo da Divulgação Estatística pode ser analisada no seguinte quadro:

Quadro 46 – Cruzamento das variáveis Região x Antiguidade x Empregados após os métodos de CDE

Classe	Frequência	Cruzamento de uma variável	Cruzamento de duas variáveis	Cruzamento de três variáveis
1	36	0	2	12
2	46	0	3	10
3	9	0	6	4
4	7	0	6	3
6	1	1	1	0
7	1	1	1	0

No Quadro 46 verifica-se a existência 19 variáveis inseguras no cruzamento de duas variáveis, sendo uma delas a região e 29 combinações inseguras no cruzamento das três variáveis. O

Quadro 47 e o Quadro 48 apresentam de forma mais detalhada as células inseguras, nos quais se constata a existência de 7 combinações inseguras no cruzamento das variáveis Região e Antiguidade, assinaladas no Quadro 48 a sombreado.

Quadro 47 - Cruzamento das variáveis após os métodos de CDE – Variável Região

Células inseguras	Variável 1	Variável 2	Variável 3
2	Região		
7	Região	Antiguidade	
12	Região	Empregados 07	
29	Região	Antiguidade	Empregados 07

Quadro 48 – Região x Antiguidade após aplicação dos métodos CDE

		ANTIGUIDADE C							Total
		.	1	2	3	4	5	6	
REGIÃO	.	0	5	11	11	4	8	8	47
	1	0	8	4	10	3	7	4	36
	2	1	8	9	2	13	8	5	46
	3	0	2	0	0	3	0	4	9
	4	1	2	0	0	2	0	2	7
	6	1	0	0	0	0	0	0	1
	7	1	0	0	0	0	0	0	1
Total		**4**	**25**	**24**	**23**	**25**	**23**	**23**	**147**

O Quadro 49 identifica as 12 células inseguras no cruzamento das variáveis região e empregados, que são todas as células com valores iguais ou inferiores a 3.

Quadro 49 – Região x Empregados após aplicação dos métodos CDE

		EMPREGADOS 07C							Total
		.	1	2	3	4	5	6	
REGIÃO	.	0	10	6	5	9	9	8	47
	1	0	3	5	6	7	4	11	36
	2	1	10	10	12	2	8	3	46
	3	1	2	0	2	2	0	2	9
	4	0	0	2	3	0	2	0	7
	6	0	1	0	0	0	0	0	1
	7	0	0	0	0	0	1	0	1
Total		**2**	**26**	**23**	**28**	**20**	**24**	**24**	**147**

2) *Antiguidade*

A combinação da variável antiguidade com as restantes variáveis resulta nas combinações inseguras dos seguintes quadros:

Quadro 50 – Antiguidade x Região x Empregados após aplicação dos métodos CDE

Classe	Frequência	Cruzamento de uma variável	Cruzamento de duas variáveis	Cruzamento de três variáveis
1	25	0	5	8
2	24	0	1	2
3	23	0	1	4
4	25	0	6	5
5	23	0	3	5
6	23	0	3	5

Quadro 51 – Antiguidade após aplicação dos métodos CDE

Células inseguras	Variável 1	Variável 2	Variável 3
7	Região	Antiguidade	
12	Antiguidade	Empregados 07	
29	Região	Antiguidade	Empregados 07

Quadro 52 – Antiguidade x Empregados após aplicação dos métodos CDE

		EMP07C							Total
		.	1	2	3	4	5	6	
ANTIGC	.	0	1	1	1	0	1	0	4
	1	1	8	5	3	3	4	1	25
	2	0	7	6	0	2	5	4	24
	3	0	5	4	6	4	0	4	23
	4	1	3	4	13	3	1	0	25
	5	0	0	3	3	3	9	5	23
	6	0	2	0	2	5	4	10	23
Total		**2**	**26**	**23**	**28**	**20**	**24**	**24**	**147**

No Quadro 51 constata-se a existência de 12 combinações inseguras no cruzamento das variáveis antiguidade e empregados, devidamente identificadas no Quadro 52.

As 7 combinações inseguras entre antiguidade e região foram identificadas na análise da variável região.

3) *Variável Empregados*

Relativamente à variável empregados a tabela abaixo indica as combinações inseguras existentes entre as três variáveis, conforme se segue:

Quadro 53 – Empregados x Antiguidade x Região após aplicação dos métodos CDE

Classe	Frequência	Cruzamento de uma variável	Cruzamento de duas variáveis	Cruzamento de três variáveis
1	26	0	5	4
2	23	0	2	5
3	28	0	5	5
4	20	0	6	5
5	24	0	3	5
6	24	0	3	5

Quadro 54 – Empregados após aplicação dos métodos CDE

Células inseguras	Variável 1	Variável 2	Variável 3
12	Região	Empregados 07	
12	Antiguidade	Empregados 07	
29	Região	Antiguidade	Empregados 07

As células inseguras que resultam do cruzamento das variáveis empregados, região e antiguidade foram identificadas anteriormente na análise da variável região e da variável antiguidade respectivamente.

O cruzamento das três variáveis, região, antiguidade e empregados resulta em 29 combinações inseguras, que são assinaladas a sombreado no Quadro 55.

