Seminário
Princípio do Segredo Estatístico

Segurança nas Bases de Dados

Cautelas na Obtenção de Dados Estatísticos Através de "Interfaces" Automáticos

INSTITUTO NACIONAL DE ESTATÍSTICA
PORTUGAL

Pedro Corte Real

Departamento de Metodologia Estatística

Agenda

- Introdução
- Métodos Automáticos de Obtenção de Informação Estatística
- Tornar Confidencial
- Conclusões

Introdução

- **Sociedade ávida de informação <u>tratada:</u>**
  - "Fast" + {food, lane,..., info*rmation*}

- **O termo "Fast" tem dois sentidos:**
  - Disponibilizada rapidamente, ou seja, quando necessária
  - Actual, ou seja, útil, isto é, pronta a consumir

Introdução

- **Preocupação actual...**
  - Que uso fazem da informação **confidencial**, mantida em bases de dados, que muitas entidades dispõem?
  - Quem, dentro das entidades que gere esta informação, lhe tem **acesso**?

Introdução

- **...Com potência para crescer**
  - Quantas destas entidades nos são completamente estranhas?
    - ?? Como é que sabem tanto de mim?? Poderá ter sido a informação que eu dei:
      - » Supermercado?
      - » Biblioteca?
      - » Clube de Vídeo?
      - » Cartão de Cliente da Loja 1, Loja 2,...
      - » ...
      - » **De uma coisa tenho a certeza, não foi o INE que permitiu o acesso a esta informação!**

Introdução

- **Temos de encontrar um compromisso entre segurança e facilidade de acesso**

  – Pressão crescente de várias entidades para acederem a informação privilegiada:

    - Comunidade científica
    - Empresas de estudos de mercado
    - Negócios que dependem de hábitos de consumo
    - Organismos fora da UE

Introdução

- **Pressão crescente de várias entidades para acederem a informação confidencial:**

  – **Comunidade científica** (Reg. CE Nº 831/2002, 17 Maio)

(1) O acesso, para fins científicos, a dados confidenciais transmitidos à autoridade comunitária é objecto de uma procura crescente dos investigadores e da comunidade científica em geral.

(2) O acesso, para fins científicos, a dados confidenciais pode ser concedido, quer autorizando a sua consulta nas instalações da autoridade comunitária, quer pondo os dados tornados anónimos à disposição dos investigadores em condições específicas (acesso controlado).

Introdução

- **Pressão crescente de várias entidades para acederem a informação confidencial:**
  - **Comunidade científica** (Reg. CE Nº 831/2002, 17 Maio)

Artigo 5.º

Acesso nas instalações da autoridade comunitária

1. A autoridade comunitária pode conceder o acesso nas suas instalações a dados confidenciais obtidos a partir dos seguintes inquéritos ou fontes de dados estatísticos:

— painel de agregados domésticos privados da União Europeia,

— inquérito às forças de trabalho,

— inquérito comunitário à inovação,

— inquérito à formação profissional contínua.

No entanto, mediante pedido da autoridade nacional que forneceu os dados, o acesso aos dados dessa autoridade nacional não será concedido para um projecto específico de investigação.

2. Sujeito à aprovação prévia e explícita da autoridade nacional pertinente, a autoridade comunitária pode conceder o acesso nas suas instalações a outros dados confidenciais além dos referidos no n.º 1.

Introdução

- **Pressão crescente de várias entidades para acederem a informação confidencial:**
  - Empresas de estudos de mercado
  - Negócios que dependem de hábitos de consumo
  - Organismos fora da UE

Introdução

- **Pressão crescente de várias entidades para acederem a informação confidencial:**
  - **Painel de Agregados Domésticos Privados da UE (ECHP):**
    - O requerente faz o pedido, recebe uma resposta no mesmo dia, com um conjunto de questões, cujas respostas serão usadas no contrato a celebrar.
    - **1 semana**: Tempo médio de resposta do requerente
    - **1 - 3 dias**: Até a consulta aos EM se iniciar e se pedir à Unidade D2 do Eurostat que verifique se a UDB-ECHP é adequada ao projecto
    - **6 semanas**: Tempo de consulta aos EM (Regulamento)

Introdução

- **Pressão crescente de várias entidades para acederem a informação confidencial:**
  - **Painel de Agregados Domésticos Privados da UE (ECHP):**
    - **2 semanas:** Sem objecções dos EM, o contrato segue o circuito hierárquico usual. É enviado para o requerente que deverá propor as últimas alterações e reenviá-lo à unidade A5 do Eurostat.
    - **3 semanas:** Envio do contrato assinado pelo requerente para o Eurostat.
    - **1 semana:** Reenvio do contrato assinado e da informação em CD-ROM para o requerente.

