

Utilização de bases de microdados na investigação em ciências sociais

Autores:

Rosário Mauritti

Susana da Cruz Martins

Ana Simões Antunes

António Firmino da Costa

VOLUME II

2º QUADRIMESTRE DE 2002

UTILIZAÇÃO DE BASES DE MICRODADOS NA INVESTIGAÇÃO EM CIÊNCIAS SOCIAIS

THE UTILIZATION OF MICRO-DATABASES IN SOCIAL SCIENCES RESEARCH

Autores: Rosário Mauritti

- Investigadora do Centro de Investigação e Estudos de Sociologia do Instituto Superior de Ciências do Trabalho e da Empresa (CIES/ISCTE).

Susana da Cruz Martins

- Investigadora do CIES/ISCTE.

Ana Simões Antunes

- Departamento de Estatísticas Sociais, Coordenadora do Núcleo de Estatísticas do Emprego do Serviço de Estatísticas do Trabalho, Instituto Nacional de Estatística.

António Firmino da Costa

- Professor Auxiliar do ISCTE. Presidente da Direcção e Coordenador Científico do CIES/ISCTE.

RESUMO:

- Este artigo procura analisar as possibilidades e limites da utilização de microdados estatísticos produzidos pelo INE na investigação em ciências sociais. Evocam-se a este propósito questões centrais que se colocam no uso desta informação, tendo em consideração, por um lado, o princípio fundamental de salvaguarda do segredo estatístico e, por outro, as potencialidades inscritas no acesso às bases de microdados. Reflete-se assim sobre os procedimentos de anonimização, as condições de disponibilização da metainformação, a validação e o controlo de qualidade dos dados, designadamente no que respeita à harmonização e à comparabilidade, e sobre a concertação de critérios de divulgação da informação.

PALAVRAS-CHAVE:

- *Microdados, anonimização, metainformação, harmonização, comparabilidade e divulgação dos dados.*

ABSTRACT:

- The present article analyses the possibilities and limitations of the utilization of statistical microdata (produced by the Portuguese National Statistics Institute) in social sciences research. In order to do this, we aim to consider the fundamental principle of statistic privacy, on one hand, and the possibilities to research offered by the access to micro-databases, on the other. Thus, we reflect on the anonimiza-

tion procedures, on the conditions of reception of metadata, on the validation and control of data, namely harmonization and comparability, and on the criteria of information's diffusion.

KEY-WORDS:

- *Micro-data, anonimization, meta-information, harmonization, comparability and diffusion of data.*

1. ACESSO A BASES DE MICRODADOS ESTATÍSTICOS: POTENCIALIDADES E PROBLEMAS

A utilização de dados estatísticos produzidos institucionalmente é uma das peças-chave da investigação em ciências sociais. Foi assim desde os primeiros passos do desenvolvimento sistemático destas ciências, ainda no século XIX.

Mas, hoje, o recurso a indicadores estatísticos tornou-se ainda muito mais importante, nomeadamente em áreas científicas como a sociologia, a economia, a demografia e a geografia. Isso acontece, aliás, tanto na investigação fundamental, onde tem permitido análises decisivas para o avanço das ciências sociais, como na investigação aplicada, nomeadamente na que tem visado apoiar processos de planeamento, de decisão e de avaliação, em diversos domínios das políticas públicas.

É bem conhecido nas ciências sociais que nem só a informação estatística – nem sequer, de um modo mais geral, só a informação empírica quantitativa – é susceptível de apoiar a produção de conhecimento substantivo sobre as relações sociais, as suas configurações e as suas dinâmicas. Também a informação qualitativa decorrente de observação controlada, nas variadas formas em que as ciências sociais a utilizam, proporciona base empírica relevante para esse conhecimento.

Reconhecê-lo não equivale, porém, a qualquer subalternização dos dados quantitativos na caracterização da realidade social, nem desse género específico de dados quantitativos que são os indicadores estatísticos institucionais.

Por vezes, alguns investigadores têm contraposto, em regime de exclusão recíproca, a pertinência da informação quantitativa ou da informação qualitativa para a produção de enunciados cognitivos válidos e significativos sobre a realidade social.

No entanto, pelo menos na sociologia, as posições mais consistentes analiticamente, e mais cognitivamente produtivas, têm adoptado uma perspectiva de pluralismo metodológico, para a qual as diferenças entre informação empírica de carácter quantitativo e qualitativo não são de essência mas de procedimento. O diálogo reflectido entre métodos e a integração judiciosa de fontes diversas de informação empírica tem-se revelado crucial para o avanço das ciências sociais.

Como referido, a informação estatística institucionalmente produzida, se sempre foi importante para a investigação em ciências sociais, adquiriu hoje uma presença muito maior e mais diversificada. E, do mesmo passo, adquiriu também níveis de exigência muito mais elevados. Isso aconteceu em correspondência, por um lado, com o desenvolvimento interno destas ciências, e, por outro, com os desenvolvimentos da própria sociedade.

