INVESTIGAÇÃO OPERACIONAL

Junho 1997
Número 1
Volume 17

Publicação Científica da

Associação Portuguesa para o Desenvolvimento
da Investigação Operacional

INVESTIGAÇÃO OPERACIONAL

Propriedade:
APDIO — Associação Portuguesa para o Desenvolvimento da Investigação Operacional

ESTATUTO EDITORIAL

«Investigação Operacional», órgão oficial da APDIO cobre uma larga gama de assuntos reflectindo assim a grande diversidade de profissões e interesses dos sócios da Associação, bem como as muitas áreas de aplicação da I. O. O seu objectivo primordial é promover a aplicação do método e técnicas da I.O. aos problemas da Sociedade Portuguesa.
A publicação acolhe contribuições nos campos da metodologia, técnicas, e áreas de aplicação e software de I. O. sendo no entanto dada prioridade a bons casos de estudo de carácter eminentemente prático.

Distribuição gratuita aos sócios da APDIO

INVESTIGAÇÃO OPERACIONAL

Volume 17 - nº 1 - Junho 1997 Publicação semestral

A Revista "INVESTIGAÇÃO OPERACIONAL" está registada na Secretaria de Estado da Comunicação Social sob o nº 108335.
Esta Revista é distribuida gratuitamente aos sócios da APDIO. As informações sobre inscrições na Associação, assim como a correspondência para a Revista devem ser enviadas para a sede da APDIO - Associação Portuguesa para o Desenvolvimento da Investigação Operacional - CESUR, Instituto Superior Técnico, Av. Rovisco Pais, 1000 Lisboa.

Este Volume foi subsidiado por :

Junta Nacional de Investigação Científica e Tecnológica (JNICT)

Fundação Calouste Gulbenkian

Para efeitos de dactilografia e composição, foram utilizados equipamentos gentilmente postos à disposição pelo Centro de Investigação Operacional (DEIO - FCUL).

Assinatura: 5.000$00

INVESTIGAÇÃO OPERACIONAL, vol. 17 (Junho 1997) 3-21

ANÁLISE PROBABILÍSTICA DA CASUALIDADE SÍSMICA EM PORTUGAL CONTINENTAL

M.L. Sousa

Centro de Estudos e Equipamento de Engenharia Sísmica
Laboratório Nacional de Engenharia Civil
Av. do Brasil, 101
1799 Lisboa Codex - Portugal

R.C. Oliveira

C.S. Oliveira

Departamento de Engenharia Civil
Instituto Superior Técnico
Av. Rovisco Pais
1096 Lisboa - Portugal

Abstract

In this paper probabilistic models to be applied to seismic hazard analysis in Portugal are evaluated.

To characterize the seismic process of occurrence, data on instrumental and historical earthquake catalogue, for Portuguese region, were considered. The geographical area of analysis was subdivided into seismic source zones, taking into consideration the earthquake history, and the knowledge about tectonic processes causing seismic activity. In each zone the instrumental earthquake time sequence, without aftershocks, was modelled as a Poisson process. The Gutenberg-Richter law was the magnitude-recurrence relationship used to describe the frequency of earthquakes within a range of different sizes or magnitudes over a period of time.

To characterize the strong ground motion process the Data Base on Macroseismic Information of Continental Portugal was used. The macroseismic intensity was selected as the ground motion parameter to be used as dependent variable in attenuation relationships. The segmentation of data considering source mechanisms and site soil conditions was attempted. The attenuation models were estimated applying multiple regression techniques.

Finally the evaluated models were applied to assess the seismic hazard in Portugal and the results were presented as seismic hazard curves and maps.

Resumo

Apresentam-se neste artigo modelos probabilísticos para serem aplicados à análise da casualidade sísmica em Portugal.

Para caracterizar o processo de ocorrência consideraram-se os sismos históricos e instrumentais do Catálogo da região portuguesa. A região em análise foi subdividida em zonas de geração sísmica, com base na sua história sísmica e em informação geotectónica. Em cada zona de geração modelou-se a sucessão cronológica dos sismos, sem réplicas, através de um processo de Poisson. Utilizou-se a relação frequência-magnitude de Gutenberg-Richter para descrever a frequência com que ocorrem sismos para as várias gamas de magnitude num período de tempo especificado.

Para caracterizar o processo dos movimentos sísmicos intensos recorreu-se à Base de Dados de Informação Macrossísmica de Portugal Continental. Escolheu-se a intensidade macrossísmica para parâmetro do movimento do solo a usar como variável dependente nos modelos de atenuação. Ensaiou-se a segmentação dos dados considerando o mecanismo do sismo e o solo do local em análise e estimaram-se os modelos de atenuação recorrendo a técnicas de regressão múltipla.

Aplicaram-se os modelos obtidos para a avaliação da casualidade sísmica em Portugal continental, apresentando-se os resultados sob a forma de curvas e mapas de casualidade sísmica.

Keywords

Probabilistic Analysis, Seismic Hazard, Attenuation Relationships, Seismic Occurrence Process, Linear Regression.

1. Introdução

A *análise probabilística da casualidade sísmica* é o domínio científico em que se analisam as distribuições de probabilidade de um dado efeito de um sismo num local, ou conjunto de locais, durante um período de tempo especificado. Os efeitos do sismo usualmente considerados nesta análise são os movimentos do solo, por serem as principais causas de danos durante um sismo. Estes efeitos podem ser traduzidos por diversas grandezas; entre elas tem-se os valores máximos da aceleração, velocidade e deslocamento do solo, ou a intensidade macrossísmica.

O termo *casualidade sísmica* é o termo recentemente empregue em Portugal para designar o conceito de *seismic hazard*, reservando-se o termo *risco sísmico*, para traduzir o conceito de *seismic risk*.

Entende-se por *análise do risco sísmico* o cálculo da probabilidade de uma perda específica (por exemplo número de mortos, valores monetários, etc.) exceder um dado valor quantificável durante um período de exposição especificado. Esta análise engloba três elementos fundamentais: **(i)** a análise probabilística da casualidade sísmica e período de exposição para o qual foi calculada, **(ii)** o valor do "portfolio" analisado e **(iii)** a vulnerabilidade do mesmo.

Este trabalho tem o objectivo de estabelecer modelos probabilísticos que sirvam de base à análise da casualidade sísmica em Portugal e assim contribuir para a mitigação do risco sísmico no país. Neste sentido, a modelação incidiu sobre dois aspectos fundamentais para a análise da casualidade sísmica: o processo sísmico de ocorrência e o processo dos movimentos sísmicos intensos.

Entre os diversos métodos existentes, determinísticos e probabilísticos, de análise da casualidade sísmica, adoptou-se o método de Cornell (Cornell, 1971) que é o de utilização mais frequente a nível mundial. Segundo este autor, o estudo da casualidade sísmica de uma região inserida num ambiente sismotectónico visa a determinação, em vários locais dessa região, das probabilidades de excedência de um dado nível de intensidade dos movimentos sísmicos. Esta metodologia será apresentada resumidamente na secção seguinte.

2. Metodologia para a Avaliação da Casualidade Sísmica

2.1 Zonas de Geração Sísmica

Considere-se um dado local que pode ser afectado por sismos que ocorrem numa dada região sísmica. O primeiro passo da metodologia consiste em dividir essa região em *zonas de geração sísmica*, as quais delimitam regiões que partilham as mesmas características sismológicas, tectónicas e geológicas. Neste contexto, uma zona de geração sísmica representa uma região da crusta terrestre nas quais são identificadas, ou não, falhas activas onde os sismos podem ser gerados, sendo zonas aproximadamente homogéneas no que respeita às distribuições que caracterizam a sua actividade sísmica. Os limites das zonas de geração não deverão segmentar estruturas neotectónicas identificadas e deverão basear-se na correlação de dados geológicos, fundamentalmente tectónicos, com dados de sismicidade.

Uma vez construído o modelo de zonas de geração, a exposição de determinado local à acção dos sismos resultará da soma da casualidade associada a cada uma das zonas.

2.2 As Ocorrências no Espaço

Nas regiões de sismicidade marcadamente difusa, como é o caso da maior parte das possíveis zonas de geração identificadas na região continental portuguesa, em que é difícil estabelecer uma associação entre as estruturas tectónicas activas e os epicentros, admite-se que a sismicidade se distribui uniformemente dentro de cada zona. Desta forma, assume-se que a localização dos epicentros é equiprovável em qualquer ponto de uma dada zona de geração sísmica, utilizando-se a distribuição uniforme bidimensional para modelar a distribuição de ocorrências no espaço.

2.3 As Ocorrências no Tempo

Entre os modelos estocásticos utilizados para modelar a ocorrência temporal dos sismos os mais frequentes são os modelos de Poisson, os de Poisson não homogéneos, e os do tipo Markoviano.

Optou-se pelo modelo de Poisson homogéneo devido à simplicidade da sua formulação e aplicação e ainda ao facto do algoritmo computacional de avaliação da casualidade sísmica a que se recorre neste trabalho assumir este modelo.

Num processo de Poisson homogéneo as ocorrências são independentes ao longo do tempo e a taxa média de ocorrências, λ, é constante. A probabilidade de ocorrência de um sismo em qualquer intervalo de tempo de comprimento fixo é a mesma independentemente do instante da última ocorrência.

2.4 A Distribuição de Magnitudes

A análise dos catálogos sísmicos de uma determinada região permite obter a relação entre a frequência com que ocorrem os sismos e as respectivas magnitudes, para um período de tempo dado.

A relação de frequência-magnitude que mais se utiliza a nível mundial foi desenvolvida por Gutenberg e Richter em 1944 (Gutenberg *et al.*, 1944). O modelo proposto por estes autores estabelece uma dependência linear entre o logaritmo da frequência de ocorrências de sismos e as suas magnitudes, para uma dada região, k, que se presume homogénea em termos de sismicidade:

$$\log N_k(m) = a_k + b_k \cdot m, \qquad (1)$$

em que $N_k(m)$ é o número de sismos com magnitude maior ou igual a m que ocorre na região analisada, para um dado período de observação. O coeficiente a_k é conhecido por *actividade sísmica* da região e está relacionado com a taxa de ocorrência total no período de observação considerado. O coeficiente b_k descreve a taxa relativa de ocorrência entre sismos de maior e menor magnitude.

A expressão (1) escreve-se usualmente na forma:

$$N_k(m) = e^{\alpha_k - \beta_k \cdot m}, \quad (2)$$

sendo portanto $\alpha_k = \ln(10) \cdot a_k$ e $\beta_k = -\ln(10) \cdot b_k$.

Assumindo que a grandeza de eventos sucessivos de uma dada zona de geração são independentes, e que os sismos com magnitude inferior a m_0, que ocorrem na zona de geração k, contribuem pouco para os valores finais das intensidades dos movimentos sísmicos no local, então a distribuição cumulativa de probabilidades das magnitudes para essa zona, $F_{M(k)}(m)$, que se adequa à relação de Gutenberg-Richter truncada inferiormente escreve-se (Araya *et al.*, 1988):

$$F_{M(k)}(m) = P(M < m \mid M \geq m_0) = \frac{N_k(m_0) - N_k(m)}{N_k(m_0)} = 1 - e^{-\beta_k(m-m_0)}, \; m \geq m_0, \quad (3)$$

em que m_0 é o limiar mínimo de magnitude abaixo do qual se considera que um sismo não causa estragos do ponto de vista da engenharia.

A respectiva função densidade de probabilidade vem dada por:

$$f_{M(k)}(m) = \frac{d}{dm} = F_{M(k)}(m) = \beta_k e^{-\beta_k(m-m_0)}, \; m \geq m_0. \quad (4)$$

A relação de frequência-magnitude utilizada usualmente em estudos de casualidade sísmica é a relação de Gutenberg-Richter truncada superiormente. Esta tem o objectivo de incorporar no modelo uma magnitude máxima designada por m_1, ou de magnitude do *sismo máximo provável*. A existência de um valor de magnitude que não pode ser ultrapassado é fundamentada em estudos de sismotectónica e paleosismicidade. Quando se impõe um limite máximo para a magnitude de dada zona de geração, a distribuição de magnitudes acumulada correspondente à relação de Gutenberg-Richter, é modificada para:

$$F_{M(k)}(m) = P(M < m \mid m_0 \leq M \leq m_1) = \frac{1 - e^{-\beta_k(m-m_0)}}{1 - e^{-\beta_k(m_1-m_0)}}, \; m_0 \leq m \leq m_1 \quad (5)$$

e a função densidade de probabilidade correspondente vem dada por:

$$f_{M(k)}(m) = \frac{\beta_k e^{-\beta_k(m-m_0)}}{1 - e^{-\beta_k(m_1-m_0)}}, \; m_0 \leq m \leq m_1. \quad (6)$$

2.5 Atenuação da Intensidade do Movimento Sísmico

A intensidade do movimento do solo aumenta com a energia libertada na fonte (magnitude do sismo) e diminui geralmente com a distância ao epicentro, uma vez que as ondas sísmicas são atenuadas no meio em que se propagam.