Introdução

- **Pressão crescente de várias entidades para acederem a informação confidencial:**
  - **Painel de Agregados Domésticos Privados da UE (ECHP):**
  - **Tempo Médio de Resposta: 13.5 semanas.**
  - Situações mais simples. Se a unidade D2 tiver quaisquer dúvidas relativamente ao projecto, todo o processo fica adiado.
  - Situação típica para Centros de Investigação ou Universidades dentro da UE. Todos os outros casos precisam de aprovação prévia do Comité de Confidencialidade Estatística, o que demora o processo.

- **Esforço de sensibilizar as várias entidades públicas que produzem informação administrativa:**

  – Boas Práticas
  – Celeridade e Segurança na Recepção da Informação
  – Qualidade da Informação
  – Custos

- Esforço de sensibilizar as várias entidades públicas que produzem informação administrativa:

  – **Boas Práticas**

  Boa gestão da carga estatística não usando, para além do indispensável, o tempo dos indivíduos, famílias ou empresas no fornecimento de informação, ou seja, quando esta não pode ser obtida por via administrativa ou derivada a partir de outros inquéritos.

- Esforço de sensibilizar as várias entidades públicas que produzem informação administrativa:

– Celeridade e Segurança na Recepção da Informação

Os sistemas informáticos existentes, em particular as redes de Internet, Intranet e as Extranet, permitem a troca de informação, como sejam registos, se necessário for, em tempo real e com níveis de segurança iguais ao das instituições mais exigentes, como por exemplo, as bancárias.

- Esforço de sensibilizar as várias entidades públicas que produzem informação administrativa:

  – **Qualidade da Informação**

  Como a informação obtida provém de instituições que recolhem informação administrativa e que dispõem de processos de controlo da qualidade e fiabilidade, espera-se que a qualidade dos dados recebidos seja superior àquela que teriam se fossem obtidos por inquirição amostral.

- Esforço de sensibilizar as várias entidades públicas que produzem informação administrativa:

  – **Custos**

  De acordo com as estimativas, os custos de tratamento da informação administrativa é significativamente inferior à de outros métodos de recolha, mesmos os menos dispendiosos que possam ser usados para actualizar a informação relevante.

- **Garantias à Sociedade e aos Fornecedores de Informação Administrativa:**

a) A transferência electrónica da informação para o INE é feita através de uma ligação segura e com mecanismos de autenticação apropriados;

b) O transporte em suporte informático, com os dados sob a forma encriptada, é efectuado em mão e por pessoal credenciado, com utilização de protocolo para registo da identificação do expedidor e do destinatário, que estão devidamente referenciados;

c) O processamento e armazenamento dos dados é efectuado fisicamente nas instalações do INE, protegidas por um mecanismo de controlo de acessos;

- **Garantias à Sociedade e aos Fornecedores de Informação Administrativa:**

d) O tratamento e análise interna dos dados é efectuado pelos trabalhadores do INE que estão sujeitos às normas do Segredo Estatístico, nomeadamente a alínea b) do n.º 2. do art.º 5º da Lei 6/89, de 15 de Abril;

e) O acesso ao sistema informático e aos programas de gestão de dados é efectuado somente com identificação e passwords individuais, existindo uma listagem dos acessos efectuados;

f) A atribuição dos acessos ao sistema informático é efectuada nominalmente e por um responsável referenciado, e com níveis de permissão que identificam claramente quais os tipos de acessos e operações sobre os dados a que estão autorizados;

- **Garantias à Sociedade e aos Fornecedores de Informação Administrativa:**

g) o armazenamento da informação é efectuado de modo que os dados económicos e financeiros ficam separados fisicamente das variáveis de identificação, e em que a sua ligação é efectuada recorrendo a algoritmos específicos para o efeito;

h) durante o tempo de retenção, os dados são armazenados sob a forma encriptada e a respectiva chave fica na posse de funcionários do INE devidamente referenciados.