Quadro 55 – Região x Antiguidade x Empregados após aplicação dos métodos CDE

REGIÃO			EMPREGADOS 07C							Total
			.	1	2	3	4	5	6	
1	ANTIGC	1		3	0	2	3	0	0	8
		2		0	0	0	0	2	2	4
		3		0	2	4	2	0	2	10
		4		0	3	0	0	0	0	3
		5		0	0	0	2	2	3	7
		6		0	0	0	0	0	4	4
		Total		**3**	**5**	**6**	**7**	**4**	**11**	**36**
2	ANTIGC	.	0	0	1	0	0	0	0	1
		1	1	3	2	0	0	2	0	8
		2	0	4	5	0	0	0	0	9
		3	0	0	0	2	0	0	0	2
		4	0	3	0	8	2	0	0	13
		5	0	0	2	2	0	4	0	8
		6	0	0	0	0	0	2	3	5
		Total	**1**	**10**	**10**	**12**	**2**	**8**	**3**	**46**
3	ANTIGC	1	0	2		0	0		0	2
		4	1	0		2	0		0	3
		6	0	0		0	2		2	4
		Total	**1**	**2**		**2**	**2**		**2**	**9**
4	ANTIGC	.			0	1		0		1
		1			2	0		0		2
		4			0	2		0		2
		6			0	0		2		2
		Total			**2**	**3**		**2**		**7**

Após a aplicação de qualquer um dos métodos de Controlo da Divulgação Estatística, é necessário proceder-se à avaliação do risco, de forma a verificar se o risco individual do ficheiro tratado está dentro do valor estabelecido atrás como aceitável (20%). Para isso recorreu-se à análise gráfica do risco individual, na qual se determinou um risco individual de 17,1%, o que significa que o ficheiro alterado pode ser divulgado.

6.3.9. Qualidade dos dados

A aplicação de diferentes métodos de Controlo da Divulgação Estatística leva à obtenção de diferentes ficheiros de dados para divulgação. Há que escolher o melhor método, essa escolha foi baseada qualidade dos dados, uma vez que todos eles têm a mesma quantidade de células inseguras.

Para analisar a qualidade dos dados foram utilizadas três medidas para as variáveis contínuas, o Erro Quadrático Médio, o Erro Absoluto Médio e a Variação Média, todas elas aplicadas a variáveis contínuas e todas elas referentes à matriz de dados originais ***X*** quando comparada com a matriz ***X'***, como se referiu anteriormente.

Relativamente às variáveis categóricas, a qualidade dos dados não será analisada uma vez que não é comparável com as medidas utilizadas em variáveis contínuas.

1) *Erro Quadrático Médio (EQM)*

O EQM é uma medida de qualidade de dados que mede o somatório da diferença, dos quadrados, entre a variável original e a variável modificada dividindo este resultado pelo produto do número de registos com o número de variáveis.

Assim tem-se a seguinte expressão para o EQM:

$$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij} - x'i_{j}\right)^2}{np} \qquad (6.1)$$

2) *Erro Absoluto Médio (EAM)*

O EAM mede o somatório da diferença em termos absolutos entre a variável original e a variável modificada, dividido pelo produto do número de registos com o número de variáveis.

O EAM é calculado de acordo com a seguinte expressão:

$$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\left|x_{ij} - x'_{ij}\right|}{np} \qquad (6.2)$$

3) *Variação média*

A VM é dada pela seguinte expressão:

$$\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\frac{\left|x_{ij} - x'_{ij}\right|}{\left|xij\right|}}{np} \qquad (6.3)$$

O Quadro 56 contém os resultados das medidas de qualidade dos dados após a aplicação dos métodos de Controlo da Divulgação Estatística.

Quadro 56 – Medidas de qualidade dos dados

Medida de Qualidade	Microagregação Numérica	Numeric Variables Codificação Superior	Rank Swapping	Arredondamento	Dados Hibridos
Erro Quadrático Médio	0,0487	0,004927	0,2367	0,000096	0,0186
Erro Absoluto Médio	0,1000	0,0145	0,1838	0,0026	0,0246
Variação Média	0,0221	0,0020	0,0370	0,0005	0,0077

Após a análise do Quadro 56 constata-se o seguinte:

1) Se for utilizada como medida de qualidade dos dados o EQM, o método que apresenta melhor resultado é o método do Arredondamento, uma vez que é o que apresenta o menor valor;
2) Utilizando como medida o EAM, optava-se pelo método Arredondamento;
3) Utilizando a Variação Média como medida de qualidade dos dados, o método escolhido também seria o do arredondamento.

A escolha do melhor método deve recair sobre aquele que origina menor perda de dados, isto é o que apresenta menor valor na medida de qualidade dos dados.

Pode-se então concluir que o método que origina menor perda de dados é o método do arredondamento, uma vez que aplicando qualquer uma das medidas de qualidade dos dados, é o que apresenta melhor resultado. O método que apresenta maior perda de informação é o método rank swapping.

6.4.Análise da base de dados familiares

Recorreu-se a uma segunda base de dados para testar de novo os métodos de controlo da divulgação estatística. Esta base de dados refere-se a dados simulados semelhante aos obtidos através de um inquérito às famílias realizado pelo INE. Como variáveis categóricas tem-se: a região, profissão e número de pessoas em alojamentos. A única variável contínua neste ficheiro é a variável remunerações.

6.4.1. Etapas para a divulgação de um ficheiro de dados seguro

Neste capítulo são identificadas e descritas a etapas a percorrer para se divulgar um ficheiro de dados em segurança.