Deste ponto de vista, multiplicou-se exponencialmente o pedido social de informação estatística, sobre os mais diversos aspectos. Tipicamente, essa procura visa a produção de caracterizações e a análise de distribuições, relações e evoluções, respeitantes a indivíduos e grupos, a entidades e agregados, a atributos e actividades, a comportamentos e opiniões.

Mas para além disso, a informação estatística passou, ela própria, a fazer parte constitutiva importante de vários processos sociais. É um dos aspectos em que se manifesta a reflexividade cognitiva e institucional das sociedades contemporâneas, hoje em dia objecto de análise incontornável das ciências sociais, e tematizada publicamente sob expressões como “sociedade da informação” ou “sociedade do conhecimento”, associada a processos como os de “globalização” ou de “inovação tecnológica”, de “regulação” ou de “governança”.

Com efeito, a presença social da informação estatística institucional verifica-se, hoje, de forma intensiva, no plano de sistemas como o político-administrativo, o económico-empresarial, o educacional-formativo ou o comunicacional-mediático. Mas repercute-se, também, no plano dos modos de vida quotidiana, no dos processos de acção colectiva e no dos direitos, liberdades e garantias.

A informação estatística surge, pois, na intersecção de duas vertentes decisivas da constituição das sociedades contemporâneas:

- mecanismos importantíssimos de reflexividade cognitiva e instrumentalidade social;

mas, igualmente,

- preocupações cívicas com os direitos e as liberdades, nomeadamente no que concerne à defesa da privacidade e da autonomia individuais.

Do ponto de vista dos desenvolvimentos das ciências sociais, a evolução do seu objecto de estudo, isto é, da sociedade, com as tendências apontadas, bem como a crescente informação estatística institucionalmente produzida e, ainda, a disponibilidade de novas tecnologias de armazenagem e tratamento de informação, incomparavelmente mais poderosas do que há alguns anos atrás, confluem com a recolocação dos objectivos de investigação e dos procedimentos de análise em patamares muito mais exigentes.

As análises convencionais de dados agregados, habitualmente fornecidas pelas instituições oficiais de produção de informação estatística, sem que tenham perdido utilidade, tornaram-se claramente insuficientes.

Muitas das interrogações analíticas pertinentes nas ciências sociais actuais, e certamente alguns dos problemas de investigação mais interessantes e que mais podem fazer avançar o conhecimento científico sobre as configurações de sociedade e as dinâmicas sociais contemporâneas, requerem hoje, para que se possa investigar com vista à respectiva elucidação, a utilização de bases de microdados estatísticos institucionalmente produzidos.

O trabalho de investigação conduzido com recurso directo aos microdados possibilita:

- *análises de dados mais finas;*
- a construção de *variáveis compostas de tipos diversificados*, mais directamente ajustadas a parâmetros teóricos e temas de investigação;

- a realização de *análises multivariadas*, estabelecendo relações entre um grande número de indicadores;
- um modo de *relação experimental com as bases de dados*, permitindo um número alargado de tentativas sucessivas de análise, conduzidas segundo um processo de interação entre parâmetros teóricos, ensaios de operacionalização e tentativas de interpretação de resultados.

Em simultâneo, as ciências sociais não podem deixar de se colocar na primeira linha das preocupações cívicas e éticas com a defesa dos direitos, as liberdades e as garantias dos cidadãos, tal como se configuram numa sociedade democrática contemporânea.

O conhecimento aprofundado que têm dos processos sociais atribui-lhes responsabilidades acrescidas, mais ainda quando está em causa algo que se situa no seu próprio âmago, isto é, os dispositivos de recolha de informação sistemática e produção de conhecimento analítico sobre a realidade social. Os investigadores destas áreas, não só têm obrigação de respeitar as disposições legalmente estabelecidas a este propósito, como estão provavelmente melhor colocados do que muitos outros para detectar riscos e sugerir soluções aperfeiçoadas a este respeito.

É também neste sentido que ganha particular importância o carácter institucional da informação estatística aqui em causa. O Sistema Estatístico Nacional e, em particular, o Instituto Nacional de Estatística (INE) têm aqui uma responsabilidade particular quanto à possibilidade de supervisão técnica e responsabilização pública, sob tutela com legitimidade democrática, destes processos de recolha e disponibilização de informação estatística fina, sistemática e abrangente.

2. FONTES DE INFORMAÇÃO ESTATÍSTICA: NOVAS POSSIBILIDADES DO TRATAMENTO DE MICRODADOS

Está em desenvolvimento uma nova fase de interligação entre os produtores institucionais de informação estatística (em especial, o INE) e a comunidade científica, nomeadamente as equipas de investigadores em ciências sociais.