Uma *lei de atenuação* é uma relação empírica que exprime a dependência entre a intensidade do movimento do solo e uma série de variáveis explicativas, nomeadamente a distância à fonte e a energia nela libertada. A forma geral de um modelo de atenuação é:

$$Y = m_Y(M,R,w) + \varepsilon_Y, \quad (7)$$

em que Y é a intensidade do movimento do solo que se pretende prever, $m_Y(M,R,w)$ é o valor estimado pelo modelo matemático; R é a distância entre a fonte e o local em análise; M é a variável que descreve a grandeza do sismo que pode ser a magnitude ou a intensidade epicentral;

w é um vector de variáveis que pode caracterizar a fonte, a propagação das ondas no meio, ou mais usualmente as condições do local onde é medida a intensidade; ε_Y é uma variável aleatória que representa as flutuações em Y não explicadas pelas variáveis do modelo, bem como os erros de medição da variável dependente. Admite-se que os valores de ε_Y se distribuem independentemente uns dos outros, são independentes das variáveis explicativas do modelo e seguem uma distribuição normal de média nula e variância constante: $\varepsilon_Y \sim N(0,\sigma_{\varepsilon_Y})$. Desta forma assume-se que a distribuição da variável Y em torno do valor central, $\bar{Y} = m_Y(M,R,w)$, é gaussiana e independente da grandeza do sismo e da distância do seu epicentro ao local analisado.

Devido à carência de registos instrumentais para a região em estudo, optou-se pela utilização da intensidade macrossísmica, ($Y \equiv I$) como variável dependente nos modelos de atenuação.

A maioria das leis de atenuação da intensidade macrossísmica habitualmente utilizadas são equivalentes ou são casos especiais da equação:

$$I = c_1 + c_2 \cdot M + c_3 \cdot \ln(R) + c_4 \cdot R, \tag{8}$$

em que, pelas razões físicas atrás indicadas, o coeficiente c_2 deverá ser positivo e os coeficientes c_3 e c_4 deverão ser negativos.

2.6 Modelo Matemático para a Análise Probabilística da Casualidade Sísmica

O modelo matemático desenvolvido por Cornell (Cornell, 1971), para o cálculo da casualidade sísmica baseia-se no *teorema da probabilidade total*:

$$P(A) = \int_X P(A|x) f_X(x) dx, \tag{9}$$

em que A é o acontecimento cuja probabilidade se pretende calcular e X é um vector de variáveis aleatórias contínuas das quais A depende.

No caso particular do cálculo da casualidade sísmica, o acontecimento A representa o facto da intensidade de um efeito qualquer do sismo, designada por Y, exceder um dado nível de intensidade y, num dado local durante um sismo, ou seja, $A \equiv Y > y$, e a integração é feita para todos os valores de X para os quais a intensidade Y excede y.

As variáveis aleatórias do vector X, do qual Y depende, são aquelas que caracterizam o sismo, desde a sua origem até ao local em estudo, bem como a sua interacção com o local em estudo, ou seja, o vector X inclui todas as variáveis explicativas contabilizadas no modelo geral de atenuação (expressão 7). Em geral, a escolha das variáveis aleatórias contidas no vector X, recai, exclusivamente, sobre a magnitude M e a distância hipocentral R.

Considere-se que os acontecimentos capazes de afectar o local em análise ocorrem numa região sísmica constituída por n zonas de geração, sendo cada uma delas designada arbitrariamente pela zona k. Considere-se ainda que um sismo, com origem nessa zona k, é capaz de causar o efeito genérico traduzido pela intensidade macrossísmica I. Assumindo que as variáveis aleatórias do vector X são estatisticamente independentes, então o teorema da

probabilidade total permite calcular a probabilidade de excedência de um nível I_0 de intensidade de referência:

$$P(I > I_0)_{(k)} = \int_R \int_M P(I > I_0|m, r)_{(k)} f_{M(k)}(m) f_{R(k)}(r) dm dr. \quad (10)$$

Para se calcular a taxa média, ω_k, de ocorrência de sismos na zona de geração k que originam no local intensidades superiores ou iguais a um determinado nível de referência I_0, bastará multiplicar a probabilidade dada pela expressão anterior por λ_k, que representa o número médio de ocorrências, na unidade de tempo, nessa zona de geração sísmica, ou seja:

$$\omega_k = \lambda_k \int_R \int_M P(I > I_0|m, r)_{(k)} f_{M(k)}(m) f_{R(k)}(r) dm dr. \quad (11)$$

Saliente-se que nestas expressões, a probabilidade condicional dada por $P(I > I_0|m, r)_{(k)}$ depende, exclusivamente, dos diferentes tipos de leis de atenuação das intensidades dos movimentos sísmicos, em cada zona de geração, e das incertezas a elas associadas (equação 7). A função densidade de probabilidade $f_{M(k)}(m)$, ou distribuição da grandeza do sismo, é deduzida a partir das relações de frequência-magnitude, determinadas para cada zona de geração (equação 6). A função $f_{R(k)}(r)$ é a função densidade de probabilidade da distância hipocentral obtida a partir da distribuição espacial dos epicentros da zona k e da relação desta zona com o local em análise.

O cálculo final da distribuição de probabilidade de serem excedidos, pelo menos uma vez, determinados níveis de intensidade I_0, na unidade de tempo, devido à ocorrência aleatória de sismos em qualquer das n zonas de geração que contribuem para a sismicidade sentida no local, baseia-se na propriedade do processo de Poisson de não ser afectado pela agregação de processos de Poisson independentes.

Assim, supondo que o processo de ocorrências no tempo é independente de zona para zona, isto é, o facto de ter ocorrido um sismo numa determinada zona de geração não condiciona o processo de ocorrências em qualquer outra zona de geração, então, as taxas de ocorrência poderão ser somadas para obter a taxa de ocorrência de um processo geral de Poisson do acontecimento $I > I_0$:

$$P(I > I_0) = 1 - e^{-\sum_{k=1}^{n} \omega_k}. \quad (12)$$

Neste contexto, o acontecimento $I > I_0$ corresponde a verificar-se, pelo menos uma vez, no local em análise, uma intensidade superior ao nível de referência I_0.

Quando a unidade de tempo é o ano a expressão anterior fornece a probabilidade anual de excedência. Finalmente, define-se *período de retorno*, RP_0, como sendo o inverso da probabilidade anual de ser excedido, pelo menos uma vez, no local em análise um determinado nível de intensidade, I_0, sendo aquela probabilidade dada pela expressão (12).

Na figura 1 apresentam-se de forma esquemática as diversas etapas descritas ao longo desta secção que conduzem ao cálculo da casualidade sísmica por métodos probabilísticos.

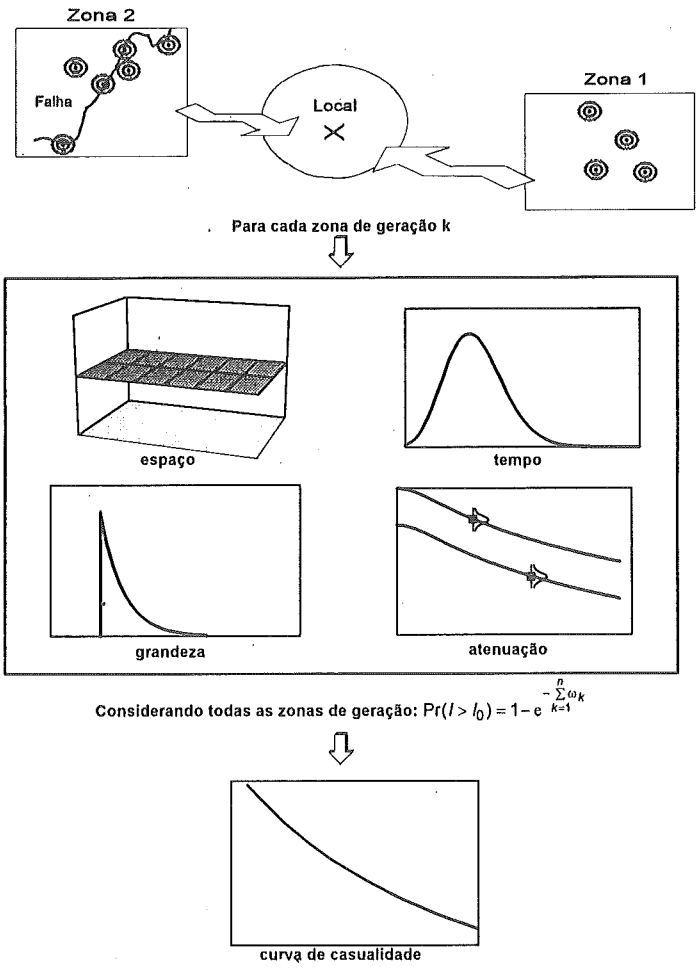

Figura 1 - Esquema do procedimento geral para a determinação da casualidade sísmica por métodos probabilísticos.

3. Modelação do Processo Sísmico de Ocorrência

3.1 Os Dados de Base

A caracterização da sismicidade de uma dada região, passa pelo conhecimento da história sísmica da mesma, nomeadamente do instante de ocorrência de cada sismo, da sua localização, intensidade e outras características. Esta informação encontra-se compilada nos catálogos sísmicos locais, regionais e mundiais. Para completar a informação de natureza sísmica deverá ser possível identificar as falhas activas da região em estudo, mesmo aquelas de actividade reduzida, uma vez que também estas poderão contribuir, embora com uma probabilidade menor, para a ocorrência de sismos. Com esta informação deverá ser possível delinear, na região em estudo, as zonas de geração sísmica e calcular os parâmetros das distribuições que as caracterizam.

Para caracterizar o processo de ocorrência foi utilizado o Catálogo Sísmico da Região Ibérica (Sousa *et al.*, 1992) e informação sobre a neotectónica da região de Portugal continental (Cabral, 1993). Este tipo de informação permite concluir que os sismos que afectam o território português possuem dois tipos de mecanismos bem diferenciados: os *sismos interplacas* que resultam da interacção da fronteira oceânica das placas Africana e Euroasiática e os *sismos intraplaca* associados à actividade sísmica em falhas no interior da placa Euroasiática. O Catálogo seleccionado para estudar a casualidade sísmica de Portugal continental contem 5394 sismos e o primeiro evento nele registado ocorreu no ano 33 d.C., embora a sua magnitude e epicentro sejam desconhecidos. Na figura 2 apresenta-se uma mapa de epicentros da selecção efectuada para o presente estudo.

Para assegurar que a distribuição de magnitudes do Catálogo não se encontra enviesada quer pelo desconhecimento da magnitude dos sismos das épocas históricas, quer pelo facto do Catálogo não ser completo na gama de magnitudes baixas nas épocas antigas, recorreu-se a um processo de conversão de intensidades macrossísmicas em magnitudes (Costa, 1989) e à filtragem do Catálogo para magnitudes baixas, respectivamente. Assim, os 3735 registos do Catálogo que possuíam informação de magnitude (ou magnitude convertida a partir de intensidade macrossísmica) foram reduzidos para 2351 em consequência do processo de filtragem das magnitudes baixas. De salientar que, entre os registos eliminados, 1525 possuíam magnitude especificada inferior a 3.05, sendo por isso de contribuição pouco importante para a casualidade sísmica do país.

3.2 Modelo de Zonas de Geração Sísmica

Para desenvolver o modelo actual de zonas de geração confrontou-se o ambiente neotectónico e informação geológica diversa, com os dados de sismicidade histórica e instrumental. Além disso, foram analisados diversos modelos já publicados, tendo em atenção que dadas as incertezas existentes nos conhecimentos geotectónicos e nos dados de sismicidade, o traçado de um modelo de zonas de geração constitui um problema sem solução única, verificando-se modelos muito díspares de autor para autor.

O modelo de geração adoptado no presente trabalho (figura 2) é constituído por doze zonas, sendo dez de grande expressão geográfica e duas de menor dimensão. A individualização destas duas zonas, a 5A e a 6A, prende-se com o facto de apresentarem uma grande concentração espacial de sismos e de estarem associadas às principais estruturas que historicamente têm produzido os maiores sismos sentidos no continente.