Quadro 57 – Guia para a divulgação do ficheiro da base de dados familiar

Etapas para o processo de divulgação	**Análises a efectuar/Problema resolvido** ↓ **Resultados esperados**
1. Porque é que a protecção da confidencialidade é necessária?	**Os dados referem-se a pessoas singulares ou colectivas?** ↓ Os dados referem-se a pessoas individuais, em agregados familiares.
2. Quais são as principais características e utilização dos dados?	**Análise dos dados/Estrutura dos dados** **Os dados apresentam uma estrutura específica?** ↓ Os dados referem-se a informações gerais de indivíduos. O capítulo 6.4.2 faz uma análise preliminar dos dados
	Análise das metodologias de pesquisa ↓ (Não se aplica)
	Análise dos objectivos dos Institutos de Estatística **Que tipo de divulgação?** ↓ Na realidade não vai haver divulgação dos dados. Os dados são para efectuar um estudo da aplicação de alguns métodos CDE e a comparação dos mesmos.
	Análise das necessidades dos utilizadores ↓ (Não se aplica)
	Análise de questionários **Listagem das variáveis a ser removidas, variáveis a ser incluídas** ↓ As variáveis incluídas na base de dados e em estudo são: região; profissão; nº de pessoas em alojamentos e remunerações.
3. Riscos de divulgação	**Cenário de divulgação - Lista de variáveis identificadoras indirectas** ↓ Região, Profissão e Número de pessoas em alojamentos
	Definição do risco ↓ O Eurostat recomenda um risco individual máximo de 4%. No entanto este nível de rico justifica-se apenas em grandes bases de dados (com milhões de registos). Neste caso assume-se um risco individual máximo de 20% dada a pequena dimensão do ficheiro de dados.
	Avaliação do risco ↓ A avaliação do risco é feita no capítulo 6.4.3 e no qual se pode verificar que os dados são inseguros para divulgação dado o risco definido no ponto anterior.

Etapas para o processo de divulgação (Continuação)	Análises a efectuar/Problema resolvido ↓ Resultados esperados
4. Métodos de controlo da divulgação	**Análise do tipo de dados envolvidos, políticas dos Serviços de Estatística e necessidades dos utilizadores.** **Identificação dos métodos de limitação da divulgação** ↓ Recodificação Global, microagregação, rank swapping, arredondamento e microagregação híbrida (capítulo 6.4.5).
	Análise da perda de informação ↓ Análise da perda de informação é realizada no capítulo 6.4.7 através do SSE; EAM e VA.
5. Implementação	Escolha do software, parâmetros e limites dos diferentes métodos ↓ µ-Argus e SPSS

6.4.2. Análise preliminar das variáveis

Inicialmente foi calculada a média, mediana e desvio padrão da variável contínua, remunerações e as frequências das variáveis identificadoras indirectas, também referidas neste documento como variáveis categóricas, região, profissão e número de pessoas em alojamentos, como se pode verificar de seguida.

1) Variáveis contínuas

O Quadro 58 apresenta o valor da média, mediana, desvio padrão e variância da variável remunerações.

Quadro 58 – Análise descritiva da variável remunerações

	Remunerações
Valores válidos	1067
Missing	0
Média	490,41
Mediana	483,00
Desvio padrão	287,63

Para verificar a existência de indivíduos com valores muito diferentes dos normais, isto é, indivíduos considerados outliers, recorreu-se à análise gráfica da caixa de bigodes. Como se pode constatar na Figura 10 verifica-se a inexistência de indivíduos outliers na base de dados familiares.

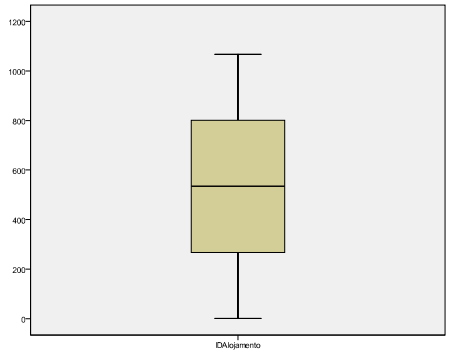

Figura 10 – Indivíduos Outliers

2) Variáveis categóricas

A análise das variáveis categóricas região, profissão e número de pessoas em alojamentos é feita na tabela de frequências, conforme se segue:

Quadro 59- Tabela de frequências das variáveis categóricas

Classes	Região	Profissão	Nº Pessoas
1	156	86	273
2	177	98	267
3	157	117	265
4	157	118	262
5	143	104	___
6	151	138	___
7	126	107	___
8	___	100	___
9	___	88	___
10	___	111	___
Total	**1 067**	**1 067**	**1 067**

6.4.3. Avaliação do risco individual

A avaliação do risco é uma das etapas mais importantes, no Controlo da Divulgação Estatística. Uma das formas de fazer essa avaliação é recorrendo à análise gráfica do risco individual, realizada às variáveis categóricas. Inicialmente procedeu-se à avaliação do risco do ficheiro de dados original, conforme se segue:

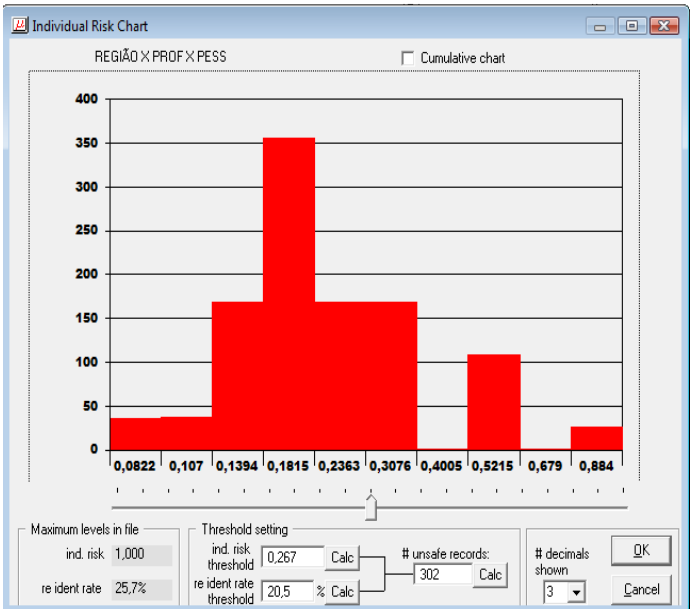

Figura 11 – Risco individual dos dados familiares originais

Como se pode verificar pela análise à Figura 11, a probabilidade de identificação dos dados é de 25,7%, traduzindo-se num número esperado de 302 de identificações. O risco individual é de 26,7%, existindo uma elevada probabilidade de ocorrerem identificações. Isto risco fica acima do risco individual definido anteriormente (20%), sendo necessário recorrer aos métodos de controlo da divulgação estatística para diminuir o risco de identificação.