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**

  – Permite a investigação de questões mais complexas, incluindo efeitos marginais ou mais específicos, que não são os objectivos do INE's

  – A comunidade científica tem um conhecimento específico que é importante colocar ao serviço da sociedade e que será difícil (se útil) duplicar nos INE's

  – Acresce a importância dos INE's, mostrando que a informação (cuja recolha é custosa, financeiramente e não só) é usada com ganhos sociais evidentes

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**

  – Aumentar a confiança nas interpretações e resultados difundidos pelos INE's, através de um diálogo com a comunidade científica, fundido a comunidade de estaticistas oficiais e investigadores

  – O diálogo entre INE's e a comunidade científica permitirá fazer um uso profundo dos microdados, potenciando uma melhoria da qualidade (instrumentos de recolha, processo de recolha, imputação, etc.)

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - **Risco** de Identificação das Unidades Estatísticas
  - **Risco** permanece com os INE's que permitiram o acesso à informação
  - **Risco** de ao Proteger a Confidencialidade se introduzir ruído que torne os dados impróprios para os objectivos da investigação

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - **Custos** associados ao acesso
  - **Custos** associados à reputação
  - **Custos associados às** taxas de respostas

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**

Uma aposta na **METODOLOGIA** é fundamental já que nos permitirá controlar a aversão (total ?) ao risco, passando a gestores de risco.

Metodologias estatísticas para divulgação de microdados (Statistical disclosure control - SDC) são parte da solução para uma divulgação metódica e com o risco medido e controlado.

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**

  – Como devemos alterar um ficheiro de microdados de modo a garantir que o risco de identificação é aceitável e com aumento de confusão mínimo (perda de informação mínima)?

  – Como podemos definir risco de identificação?

  – Como podemos medir perda de informação?

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**

  – Precisamos de propor um cenário relativamente aos meios e às capacidades de um utilizador do ficheiro a ser disponibilizado ("intruder").

  – Um registo será considerado pouco seguro ou em risco de identificação se a probabilidade (estimada com base no modelo e cenário acima proposto) de ser identificado como o elemento *X* da população, for superior a um determinado limiar.

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - Microdados:

Registo	Nome	Idade	Profissão	Salário
1	Carvalho	44	Construção Civil	45.500 €
2	Silva	44	Medicina	37.900 €
3	Pereira	55	Construção Civil	67.000 €
4	Ferreira	44	Medicina	21.000 €
5	Branco	55	Medicina	90.000 €
6	Jacob	45	Advocacia	48.000 €
7	Coelho	25	Advocacia	49.000 €
8	Santos	35	Construção Civil	66.000 €
9	Rosa	55	Construção Civil	69.000 €
10	Lopes	45	Física Nuclear	34.000 €

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - Microdados:

1. Remover todos as variáveis que permitam uma identificação individual. Nome, Morada, Identificações Administrativas (BI, NPC, NIF, etc.)

Registo	Nome	Idade	Profissão	Salário
1	-	44	Construção Civil	45.500 €
2	-	44	Medicina	37.900 €
3	-	55	Construção Civil	67.000 €
4	-	44	Medicina	21.000 €
5	-	55	Medicina	90.000 €
6	-	45	Advocacia	48.000 €
7	-	25	Advocacia	49.000 €
8	-	35	Construção Civil	66.000 €
9	-	55	Construção Civil	69.000 €
10	-	45	Física Nuclear	34.000 €

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - Microdados:

2. Verificar que outras células contém informação susceptível de permitir uma identificação directa ou indirecta

Count of Profiss	
Profissão	Total
Física Nudear	1
Advocacia	2
Medicina	3
Construção Civil	4
Grand Total	10