Neste contexto, procura-se a implementação de um novo quadro institucional que possibilite a divulgação regulamentada de microdados estatísticos. O protocolo entre o Ministério da Ciência e da Tecnologia (MCT) e o Instituto Nacional de Estatística, em 1999, estabelece esse novo quadro.

Antes de passar à análise detalhada do protocolo, importa ainda explicitar alguns dos vectores que poderão contribuir para o desenvolvimento de uma prática testada, consolidada e produtiva desses novos moldes de interligação. Uma forma de o

concretizar será certamente através da implementação de projectos de investigação científica.¹¹ Estes poderão contribuir para isso em termos de:

- *procedimentos técnicos* (suportes tecnológicos e processamento informático das bases de dados);
- *metainformação estatística* (sistematização, actualização e aperfeiçoamento permanentes, susceptíveis de serem conseguidos através da respectiva utilização na investigação);
- *processos de anonimização* (soluções exigentes quanto a garantias de anonimato e confidencialidade e, simultaneamente, viabilizadoras de aprofundamentos analíticos);
- *controlo dos dados e resultados a divulgar* (quanto a aspectos como validade, representatividade, fiabilidade e significância);
- *comparabilidade* (entre indicadores nacionais e internacionais e entre fontes estatísticas);
- *aperfeiçoamento* das operações estatísticas e aproveitamento analítico das suas potencialidades (testes e sugestões a partir das experiências de investigação);
- *processos de trabalho* (entre técnicos do INE e equipas de investigadores em ciências sociais).

A possibilidade de tratamento de microdados recolhidos e processados para fins estatísticos por uma instituição pública como o INE, constitui uma novidade de grande importância para o desenvolvimento das actividades de investigação em ciências sociais. Quais as razões que fundamentam a disponibilização destes dados para fins de investigação científica?

No quadro convencional tradicional, quando pretendem desenvolver uma investigação com apoio de estatísticas produzidas institucionalmente, os investigadores têm ao seu dispor:

- publicações e relatórios disponíveis ao público em geral;
- informação divulgada em linha na Internet;
- a possibilidade de solicitar dados específicos, não publicados, cedidos em suporte papel ou magnético já sob a forma de *output*.

¹¹ Enquadra-se neste âmbito o projecto de investigação *Padrões de vida: perfis e tendências na sociedade portuguesa contemporânea (Desenvolvimentos conceptuais e exploração analítica de microdados estatísticos)*, desenvolvido pelos autores do presente artigo. Trata-se de um projecto em curso no Centro de Investigação e Estudos de Sociologia (CIES), centro associado do Instituto Superior de Ciências do Trabalho e da Empresa (ISCTE), apoiado pela Fundação para a Ciência e Tecnologia (FCT), do Ministério da Ciência e da Tecnologia (MCT). A contribuição do INE para este projecto é também fundamental na constituição da equipa e na disponibilização das bases de microdados estatísticos.

As estatísticas divulgadas nestes três modelos são fornecidas, geralmente, sob a forma de tabelas de frequências ou cruzamentos simples de indicadores agregados. Indicadores esses construídos, muitas vezes, a partir de procedimentos estandardizados, os quais, sendo comuns aos diversos organismos nacionais e internacionais, apresentam a vantagem de permitir a comparação dos dados.

Não obstante, esta informação é, muitas vezes, insuficiente para o desenvolvimento das análises que os investigadores se propõem realizar. Nestes três modelos de disponibilização de dados é muito difícil, ou mesmo impossível, o desenvolvimento de um conjunto de operações muito importantes na actual investigação em ciências sociais. Em concreto, não é em geral possível proceder ao ensaio de categorizações e agregações alternativas, à construção e teste de novas variáveis derivadas, à realização de análises multivariadas de diversos tipos. Tudo isso precisando de ser conduzido, por um lado, em função de orientações teórico-conceptuais muito complexas ou específicas e, por outro, em regime de vai-vem experimental entre sucessivas variantes operatórias e análises dos respectivos resultados.

Em Portugal, os condicionalismos legais respeitantes à necessidade de preservação do segredo estatístico impõem restrições rigorosas na disponibilização da informação sobre a forma de microdados (CSE, 1998). Isto, mesmo quando as pretensões de manuseamento dessa informação se circunscrevem à investigação científica. O acesso a este tipo de informação é uma possibilidade apenas disponível com a celebração do referido *Protocolo entre o Ministério da Ciência e da Tecnologia e o Instituto Nacional de Estatística*.