Antes da aplicação dos métodos de CDE fez-se uma análise das variáveis no Argus e no SPSS para verificar o número de células inseguras.

6.4.4. Análise das variáveis no µ-Argus

Recorreu-se ao Argus para verificar a existência ou não de células inseguras na base de dados familiares. O cruzamento das variáveis categóricas região, profissão e número de pessoas em alojamentos indica a existência de 136 células inseguras, de referir que o cruzamento de apenas 2 variáveis não contém células inseguras, conforme se pode verificar no Quadro 60.

Quadro 60 - Cruzamento das variáveis região x profissão x número de Pessoas no µ-Argus

Código	Frequência	Cruzamento de duas variáveis	Cruzamento de três variáveis
1	156	0	18
2	143	0	22
3	177	0	13
4	129	0	26
5	143	0	19
6	152	0	21
7	167	0	17

As células inseguras são todas as que apresentam valores iguais ou inferiores a 3, marcadas a sombreado no Quadro 61.

Quadro 61 – Cruzamento das variáveis região x profissão x número de pessoas

Nº pessoas alojamento			Profissão										Total
			1	2	3	4	5	6	7	8	9	10	
1	Região	1	5	2	7	3	6	4	1	3	0	2	33
		2	5	1	2	2	1	1	2	4	4	3	25
		3	4	5	4	7	4	0	1	6	6	3	40
		4	2	4	2	3	5	2	5	2	3	1	29
		5	4	5	5	2	5	2	6	2	6	3	40
		6	4	4	3	1	1	6	2	7	7	8	43
		7	2	2	3	5	4	4	6	1	2	4	33
	Total		**26**	**23**	**26**	**23**	**26**	**19**	**23**	**25**	**28**	**24**	**243**
2	Região	1	3	4	10	5	5	6	4	1	5	2	45
		2	2	6	3	1	4	5	2	3	1	4	31
		3	4	4	5	3	3	2	5	6	3	9	44
		4	4	2	5	1	2	2	3	2	2	7	30
		5	4	2	5	5	2	7	1	2	6	7	41
		6	6	3	4	3	2	5	3	5	3	2	36
		7	1	8	6	7	6	3	1	6	6	3	47
	Total		**24**	**29**	**38**	**25**	**24**	**30**	**19**	**25**	**26**	**34**	**274**
3	Região	1	6	2	1	1	5	5	5	3	2	11	41
		2	5	7	2	2	3	1	5	3	11	3	42
		3	7	3	5	7	5	0	6	3	3	8	47
		4	2	3	7	3	3	2	6	3	3	2	34
		5	1	6	4	3	0	5	6	3	5	3	36
		6	4	2	5	5	5	4	3	2	4	2	36
		7	1	3	5	8	8	2	2	5	3	9	46
	Total		**26**	**26**	**29**	**29**	**29**	**19**	**33**	**22**	**31**	**38**	**282**
4	Região	1	7	2	2	4	3	2	5	4	3	5	37
		2	4	2	3	3	6	5	9	4	5	4	45
		3	4	3	4	0	3	3	14	6	1	8	46
		4	8	1	2	6	5	2	5	4	0	3	36
		5	4	4	3	1	3	4	2	1	2	2	26
		6	2	3	4	7	3	2	7	3	3	3	37
		7	4	4	1	3	4	8	6	2	4	5	41
	Total		**33**	**19**	**19**	**24**	**27**	**26**	**48**	**24**	**18**	**30**	**268**

6.4.5. Aplicação dos métodos de Controlo da Divulgação Estatística

A análise da base de dados originais indicou a existência de células inseguras para divulgação e um risco individual superior ao aceitável. É necessário aplicar métodos de Controlo da Divulgação Estatística de forma a tornar os dados mais seguros para poderem ser divulgados. De seguida são analisados os resultados da aplicação de alguns métodos, nomeadamente a recodificação global, a microagregação, rank swapping, o arredondamento e a microagregação híbrida.

6.4.5.1. *Recodificação global*

Como referido anteriormente o método da recodificação global consiste na criação de novas classes nas variáveis categóricas, as quais podem ser consultadas no Quadro 62.

Quadro 62 – Novas classes para as variáveis região, profissão e número de pessoas

	Classes originais		
Novas classes	**Região**	**Profissão**	**Número de pessoas**
1	1 e 2	1 e 2	1 e 2
2	3e 4	3 e 4	3 e 4
3	5 e 6	5 e 6	+ 46
4	7	7 e 8	_____
5	_____	9 e 10	_____

A aplicação do método da recodificação global resultou na anulação de todas as células inseguras, como se pode verificar no Quadro 63. O ficheiro depois da recodificação apresenta um risco individual de 4,2%.