Grupo Etário	Data	Total
[20;30[	Sum of Salário	49.000 €
	Count of Grupo Etário	1
[30;40[	Sum of Salário	66.000 €
	Count of Grupo Etário	1
[40;50[	Sum of Salário	186.400 €
	Count of Grupo Etário	5
[50;60[	Sum of Salário	226.000 €
	Count of Grupo Etário	3
Total Sum of Salário		527.400 €
Total Count of Grupo Etário		10

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - Microdados:

2. Verificar que outras células contém informação susceptível de permitir uma identificação directa ou indirecta

Registo	Nome	Idade	Profissão	Salário
1	-	44	Construção Civil	50.000 €
2	-	44	Medicina	40.000 €
3	-	55	Construção Civil	70.000 €
4	-	44	Medicina	20.000 €
5	-	55	Medicina	90.000 €
6	-	45	Advocacia	50.000 €
7	-	25	Advocacia	50.000 €
8	-	35	Construção Civil	70.000 €
9	-	55	Construção Civil	70.000 €
10	-	45	Física Nuclear	30.000 €

Suprimir e Classificar

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - Microdados:

2. Verificar que outras células contêm informação susceptível de permitir uma identificação directa ou indirecta

Registo	Nome	Idade	Profissão	Salário
1	Carvalho	44	Construção Civil	45.500 €
2	Silva	44	Medicina	37.900 €
3	Pereira	55	Construção Civil	67.000 €
4	Ferreira	44	Medicina	21.000 €
5	Branco	55	Medicina	90.000 €
6	Jacob	45	Advocacia	48.000 €
7	Coelho	25	Advocacia	49.000 €
8	Santos	35	Construção Civil	66.000 €
9	Rosa	55	Construção Civil	69.000 €
10	Lopes	45	Física Nuclear	34.000 €

Registo	Idade	Profissão	Salário
1	44	Construção Civil	50.000 €
2	44	Medicina	40.000 €
3	55	Construção Civil	70.000 €
4	44	Medicina	20.000 €
5	55	Medicina	90.000 €
6	45	Advocacia	50.000 €
7	-	Advocacia	50.000 €
8	-	Construção Civil	70.000 €
9	55	Construção Civil	70.000 €
10	45	-	30.000 €

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - Microdados:

2. Verificar que outras células contém informação susceptível de permitir uma identificação directa ou indirecta

Count of Profissão	
Profissão	Total
-	1
Advocacia	2
Medicina	3
Construção Civil	4
Grand Total	10

Grupo Etário	Data	Total
-	Sum of Salário	120.000 €
	Count of Grupo Etário	2
[40;50[	Sum of Salário	190.000 €
	Count of Grupo Etário	5
[50;60[	Sum of Salário	230.000 €
	Count of Grupo Etário	3
Total Sum of Salário		540.000 €
Total Count of Grupo Etário		10

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - **–Exemplo ECHP**

É um painel realizado em 15 países da UE, desde 1994. Procura compilar informação sobre o rendimento e as condições de vida das famílias ao longo de 8 anos. O painel compreende, aproximadamente, 400 variáveis.

Os microdados são transmitidos ao Eurostat sem identificadores directos.

Uma base de dados tornada anónima foi preparada para efeitos de investigação científica.

A informação é usada para determinar os indicadores de Laeken relativos ao fenómeno da pobreza e para monitorar a "Estratégia Lisboa".

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - **Exemplo ECHP**

– **Idade:** As idades são classificadas com um limite superior ("top coded"), classificando como tendo nascido em 1909 todos os nascidos anteriormente, com a excepção da Alemanha que usa como limite superior 1924.

– **Região (NUTS)**: A maioria dos países fornece informação ao nível de NUTS 1 ou 2, com excepção da Alemanha, Dinamarca e Holanda que não fornecem qualquer tipo de informação regional.

– **Dimensão do Agregado:** Codificado com um máximo de 7.

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - **Exemplo ECHP**

  – **Número de Divisões:** Classificado com um máximo de 10.

  – **Nacionalidade**: Classificado como "**Nacional**", "**Não nacional mas da UE**", "**Não nacional e não da EU**", "**Não nacional e desconhecido**".