Neste protocolo reconhece-se a necessidade de disponibilizar mais e melhor informação para uso na investigação científica, ao mesmo tempo que se regulamenta a sua utilização. É prestada especial atenção à necessidade de protecção dos direitos individuais, sobretudo no que diz respeito ao anonimato e privacidade. Neste sentido, o protocolo sugere que o INE, a instituição detentora dos dados, proceda à anonimização prévia de todas as bases de dados com informação de natureza individual. Cabe, pelo seu lado, ao MCT a credenciação das instituições de investigação científica e desenvolvimento tecnológico, cujos investigadores pretendam aceder a este tipo de informação. Adicionalmente, há uma responsabilização individualizada dos investigadores que irão utilizar a informação. A cada um é exigida a assinatura de um termo de responsabilidade que o obriga a respeitar um conjunto de regras, entre as quais se destacam:

- utilizar os dados estatísticos apenas para os fins da investigação científica especificados num resumo do projecto previamente apresentado;
- não procurar identificar indivíduos ou instituições envolvidos nas bases de dados;
- não comercializar ou ceder os dados a terceiros;
- referenciar o INE em qualquer publicação ou comunicação baseada nos dados fornecidos;
- informar o INE de quaisquer problemas de natureza técnica, metodológica, de registo de dados ou de erros detectados nas bases de dados utilizadas.

3. MODELOS DE ANONIMIZAÇÃO

A obrigação de anonimização de bases de dados colectados por organismos públicos está inscrita nos regulamentos do Conselho Europeu (CE) relativos às estatísticas comunitárias. Como salienta o regulamento do CE n. 322/97, “qualquer utilização dessa informação deve assegurar a impossibilidade prática de identificação das unidades estatísticas de registo envolvidas”, como sejam indivíduos, famílias, empresas, etc. Reconhece-se, assim, a necessidade de anonimização, ou seja, a necessidade de “anulação de todos os meios que possam razoavelmente ser utilizados por uma terceira parte para identificar a dita unidade estatística”.

Existem presentemente diversos métodos disponíveis para prevenir o risco de identificação de uma unidade de registo que conste na base de microdados (Dale, 2001). Importa sublinhar, antes de mais, que os procedimentos de anonimização não são independentes das características dos microdados em causa. Por exemplo, as soluções de anonimização podem ser, em parte, diferentes consoante se trate de dados censitários ou de dados recolhidos através de processos de amostragem. Essas diferenças podem ter a ver também com outros aspectos, designadamente com o tipo de unidades de registo – se se trata de empresas ou de edifícios, de agregados familiares ou de indivíduos ou de outras entidades.

Quando estão em causa universos de grande dimensão e dados recolhidos por amostragem, a eliminação das variáveis de identificação das unidades de registo constitui com frequência, só por si, um procedimento eficaz de anonimização. Ainda assim, pode ser conveniente activar outros procedimentos de anonimização e, por maioria de razão, quando o número de casos é reduzido para certas categorias ou combinações de categorias.

Os modelos de anonimização, enquanto componentes de estratégias visando a disponibilização de bases de microdados para a investigação científica, podem ser agrupados como segue.

A) A CONSTRUÇÃO DE SAFE VARIABLES: AGREGAÇÃO DE CATEGORIAS

A construção de *safe variables* constitui uma das metodologias mais utilizadas. Esta consiste na selecção de variáveis contidas nas bases de dados originais – as variáveis que poderiam conduzir à identificação de unidades estatísticas –, as quais são sujeitas a um processo de *anonimização por agregação de categorias*. Este método pode ser utilizado quando não é possível assegurar a confidencialidade dos dados originais, por exemplo, devido ao nível de detalhe de uma determinada variável (unidade geográfica, profissional, etária, etc.).

A operacionalização destes procedimentos de anonimização permite soluções diversas consoante o tipo de utilização que se pretende dar aos dados. Na selecção das variáveis a agregar procura-se ter em conta as perspectivas de análise e os temas específicos de investigação. Ou seja, se, por exemplo, em algumas pesquisas pode ser vantajoso construir a anonimização com base numa forte agregação do âmbito geográfico, permitindo maiores especificações de variáveis educacionais ou profissionais,

por exemplo, noutras, onde os indicadores de caracterização espacial fina correspondam a vectores-chave da investigação, o balanço mais útil pode ir no sentido inverso, passando pela agregação de outros indicadores, nomeadamente níveis etários, grupos profissionais, etc.

B) A CONSTRUÇÃO DE SAFE SETTINGS: AGREGAÇÃO DE CASOS

Uma alternativa ao procedimento anterior é *agregar*, não as categorias de variáveis, mas *os casos em subconjuntos*, atribuindo a cada caso o valor de caracterização do subconjunto.

O acesso em linha aos microdados apoiado por um *safe setting* foi desenvolvido, designadamente no Luxemburgo, no âmbito do CEPS/INSTEAD.¹² Neste modelo, os investigadores munidos de uma *password*, acedem, também, a uma subamostra representativa da base de sondagem. As variáveis dos vários inquéritos são previamente harmonizadas para permitir a comparabilidade, o que constitui por si só um grande benefício para os investigadores (Förster, Helliesen e Kolberg, 1996; Christos, 2000; Dale, 2001).