A maioria das zonas foram modeladas por áreas de grandes dimensões, devido à dificuldade em relacionar os epicentros com as falhas cartografadas na carta neotectónica. As zonas de geração assim delineadas podem classificar-se em duas grandes categorias: as que originam sismos maioritariamente na fronteira de placas, designadas de zonas interplacas (zonas 1, 6-6A, 6A, 8 e 9), e as que originam sismos predominantemente no interior da placa Euroasiática, designadas por zonas intraplaca (zonas 2, 3, 4, 5-5A, 5A, 7 e 10).

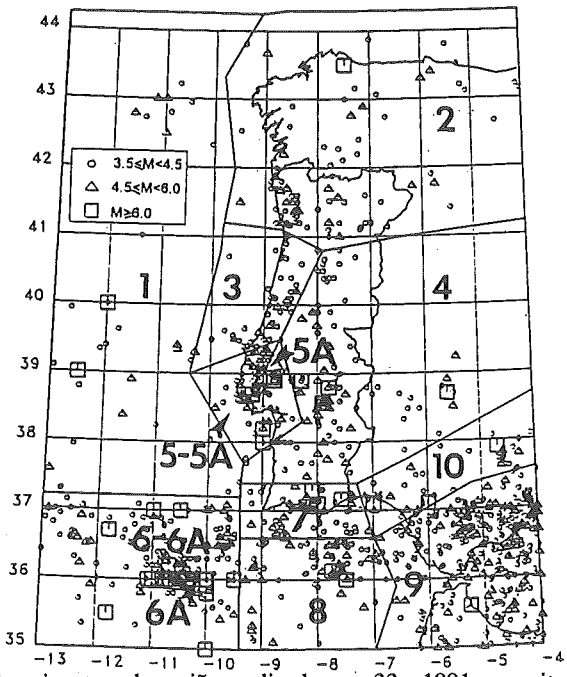

Figura 2 - Mapa de epicentros da região analisada, ano 33 a 1991, magnitudes superiores ou iguais a 3.5 e modelo de zonas de geração sísmica adoptado no presente trabalho.

3.3 Estudo da Ocorrência no Tempo

Apesar do processo de filtragem das magnitudes baixas ter sido aplicado na tentativa de obter um catálogo "completo", o número de sismos no período instrumental (período posterior a 1909) é ainda consideravelmente superior ao do período histórico. Por essa razão o processo de ocorrência no tempo foi estudado exclusivamente para o período instrumental, e para magnitudes superiores a 3.45, sendo depois extrapolado para o histórico. Desta forma, está a admitir-se a estacionaridade do processo temporal de ocorrência.

3.3.1 Eliminação de Réplicas e Premonitores

A caracterização da ocorrência temporal de sismos de acordo com o processo de Poisson pressupõe que a sequência de acontecimentos num catálogo sísmico seja aleatória, e que nenhuma ocorrência seja afectada pelos acontecimentos passados nem venha a afectar os acontecimentos futuros. Assim, as réplicas e os fenómenos premonitores deverão ser identificados e eliminados do Catálogo, admitindo-se por isso a simplificação de modelar apenas a sucessão de choques principais.

O processo adoptado para eliminar réplicas é empírico, embora calibrado para a situação portuguesa. Deste modo, um sismo é considerado uma réplica de um sismo principal, se obedecer simultaneamente às seguintes condições (implementadas num programa de cálculo automático):

- A magnitude da réplica deverá ser inferior à magnitude *M* do sismo principal que a precede.

- O intervalo de tempo entre ocorrências deverá ser inferior a T, sendo esse intervalo dependente da magnitude do sismo principal, $T \equiv T(M)$.
- A distância entre os epicentros da réplica e do sismo principal, não deve exceder D. Essa distância também depende da magnitude do sismo principal, $D \equiv D(M)$.

As "janelas" de magnitude, tempo e distância que permitem identificar as réplicas apresentam-se no quadro 1, tendo-se diferenciado as "janelas" da distância entre as zonas intraplaca e interplacas.

Ainda com o propósito de procurar garantir a aleatoriedade da sequência de acontecimentos no catálogo sísmico deveriam ser também eliminados os sismos premonitores. Os premonitores foram eliminados manualmente dos subcatálogos das zonas que, após a eliminação das réplicas, se verificaram ser não Poissonianas (ver secção 3.3.2). Essa eliminação apenas foi efectuada nos casos em que não existiam dúvidas de que os sismos eram de facto premonitores. Por exemplo, na zona 2 foram eliminados 8 premonitores.

Magnitude do sismo principal	Intervalo de tempo (dia)	Distância (km) Zonas intraplaca	Distância (km) Zonas interplacas
$0.0 \leq M < 3.5$	6	15.0	25.0
$3.5 \leq M < 4.5$	12	25.0	35.0
$4.5 \leq M < 5.0$	30	35.0	45.0
$5.0 \leq M < 5.5$	120	40.0	50.0
$5.5 \leq M < 6.0$	280	47.0	57.0
$6.0 \leq M < 6.5$	495	54.0	64.0
$6.5 \leq M < 7.0$	730	61.0	71.0
$7.0 \leq M < 7.5$	830	70.0	80.0
$7.5 \leq M < 8.0$	860	81.0	91.0
$M \geq 8.0$	890	94.0	104.0

Quadro 1 - Critério adoptado para eliminação de réplicas.

3.3.2 Ajustamento do Processo de Poisson ao Processo de Ocorrências no Tempo

Como é sobejamente conhecido, se num dado intervalo de tempo as ocorrências seguem um processo de Poisson, então os intervalos de tempo entre duas ocorrências consecutivas são independentes e distribuem-se exponencialmente. Para modelar o processo de ocorrências no tempo optou-se pela variável aleatória intervalo de tempo entre ocorrências sucessivas, pois do ponto de vista analítico é mais fácil lidar com a distribuição exponencial, do que com a de Poisson.

Recorreu-se aos testes de ajustamento do χ^2 e de Kolmogorov-Smirnov, para averiguar, em cada zona de geração, a validade da hipótese de que *a amostra dos intervalos de tempo entre sismos consecutivos foi extraída de uma população que segue a distribuição exponencial, com média igual à da amostra.*

Para um nível de significância de 5%, a aplicação das "janelas" do quadro 1 e a eliminação de premonitores conduziu à rejeição da hipótese nas zonas 5A, 6-6A, 8 e 9, com ambos os testes.

Ensaiaram-se sucessivamente diversas "janelas" para eliminação de réplicas sobre estas quatro zonas 5A, 6-6A, 8 e 9, sem fenómenos premonitores, na tentativa de obter catálogos residuais cujos intervalos de tempo entre ocorrências seguissem uma distribuição exponencial. Os diversos ensaios, relatados em pormenor em Sousa (Sousa, 1996), foram mal sucedidos.

No quadro 2 apresentam-se as taxas anuais de ocorrência de sismos, λ_k, em cada zona de geração.

Zona	λ_k	Zona	λ_k	Zona	λ_k
1	0.50	5-5A	0.70	7	0.77
2	1.14	5A	0.26	8	0.86
3	0.54	6-6A	1.59	9	3.97
4	1.37	6A	0.71	10	0.57

Quadro 2 - Valores do parâmetro λ_k (nº de sismos/ano) necessários à quantificação do processo temporal de ocorrência em cada zona de geração.

Para que um processo estocástico seja Poissoniano, além da distribuição dos intervalos de tempo entre ocorrências consecutivas ser exponencial, é necessário que se verifique a independência entre intervalos.

Foram realizados dois testes de *runs* para verificar se a sequência ordenada de intervalos de tempo é aleatória. Como resultado dos testes realizados, para um nível de significância de 5%, concluiu-se pela não rejeição da hipótese de aleatoriedade, com excepção das zonas 9, 10, todas as zonas e sismicidade remanescente[1].

Em conclusão e face aos resultados obtidos, verifica-se que o processo de ocorrências no tempo só pode ser considerado Poissoniano em 7 das 12 zonas analisadas.

3.4 Modelo de Frequência-Magnitude

Nesta secção estima-se a frequência com que ocorrem sismos para as várias gamas de magnitude, calculando-se a_k e b_k, parâmetros da expressão 1, para cada zona de geração. Para todas as zonas de geração, escolheu-se o valor de magnitude 3.45 para o parâmetro m_0 das expressões 3 a 6.

Para além dos limites inferiores de truncatura, é possível identificar, em cada zona de geração, a magnitude máxima que um sismo, ocorrendo nas estruturas activas da região, poderá originar. Esse evento, atrás designado de *sismo máximo provável*, é utilizado para a truncatura superior, m_1, da distribuição de probabilidade da magnitude (equações 5 e 6). Para estabelecer o valor das magnitudes máximas, m_1, que cada zona tem o potencial de gerar, adoptou-se a

[1] Sismicidade remanescente é a sismicidade que não pode ser atribuída a uma falha específica ou a uma zona de geração. Normalmente corresponde a sismos de baixa magnitude e neste trabalho corresponde a sismos de magnitude inferior a 3.45.

magnitude do sismo mais intenso que ocorreu no passado, com epicentro na zona analisada. É óbvio que este procedimento avalia por defeito o valor da magnitude do sismo máximo provável, pois no passado poderão ter ocorrido sismos de maior magnitude não documentados.

Conhecidos os limites de truncatura e seleccionando do Catálogo filtrado, histórico e instrumental, os sismos com magnitude superior a 3.45, estimaram-se pelo método dos mínimos quadrados, para cada zona de geração, os parâmetros a_k e b_k da relação de frequência-magnitude de Gutenberg-Richter (expressão 1), ou o parâmetro β_k da distribuição de probabilidade da grandeza do sismo truncada superiormente, (expressões 5 e 6).

Resumindo os resultados dos testes estatísticos associados a estas estimativas, verifica-se que existe uma associação linear muito forte entre o logaritmo do número acumulado de sismos e a magnitude, estando estas variáveis correlacionadas negativamente, como seria de esperar. Os testes F à qualidade global do modelo são significativos a 5%, para todas as zonas, verificando-se o mesmo para os testes t à significância dos parâmetros a_k e b_k do modelo de regressão linear. A proporção da variação total da variável dependente explicada pela recta de regressão é sempre superior a 90%, com excepção das zonas 6A e 8 em que é cerca de 88% (quadro 3).

No quadro 3 reúnem-se os parâmetros estimados para caracterizar as distribuições de magnitude de cada zona de geração, à excepção dos valores das truncaturas inferiores das mesmas que se estabeleceu serem de 3.45. A actividade apresentada, $\hat{a}_k$, corresponde a esta magnitude.

Zona	M_{max}	Data do sismo	$\hat{a}_k$ (M=3.45)	$\hat{b}_k$	R^2
1	7.0	1724.10.12	0.4908	-0.6636	0.9776
2	6.0	1916.12.03	1.6329	-0.8415	0.9702
3	5.6	1940.10.03	0.5825	-0.8940	0.9510
4	7.0	1504.04.05	1.7833	-0.8370	0.9573
5-5A	7.2	1858.11.11	0.9388	-0.9497	0.9664
5A	7.0	1531.01.26	0.3148	-0.7585	0.9175
6-6A	6.6	1915.07.11	1.6352	-0.6442	0.9769
6A	8.5	1755.11.01	0.1471	-0.3373	0.8839
7	7.8	1722.12.17	1.3004	-0.9213	0.9147
8	7.1	1964.03.15	0.8735	-0.6431	0.8811
9	6.2	1909.01.21	9.7550	-1.2233	0.9304
10	7.0	1719.03.06	0.7180	-0.8664	0.9619
Sismic. Reman.	3.5	--	12.0413	-1.3879	0.9586

Quadro 3 - Magnitude máxima de cada zona, data do sismo que lhe corresponde, valores de $\hat{a}_k$ e $\hat{b}_k$ da relação de Gutenberg-Richter e coeficiente de determinação.

A relação de Gutenberg-Richter da zona de geração 9 corresponde à recta de maior inclinação ou, por outras palavras, ao maior valor absoluto de $\hat{b}_k$, excepção feita para o valor da sismicidade remanescente. É também na zona 9 que se verifica a maior taxa de actividade sísmica, excluindo-se mais uma vez a actividade correspondente à sismicidade remanescente.

Por outro lado, a zona 6A é aquela que tem o valor de $\hat{b}_k$ mais reduzido. Isto significa que a proporção entre choques de magnitude elevada relativamente aos de menor magnitude é maior, do que quando a relação de Gutenberg-Richter corresponde à recta mais inclinada. É natural que o menor valor de $\hat{b}_k$ surja na zona do Gorringe (6A), pois esta zona de geração engloba as estruturas que produzem os maiores sismos que afectaram o continente português.

4. Modelação do Processo dos Movimentos Sísmicos Intensos

4.1 Os Dados de Base

Os dados utilizados para estabelecer as leis de atenuação encontram-se armazenados na *Base de Dados de Informação Macrossísmica de Portugal Continental* (Oliveira *et al.*, 1995 e Paula, 1994).