  – **Percurso Migratório e Ocupação:** É codificado com diferente detalhe pelos vários países.

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - **Exemplo ECHP**

  – Para medir o risco de identificação, precisamos de formular um conjunto de pressupostos que usaremos em conjunto com um modelo matemático (em princípio probabilístico) e que incorpora o grau de conhecimento que um possível utilizador (mal intencionado ?) possa ter.

  – Evidentemente, podem surgir identificações espontâneas, pois combinações raras de variáveis (se não existiu supressão de valores) possibilitam a identificação de algumas unidades estatísticas.

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - **Exemplo ECHP**

  – Com Grande Potencial de Identificação:
    - **Região (NUTS)**

  – Com Potencial de Identificação:
    - **Meio Rural ou Urbano**
    - **Dimensão do Agregado**
    - **Sexo**
    - **Estado Civil**

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - **Exemplo ECHP**
  - **País 1, amostra de dimensão 3.763, sem apresentar informação relativa à região**

Número de células só com uma observação

Idade	Profissão	Dimensão Agregado	1.318
Idade	Profissão	Divisões	1.619
Idade	Profissão	Migração	907

- Resultante da idade não ser classificada.

Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - **Exemplo ECHP**
  - **País 2, amostra de dimensão 10.120**

Número de células só com uma observação

Idade	Região	Sexo	88
Idade	Região	Estado Civil	426
Idade	Região	Dimensão Agregado	648
Idade	Região	Profissão	2.278
Idade	Região	Migração	633

  - Neste caso Região e Ocupação são bastante identificativas.

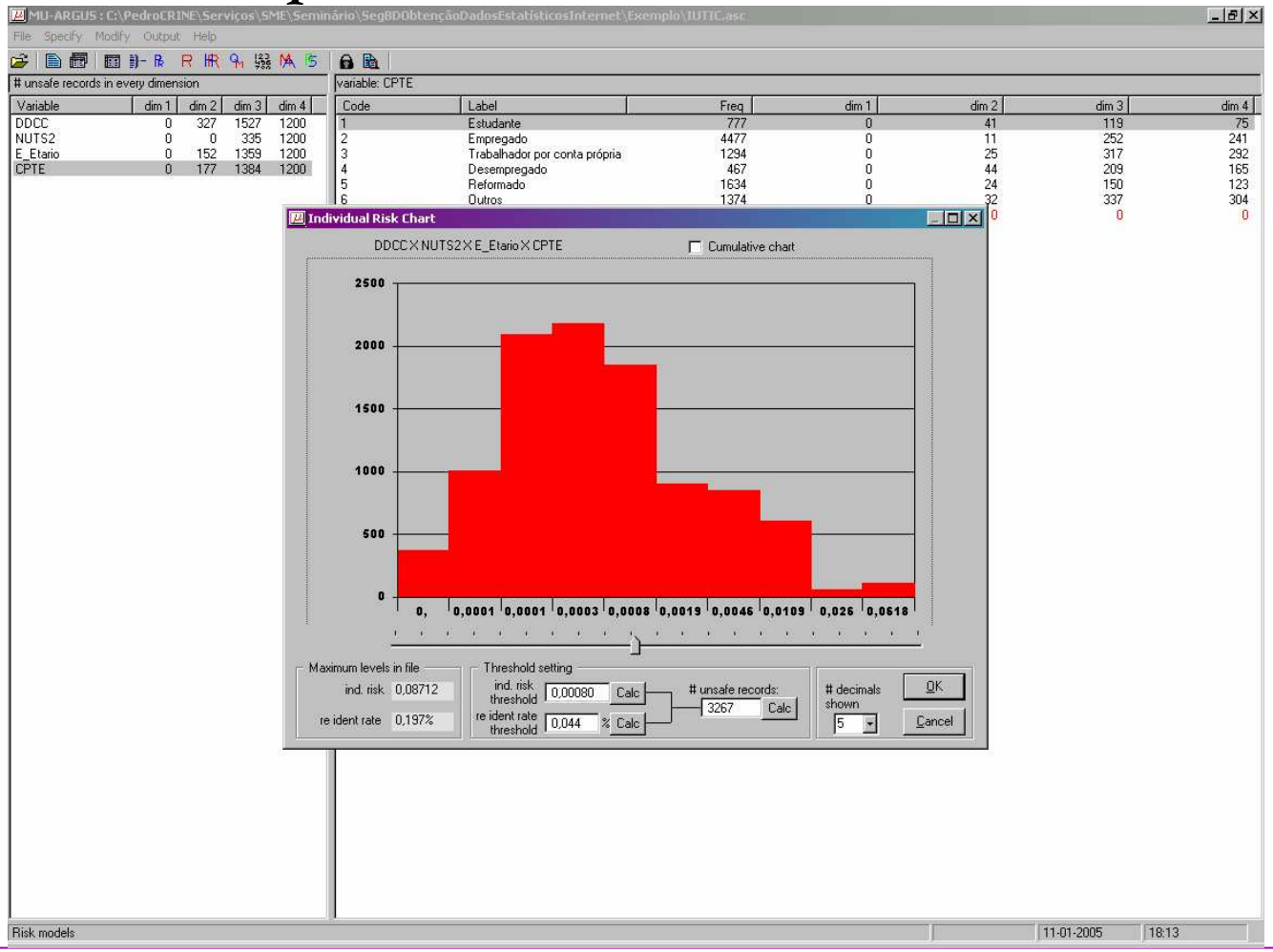
Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - **Exemplo IUTIC (n = 10.023)**