No entanto, este método de trabalho, comparativamente à metodologia anterior, apresenta algumas desvantagens. Entre elas, o facto de o processo de agregação ser mais complexo e introduzir maior indeterminação na análise. Para além disso o nível de interacção com os dados é limitado, aumentando a barreira à construção e teste de novas variáveis.

C) A CONSTRUÇÃO DE SUBAMOSTRAS DAS BASES DE DADOS ORIGINAIS

A construção de subamostras dos dados originais, como forma de facilitar o acesso aos microdados em formato digital, é utilizada, por exemplo no Reino Unido, desde 1993, para as *estatísticas censitárias* (Dale, 2001). Parte-se do princípio de que, com a construção destas subamostras, fica assegurado o anonimato, com a impossibilidade de identificação dos dados. A informação assim disponibilizada é acessível tanto para fins de investigação como para o ensino. Cada universidade foi convidada a nomear um membro que assume a responsabilidade pelo uso dos dados no quadro da respectiva instituição. Adicionalmente, professores ou investigadores que queiram utilizar a informação estatística assinam um compromisso que os obriga pessoalmente a manterem a segurança dos dados.

D) A DISPONIBILIZAÇÃO DE MICRODADOS EM CENTROS AUTORIZADOS

Um outro modelo alternativo, também disponível, por exemplo no, já referido, CEPS/INSTEAD, conjuga o *safe setting* com a disponibilização das bases de dados em centros autorizados onde os investigadores podem permanecer durante determinado período para o desenvolvimento de análises específicas. O *safe setting* assume, neste caso, a forma de um cálculo laboratorial seguro, onde é também rigorosamente impedida a alteração dos registos originais. Este método faculta a oportunidade de tratamento de dados de todos os casos colectados e não apenas de uma subamostra, como acontecia no caso anterior. A implementação deste processo de anonimização é,

¹² Site: <http://lisweb.ceps.lu/lis>

no entanto, muito dispendiosa, implicando a existência de uma equipa permanente que dê apoio e supervisione os investigadores. Além disso, custos e inconvenientes são também impostos aos investigadores, que têm de se deslocar para trabalharem com todos os dados, podendo implicar que tenham de aí permanecer algum tempo até concluírem a pesquisa.

E) *O ACESSO INDIRECTO AOS DADOS*

Outro processo que não envolve microdados, mas que em muitos casos resolve as necessidades dos utilizadores, consiste no acesso indirecto aos dados através da elaboração de pedidos específicos, dirigidos à instituição produtora/gestora das bases de dados, de acordo com as análises que se pretendem conduzir. Este sistema permite solicitar frequências ou cruzamentos entre indicadores. Porém, não permite, por exemplo, a criação de novas variáveis ou a elaboração de testes estatísticos mais sofisticados, designadamente de estatística multivariada, nem possibilita experimentação indirecta de sucessivos tratamentos alternativos, o que o torna menos atractivo para alguns tipos de investigação.

F) *O ACESSO ATRAVÉS DA INTERNET*

Uma situação intermédia entre a referida anteriormente e os vários modelos de acesso directo aos microdados, consiste na disponibilização de bases de dados standardizadas na Internet. Este modelo possibilita a realização de apuramentos simples de frequências, a selecção de categorias de variáveis e/ou segmentos da população, bem como, pedidos de cruzamentos com duas ou mais variáveis. Para além de facultar a experimentação das bases de dados, tem a vantagem de integrar na informação disponibilizada todo um conjunto de variáveis derivadas de uso comum (como a condição perante o trabalho, diversas taxas de emprego, etc.), cuja construção implicaria o domínio de algoritmos específicos. A nível nacional, esta poderá ser a orientação dominante na disponibilização de informação estatística por parte das instituições produtoras de microdados.

4. MICRODADOS ESTATÍSTICOS E META-INFORMAÇÃO

No uso de microdados estatísticos, outra questão relevante, também ela respeitante às condições de disponibilização deste tipo de informação, tem a ver com a meta-informação associada a cada projecto do INE (veja-se a este respeito Fonseca, Valente e Lima, 1995). Esta refere-se à informação conceptual e metodológica sobre cada base de dados e sobre as operações estatísticas que a ela conduziram.