A Base de Dados armazena, até à presente data, a informação macrossísmica relativa a 199 sismos, históricos e instrumentais, que foram sentidos em Portugal e dos quais se conhece o epicentro. Dos 199 sismos, a grande maioria (194) ocorreu no período delimitado pelos anos de 1947 e 1993. Os sismos restantes são alguns dos sismos históricos mais importantes que afectaram Portugal (por exemplo os sismos de 1531, 1755 e 1909) tendo sido incorporados na Base com o objectivo de aumentar a gama de magnitudes por ela coberta.

A Base de Dados assenta em duas tabelas básicas, SISMOS e LOCALIDADES, ligadas entre si através da tabela INTENSIDADES, por relações do tipo '*um para muitos*': os efeitos de cada **sismo** são traduzidos por diversas **intensidades** cada uma delas sentida numa dada **localidade**.

A informação constante da tabela SISMOS permite concluir que a sismicidade de Portugal continental no período 1947-1993 se caracteriza maioritariamente por sismos pouco intensos, com magnitudes entre de 3.0 e 5.0.

A tabela INTENSIDADES armazena 3209 intensidades sentidas em várias localidades relativas a todos os sismos registados. Sem ter em conta os sismos históricos, os efeitos produzidos traduziram-se por intensidades que não excederam o grau IV da escala EMS-92 em 78% dos sismos. Os restantes 22% produziram danos em construções de uma ou mais localidades.

A tabela LOCALIDADES inclui 1194 *registos* correspondentes a outras tantas localidades onde foram avaliadas intensidades macrossísmicas.

Para classificar de forma simplificada a formação geológica superficial de cada localidade adoptaram-se três categorias de solos, o *brando*, o *intermédio* e o *rijo*. Grosso modo, o solo brando corresponde a depósitos aluvionares e o rijo corresponde a rocha.

4.2 Estimação dos Modelos de Atenuação

Os parâmetros das leis de atenuação (expressão 8) são estimados, para cada zona de geração, por análise de regressão linear múltipla a partir dos registos macrossísmicos que constam da Base de Dados.

Os modelos a estimar possuem duas variáveis explicativas: a *magnitude*, que descreve a grandeza do sismo, M e a *distância focal* que mede a distância entre a fonte e o local em análise, R.

Os modelos de atenuação foram estimados com a informação constante da Base de Dados, mas são utilizados para a gama de magnitudes que ocorreram no Catálogo Sísmico, mais precisamente para o intervalo compreendido entre a magnitude 3.45 e a magnitude do sismo máximo provável de cada zona. Embora se tivesse acrescentado à Base as observações relativas a 5 sismos históricos severos, as magnitudes máximas observadas no Catálogo, são, em geral, bastante superiores às observadas na Base, exceptuando-se os casos das zonas 5, 6 e 8. Este facto advém do Catálogo cobrir um período de tempo de cerca de 2000 anos, enquanto a Base apenas cobre exaustivamente um período de cerca de 50 anos.

Face à carência de dados de magnitude nas gamas elevadas e para evitar extrapolar os modelos para magnitudes não observadas, tomaram-se as seguintes opções:

- **(i)** O dados das zonas 5-5A e 5A e das zonas 6-6A e 6A foram agrupados apenas em duas zonas, a 5 e a 6. Assim para as zonas 5-5A e 5A estima-se um único modelo, com os dados das duas zonas agregadas. Analogamente para as zonas 6-6A e 6A.
- **(ii)** Nas zonas 2, 3, 4, 7 e 10, utilizaram-se os dados disponíveis do conjunto das zonas intraplaca, ou seja, do conjunto das zonas 2, 3, 4, 5 e 7.
- **(iii)** Nas zonas 1 e 9, utilizaram-se os dados disponíveis do conjunto das zonas interplacas, ou seja, do conjunto das zonas 1, 6 e 8.
- **(iv)** Nas zonas 5, 6 e 8, utilizaram-se, exclusivamente, os dados das zonas respectivas.
- **(v)** Relativamente à sismicidade remanescente, na qual foram integrados todos os sismos com magnitude inferior a 3.45, tal como na caracterização do processo de ocorrência, utilizaram-se exclusivamente os sismos com origem em zonas intraplaca.

Para além da segmentação dos dados devida ao mecanismo do sismo (zonas intraplaca e interplacas), ensaiou-se outro tipo de segmentação, a segmentação nos solos. Na Base de Dados existem 3209 observações de intensidades macrossísmicas, sendo a grande maioria em solo rijo (69%) e as restantes distribuídas pelo solo intermédio (14%) e solo brando (17%). A amostra foi segmentada em três categorias, correspondentes aos três tipos de solo, e ajustaram-se três modelos a cada conjunto de dados e um modelo adicional não considerando a segmentação de solos.

No quadro 4 apresentam-se os coeficientes do modelos de atenuação obtidos por regressão linear múltipla. No mesmo quadro apresentam-se o número de observações de magnitude e o

número de pares distância hipocentral - intensidade utilizados na regressão. Apresentam-se ainda, os erros globais das estimativas e os coeficientes de determinação múltipla ajustados.

Neste quadro a abreviatura "ns" indica que, de acordo com o teste t, o coeficiente da variável respectiva não é significativamente diferente de zero. Nesses casos a regressão foi efectuada de novo, sem entrar em consideração com a referida variável.

Dados zona	solo	nº obs. magn.	nº obs. int.-dist.	c_1	c_2	c_3	c_4	erro estim.	R^2 ajust.(%)
5	todos	11	854	0.9824	0.8554	-0.2335	-0.0064	1.09	35.7
	rijo	10	587	1.6983	0.7700	-0.2999	-0.0069	0.71	44.2
	interm.	9	132	2.5552	0.8978	-0.8404	ns	0.91	55.5
	brando	8	135	ns	1.2102	-0.4700	-0.0064	1.01	66.1
6	todos	11	578	7.7988	1.3376	-2.0167	ns	0.76	74.8
	rijo	10	401	8.2329	1.2529	-1.9995	ns	0.75	73.0
	interm.	8	72	4.1466	1.4592	-1.5493	ns	0.63	82.8
	brando	6	105	7.1649	1.4794	-2.0300	ns	0.79	78.5
8	todos	10	213	3.7374	0.7967	-0.8671	ns	0.61	50.8
	rijo	9	132	3.9522	0.7811	-0.8928	ns	0.59	53.3
	interm.	6	20	ns	0.6141	ns	ns		
	brando	8	61	3.4291	0.8317	-0.8306	ns	0.68	40.2
intra.	todos	19	1753	1.8819	0.7613	-0.3509	-0.0047	0.84	48.1
	rijo	19	1227	1.8479	0.6384	-0.1870	-0.0053	0.80	36.0
	interm.	18	255	2.9206	0.8010	-0.7568	ns	0.84	58.2
	brando	18	301	ns	1.0583	-0.1672	-0.0080	0.89	66.1
inter.	todos	19	917	ns	1.1625	-0.2682	-0.0035	0.85	70.5
	rijo	18	615	-0.8703	1.0826	ns	-0.0041	0.83	67.7
	interm.	14	111	ns	1.2827	-0.3608	-0.0043	0.80	77.8
	brando	16	181	-2.2724	1.3566	ns	-0.0046	0.85	76.7
sismic.reman.		3	255	4.3817	--	-0.3495	--	0.69	30.0

Quadro 4 - Modelos de atenuação das zonas 5, 6, 8, zonas intraplaca, zonas interplacas e sismicidade remanescente, com e sem segmentação de solos.

No mesmo quadro o símbolo "--" indica que o coeficiente correspondente tem um sinal fisicamente incorrecto, ou seja, tem um sinal negativo caso seja o coeficiente de M e tem um sinal positivo caso seja coeficiente de R ou de $\ln(R)$. Como a intensidade não cresce quando a magnitude decresce, nem quando a distância focal aumenta, a regressão foi efectuada de novo sem entrar em consideração com a variável respectiva. O modelo que possui maior capacidade explicativa é o da zona 6, solo intermédio, e o que possui menor poder explicativo é o relativo à

sismicidade remanescente, seguido dos modelos da zona 5, todos os solos e dos sismos intraplaca, solo rijo.

No que toca aos erros globais das estimativas, o menor erro é o do modelo da zona 8, solo rijo, atingindo cerca de meia unidade de intensidade macrossísmica, enquanto que os maiores erros são os dos modelos da zona 5, todos os solos e solo brando, atingindo cerca de uma unidade de intensidade macrossísmica. Em todos os modelos analisados, verifica-se que os do solo brando exibem sempre maior erro que os dos outros dois solos.

Na zona 8, solo intermédio, os coeficientes da variável distância hipocentral, ou do seu logaritmo, não se mostraram significativamente diferentes de zero, segundo o teste t para um nível de significância de 5%, não sendo por isso possível estabelecer modelos de atenuação com a distância. Neste caso o modelo de atenuação utilizado foi o da zona 8, todos os solos.

A representação dos modelos de atenuação para as diversas zonas e para algumas magnitudes fixas apresentada em Sousa (Sousa, 1996), permitiu concluir que os modelos confirmam as expectativas de que a intensidade macrossísmica cresce do solo rijo para o brando.

Uma outra forma de ajuizar da justeza destas leis é através da comparação das isossistas previstas pelos modelos com as isossistas de sismos reais da mesma magnitude. Tal comparação é ilustrada nas figuras 3 e 4 em que se representam os pares das isossistas reais dos sismos de 1909 e 1755 e das isossistas previstas pelos modelos de atenuação das zonas e magnitudes correspondentes. Em cada local onde existem valores reais de intensidades, o modelo de atenuação tomado é o correspondente ao tipo de solo que lhe foi atribuído.

No caso do sismo de 1909 (figura 3), existe uma razoável concordância entre o modelo proposto e as isossistas que se verificaram na realidade.

O modelo de atenuação da zona 6, conforme se mostra na figura 4, não consegue prever, no sul do país, intensidades tão elevadas quanto as reais do sismo de 1755, enquanto que, no norte, o modelo atenua menos rapidamente que as isossistas reais.

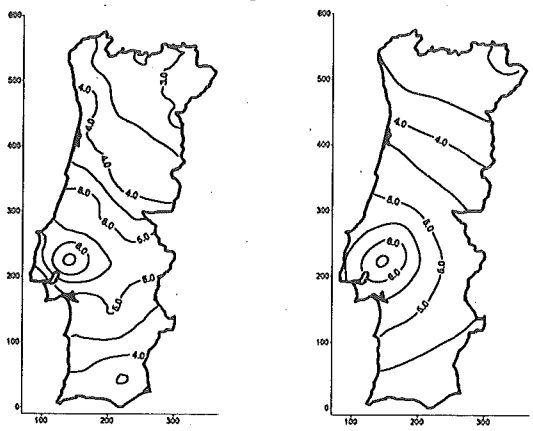

Figura 3 - a) Isossistas reais do sismo de 23 de Abril de 1909; b) Isossistas previstas pelos modelos de atenuação da zona 5, com segmentação de solos, magnitude 6.9.

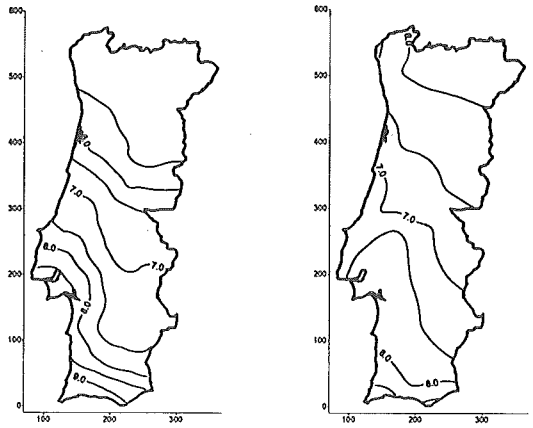

Figura 4 - a) Isossistas reais do sismo de 1 de Novembro de 1755; b) Isossistas previstas pelos modelos de atenuação da zona 6, com segmentação de solos para a magnitude 8.5.

5. Avaliação da Casualidade Sísmica em Portugal Continental

A avaliação da casualidade sísmica em Portugal continental decorre da aplicação do programa EQRISK, desenvolvido por McGuire (McGuire, 1976), que permite implementar computacionalmente o modelo matemático para a quantificação probabilística da casualidade sísmica, exposto na secção 2, com as estimativas dos parâmetros dos modelos probabilísticos obtidos nas secções 3 e 4.

Os resultados foram obtidos sob a forma de curvas de casualidade sísmica[2] para várias localidades do território continental e mapas de casualidade sísmica[3] para vários períodos de retorno.

