RESUMO

Os grandes volumes de dados requerem métodos automáticos de tratamento que permitam deles extrair informação adequada ao nível de exigência do utilizador. Refere-se neste trabalho dois tipos de instrumentos de software existentes no LNEC para abordar tais problemas: um package de apuramento de Inquéritos desenvolvido no CI do LNEC é um conjunto de packages, operacionais ou em vias de o estarem, sobre Analise estatística de Dados Multidimensionais
(AED).

Apresenta-se um exemplo de aplicação em ambos os domínios e, sobre a AED apresentam-se sumariamente alguns fundamentos teóricos. Reterem-se alguns trabalhos em curso no LNEC e mencionam-se os. packages disponíveis.

1 - INTR0DUÇAO

Os grandes volumes de dados representativos de certa informação que se pretende conhecer exigem inevitavelmente o recurso ao computador e o uso de modernas técnicas Informáticas. Problemas próprios e com certa especificidade podem identificar-se no que respeita à recolha da informação, ao seu armazenamento (recorrendo ou não a Bases de Dados), à sua utilização. Para além da pesquisa de fontes publicadas é muitas vezes necessário lançar mão de inquéritos, fundamentar estatisticamente o seu lançamento, representar e tratar os seus resulta dos. Mas, devido à complexidade da estrutura dos dados obtidos, à dissimulação da Informação essencial que se pretende extrair, as múltiplas inter-relações das variáveis representativas dos dados, desenhou-se um corpo autónomo de conhecimentos na Estatística, a Análise Estatística de Dados Multidimensionais (AED).

É sobre o apuramento de Inquéritos (tratamento e representação de resultados) e sobre aquela ultima matéria que o presente trabalho se desenvolve, descrevendo e posicionando, em linhas gerais, os problemas e métodos envolvidos e referindo a experiência (ainda curta) do LNEC no que respeita a meios Informáticos disponíveis ou em desenvolvimento: um package de apuramento de Inquéritos produzido pelo Centro de Informática (CI) do LNEC e um Conjunto de packages adguiridos ou obtidos do exterior sobre AED.

2 – Apuramento DE INQUERITOS

2.1- Considerações gerais

O apuramento de inquéritos é exemplo típico de uma classe de problemas em que a dificuldade resulta do

numero de operações a efectuar e não da complexidade de cada uma. Esta situação é semelhante a que se encontra em muitos domínios de gestão e a contribuição da Informática num e noutro caso é também semelhante. Os computadores são a máquina Ideal para efectuar repetitivamente operações do mesmo tipo sendo imunes, ao contrário do homem, ao cansaço e aos erros que lhe são inerentes. No entanto, os computadores têm de ser programados e os programas, depois de concebidos e escritos, têm de ser testados para se obter a certeza de que fazem a tarefa para que foram elaborados. E é sabido que só uma longa utilização permite adquirir a certeza de que um programa está correcto. Por outras palavras, o problema da prova matemática da correcção de um programa é, ainda hoje, um tema de investigação académica.

E com estes dados que se tem de responder i pergunta : é ou não rentável recorrer ao computador para a resolução de um problema do tipo do apuramento de inquéritos? No caso de ser necessário escrever e testar um programa total mente novo a resposta e, normalmente. não, a menos que o volume de dados (respostas) torne inviável uma solução manual. Isto porque o tempo e o custo envolvidos para a obtenção de um programa que se possa considerar testado são muito elevados. O problema e análogo ao da produção industrial: a produção em serie só se justifica para um numero muito elevado de unidades.
Postos perante a ocorrência frequente de solicitações para o apuramento automático de inquéritos (e também para outros tipos de trabalho aos quais estas considerações se aplicam plenamente) fomos levados a conclusão de que seria extremamente útil o desenvolvimento de um programa genérico para o apuramento de inquéritos (a exemplo do que já fora feito para a formatação de textos (LNEC, 1978). Por programa genérico entende-se um programa capaz de mediante um numero reduzido e bem determinado de alterações, dar resposta a uma gama m uito larga de problemas do mesmo tipo.
A elaboração de um tal programa passa, obrigatoriamente, pela caracterização da classe de problemas que se pretende resolver. E aqui um compromisso é necessário entre o querer ser absolutamente genérico - e criar uma ferramenta Inútil por: demasiado complexa - e o pretender ser expedito - e deixar de fora uma percentagem excessiva de aplicações. A nossa experiência nesta matéria indica que é razoável apontar para a cobertura de cerca de 90% das necessidades Os Custos de desenvolvimento para a cobertura de mais alguns : crescem muito rapidamente à medida que nos aproximamos de 100% - e que o amadurecimento da concepção e extremamente importante, mesmo essencial.
No Caso do apuramento de inquéritos definiu-se uma classe de problemas ~ base nos seguintes conceitos:

Átomo-Entidade elementar (dai o seu nome) que contém a resposta do Inquirido a uma dada pergunta.
Há 4 tipos de Átomos, consoante a natureza da pergunta:
Inteiro- Idade, salário, etc.
Real - Qualquer valor numérico não representável por um inteiro
Carácter- Sexo (M/F), etc.
Texto - Qualquer informação descritiva

Boletim -Conjunto de Átomos. Cada Átomo Contém a resposta a uma dada pergunta

Inquérito - Um ou mais Boletins contendo respostas a perguntas. A cada entidade Inquirida (indivíduo, família, empresa, etc.) corresponde um Inquérito.
Apuramento -
O Apuramento consiste em percorrer todos os Inquéritos e construir um certo número de Quadros que, de acordo com o estabelecido previamente, resumem as características relevantes da população inquirida.

Quadro - Um Quadro e a representação da dependência entre varias Entidades (1). As Entidades que figuram num Quadro tomam obrigatoriamente um número finito e geralmente pequeno de valores. Corno caso particular as Entidades podem tomar apenas um valor e o Quadro reduzir-se a um numero.

Entidade
As Entidades podem ser simples (Átomos) ou complexas Uma Entidade complexa é o resultado da combinação de uma ou mais Entidades simples ou complexas As Entidades complexas deu-se o nome de Moléculas. O caso particular em que uma Entidade complexa tem um valor que depende da frequência de repetição de um valor particular de uma outra Entidade num Inquérito deu origem a noção de Contador.

Átomos, Moléculas e Contadores são pois os 3 tipos de Entidades.
O exemplo seguinte poderá esclarecer melhor estes conceitos.
Exemplo:

Num Inquérito (hipotético) aos trabalhadores de uma empresa figuram as seguintes perguntas.

A cada pergunta corresponderá um Átomo. Todos os Átomos são Inteiros com excepção do 2º que é do tipo Caracter (M/F).

Pretendem-se os seguintes Quadros:

Q.1 - Distribuição dos trabalhadores por Sexo e Escalão de Vencimento

Q.2 - Distribuição dos agregados por Escalão de Rendimento e (Sexo x % de horas extraordinárias).

(1)Na prática consideram-se apenas Quadros bidimensionais (uma entidade representada em função de 2 outras) por se considerar que o caso geral e sempre redutível a este e que tal acaba sempre por ser feito dado o caracter bidimensional do meio de comunicação usual - o papel.

Q.3 - Distribuição dos agregados por nº de filhos e Escalão de Rendimento.

Para realizar o apuramento há que definir varias Entidades complexas:

Escalão de Vencimento -Molécula que depende apenas do Vencimento.

Escalão de Rendimento -Molécula que depende dos Vencimentos do funcionário e do cônjuge

Sexo x % de Horas Ext. - Molécula que depende do Sexo do funcionário e da % de horas extraordinárias realizadas em média.

Nº de filhos -Contador que depende do número de vezes que a pergunta Idade do filho é respondida.

Com base nos conceitos que acima se apresentam de forma muito sumária elaborou-se um programa que serviu já para o apuramento de 4 conjuntos de inquéritos de características e complexidade bastante diferentes. A experiência adquirida permitiu o amadurecimento da concepção que se considerou essencial e a passagem à fase seguinte que consiste em dar corpo a um "package" de apuramento de Inquéritos. Numa primeira fase a utilização do "package" para um caso concreto exigirá a programação correspondente à definição das Entidades simples e complexas envolvidas, dos Quadros pretendidos e do formato dos dados.

Simultaneamente investe-se, com a colaboração do grupo de Lógica Computacional, na escrita de um módulo dialogante que escreva ele próprio a totalidade ou a quase totalidade do código necessário para uma dada aplicação, tornando o "package" utilizável quase sem apoio de pessoal especializado.

2.2 - Exemplo de aplicação

Como exemplo, apresenta-se uma aplicação feita para o estudo do ruído em redor do aeroporto de Faro. Demarcadas que foram quatro zonas procedeu-se à recolha de inquéritos tendo-se obtido duas amostras de 206 e 90 exemplares, respectivamente.

Cada inquérito englobava um conjunto de 279 Átomos dos tipos Inteiros e caracter definidores da população em causa. Esta Informação foi perfurada em conjuntos de 6 cartões de 80 colunas por Inquérito.