					Número
	Distrito Concelho				0
	NUTS2				0
	Escalão Etário				0
1	Situação Emprego				0
	Distrito Concelho	NUTS2			0
	Distrito Concelho	Escalão Etário			151
	Distrito Concelho	Situação Emprego			176
	NUTS2	Escalão Etário			0
	NUTS2	Situação Emprego			0
2	Escalão Etário	Situação Emprego			1
	Distrito Concelho	NUTS2	Escalão Etário		151
	Distrito Concelho	NUTS2	Situação Emprego		176
	Distrito Concelho	Escalão Etário	Situação Emprego		1200
3	NUTS2	Escalão Etário	Situação Emprego		8
4	Distrito Concelho	NUTS2	Escalão Etário	Situação Emprego	1200

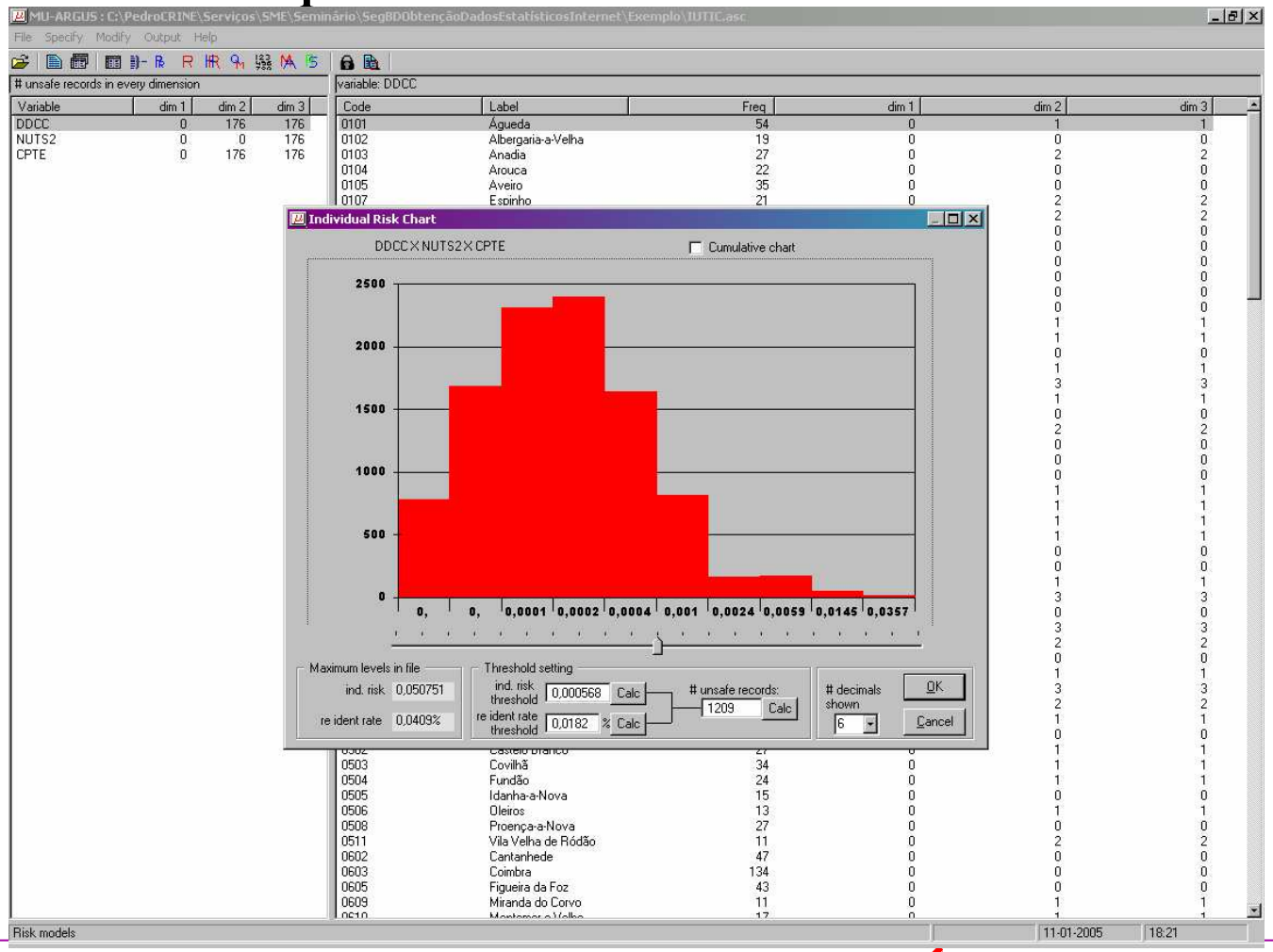
Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - **Exemplo IUTIC**



Tornar Confidencial

- **Permitir Acesso aos Microdados Tornados Anónimos:**
  - **Exemplo IUTIC**



Conclusões

- **A divulgação dos microdados trará à comunidade científica um peso acrescido nas estatística oficiais.**

- **É importante desenvolver metodologias de gestão do risco associado à disponibilização dos microdados.**

- **É importante conhecer a capacidade dos Centros de Investigação de proteger a informação confidencial.**

Conclusões

- **A divulgação dos microdados trará à comunidade científica um peso acrescido nas estatística oficiais.**
  - **Pertinência**
  - **Precisão**
  - **Actualidade**
  - **Acessibilidade**