Ilustram-se os resultados apresentando a curva de casualidade sísmica para a localidade de Lisboa (figura 5a) e o mapa de casualidade sísmica para o período de retorno de 1000 anos (figura 5b).

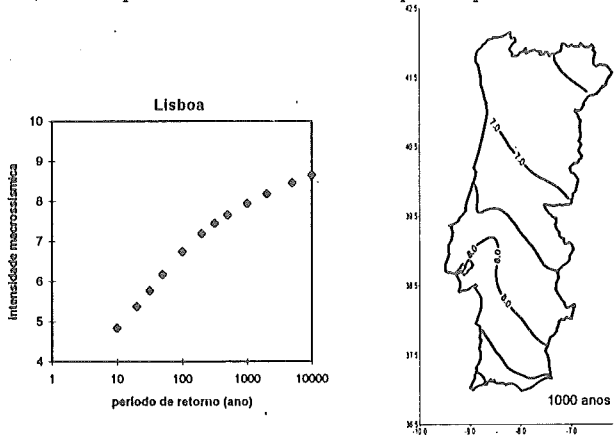

Figura 5 - a) Curva de casualidade sísmica para a localidade de Lisboa. b) Mapa de casualidade sísmica para o período de retorno de 1 000 anos.

2 *Curva de casualidade sísmica* - Função de distribuição da grandeza que traduz o efeito do sismo num dado local, por exemplo a representação das probabilidades de excedência de diversos níveis de intensidades macrossísmicas num dado local.

3 *Mapa de casualidade sísmica* - Mapa de isolinhas do efeito estudado para um período de retorno fixo.

6. Conclusões

6.1 Sobre o Processo Temporal de Ocorrência

Na maioria das zonas de geração é estatisticamente válido modelar o processo temporal de ocorrência aplicando a distribuição exponencial. Apenas nas zonas 5A, 6-6A, 8 e 9, e de acordo com os testes do χ^2 e de Kolmogorov-Smirnov, se conclui que as frequências observadas não são provenientes de uma população que se distribui exponencialmente.

Face a estes resultados três hipóteses se levantam: **(i)** ou as zonas de geração não foram delineadas correctamente e por isso não delimitam regiões que partilham as mesmas características sismológicas, **(ii)** ou as réplicas e premonitores não foram correctamente eliminados **(iii)** ou o processo não é efectivamente Poissoniano em algumas das zonas de geração.

Testando a aleatoriedade dos intervalos de tempo entre ocorrências sucessivas, concluiu-se que a hipótese de aleatoriedade não se verifica nas zonas 9, 10, todas as zonas e sismicidade remanescente.

Desta forma, conclui-se ser duvidoso que o processo temporal de ocorrência seja Poissoniano para o Catálogo filtrado, instrumental, sem réplicas e com magnitudes superiores a 3.45, pelo que os resultados obtidos deverão ser encarados com as devidas reservas face ao não cumprimento de algumas das hipóteses de base do modelo adoptado.

6.2 Sobre os Modelos de Atenuação

Os modelos de atenuação foram obtidos com e sem segmentação de solos. Apesar de não se verificarem grandes melhorias, em termos dos erros globais das estimativas, devido à introdução da segmentação de solos, foram estes os modelos utilizados no cálculo da casualidade sísmica, pois permitem obter melhores aproximações das isossistas previstas às reais. De facto, as isossistas previstas são bastante sensíveis à classificação macroscópica de solos existente na Base de Dados. Caso esta não fosse considerada, as isossistas previstas reduzir-se-iam a circunferências centradas no epicentro do sismo.

Em todos os modelos analisados, verifica-se que os do solo brando exibem sempre maior erro da estimativa que os dos restantes solos. Possivelmente esta categoria de solos apresenta em si maior variabilidade que as restantes.

Os principais resultados que ressaltam do estudo de atenuação são: **(i)** a segmentação em solos confirma as expectativas de que a intensidade cresce do solo rijo para o brando e **(ii)** que os modelos atenuam em geral menos do que seria expectável, à excepção dos sismos de magnitude elevada em que a atenuação é bastante mais acentuada. Foram ensaiados, em algumas zonas de geração, modelos de atenuação não lineares, tendo-se obtido algumas melhorias em termos dos erros das estimativas, pelo que este é um domínio que importa investigar.

6.3 Sobre os Resultados da Casualidade Sísmica

Os mapas de casualidade sísmica têm uma forma muito semelhante à das isossistas previstas para o sismo de 1755 (ver figura 4b). Daqui se conclui que a casualidade é controlada, em praticamente todo o país, pela sismicidade decorrente da zona 6, notando-se a primordial influência do sismo de 1755 que é o de magnitude mais elevada.

Referências

[1] Araya, R. e Der Kiureghian, A., *Seismic Hazard Analysis Improved Models, Uncertainties and Sensitivities, Report No.* UCB/EERC 90/11, University of California, Berkeley (1988).

[2] Cabral, J., *Neotectónica de Portugal Continental*, Tese de doutoramento em geologia, Universidade de Lisboa (1993).

[3] Cornell, C. A., *Probabilistic Analysis of Damage to Structures under Seismic Load*, Howell, D.A. Haigh, I. P. e Taylor, C., Dynamic Waves in Civil Engineering, London, Interscience (1971) 473-488.

[4] Costa, R., *Modelação do Processo Estocástico Sísmico na Península Ibérica*, Tese de Doutoramento em Engenharia de Sistemas, Universidade Técnica de Lisboa, Instituto Superior Técnico, Lisboa (1989).

[5] Gutenberg, B. e Richter, C.F., *Frequency of Earthquakes in California*, Bulletin of the Seismological Society of America 34 (1944) 185-188.

[6] McGuire, R. K., *EQRISK, Evaluation of Earthquake Risk to Site. Open File Report 76-67*, United States Department of the Interior Geological Survey (1976).

[7] Oliveira, C.S., Campos-Costa, A., Sousa, M.L., Martins, A., Paula, A., Guedes, J. e Lucas, A., *Estimativa dos Danos Causados por Sismos no Parque Habitacional do Continente Português. Contribuição para a Definição de uma Política de Seguros*, Estudo realizado para Associação Portuguesa de Seguradores, Lisboa (1995).

[8] Paula, A., *Avaliação de Informação Macrossísmica dos Sismos Sentidos em Portugal Continental no Período 1947-1993. Estudos de Atenuação*, Relatório de Estágio Profissionalizante da Licenciatura em Ciências Geofísicas. Universidade de Lisboa, Lisboa (1994).

[9] Sousa, M.L., *Modelos Probabilísticos para a Avaliação da Casualidade Sísmica em Portugal Continental*, Tese de Mestrado em Investigação Operacional e Engenharia de Sistemas. Instituto Superior Técnico. Universidade Técnica de Lisboa (1996).

[10] Sousa, M.L., Martins, A. e Oliveira, C.S., *Compilação de Catálogos Sísmicos da Região Ibérica*, Relatório 36/92 - NDA, Laboratório Nacional de Engenharia Civil, Lisboa (1992).

INVESTIGAÇÃO OPERACIONAL, vol. 17 (Junho 1997) 23-40

ESTUDO COMPARATIVO DA INFLUÊNCIA DAS CONDIÇÕES INICIAIS NUM MODELO DE SIMULAÇÃO DO PROCESSO DE OCORRÊNCIAS SÍSMICAS NA PENÍNSULA IBÉRICA

Maria Cecília Marques Rodrigues

Ruy Araújo da Costa

Departamento de Matemática
FCT - UNL
Quinta da Torre
2825 Monte Caparica - Portugal

Abstract

In this paper we present an analysis of the initial conditions of a seismic simulation model. The model admits that:

- the time interval between two occurrences (Dt_i) depends on the two proceding time intervals (Dt_{i-1}, Dt_{i-2})
- the magnitude (G_i) depends on the two preceding magnitudes (G_{i-1}, G_{i-2}), and the time interval between this occurrence and the preceding occurrence (Dt_i)
- the location of an occurrence (E_i) depends on the preceding location (E_{i-1}), the magnitude of the last occurrence (G_i) and the time interval between this occurrence and the last (Dt_i)

To start the simulation process in this model we need to assign values to Dt_{i-1}, Dt_{i-2}, G_{i-1}, G_{i-2} and E_{i-1}.

In this paper will assess the influence of the initial conditions in the simulations. To do this we considered two regions (Lisbon and South-Spain) and four "time windows": 7, 15, 30 and 90 days.

Results were computed separately for two levels of Richter magnitude (between 4 and 5, and greater than 5).

The AHP methodology has been used to analyse the influence of the three factors: Time, Magnitude and Space (location). This methodology allows us to analyse the global influence of each factor, using pair comparisons between factors.

Resumo

Apresenta-se uma análise da influência das condições iniciais num Modelo de Simulação do Processo de Ocorrências Sísmicas na Península Ibérica.

O modelo estudado admite que:

- o intervalo de tempo entre duas ocorrências sísmicas (Dt_i) é dependente dos dois intervalos precedentes (Dt_{i-1}, Dt_{i-2}),
- a grandeza (magnitude) de uma ocorrência sísmica (E_i) dos valores das grandezas dos dois sismos precedentes (G_{i-1}, G_{i-2}), bem como do valor do intervalo de tempo entre essa ocorrência e a precedente (Dt_i), e
- a localização (espaço) de uma ocorrência sísmica (E_i) é dependente do respectivo valor precedente (E_{i-1}) bem como da grandeza da ocorrência (G_i) e do intervalo de tempo entre essa ocorência e a precedente (Dt_i).

Assim, para se iniciar a simulação de ocorrências com o modelo referido, é necessário tomar como condições iniciais os valores de Dt_{i-1}, Dt_{i-2}, G_{i-1}, G_{i-2} e E_{i-1}.

Este trabalho incidirá especialmente na avaliação da influência das condições iniciais na geração de processos de ocorrências.

Para tal foram consideradas duas regiões em estudo: a de Lisboa e a da Andaluzia.

Foram testados quatro horizontes temporais (7, 15, 30 e 90 dias) e estudaram-se separadamente os sismos com magnitude (Richter) entre 4 e 5 (sismos médios) e com magnitude superiores a 5 (sismos fortes).

Foi utilizada a metodologia AHP (Analytic Hierarchy Process) para proceder ao estudo da influência dos factores Tempo, Grandeza e Espaço nas condições iniciais, já que esta metodologia permite estudar a influência global de cada factor a partir de comparações entre pares de factores.

O estudo efectuado permite verificar, de entre os cenários sísmicos estudados, qual o que torna mais provável a ocorrência de um sismo de forte magnitude, a curto prazo.

Keywords

Simulation, Earthquake, Initial Conditions, AHP.

1. Introdução

O objectivo deste trabalho é analisar a influência das *condições iniciais* num modelo de simulação do processo de ocorrências sísmicas na Península Ibérica.

O modelo utilizado (Costa, 1989 - [2]) foi construído com base no catálogo de sismos da Península Ibérica (LNEC, 1992 - [6]).

No referido modelo cada ocorrência é caracterizada por três factores: **Tempo, Grandeza** e **Espaço.** O factor **Tempo** é caracterizado pelos valores dos intervalos de tempo entre sismos consecutivos (**Dt**); a **Grandeza** (G) é caracterizada pelos valores de magnitude de Richter associados às ocorrências; o **Espaço** (E) é caracterizado pelo número da zona sísmica correspondente ao epicentro da ocorrência (ver figura 1).

Seja Dt_i intervalo de tempo que decorre a (i-1)-ésima e a i-ésima ocorrência, e G_i e E_i, respectivamente, os valores da Grandeza e Espaço associados à i-ésima ocorrência.

De acordo com o modelo adoptado, a geração da i-ésima ocorrência é feita da seguinte forma:

- o valor do intervalo de tempo Dt_i é determinado com base nos valores dos intervalos de tempo Dt_{i-2} e Dt_{i-1}, i.e.,
  $$(Dt_{i-2}, Dt_{i-1}) \rightarrow Dt_i$$
- o valor de Grandeza G_i é calculado com base em Dt_i, já determinado, e nos valores de Grandeza G_{i-1} e G_{i-2}, i.e.,
  $$(Dt_i, G_{i-2}, G_{i-1}) \rightarrow G_i$$
- o valor de Espaço E_i é determinado com base em Dt_i, G_i e E_{i-1}, i.e.,
  $$(Dt_i, G_i, E_{i-1}) \rightarrow E_i$$

Assim, para se iniciar a geração do processo de ocorrência sísmicas é necessário caracterizar as ***condições iniciais***, ou seja, conhecer os valores de

$$Dt_{i-2}, Dt_{i-1}, G_{i-2}, G_{i-1} \text{ e } E_{i-1}$$

Por uma questão de simplificação de linguagem, às grandezas Dt, G e E chamar-se-á ***factores*** e qualquer conjunto cujos elementos sejam factores chamar-se-á ***combinação de factores.***

Adoptou-se a caracterização do factor Espaço proposta (Costa, 1989 - [2]); Costa e Oliveira, 1991 - [3]), que se traduz na definição de 21 ***zonas sísmicas*** na Península Ibérica (Longitude 15° W - 4° E, latitude 34,5°, N - 44° N), que se representa na figura 1.