Quadro 63 – Cruzamento das variáveis após recodificação global – Variável Região

Classes	Frequência	Cruzamento de duas variáveis	Cruzamento de três variáveis
1	299	0	0
2	306	0	0
3	295	0	0
4	167	0	0

6.4.5.2. *Microagregação numérica*

A microagregação foi aplicada à única variável contínua da base de dados familiares, a variável remunerações. Tal como no ficheiro da base de dados SABI, foram criadas novas classes, para as quais foram calculadas as médias, valor que é dado a cada célula da classe a que pertence. Após a aplicação do método foi realizada a análise descritiva das remunerações, conforme se segue:

Quadro 64 – Análise descritiva após a microagregação

	Remunerações	
	Original	Microagregação
Valores válidos	1 067	1 067
Missing	0	0
Média	490,41	490,91
Mediana	483,00	483,00
Desvio padrão	287,63	287,64

O Quadro 64 contém a análise descritiva da variável remunerações quer dos valores originais, quer dos valores microagregados. Pode-se concluir que não existem diferenças (apenas de cêntimos) na média, mediana e desvio padrão, o que significa que as alterações introduzidas nos dados originais não conduzem a uma ideia errada da realidade.

6.4.5.3. Arredondamento

Foi aplicado o método do arredondamento à base de dados familiares de forma a diminuir o número de células inseguras. A base de arredondamento utilizada neste ficheiro foi de 50 €, assim o valor das remunerações de cada registo foi está expresso em múltiplos de 50 €. Após o arredondamento foram calculados os novos valores para a média, mediana e desvio padrão, que podem ser consultados no quadro abaixo.

Quadro 65 – Análise descritiva após o arredondamento

	Remunerações	
	Original	Arredondado
Valores válidos	1 067	1 067
Missing	0	0
Média	490,41	491,85
Mediana	483,00	500,00
Desvio padrão	287,63	288,82

O método do arredondamento provocou ligeiras alterações nos valores da média, desvio padrão e ligeiramente maior na mediana, no entanto estas alterações não conduzem a uma análise errada do ficheiro de dados.

6.4.5.4. Rank Swapping

A variável original remunerações foi ordenada por ordem crescente, os seus valores foram trocados por outros dentro de um intervalo de 15% do total de linhas do ficheiro, o que corresponde a 160 linhas.

Por exemplo, o valor das remunerações do ficheiro original do registo 1 é de 881 € posicionada na linha 127 do ficheiro original ordenado, após o rank swapping, esse valor foi substituído por 997 € pertencente ao registo 99 e posicionado na linha 4.

O Quadro 66 contém a análise descritiva da variável remunerações, podendo-se verificar que os valores da média, mediana e desvio padrão não sofreram alterações relativamente aos valores originais.

Quadro 66 – Análise descritiva após o rank swapping

	Remunerações	
	Original	"Rank Swapping"
Valores válidos	1 067	1 067
Missing	0	0
Mean	490,41	490,92
Median	483,00	483,00
Std. Deviation	287,63	287,65

6.4.5.5. Microagregação Híbrida

Os dados híbridos aplicados a esta base de dados consiste na alteração dos dados originais através da média dos valores das classes da variável remunerações e da aplicação de um factor multiplicativo à variável original e à média da variável da sua classe.

Após a sua aplicação foi realizada a análise descritiva para calcular a média, mediana e desvio padrão da variável remunerações, no ficheiro de dados híbridos, na qual se verifica uma alteração mais significativa na mediana, como se pode constatar no Quadro 67.

Quadro 67 – Análise descritiva nos dados híbridos

	Remunerações	
	Original	Dados híbridos
Valores válidos	1 067	1 067
Missing	0	0
Média	490,41	490,92
Mediana	483,00	445,50
Desvio padrão	287,63	284,55

6.4.6. Análise global do ficheiro seguro

A protecção dos dados não termina com a aplicação dos métodos de Controlo da Divulgação Estatística, é necessário gerar ficheiros seguros após cada método. A criação de ficheiros seguros consiste em suprimir a totalidade ou parte das células inseguras para divulgação. Foram criados ficheiros seguros após a aplicação dos métodos de CDE à variável contínua.

É apresentado de seguida o Quadro 68 que indica quais e quantas células foram suprimidas.

Quadro 68 – Supressão de células

Variável	Supressões
Região	26
Profissão	0
Número de pessoas	0
Total	**26**

6.4.6.1. Conclusão da aplicação dos métodos de CDE nas variáveis contínuas

O objectivo dos métodos de Controlo da Divulgação Estatística é, quando possível, eliminar as células, caso contrário diminuir o seu número. A aplicação de qualquer um dos métodos às variáveis contínuas (remunerações) contribuiu para uma pequena redução do número de células inseguras, verificou-se uma redução de vinte e seis células.

De seguida são apresentados os resultados do cruzamento das variáveis região, profissão e número de pessoas em alojamentos após a aplicação de qualquer um dos métodos à variável remunerações.

O quadro abaixo identifica as 110 células inseguras, que são todas as células com valores iguais ou inferiores a 3, marcadas a sombreado.