A utilização informada de bases de microdados estatísticos requer, nomeadamente, informação especificada sobre:

- a *metodologia do inquérito* (método de recolha, questionário, características da amostra, unidades estatísticas, método de tratamento de não-respostas, nomenclaturas);

- as *metodologias de estimação* (cálculo dos ponderadores e sua utilização); estas podem assentar em perspectivas diversas, atendendo a objectivos diferenciados: uma, como forma de corrigir pesos relativos para trabalhar a amostra com probabilidades corrigidas; outra, é a sua utilização como critério de extrapolação, quando se quer trabalhar com valores absolutos;
- as *variáveis e seus conceitos* (descritivos das variáveis, incluindo as variáveis derivadas, tabelas de codificação e condições de resposta ou filtros de selecção de casos);
- a *informação técnica* relativa às bases de dados (desenho de registo e *software* utilizado);
- os critérios de *interpretação e divulgação* (metodologia de cálculo dos erros de amostragem associados aos resultados e sua utilização).

5. VALIDAÇÃO E CONTROLO DA QUALIDADE: CONCEPÇÕES E PROCEDIMENTOS PARA A UTILIZAÇÃO DOS MICRODADOS

A utilização adequada de microdados não pode deixar de ser acompanhada de uma garantia de qualidade e validação da informação produzida. Neste sentido, torna-se crucial que se conheçam as metodologias e procedimentos técnicos de recolha e tratamento da informação, dos quais se podem destacar:

- em primeiro lugar, os *métodos de amostragem* e possibilidades de representatividade das amostras construídas;
- em segundo lugar, as *formas de cálculo dos ponderadores*;
- em terceiro lugar, os procedimentos de *validação* sistemática de recolha, tratamento e análise dos dados.

Em relação aos *métodos de amostragem*, o controlo dos processos, com a consciência dos limites e potencialidades associados a esses métodos, supõe que o investigador tenha informação, nomeadamente no que diz respeito à população seleccionada, à base de sondagem, à dimensão e métodos de selecção da amostra e critérios de estratificação que lhe estão associados, entre outros (Azorín e Sánchez-Crespo, 1986; Ghigliione e Matalon, 1993; Vicente, Reis e Ferrão, 1996).

Outra questão de relevo no uso das bases de microdados prende-se com o *cálculo dos ponderadores* e sua utilização. A opção por uma determinada metodologia pode estar associada a objectivos de utilização diversificados:

- na investigação em ciências sociais, a sua utilização tem muitas vezes como objectivo a correcção da proporcionalidade no cálculo das relações entre indicadores. Aqui os valores ponderados apresentam-se, regra geral, em termos relativos ou em médias;

- outro objectivo de utilização frequentemente operacionalizado pelo INE, pondera os dados da amostra para estimar os valores absolutos do universo. Neste caso, uma vez o ponderador activo, os dados apurados reflectem em termos absolutos as ordens de grandeza da população.

No que se refere à *validação da informação*, um dos aspectos a ter em conta prende-se com a eventual transposição da base de microdados de uma aplicação informática para outra, um processo que muitas vezes requer o apoio técnico do INE. Em seguida, quando da recepção dos dados, é de grande importância, como se referiu, a familiarização com a metainformação. O questionário e os desenhos de registos são conteúdos incontornáveis deste instrumento, nomeadamente para a validação da estrutura da base de dados e para o reconhecimento das variáveis.

Um outro procedimento de controlo no uso de microdados é a comparação entre as questões (e respectivas modalidades de resposta) que constam no instrumento de recolha de informação, e as variáveis presentes na base de dados recebida.

Para o utilizador externo à comunidade produtora dos dados, o controlo de qualidade e consequente validade dos resultados poderá passar, numa primeira fase e sempre que possível, pela comparação desses apuramentos com estatísticas já publicadas pelo organismo responsável pela sua recolha e difusão, e que tenham por base as mesmas definições das variáveis. Uma análise crítica da coerência e consistência dos apuramentos realizados, nomeadamente através de uma perspectiva longitudinal, poderá ainda consolidar essa validação.

A validação engloba ainda o conhecimento sobre mudanças de séries dos questionários e das respectivas aplicações (Fonseca, Lourenço e Lima, 1995). Tais modificações podem decorrer de melhoramentos na cobertura do inquérito em causa, correcções sazonais, aperfeiçoamentos metodológicos, etc. (Suominen, 2001).

6. CONDIÇÕES DE DIVULGAÇÃO DE ANÁLISES APOIADAS EM MICRODADOS

As exigências de utilização de microdados alargam-se também às condições de divulgação e difusão de análises apoiadas em informação estatística.

Os procedimentos associados à divulgação dos dados obrigam a que sejam cumpridos os requisitos acordados entre INE e investigadores, no sentido de garantir rigor tanto na informação como na sua referenciação. Nesta medida, é também importante que se estabeleçam alguns compromissos de divulgação de informação e análise estatística por parte dos utilizadores.

Um desses compromissos, que se pode designar como *ético*, prende-se com as garantias de anonimização que o investigador deve dar quando publica ou difunde informação deste tipo. Nesta perspectiva, o rigor que pauta a produção de conhecimento e respectiva divulgação está profundamente associado a preocupações cívicas, no sen-

tido da garantia dos direitos e das liberdades, sobretudo no que respeita à defesa da privacidade e da autonomia individuais.