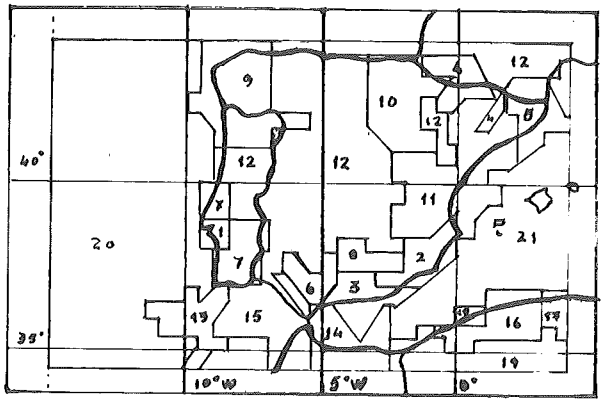

Figura 1 - Zonas sísmicas da Península Ibérica

Pretende-se saber se as *condições iniciais*, e consequentemente, as combinações de factores nelas envolvidas, influem de forma significativa no processo de ocorrências sísmicas.

2. Metodologia

Admita-se, a título de exemplo, que se pretende saber qual dos seguintes três cenários torna mais provável a ocorrência de um sismo de magnitude maior ou igual a um valor previamente fixado - **G0**, na região X, durante um determinado período de tempo:

- as duas últimas ocorrência tiveram forte magnitude, ocorreram há pouco tempo e o epicentro da última delas situou-se fora da região X
- as duas últimas ocorrências tiveram lugar há pouco tempo, com fraca magnitude, mas a última ocorreu na região X
- o epicentro da última ocorrência situou-se na região X, e tanto esta ocorrência como a que a precedeu tiveram forte magnitude mas ocorreram há já algum tempo.

Ao primeiro cenário referido pode-se associar a combinação de factores

Dt, G.

Neste cenário o epicentro da última ocorrência situou-se **fora** da região X, pelo que não se considera interveniente o factor Espaço E. Por outro lado, como as duas últimas ocorrências tiveram **forte magnitude**, então a Grandeza G é um factor interveniente nesse cenário. Adicionalmente, se as duas últimas ocorrências são **recentes**, então o factor Tempo Dt é interveniente nesse cenário. A propósito deve referir-se que, estudos prévios indicaram que se as duas últimas ocorrência tiveram lugar há muito tempo, é menor a probabilidade de ocorrer um sismo de forte magnitude do que se as duas últimas ocorrências tivessem tido lugar há pouco tempo. Esta conclusão foi tirada com base em experiências efectuadas antes da realização do presente trabalho, tendo-se efectuado um número elevado de simulações do processo de ocorrências sísmicas na Península Ibérica, em que se comparavam condições iniciais que só diferiam nos valores de Dt. Poder-se-ia pensar que estes resultados contrariam a convicção de

que "quanto maior for o intervalo de tempo entre ocorrências, maior é a energia acumulada". No entanto, como adiante se verá, os horizontes temporais estudados neste trabalho não excedem três meses, e a convicção de que quanto maior for o intervalo de tempo mais elevada será a magnitude, refere-se normalmente a períodos de tempo bem maiores do que os estudados. Além disto, no modelo de simulação utilizado, Dt representa o intervalo entre ocorrências consecutivas na Península Ibérica ao passo que na convicção referida, o intervalo de tempo se refere a ocorrências consecutivas na mesma região.

Seguindo um raciocínio análogo, podem associar-se aos dois últimos cenários, as seguintes combinações de factores:

Dt, E

G, E

Se se estudarem todas as possibilidades de combinações de factores (isoladamente, dois a dois e os três) obtém-se um total de sete combinações:

Dt

G

E

Dt, G

Dt, E

G, E

Dt, G, E

O estudo da influência das *condições iniciais* no processo de ocorrências sísmicas na Península Ibérica vai ser feito com base nestas 7 combinações de factores.

Note-se que a metodologia que se vai expor poder ser utilizada noutras combinações de factores, nomeadamente em combinações dos parâmetros de Dt e G (Dt_{i-2}, Dt_{i-1}, G_{i-2}, G_{i-1}), no entanto, o principal objectivo é apresentar uma metodologia que permita comparar a influência de condições iniciais quaisquer.

Dado que o modelo utilizado na simulação do processo de ocorrências sísmicas abrange toda a Península Ibérica, e esta é uma região demasiado extensa para ser estudada globalmente, seleccionam-se duas regiões sísmicas a estudar. Optou-se por escolher, a título de exemplo, a região de Lisboa (a que correspondem as *zonas sísmicas* números 1 e 7) e a região da Andaluzia (a que correspondem as *zonas sísmicas* números 2 e 3).

O estudo de factores fez-se para 4 ***horizontes temporais*** - **T0**: 7 dias, 15 dias, 30 dias e 90 dias e para 2 níveis de magnitude:

$$4 \leq G < 5 \quad \text{e} \quad G \geq 5.$$

Uma vez definidas as regiões a estudar, os períodos de tempo e os níveis de magnitude, é necessário encontrar um processo que permita contabilizar "quantas vezes" uma *condição inicial* é "mais importante" do que outra.

Como se pretende estudar as sete combinações de factores referidos, é difícil compará-los globalmente sem os comparar primeiramente dois a dois.

Para estudar a influência dos factores (ou combinação deles) nas *condições iniciais* utilizou-se o **AHP - Analytic Hierarchy Process** (Saaty, 1990 - [9]), já que este método permite estudar a influência global de cada combinação de factores utilizando comparações entre pares de combinação deles.

Para se utilizar este método, tem de se construir a ***matriz de julgamentos*** para as sete combinações de factores envolvidas: Dt; G; E; Dt,G; DT,E; G,E e Dt,G,E.

Nesta matriz cada elemento a(i, j) tem o seguinte significado: se a(i, j) = k, então a combinação de factores da linha i é k vezes "mais importante" do que a combinação de factores da coluna j.

Os elementos da diagonal valem 1, já que cada a combinação de factores vale tanto quanto ela própria. Os restantes elementos da matriz são recíprocos, ou seja, $a(i, j) = 1/a(j, i)$, $i \neq j$.

Para se saber quantas vezes uma combinação de factores é "mais importante" do que outra pode procede-se do seguinte modo:

Partindo de duas *condições iniciais*, cada uma delas relativa a uma combinação de factores, sejam elas a *condiçãoinicial A* e a *condição inicial B*, efectuam-se ***n*** simulações do processo de ocorrências sísmicas, durante um período de tempo T0. Contabiliza-se o número de simulações em que se verificou pelo menos uma ocorrência de magnitude superior ou igual a G0 (ou entre dois valores estabelecidos de magnitude) na região X. Seja *r1* esse valor para a *condição inicial A* e *r2* para a *condição inicial B*.

Se por exemplo $r1 = 200$, $p1 = r1/n$ é uma estimativa da probabilidade de se verificar pelo menos uma ocorrência de magnitude superior ou igual a G0 na região X, durante o período de tempo T0, **sabendo-se que** se partiu de um cenário correspondente à *condição inicial A*.

Analogamente, se $r2 = 100$, $p2 = r2/n$ é uma estimativa da probabilidade de se verificar pelo menos uma ocorrência de magnitude superior ou igual a G0 na região X, durante o período de tempo T0, **sabendo-se que** se partiu de um cenário correspondente à *condição inicial B*.

Parece razoável admitir que a *condição A* é duas vezes "mais importante" do que a *condição B*, já que $p1/p2 = 200/100 = 2$.

Se para apenas duas *condições iniciais* é "fácil" determinar quantas vezes uma *condição inicial* é "mais importante" do que a outra, para sete *condições iniciais* o processo de comparação é complicado.

Uma vez que a metodologia AHP permite estudar a influência global de cada factor (ou combinações de factores), optou-se por utilizá-la.

Para aplicar a AHP ao estudo das sete combinações de factores que se pretendem estudar é necessário construir uma matriz de 7×7, e efectuar 21 comparações, já que, numa matriz de julgamentos é necessário efectuar $m(m-1)/2$ comparações, sendo ***m*** o número de características a comparar. De lembrar que os elementos da diagonal valem 1, e a matriz é recíproca.

Para cada combinação de factores efectuaram-se n simulações do processo de ocorrências sísmicas e contabilizou-se, para cada caso, o número de simulações em que se obteve pelo menos uma ocorrência de magnitude maior ou igual a G0. Designem-se esses valores por:

- r_A - para a *condição inicial* A relativa a Dt
- r_B - para a *condição inicial* B relativa a G
- r_C - para a *condição inicial* C relativa a E
- r_D - para a *condição inicial* D relativa a Dt, G
- r_E - para a *condição inicial* E relativa a Dt, E
- r_F - para a *condição inicial* F relativa a G, E
- r_G - para a *condição inicial* G relativa a Dt, G, E

A matriz de julgamentos é então construída da seguinte forma:

	Dt	**G**	**E**	**Dt,G**	**Dt,E**	**G,E**	**Dt,G,E**
Dt	1	r_A/r_B	r_A/r_C	r_A/r_D	r_A/r_E	r_A/r_F	r_A/r_G
G	r_B/r_A	1	r_B/r_C	r_B/r_D	r_B/r_E	r_B/r_F	r_B/r_G
E	r_C/r_A	r_C/r_B	1	r_C/r_D	r_C/r_E	r_C/r_F	r_C/r_G
Dt,G	r_D/r_A	r_D/r_B	r_D/r_C	1	r_D/r_E	r_D/r_F	r_D/r_G
Dt,E	r_E/r_A	r_E/r_B	r_E/r_C	r_E/r_D	1	r_E/r_F	r_E/r_G
G,E	r_F/r_A	r_F/r_B	r_F/r_C	r_F/r_D	r_F/r_E	1	r_F/r_G
Dt,G,E	r_G/r_A	r_G/r_B	r_G/r_C	r_G/r_D	r_G/r_E	r_G/r_F	1

Uma vez formada a matriz de julgamentos é possível calcular o ***vector de prioridades*** correspondente, bastando para tal determinar o vector próprio principal da matriz de julgamentos e normalizá-lo, obtendo-se

$$[p_{Dt}\ p_G\ p_E\ p_{Dt,G}\ p_{Dt,E}\ p_{G,E}\ p_{Dt,G,E}]^T$$

O *vector de prioridades* indica a "importância" percentual de cada uma das combinações de factores. Se por exemplo o 3º elemento do vector de prioridades valer 0.12 isso significa que a contribuição no processo de ocorrências sísmicas do factor E - localização da última ocorrência é de 12%.

A partir do vector de prioridades é possível saber, de entre os cenários comparados, qual a combinação de factores que torna mais provável a ocorrência de um sismo de elevada magnitude, bastando para tal verificar qual a componente com maior valor.

3. Experiências efectuadas

Para se poder calcular o vector de prioridades, há que definir um conjunto de 7 *condições iniciais*, tais que cada uma delas possa exprimir a influência de uma combinação de factores.

Para que os resultados não sejam influenciados por um conjunto particular de *condições iniciais*, definiram-se 3 conjuntos das mesmas, ou seja, 3 conjuntos de 7 *condições iniciais*

cada, calculando-se depois a média das componentes dos *vectores de prioridades* relativos a cada combinação de factores.

O estudo da influência dos factores vai ser feito do seguinte modo:

- se se pretender estudar a influência do factor Dt, a *condição inicial* correspondente a este factor deverá possuir valores reduzidos de Dt_{i-2} e Dt_{i-1}, tendo G_{i-2} e G_{i-1} valores próximos do valor médio da magnitude dos sismos da Península Ibérica, e tendo E_{i-1} um valor correspondente a uma *zona sísmica* fora da região em estudo.
- quando se pretende estudar a influência de G, G_{i-2} e G_{i-1} devem ter valores elevados, Dt_{i-2} e Dt_{i-1} devem possuir valores próximos do valor médio de Dt e E_{i-1} deve ter um valor correspondente a uma *zona sísmica* fora da região em estudo.
- quando se pretende estudar o factor E, E_{i-1} deve ter o valor de uma das *zonas sísmicas* da região em estudo, tendo os restantes factores valores próximos dos correspondentes valores médios.