Quadro 69 – Cruzamento das variáveis região x profissão x número de pessoas após os métodos CDE

Nº pessoas em alojamentos		Profissão										Total
		1	2	3	4	5	6	7	8	9	10	
1	Região .	0	1	0	1	2	1	2	1	0	1	9
	1	5	2	7	3	6	4	0	3	0	2	32
	2	5	0	2	2	0	0	2	4	4	3	22
	3	4	5	4	7	4	0	0	6	6	3	39
	4	2	4	2	3	5	2	5	2	3	0	28
	5	4	5	5	2	5	2	6	2	6	3	40
	6	4	4	3	0	0	6	2	7	7	8	41
	7	2	2	3	5	4	4	6	0	2	4	32
	Total	26	23	26	23	26	19	23	25	28	24	243
2	Região .	1	0	0	2	0	0	2	1	1	0	7
	1	3	4	10	5	5	6	4	0	5	2	44
	2	2	6	3	0	4	5	2	3	0	4	29
	3	4	4	5	3	3	2	5	6	3	9	44
	4	4	2	5	0	2	2	3	2	2	7	29
	5	4	2	5	5	2	7	0	2	6	7	40
	6	6	3	4	3	2	5	3	5	3	2	36
	7	0	8	6	7	6	3	0	6	6	3	45
	Total	24	29	38	25	24	30	19	25	26	34	274
3	Região .	2	0	1	1	0	1	0	0	0	0	5
	1	6	2	0	0	5	5	5	3	2	11	39
	2	5	7	2	2	3	0	5	3	11	3	41
	3	7	3	5	7	5	0	6	3	3	8	47
	4	2	3	7	3	3	2	6	3	3	2	34
	5	0	6	4	3	0	5	6	3	5	3	35
	6	4	2	5	5	5	4	3	2	4	2	36
	7	0	3	5	8	8	2	2	5	3	9	45
	Total	26	26	29	29	29	19	33	22	31	38	282
4	Região .	0	1	1	1	0	0	0	1	1	0	5
	1	7	2	2	4	3	2	5	4	3	5	37
	2	4	2	3	3	6	5	9	4	5	4	45
	3	4	3	4	0	3	3	14	6	0	8	45
	4	8	0	2	6	5	2	5	4	0	3	35
	5	4	4	3	0	3	4	2	0	2	2	24
	6	2	3	4	7	3	2	7	3	3	3	37
	7	4	4	0	3	4	8	6	2	4	5	40
	Total	33	19	19	24	27	26	48	24	18	30	268

Após a aplicação dos métodos CDE à variável remunerações, o número de células inseguras é 110, com este valor, o risco individual de identificação é de 13,1%, o que significa que o ficheiro alterado é seguro para divulgação.

6.4.7. Qualidade dos dados

A aplicação dos métodos CDE leva à criação de novos ficheiro de dados, uma vez que os métodos aplicados à única variável numérica em estudo, remunerações, originam o mesmo número de células inseguras, há que verificar qual deles tem a menor perda de informação.

À semelhança da base de dados SABI, não se avalia a qualidade dos dados no método da recodificação global, uma vez que se trata da aplicação a dados categóricos, não sendo comparável com as medidas aplicadas a dados contínuos.

As medidas utilizadas para verificar a qualidade dos dados foram: o Erro Quadrático Médio (EQM), o Erro Absoluto Médio (EAM) e a Variação Média (VM). Os valores determinados para estas medidas podem ser verificados no Quadro 70.

Quadro 70 – Medidas de qualidade dos dados

Medida de Qualidade	Microagregação Numérica	Rank Swapping	Arredondamento	Dados Híbridos
Erro Quadrático Médio	2,7689	6 949,13	213,2612	590,94
Erro Absoluto Médio	1,2902	69,99	12,7255	21,1148
Variação Média	0,0086	0,45	0,0734	0,1830

Após a análise ao Quadro 70 conclui-se que o método que origina menor perda de informação, independentemente da medida de qualidade dos dados utilizada, é o método da microagregação. O método com maior perda de dados é o rank swapping.

Capítulo 7. Conclusão

A confidencialidade de dados é um tema relativamente recente, talvez devido à reduzida procura de informação estatística das últimas décadas. Com a crescente procura de informação, o segredo estatístico começou a ser visto de outra forma. Actualmente a literatura existente sobre a confidencialidade e sobre os métodos de controlo da divulgação estatística é mais diversificada.

Dada a existência de vários métodos de controlo da divulgação estatística para dados tabulares e microdados, sentiu-se a necessidade de criar um software especializado para auxiliar os responsáveis pela produção de dados seguros. O software Argus foi desenvolvido no âmbito do projecto CASC (Computational Aspects of Statistical Confidentiality). O Argus é um software utilizado pelos serviços de estatística para produzir microdados e macrodados seguros. No entanto é um software que ainda não está muito divulgado e que apresenta algumas limitações, nomeadamente no que respeita à sua integração com outros programas (Excel, SAS, etc).

O estudo realizado nesta dissertação incide sobre a aplicação de algumas técnicas de Controlo da Divulgação Estatística disponíveis neste software, nomeadamente a microagregação, codificação superior, arredondamento e rank swapping e a sua comparação. Com a utilização deste software neste trabalho pretendeu-se aprofundar o conhecimento de algumas técnicas menos utilizadas na protecção dos dados. Também foi utilizada uma nova técnica, a microagregação híbrida.

Este estudo abordou temas como a importância do segredo estatístico e a sua protecção, o quadro jurídico e histórico do segredo estatístico, os diferentes métodos de controlo da divulgação estatística, a perda de informação, a qualidade dos dados e finalmente dois casos práticos para aplicação e comparação dos métodos de controlo da divulgação estatística.

Foram analisadas duas bases de dados distintas e representativas do tipo de ficheiros utilizados pelos institutos de estatística: uma com dados provenientes de empresas e

outra com dados de famílias. Qualquer um dos ficheiros utilizado tem um número reduzido de variáveis e de registos para facilitar o estudo e manuseamento das bases de dados.

Foi calculado risco individual para cada um dos ficheiros de dados originais e comparado com o risco individual definido no inicio do trabalho, em que se considerou um risco individual máximo aceitável de 20%. Este valor está muito acima do normalmente utilizado pelo Eurostat. No entanto os ficheiros de dados utilizados nos serviços de estatística são de dimensão não comparável com as que aqui foram utilizadas, pois são bases com milhões de registos.