Como o objectivo do estudo era o do conhecimento das fontes de incomodidade sonora foram para tal elaborados 320 quadros. Muitos desses quadros envolvem o cruzamento. para conhecimento da distribuição percentual, de Entidades que não são os próprios Átomos i.e. a resposta directa do Inquirido a terminada questão (Átomos). Assim, procede-se em alguns casos à transformação do Átomo ou de Átomos. Então, o Átomo é transformado numa outra Entidade chamada Molécula. A passagem de Átomo a Molécula envolve um operador transformador que pode ser um algoritmo ou uma tabela. Muitas das vezes o algoritmo pode ser bastante complexo sendo difícil encontrar um adequado, faz-se então o recurso à tabela. Foram utilizados no caso 80 moléculas; a título de exemplo, suponhamos, como no estudo realizado, que se pretende um quadro de distribuição percentual do índice de ocupação com a zona. Em que:

Índice de ocupação número de elementos de família/ nº de divisões assoalhadas

O numero de elementos da família é fornecido em quatro átomos:
ATOMO 20- número de elementos com idade superior a 18 anos
AToMO21- numero de elementos com Idade entre 10 e 18 anos
ATOMO 22- número de elementos com idade entre 5 e 10 anos
ATOMO 23- número de elementos com idade inferior a 5 anos

Como o pretendido é a soma dos elementos temos que considerar uma primeira Molécula: a que é a transformação pelo algoritmo somados ÁTOMOS 20 21; seguidamente a molécula B formada pela soma da Molécula A com o Átomo 22 e assim sucessivamente obtínhamos C que continha o somatório total que vai ser agora operada através dum algoritmo de divisão com o Átomo 30 que contém nº de divisão assoalhadas do casa, obtendo finalmente a Molécula D (indice de ocupação). Esta Molécula D era então cruzada com a zona (associada a um Átomo Inteiro),.obtendo-se um Quadro, que é uma matriz definida pelo número de elementos em D (Molécula) e número de elementos em zona (Atomo).

Em media os Quadros realizados eram matrizes de 10 x 10.

Um outro exemplo de Molécula poderia ser a transformação em classes de amplitude definida dum Átomo do tipo idade (Inteiro).

Foram utilizados 16 algoritmos transformadores no presente estudo.

Espera-se que num futuro breve a validação dos Átomos, operação que consiste na verificação da grandeza dos valores atribuídos à questão em causa, que até este momento estava a ser feita no módulo principal, passe a ser feita separadamente limitando-se este a fazer o input já indicado.

3 - ANÁLISE ESTATÍSTICA DE DADOS MULTIDIMENSIONAIS

3.1 – Considerações gerais. Definição. Campo de aplicações. Objectivos

A Análise Estatística de Dados (AED) também designada por Análise Estatística de Dados Multdimensional ou Estatistica Descritiva Multidimensional é um ramo da Estatística que engloba um conjunto de técnicas de descrição e resumo de grandes tabelas de dados. usando determinados critérios

A AED aplica-se sobretudo em estudos que envolvem grandes quantidades de informação que representa as observações ou medidas de um conjunto p, de variáveis (u) sobre um outro conjunto n de individuos ou entidades observadas (também designados por objectos ou unidades estatísticas u.e.
Para n e p grandes, e difícil uma interpretação imediata dos dados quer devido ao elevado número de observações quer a redundâncias e interdependências que nelas existem de modo geral

Os dados podem pois agrupar-se constituindo uma matriz x= (xij)de dimensões (nxp) em que n é o n.º de u.e., p é o número de v e xij é o valor da variável j para a eu i.

O objectivo fundamental da AED é a representação resumida dos dados com um mínimo de perda de informação e
Um máximo de explicação, ou seja, de modo a facilitar a interpretação. Assim, a partir de uma tabela de dados relativa a um fenómeno em estudo, pretende-se tornar os dados mais legíveis, substituir a quantidade pela qualidade, isto e, por em evidencia estruturas e agrupamentos e se possível, identificar as variáveis e/ou u. e. mais características

Pretende-se também reduzir o conjunto das p variáveis observadas a um conjunto de dimensão inferior mais complexas construídas à custa das p iniciais Estas novas variáveis muitas vezes chamadas factores, permitem uma descrição mais clara e simultaneamente mais rica em informação dos dados de base do que as p V. iniciais

Não se trata de modelar um fenómeno nem de pretender extrair conclusões de caracter inferencial mas

sim de o descrever fazendo realçar, quer as estruturas ·existentes no conjunto das variáveis, quer semelhanças das u e observadas Não há assim hipóteses feitas, todas as conclusões surgirão dos dados O resumo estatístico obtido não pretende ser nem verdadeiro nem falso mas o sim o melhor possível adaptado ao problem a posto, a compreensão dos dados e a sua estrutura