No quadro 1 apresenta-se algumas estimativas de quatis e do valor médio das distribuições de Dt e de G. Para maior facilidade de leitura, os valores de Dt são apresentados no formato **dd:hh:mm:ss** (dia:hora:minuto:segundo).

quantil	20%	40%	60%	80%	valor médio
Dt	00:02:27:10.1	00:12:25:21.1	01:08:45:04.6	03:16:49:35.0	02:22:04:48.0
G	2.8	3.2	3.5	3.9	3.4

Quadro 1 - Estatísticas de Dt e G

Quando um factor não é objecto de estudo, o seu valor, nas *condições iniciais*, deve situar-se próximo do seu valor médio, excepto para o factor E, cujo valor deve corresponder a uma zona sísmica fora da região em estudo.

Os quadros 2 a 7 apresentam os valores das *condições iniciais* utilizadas (quadros 2 a 4 para a região de Lisboa e quadros 5 a 7 para a região da Andaluzia).

Região de Lisboa

C.I. nº	Dt i-2	Dt i-1	G i-2	G i-1	E i-1	factores em estudo
1	**00:00:05:15.4**	**00:00:15:46.1**	3.2	2.9	13	Dt
2	00:20:08:52.8	01:17:10:19.2	**4.8**	**6.0**	13	G
3	00:21:01:26.4	01:18:02:52.8	3.3	3.4	7	E
4	**00:00:47:18.2**	**00:00:31:32.2**	**6.5**	**7.0**	12	Dt, G
5	**00:00:42:28.9**	**00:00:15:46.1**	3.1	3.4	7	Dt, E
6	00:17:31:12.0	01:19:48:00.0	**4.5**	**6.0**	7	G, E
7	**00:00:05:15.4**	**00:00:15:46.1**	**4.2**	**6.0**	7	Dt, G, E

Quadro 2 - Primeiro conjunto de condições iniciais utilizado para a região de Lisboa

C.I. nº	Dt i-2	Dt i-1	G i-2	G i-1	E i-1	factores em estudo
1	**00:07:53:02.4**	**00:04:22:48.0**	3.1	3.0	20	Dt
2	01:01:24:14.4	02:04:33:36.0	**5.2**	**5.7**	12	G
3	00:17:31:12.0	01:02:16:48.0	2.8	3.3	7	E
4	**00:04:22:48.0**	**00:07:53:02.4**	**5.0**	**6.8**	13	Dt, G
5	**00:06:07:55.2**	**00:03:30:14.4**	2.2	3.0	7	Dt, E
6	01:02:16:48.0	00:17:31:12.0	**4.8**	**5.3**	7	G, E
7	**00:00:36:47.5**	**00:00:26:16.8**	**4.9**	**7.0**	7	Dt, G, E

Quadro 3 - Segundo conjunto de condições iniciais utilizado para a região de Lisboa

C.I. nº	Dt i-2	Dt i-1	G i-2	G i-1	E i-1	factores em estudo
1	**00:00:52:33.6**	**00:00:26:16.8**	2.8	2.0	12	Dt
2	00:20:08:52.8	03:06:50:24.0	**6.0**	**7.0**	12	G
3	00:21:01:26.4	02:22:04:48.0	3.3	3.4	7	E
4	**00:00:10:30.7**	**00:00:26:16.8**	**5.5**	**6.1**	12	Dt, G
5	**00:00:52:33.6**	**00:00:42:02.9**	2.7	3.4	7	Dt, E
6	01:11:02:24.0	00:08:45:36.0	**3.3**	**3.9**	7	G, E
7	**00:07:00:28.8**	**00:07:53:02.4**	**5.9**	**6.8**	7	Dt, G, E

Quadro 4 - Terceiro conjunto de condições iniciais utilizado para a região de Lisboa

Região da Andaluzia

Os valores das *condições iniciais* utilizados para esta região são iguais aos utilizados para a região de Lisboa, excepto nos casos em que está envolvido o factor espaço, já que para esta região E_{i-1} toma o valor 2 ou 3.

C.I. nº	Dt i-2	Dt i-1	G i-2	G i-1	E i-1	factores em estudo
1	**00:00:05:15.4**	**00:00:15:46.1**	3.2	2.9	13	Dt
2	00:20:08:52.8	01:17:10:19.2	**4.8**	**6.0**	13	G
3	00:21:01:26.4	01:18:02:52.8	3.3	3.4	**3**	E
4	**00:00:47:18.2**	**00:00:31:32.2**	**6.5**	**7.0**	12	Dt, G
5	**00:00:42:28.9**	**00:00:15:46.1**	3.1	3.4	**2**	Dt, E
6	00:17:31:12.0	01:19:48:00.0	**4.5**	**6.0**	**3**	G, E
7	**00:00:05:15.4**	**00:00:15:46.1**	**4.2**	**6.0**	**3**	Dt, G, E

Quadro 5 - Primeiro conjunto de condições iniciais utilizado para a região da Andaluzia

C.I. nº	Dt i-2	Dt i-1	G i-2	G i-1	E i-1	factores em estudo
1	**00:07:53:02.4**	**00:04:22:48.0**	3.1	3.0	20	Dt
2	01:01:24:14.4	02:04:33:36.0	**5.2**	**5.7**	12	G
3	00:17:31:12.0	01:02:16:48.0	2.8	3.3	2	E
4	**00:04:22:48.0**	**00:07:53:02.4**	**5.0**	**6.8**	13	Dt, G
5	**00:06:07:55.2**	**00:03:30:14.4**	2.2	3.0	3	Dt, E
6	01:02:16:48.0	00:17:31:12.0	**4.8**	**5.3**	2	G, E
7	**00:00:36:47.5**	**00:00:26:16.8**	**4.9**	**7.0**	3	Dt, G, E

Quadro 6 - Segundo conjunto de condições iniciais utilizado para a região da Andaluzia

C.I. nº	Dt i-2	Dt i-1	G i-2	G i-1	E i-1	factores em estudo
1	**00:00:52:33.6**	**00:00:26:16.8**	2.8	2.0	12	Dt
2	00:20:08:52.8	03:06:50:24.0	**6.0**	**7.0**	12	G
3	00:21:01:26.4	02:22:04:48.0	3.3	3.4	2	E
4	**00:00:10:30.7**	**00:00:26:16.8**	**5.5**	**6.1**	12	Dt, G
5	**00:00:52:33.6**	**00:00:42:02.9**	2.7	3.4	3	Dt, E
6	01:11:02:24.0	00:08:45:36.0	**3.3**	**3.9**	3	G, E
7	**00:07:00:28.8**	**00:07:53:02.4**	**5.9**	**6.8**	3	Dt, G, E

Quadro 7 - Terceiro conjunto de condições iniciais utilizado para a região da Andaluzia

Para cada região estudada e para cada *condição inicial*, efectuaram-se 10000 simulações do processo de ocorrências sísmicas, durante o período de tempo T0, tendo-se contabilizado, para cada caso, o número de simulações em que se verificou pelo menos uma ocorrência de magnitude:

- $4 \leq G < 5$ e
- $G \geq 5$.

Com estes valores determinaram-se as matrizes de julgamentos e correspondentes vectores de prioridades.

Note-se que para este estudo, foi necessário determinar 336 matrizes de julgamentos e correspondentes vectores de prioridades

nº de conjuntos de condições iniciais ×

× (nº de regiões × nº de horizontes temporais × nº de condições iniciais × nº de níveis de mag)=

$= 3 \times (2 \times 4 \times 7 \times 2) = 336$

Como se utilizaram 3 conjuntos de *condições iniciais* para estudar a influência dos factores no processo de ocorrências sísmicas, existem 3 valores da componente do vector de prioridades, correspondente a cada combinação de factores, donde é necessário efectuar a média dos três valores.

Por exemplo, para a região de Lisboa, para T0 = 7 dias e para $4 \leq G < 5$ existem 3 valores da componente do vector de prioridades correspondente ao factor Dt.

Nos quadros seguintes apresenta-se a média das componentes dos vectores de prioridades correspondentes a cada combinação de factores estudada.

T0	G	Dt	G	E	Dt, G	Dt, E	G, E	Dt, G, E
7 dias	$4 \leq G < 5$	0.114	0.142	0.105	0.172	0.133	0.154	0.180
	$G \geq 5$	0.115	0.154	0.093	0.177	0.123	0.169	0.169
15 dias	$4 \leq G < 5$	0.123	0.136	0.124	0.159	0.136	0.152	0.169
	$G \geq 5$	0.110	0.143	0.098	0.170	0.134	0.150	0.196
30 dias	$4 \leq G < 5$	0.133	0.137	0.129	0.157	0.137	0.147	0.159
	$G \geq 5$	0.128	0.152	0.135	0.155	0.131	0.148	0.154
90 dias	$4 \leq G < 5$	0.140	0.141	0.136	0.146	0.141	0.146	0.149
	$G \geq 5$	0.144	0.134	0.130	0.156	0.144	0.138	0.155

Quadro 8 - Vectores de prioridades para a Região de Lisboa

T0	G	Dt	G	E	Dt, G	Dt, E	G, E	Dt, G, E
7 dias	$4 \leq G < 5$	0.119	0.131	0.106	0.158	0.157	0.149	0.181
	$G \geq 5$	0.093	0.140	0.100	0.165	0.146	0.164	0.194
15 dias	$4 \leq G < 5$	0.124	0.137	0.124	0.154	0.148	0.149	0.164
	$G \geq 5$	0.117	0.139	0.110	0.175	0.130	0.149	0.181
30 dias	$4 \leq G < 5$	0.143	0.139	0.129	0.151	0.146	0.145	0.156
	$G \geq 5$	0.118	0.145	0.127	0.153	0.140	0.151	0.166
90 dias	$4 \leq G < 5$	0.140	0.139	0.138	0.144	0.146	0.146	0.148
	$G \geq 5$	0.134	0.145	0.132	0.151	0.145	0.144	0.149

Quadro 9 - Vectores de prioridades para a Região da Andaluza

As figuras 2 a 5 contêm a mesma informação do que estes dois últimos quadros, mas utilizou-se outro tipo de representação.

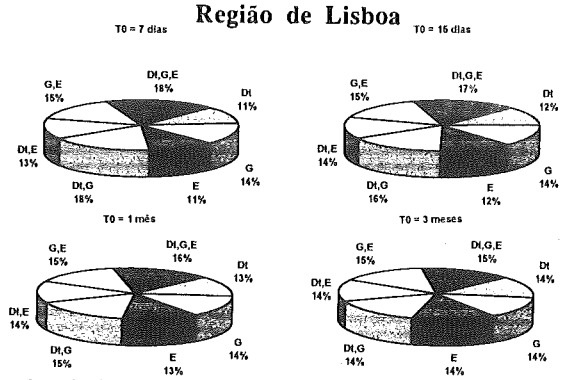

Figura 2-Vectores de prioridades para a Região de Lisboa, para sismos de magnitude $4 \leq G < 5$

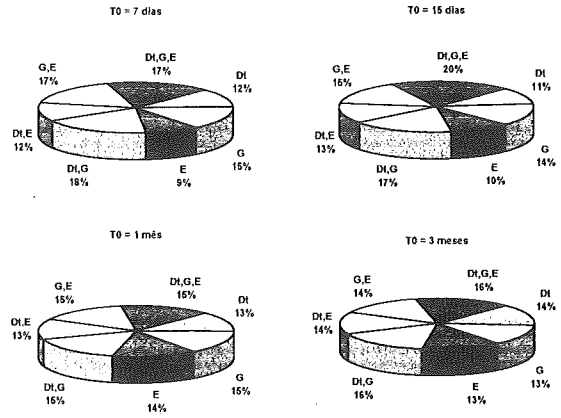

Figura 3 - Vectores de prioridades para a Região de Lisboa, para sismos de magnitude G ≥ 5

Região da Andaluzia

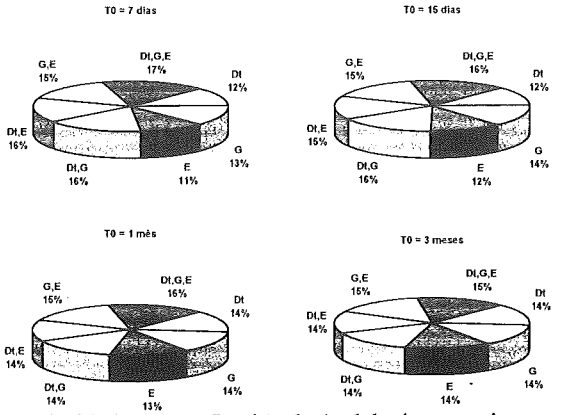

Figura 4-Vectores de prioridades para a Região da Andaluzia, para sismos de magnitude 4≤G<5

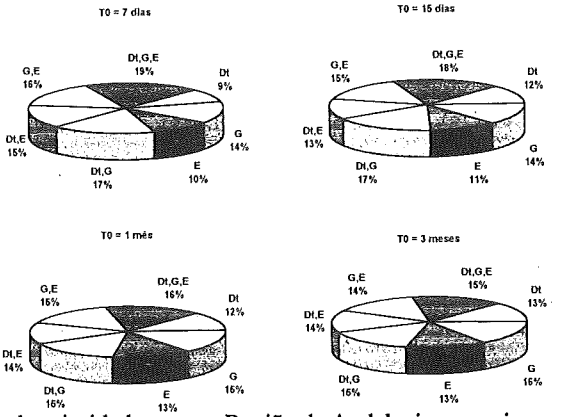

Figura 5 - Vectores de prioridades para a Região da Andaluzia, para sismos de magnitude G ≥ 5

4. Análise de resultados

Da análise das figuras 2 a 5 destaca-se o seguinte:

- nas duas regiões estudadas as combinações de factores predominantes mantêm-se, notando-se que existe semelhança na distribuição da contribuição relativa dos factores
- em ambas as regiões, a diferenciação das componentes dos vectores de prioridades vai diminuindo à medida que T0 aumenta, podendo concluir-se que a influência das *condições iniciais* diminiu com o tempo, sendo muito pequena ao fim de três meses. No entanto, para períodos de tempo de 3 meses, existe maior diferenciação (ainda que pequena) nas componentes dos vectores de prioridades para sismos fortes, do que para sismos médios, o que significa que a probabilidade de ocorrência de sismos de forte magnitude é mais sensível às condições iniciais do que a de sismos de fraca magnitude.