Na avaliação do risco individual constatou-se que nenhum dos ficheiros era seguro: a base de dados SABI tinha um risco individual de 35,4% e a base de dados familiares, um risco de 26,7%. Após a aplicação das técnicas de controlo da divulgação estatística o risco individual diminui substancialmente para 17,1% nos dados SABI e para 4,2% (se utilizar a recodificação global) e 13,1% (restantes métodos) nos dados familiares.

Relativamente à base de dados SABI, qualquer que seja a técnica utilizada, microagregação, codificação superior, arredondamento, rank swapping e dados híbridos o risco individual de divulgação é de 17,1% e o número de células inseguras também é o mesmo: 29. O número de células inseguras após a aplicação dos métodos é o mesmo porque a criação de ficheiros seguros é baseado na supressão de células, ora as células inseguras suprimidas em qualquer um deles são as mesma e, neste caso, foram 51 células. Uma vez aplicados os métodos foram calculadas as medidas de qualidade dos dados. Recorreu-se a três medidas: o Erro Quadrático Médio, o Erro Absoluto Médio e a Variação Média. Utilizando qualquer uma das medidas o método que conduziu à menor perda de informação foi o método do arredondamento, uma vez que é o que apresenta menor valor.

De referir que os valores das variáveis são valores muito elevados, o que significa que sendo a base de arredondamento de 10 000 €, é um arredondamento não muito significativo dada a grandeza dos valores das variáveis, o que implica dizer que não

causa grandes alterações nos dados originais. É importante referir que a aplicação de dados híbridos, não sendo muito utilizado habitualmente pelos institutos de estatística, foi o terceiro melhor método, logo atrás da codificação superior, apresentando melhor resultado do que a microagregação, que é um método frequentemente utilizado no controlo do segredo estatístico.

A base dados familiares apresentou resultados diferentes da base anterior. Por um lado, a aplicação da recodificação global às variáveis categóricas eliminou todas as células inseguras, resultando num risco individual de apenas 4,2%. Este seria de longe o melhor método para aplicar a este ficheiro de dados. Relativamente aos ficheiros criados após a aplicação dos métodos de CDE à variável contínua, apresentam um risco individual de 13,1% e apenas foram suprimidas vinte e seis células.

Tomando por base as medidas de qualidade dos dados referidas para a base de dados SABI, conclui-se que se opção recaísse sobre os métodos utilizados na variável contínua, o que apresenta melhor resultado é a microagregação, seguida do arredondamento e mais uma vez, o terceiro melhor método são os dados híbridos.

Não obstante a importância que se deve dar às técnicas com melhores resultados, deve ser dado maior ênfase a outras com piores resultados e melhorá-las no sentido de serem também elas uma boa opção para criar ficheiros de dados seguros para divulgação, nomeadamente os dados híbridos que não são utilizados com muita frequência. Uma vez que eles são gerados com base nos dados originais, na média dos dados originais por classes e num factor de adição, à partida parece ser o método com maior possibilidade de ser aprofundado e melhorado. É esse o trabalho que nos propomos desenvolver no futuro, através de uma combinação de outros métodos de dados híbridos.

Referências

Banks, D. L., Karr, A.F. e Sanil, A.P, 2005. Data Quality – A Statistical Perspective. *NISS, Technical Report Number 151.*

Commission Regulation (EC) No 831/2002 of 17 May 2002 implementing Council Regulation (EC) No 322/97 on Community Statistics, concerning access to confidential data for scientific purposes, Official Journal of the European Communities.

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

Council Regulation (EC) No 322/97 of 17 February 1997 on Community statistics, Official Journal of the European Union, No L 52, 22.2.1997, p. 1 - 7

Dandekar, R. A., Domingo-Ferrer, J. e Sebé, F., 2002. LHS – Based Hybrid Microdata Vs Rank Swapping and Microaggregation for Numeric Microdata Protection. *Inference Control in Statistical Databases, LNCS 2316, pp 153-162*.

Domingo-Ferrer, J. e Mateo-Sanz, J.M., 1999. On Resamplig for Statistical Confidentiality in Contingency Tables. *Computers & Mathematics with Applications, 38, pp.13-32.*

Domingo-Ferrer, J. e Torra, V., 2001. Disclosure Control Methods and Information Loss for Microdata. *Cap. 5, pp 91-110 of*: Doyle, P., Lane, J.I., Theeuwes, J.J.M. e Zayatz, L.V.(eds), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. (Elsevier).* Amesterdam.

Domingo-Ferrer, J. e Torra, V., 2005. Ordinal, Continuous and Heterogeneous k-Anonymity through Microaggregation. *Data Mining and Knowledge Discovery, 11, pp.195–212.*

Domingo-Ferrer, J. e Sebé, F., 2006. Optimal Multivariate 2 – Micraggregation for Microdata Protection: A 2 – Approximation. *Pp 129-138 of*: Domingo-Ferrer, J. e Franconi, L. (eds), *Privacy in Statistical Database. LNCS 4302. (Springer).* Rome

Duncan, G. T., Fienberg, S. E. e Krishnan, R. Padman, R. e Roehrig, S. F., 2001. Disclosure Limitation Methods and Information Loss for Tabular Data. *Cap. 2, pp 135-166 of*: Doyle, P., Lane, J.I., Theeuwes, J.J.M. e Zayatz, L.V.(eds),

Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. (Elsevier). Amesterdam.