Outro compromisso, não menos importante, é de carácter *técnico*. Este prende-se fundamentalmente com preocupações ligadas à qualidade dos dados, tais como o controlo da significância, a construção de novas variáveis e, na divulgação destas, a utilização de designações distintas das divulgadas pelo INE, para não se confundirem as respectivas definições.

7. HARMONIZAÇÃO E COMPARABILIDADE DOS DADOS

Uma outra questão que tem estado muito presente nas discussões da comunidade produtora de estatísticas é a que se refere à comparabilidade de informação entre países. Esta é concretizada, em grande parte, através da estandarização das metodologias de recolha de informação e/ou de variáveis entre os vários países membros da União Europeia, tendo necessariamente repercussões na utilização de microdados.

Através da harmonização procura-se conjugar estatísticas de fontes diversificadas, que têm como objectivo tornar os conceitos e os métodos utilizados claros e comparáveis entre si, no sentido de se apresentar um só quadro de informação estatística coerente. Tal procedimento permite, por um lado, que a informação possa ser lida cada vez mais numa perspectiva longitudinal e, por outro, que seja possível, de forma mais articulada, construir análises apoiadas em estatísticas que atravessem e, se possível, cruzem várias áreas, nomeadamente, no âmbito dos países da União Europeia.

A possibilidade de comparação constitui uma vantagem não só no âmbito das decisões económicas e políticas tomadas a nível europeu, mas também para um tipo de investigação que se pretenda situar numa escala europeia ou mundial (veja-se a este respeito Costa, Mauritti, Martins, Machado e Almeida, 2000: 9-46). Acedendo aos microdados os seus utilizadores podem, para além de reforçar a possibilidade de comparar informação construída segundo conceitos uniformizados, enriquecer as suas análises, dotando-as, de forma controlada, de especificidades de âmbito nacional, e testar os limites e a justeza dos conceitos harmonizados.

Quando o esforço de harmonização se centra essencialmente na recolha e compilação dos dados (estratégia *input*), a utilização de microdados permite controlar, de forma mais ponderada, constrangimentos que se prendem com as especificidades de jurisdição e legislação nacionais face à regulamentação da produção e uso estatísticos, bem como ao entendimento e cultura de informação que existe em cada país e Estado. As formas de ponderação e os níveis de significância, conhecendo-se as especificidades dos métodos de recolha de informação, podem aqui permitir, como referido, um leque de escolha ao utilizador que lhe possibilite um uso mais adequado às suas finalidades.

Os processos de harmonização, uma vez focalizados no tratamento de dados (estratégia *output*), estratégia de maior recorrência no âmbito das estatísticas europeias

(Jensen, 2000), colocam outros desafios aos utilizadores de microdados. Tais processos tendem a uniformizar conceitos, exigindo aos utilizadores a respectiva interpretação de forma contextualizada, apoiada tanto quanto possível no conhecimento da diversidade social e cultural, das estruturas e dinâmicas sociais próprias de cada país.

O uso de microdados permite, não se perdendo a perspectiva da comparabilidade, que se conheçam as agregações implícitas às variáveis uniformizadas e que se controle a qualidade imposta pela estandarização das variáveis. Através do acesso às bases de dados oficiais, tal comparabilidade pode ser ainda enriquecida através de especificações regionais ou nacionais, em associação ou complementarmente à identificação de perfis ou padrões construídos através da experimentação e realização de análises multivariadas, que permitem ainda assegurar a qualidade dos conceitos comparados.

As estratégias para a harmonização de dados implicam muitas vezes mudanças e interrupções dos inquéritos, o que cria constrangimentos a leituras longitudinais. Esta questão não deixa de ser relevante, também, quando se trata da utilização de bases de microdados, pois a mudança de variáveis e das suas formas de categorização torna as comparações entre as várias séries de inquéritos mais limitadas. No entanto, como referido, não se pode estancar os processos de mudança metodológica dos inquéritos e de definição de conceitos, pois, muitas vezes, vão no sentido do seu aperfeiçoamento ou exprimem dinâmicas de mudança na própria sociedade.

8. NOTAS FINAIS

A utilização de microdados estatísticos produzidos pelo INE, possibilidade aberta recentemente com o estabelecimento do protocolo entre o MCT e esse instituto, constitui um importante passo no desenvolvimento da investigação em ciências sociais em Portugal.