  - **Comparabilidade**
  - **Coerência**

Conclusões

- **É importante desenvolver metodologias de gestão do risco associado à disponibilização dos microdados.**
  - **Melhor definição de Risco**
  - **Melhores medidas de Controlo de Risco**
  - **Processos que preservem as "qualidades estatísticas da informação".**

Conclusões

- **É importante conhecer a capacidade dos Centros de Investigação de proteger a informação confidencial.**
  - **Sensibilizar os Centros de Investigação para o conceito de "Segredo Estatístico"**
  - **Sensibilizar os Centros de Investigação para o conceito de Comparabilidade e de Coerência**

Referências

- European Commission. *Monographs of Official Statistics*. 2003
- CSC 2004/B5 *item 2.1.1.bis;* CSC 2004/B5 *item 2.2.1.; CSC 2004/B5 item 2.2.3.bis*
- Fernanda Perpétuo. *Statistical Information System For Researchers*. 2001
- Júlia Cravo. *Uso de Fontes Administrativas*. INE- 2004
- Eurostat. *Manual For the Protection of Confidential Data in Eurostat*. 2004
- μ-Argus 3.2. Manual de Utilização. 2004
- Nações Unidas. *Statistical Confidentiality and Access to Microdata*. 2003
- Office of Management and Budget. *Report on Statistical Disclosure Limitation Methodology*. 1994
- Traian Marius Truta, *et al. Automatic Generation of Masked Microdata*. 2000