Eurostat, 2005, European Statistics Code of Practice 2005

Hansen, S.L. e Mukherjee, S., 2003. A Polynomial Algorithm for Optimal Univariate Microaggregation. *IEEE Transations on Knowledge and Data Engineering, Vol. 15, no. 4*.

ISI - International Statistical Institute, 1985, Declaration on Professional Ethics

Lambert, D. 1993. Measure of Disclosure Risk and Harm. *Journal of Official Statistics, Vol. 9, nº 2, pp 313-331.*

Lane, J. e Kennickell, A., 2006. Measuring the Impact of Data Protection Techniques on Data Utility: Evidence from the Survey of Consumer Finances. *Privacy in Statistical Database, LNCS 4302, pp 291-303.*

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

Liew, C. K, Choi, U.J. e Liew, C.J., 1985. A Data Distortion by Probability Distribution. *ACM Transactions on Databases Systems, vol. 10, Nº 3, pp. 395-411.*

Hundepool, A., Ramaswamy, R. e Wetering, A. V., Franconi, L, Polettini, S., Capobianchi, A., Wolf, P.P., Domingo-Ferrer, J., Torra, V., Brand, R. e Giessing, S., 2008. *μ-Argus User's Manual. Version 4.2*, (*ESS-Net project).* De Hague.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Nordholt, E.S., Seri, G. e Wolf, P.P. 2009. *Handbook on Statistical Disclosure Control. Version 1.1(ESSNet SDC)*

U.S. Census Bureau, 2006. Code Protection of Confidential Information.

Waal, t. e Willenborg, L., 1996. *Statistical Disclosure Control in Practice*, Lecture Notes in Statistics, 111. *(Springer).*

Waal, t. e Willenborg, L., 2001. *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, 155. *(Springer).*

Winkler, W.E., 2005. Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata. *Research Report Series. U.S. Census Bureau.* Washington.

Wolf, P.P., Gouweleeuw, J. M., Kooiman e P., Willenborg, L., 1998, Reflections on PRAM. Proceedings of the conference: *Statistical Data Protection.* Lisboa.

Wolf, P.P., 2006, Risk, Utility and PRAM. *Pp 189-204 of*: Domingo-Ferrer, J. e Franconi, L. (eds), *Privacy in Statistical Database. LNCS 4302. (Springer).* Rome

ANEXO

Anexo 1 – Conceitos

- ***Microdados –*** Conjunto de registos que contem informação de respondentes individuais ou de entidades económicas.
- ***Macrodados ou dados tabulares -*** Informação agregada de entidades representada em forma de tabelas.
- ***Chave –*** É uma combinação de variáveis identificadoras que identificam inequivocamente o individuo, como por exemplo o nome, o número de identificação fiscal, número do passaporte.
- ***Variáveis sensíveis –*** São variáveis em que pelo menos um dos seus valores é sensível e para as quais o protector de dados deve ser mais rigoroso na sua protecção, nomeadamente o comportamento sexual, o passado criminal.
- ***Variáveis Confidenciais -*** são variáveis que contêm informação sensível sobre o entrevistado, como o salário; religião; filiação política; estado de saúde, etc.
- ***Variáveis Identificadoras –*** São variáveis que identificam inequivocamente o individuo, como por exemplo o nome, o número de identificação fiscal, número do passaporte.
- ***Variáveis Identificadoras Indirectas –*** possibilitam deduzir as unidades estatísticas a partir de informação que não conste das variáveis identificadoras directas.
- **Variáveis qualitativas** - São as variáveis cujos seus valores, categorias, modalidades, não são números reais, às quais podem ser atribuídos códigos numéricos.
- ***Variáveis ordinais -*** apenas podem ser distinguidos diferentes graus de um atributo ou variável, existindo portanto entre eles uma relação de ordem, os valores são ordenados.
- ***Variáveis Nominais -*** se os valores não são ordenados. Geralmente estas variáveis são codificadas de 1 a m, sendo m o nº de modalidades ou categorias. Os elementos são atributos ou qualidades.
- ***Variáveis binárias*** (caso particular) - variáveis qualitativas com apenas duas modalidades, que normalmente são codificadas por 0-1 ou 1-2.
- ***Variáveis quantitativas –*** São variáveis cujo os valores são reais. Podem ser:
- ***Variáveis de escalas de intervalo*** - o uso de números para classificar os elementos é feito de forma que, a igual diferença entre os números, corresponde

à igual diferença nas quantidades do atributo medido. O zero é um valor arbitrário e não representa a ausência da característica medida.

- ***Escala rácio*** - difere de uma escala de intervalo, porque *o zero tem existência real, denotando ausência da característica medida*. Nesta escala apenas um número pode ser atribuído arbitrariamente, caso das unidades de medida ou distância, ficando os restantes completamente determinados.
- ***Discretas:*** se o conjunto de valores é finito ou infinito numerável.
- ***Contínuas:*** se o conjunto de valores é infinito não numerável.

- ***Valor limite –*** Valor abaixo do qual um registo é considerado inseguro para divulgação.
- ***Dados anonimizados –*** São dados modificados de forma a minimizar o risco de divulgação.

- ***Tabelas -*** Falando abstractamente, uma tabela consiste num conjunto de células, em que cada célula é caracterizada por um conjunto de coordenadas, de combinações de resultados de algumas variáveis categóricas.

- ***Arquivo ou ficheiro de identificação –*** É um conjunto de microdados que contem identificadores.

- ***Dados seguros –*** São dados individuais ou agregados protegidos através de métodos de controlo da divulgação de dados estatísticos.
- ***Risco de identificação*** – Probabilidade de um intruso identificar pelo menos um entrevistado nos microdados disponibilizados.