No centro da exploração deste tipo de informação está a possibilidade de combinar novas propostas teóricas com a exploração de novas alternativas de recurso à informação empírica. Esta consubstancia-se no facto de se poder construir e ensaiar modos alternativos de utilização de dados estatísticos. Entre estes, destaca-se a possibilidade de criação de novas variáveis, a oportunidade de realização de análises estatísticas multivariadas e, em parte através destas, a capacidade de construção e aperfeiçoamento de tipologias e modelos de análise. Todos estes processos podem ser ainda desenvolvidos e aprofundados através de uma perspectiva experimental no uso continuado das bases de dados.

Nem todas as situações de investigação exigem, ou mesmo aconselham, a utilização de microdados. Para muitas análises são suficientes as formas de acesso que já existiam anteriormente.

A divulgação de microdados supõe o assumir de um compromisso entre a disponibilização de mais e melhor informação e a garantia de respeito da salvaguarda de

direitos dos cidadãos, nomeadamente no que concerne à defesa de privacidade e autonomia individuais (Trivellato, 2000). Decorrem daqui alguns dos princípios básicos estabelecidos no âmbito do protocolo, como sejam a credenciação das instituições de investigação científica e tecnológica, a responsabilização dos seus investigadores e a anonimização das bases de dados por parte do INE.

Importa salientar que a utilização de microdados tem subjacente o domínio de alguns aspectos processuais de natureza técnica, transversais a todo o processo de produção, tratamento e difusão dos dados estatísticos. Nesta medida, conjugam-se aspectos cruciais ligados, entre outros, à metainformação, às formas de estimação dos resultados, à harmonização e comparabilidade.

REFERÊNCIAS BIBLIOGRÁFICAS

- ANDROVITSANEAS, CHRISTOS, “New technologies in statistics and requirements for central institutions: from a user’s perspective”, *Innovation in Provision and Production of Statistics: The Importance of New Technologies*, conferências do CEIES, Helsínquia, 20-21 de Janeiro de 2000 (policopiado).
- AZORÍN, FRANCISCO; e SANCHÉZ-CRESPO, José Luís, *Métodos e Aplicaciones del Muestreo*, Madrid, Alianza Universidad Textos, 1986.
- COSTA, ANTÓNIO FIRMINO DA; MAURITTI, ROSÁRIO; MARTINS, SUSANA DA CRUZ; MACHADO, FERNANDO LUÍS; e ALMEIDA JOÃO FERREIRA DE, “Classes sociais na Europa”, *Sociologia, Problemas e Práticas*, 34, 2000, pp. 9-46.
- CSE, *Relatório de Actividades do Conselho Superior de Estatística*, 1998, DOCT/351/CSE-2, 1998 (policopiado).
- DALE, ANGELA, “An academic perspective on the information society”, *Challenges for Community Statistics: A Users’ Perspective for the Next Ten Years*, conferências do CEIES Bruxelas, 26-27 de Abril de 2001 (policopiado).
- FONSECA, ISABEL DUARTE, VALENTE, ISABEL LOURENÇO, e LIMA, JOSÉ PEDRO, “A gestão da metainformação no sistema estatístico português”, *Sociologia, Problemas e Práticas*, 17, 1995, pp. 153-168.
- FÖRSTER, MICHAEL; HELLIESEN, ANNE; e KOLBERG, JOHN-EIVIND, “LES, the new challenge: an introduction into the concepts, documentation and background statistics”, *Luxembourg Employment Study, Working Paper*, 1, 1996, em <http://lisweb.ceps.lu/les/wps/leswp1.htm>.
- GHIGLIONE, RODOLPHE; e MATALON, BENJAMIM (1993), *Inquérito: Teoria e Prática*, Oeiras, Celta Editora.
- JENSEN, LOTHAR, “Prospects for European social statistics: from Maastricht to Lisbon”, *Sigma: The Bulletin of European Statistics*, 3/2000, pp. 2-4.

LAMEL, JOACHIM, “Feeling the pulse of society”, Sigma: The Bulletin of European Statistics, 2/2001, pp. 2-3.

PROTOCOLO ENTRE O MINISTÉRIO DA CIÊNCIA E DA TECNOLOGIA E O INSTITUTO NACIONAL DE ESTATÍSTICA PARA ACESSO A BASES DE DADOS PARA FINS DE INVESTIGAÇÃO CIENTÍFICA, 18 de Outubro de 1999.

SUOMINEM, RISTO, “Enlargement and Deepening”, Challenges for Community Statistics: A Users’ Perspective for the Next Ten Years, conferências do CEIES, Bruxelas, 26-27 de Abril de 2001 (policopiado).

TRIVELLATO, UGO, “Data Access versus privacy: an analytical user’s perspective”, Innovation in Provision and Production of Statistics: The Importance of New Technologies, conferências do CEIES, Helsínquia, 20-21 de Janeiro de 2000 (policopiado).

VICENTE, PAULA; REIS, ELIZABETH; FERRÃO, FÁTIMA, Sondagens: Amostragem Como Factor Decisivo, Lisboa, Edições Sílabo, 1996.