Para se clarificar o padrão de influência das *condições iniciais*, definiram-se 3 níveis para os valores das componentes dos vectores de prioridades - **p**:

- $p > 15\%$
- $13\% < p \leq 15\%$
- $p \leq 13\%$

e verificou-se, quais as combinações de factores cujas componentes dos vectores de prioridades se situavam em cada nível. Os resultados encontram-se nos quadros 10 a 13.

Em ambas as regiões estudadas verifica-se que são predominantes (até horizontes temporais de 30 dias) as seguintes as combinações de factores:

- Dt, G, E
- Dt, G

É ainda notória a influência das combinações de factores:

- G, E
- Dt, E
- G

A influência das restantes combinações de factores é pouco notória.

Refira-se, no entanto que, à medida que o valor de T0 aumenta, as combinações de factores vão-se concentrando no nível intermédio dos *vectores de prioridades*, o que indica que a um aumento do período de tempo corresponde uma diminuição da influência das *condições iniciais*.

Se nenhuma combinação de factores exercesse umá influência destacada, cada componente do vector de prioridades teria um valor próximo de 100/7 = 14.3. Note-se que o nível intermédio compreende este último valor, daí que quandoT0 aumenta as combinações de factores se vão concentrando no nível intermédio.

Para T0 = 3 meses e $4 \leq G < 5$, em ambas as regiões estudadas, todos os combinações de factores se situam no nível intermédio. Quer isto dizer que, para sismos médios (magnitude

entre 4 e 5), ao fim de 3 meses os factores intervenientes nas *condições iniciais* já não exercem influência visível. No entanto o mesmo não acontece relativamente aos sismos fortes (magnitude superior a 5). Com efeito na região de Lisboa (ver quadro 12), as combinações de factores, DT,G,E e Dt,G ainda são predominantes.

Sismos Médios ($4 \leq G < 5$)

	T0 = 7 dias	T0 = 15 dias	T0 = 30 dias	T0 = 90 dias
p > 15%	Dt, G, E Dt, G G, E	Dt, G, E Dt, G G, E	Dt, G, E Dt, G	
13% < p ≤ 15%	G Dt, E	Dt, E G	G, E Dt, E G Dt	Dt, G, E Dt, G G, E G Dt, E Dt E
p ≤ 13%	Dt E	Dt E	E	

Quadro 10 - Representação esquemática das componentes dos vectores de prioridades para sismos de média magnitude na Região de Lisboa

	T0 = 7 dias	T0 = 15 dias	T0 = 30 dias	T0 = 90 dias
p > 15%	Dt, G, E Dt, G Dt, E	Dt, G, E Dt, G	Dt, G, E Dt, G	
13% < p ≤ 15%	G, E G	G, E Dt, E G	G, E Dt, E G Dt	Dt, G, E Dt, G G, E G Dt, E Dt E
p ≤ 13%	Dt E	Dt E	E	

Quadro 11 - Representação esquemática das componentes dos vectores de prioridades para sismos de média magnitude na Região da Andaluzia

Sismos Médios ($4 \leq G < 5$)

	T0 = 7 dias	T0 = 15 dias	T0 = 30 dias	T0 = 90 dias
p > 15%	Dt, G, E Dt, G G, E G	Dt, G, E Dt, G	Dt, G, E Dt, G G	Dt, G, E Dt, G
13% < p ≤ 15%		G, E G Dt, E	G, E Dt, E E	G, E Dt, E Dt G
p ≤ 13%	Dt, E Dt E	Dt E	Dt	E

Quadro 12 - Representação esquemática das componentes dos vectores de prioridades para sismos de forte magnitude na Região de Lisboa

	T0 = 7 dias	T0 = 15 dias	T0 = 30 dias	T0 = 90 dias
p > 15%	Dt, G, E Dt, G G, E	Dt, G, E Dt, G	Dt, G, E Dt, G G, E	Dt, G
13% < p ≤ 15%	Dt, E G	G, E G	G, E G	Dt, G, E Dt, E G, E G Dt E
p ≤ 13%	Dt E	Dt, E Dt E	Dt E	

Quadro 13 - Representação esquemática das componentes dos vectores de prioridades para sismos de forte magnitude na Região da Andaluzia

Como se viu, a metodologia utilizada neste estudo, permite estimar a probabilidade de ocorrência de um sismo de magnitude superior a um valor fixado, durante o período de tempo T0, na região X, partindo-se de determinada *condição inicial.*

Partindo de uma *condição inicial* e efectuando um número elevado (**n**) de simulações do processo de ocorrências, durante o período de tempo T0, é possível contabilizar o número de simulações em que se verificou pelo menos uma ocorrência de magnitude superior ou igual a G0 (ou entre dois valores estabelecidos de magnitude) na região X. Seja ***r*** esse valor.

Então, $p = r/n$ é uma estimativa pontual da probabilidade de se verificar pelo menos uma ocorrência de magnitude superior ou igual a G0 na região X, durante o período de tempo T0, sabendo-se que se partiu de um cenário correspondente à *condição inicial* utilizada.

O quadro 14 apresenta os valores (em %) da probabilidade de se verificar pelo menos um sismo de magnitude maior ou igual a 5 na região de Lisboa durante o período T0, sabendo-se que se partiu de um cenário correspondente às combinações de factores estuadadas.

Região de Lisboa

T0	Dt	G	E	Dt, G	Dt, E	G, E	Dt, G, E
7 dias	0.50	0.67	0.40	0.77	0.53	0.73	0.73
15 dias	0.73	0.90	0.63	1.10	0.87	0.97	1.23
30 dias	1.27	1.50	1.30	1.53	1.30	1.47	1.53
90 dias	2.10	2.03	2.83	2.27	3.13	3.00	3.37

Quadro 14 - Probabilidade de ocorrência de um sismo forte na região de Lisboa (em %)

É interessante notar que as quatro estimativas pontuais de probabilidade, correspondentes aos quatro "horizontes temporais" permitem propor um ajustamento linear com elevado coeficiente de correlação (ρ).

Para cada combinação de factores, testou-se o ajustamento linear aos 4 valores da probabilidade correspondentes aos quatro períodos de tempo estudados, ou seja testou-se um ajustamento do tipo:

$$prob(t) = A + B.t$$

obtendo-se os resultados apresentados no quadro 15.

T0	Dt	G	E	Dt, G	Dt, E	G, E	Dt, G, E
A	0.494	0.731	0.260	0.824	0.363	0.580	0.631
B	0.018	0.015	0.029	0.017	0.031	0.027	0.031
ρ	0.977	0.939	0.995	0.970	0.999	0.999	0.996

Quadro 15 - Resultados do ajustamento linear efectuado para a região de Lisboa

Nota: *prob(t)* em percentagem e *t* em dias.

Efectuou-se o mesmo estudo para a região da Andaluzia.

Região da Andaluzia

T0	Dt	G	E	Dt, G	Dt, E	G, E	Dt, G, E
7 dias	1.73	2.60	1.87	3.07	2.73	3.07	3.63
15 dias	2.10	3.83	3.30	4.70	3.97	4.13	4.93
30 dias	5.03	6.20	5.43	6.53	6.00	6.43	7.10
90 dias	12.73	13.10	12.23	13.83	13.40	13.47	14.10

Quadro 16 - Probabilidade de ocorrência de um sismo forte na região da Andaluzia (em %)

T0	Dt	G	E	Dt, G	Dt, E	G, E	Dt, G, E
A	0.586	2.011	1.382	2.554	2.012	2.359	3.038
B	0.136	0.125	0.122	0.126	0.127	0.124	0.124
ρ	0.997	0.998	0.998	0.998	1.000	0.999	0.998

Quadro 17 - Resultados do ajustamento linear efectuado para a região da Andaluzia

Dado que o valor do coeficiente de correlação é sempre superior a 0.9, considera-se aceitável o ajustamento linear, tornando assim possível o cálculo da probabilidade de se verificar pelo menos um sismo de magnitude maior ou igual a 5, na região de Lisboa durante um período de tempo inferior ou igual a 90 dias, sabendo-se que se partiu de um cenário correspondente a cada uma das combinações de factores estudados.

5. Conclusão

5.1 Principais resultados obtidos

Comparou-se a influência das *condições iniciais* num modelo de simulação tridimensional (Tempo, Grandeza e Espaço) do processo de ocorrências sísmicas na Península Ibérica (Costa, 1989 - [2]). Para se iniciar a simulação do processo de ocorrências sísmicas, a partir do referido modelo é necessário conhecer os valores de Dt_{i-2}, Dt_{i-1}, G_{i-2}, G_{i-1} e E_{i-1}.

Estudaram-se as combinações de factores: Dt; G; E; Dt,G; Dt,E; G,E e Dt,G,E.

O estudo levado a cabo considerou duas regiões: (Lisboa e Andaluzia), quatro períodos de tempo (7, 15, 30 e 90) e dois níveis de magnitude: $4 \leq G < 5$ e $G \geq 5$.

Da análise dos resultados verifica-se que, para ambas as regiões estudadas, à medida que o horizonte temporal aumenta, diminiu a diferença nos valores das componentes dos vectores prioridades, **sendo a probabilidade de ocorrência de sismos de magnitude $G \geq 5$ mais sensível às *condições iniciais* do que a *probabilidade de ocorrência de sismos de magnitude* entre 4 e 5.**

Em ambas as regiões estudadas verifica-se que são predominantes as seguintes combinações de factores: **Dt, G, E**, significando "que as duas últimas ocorrências tiveram lugar há pouco tempo, com elevada magnitude, tendo-se a última situado na zona em estudo" e **Dt, G**, significando que "as duas últimas ocorrências tiveram lugar há pouco tempo, com elevada magnitude, mas a última não se situou na zona em estudo".

É ainda notória a influência das combinações de factores: **G, E** ("as duas últimas ocorrências tiveram elevada magnitude e a última situou-se na zona de estudo"), **Dt, E** ("as duas últimas ocorrências tiveram lugar há pouco tempo, tendo-se a última situado na zona em estudo") e **G** ("as duas últimas ocorrências tiveram elevada magnitude").

A influência das restantes combinações de factores é pouco notória.

Os resultados obtidos permitem observar uma diferenciação nas componentes dos vectores de prioridades, que se vai atenuando com o aumento do horizonte temporal.

Deve referir-se que, ao analisar-se 7 cenários sísmicos (um número relativamente elevado) não se potencia a diferenciação nas componentes dos vectores de prioridades. Com efeito, algumas análises estatísticas clássicas levaram à sobreposição parcial dos vários intervalos de confiança a 95% para a proporção de ocorrências correspondentes aos diferentes cenários. Estes resultados estão na base da geração de condições iniciais psuedo aleatórias (ver 5.4) que permitem antever uma maior diferenciação nas componentes do vector de prioridades.

Ressalva-se, no entanto, a relevância da metodologia seguida quando se pretende comparar cenários sísmicos **reais**.

5.2 Limitações da abordagem seguida

- Deve referir-se que uma óvia limitação da abordagem seguida tem como génese o próprio modelo de geração do processo de ocorrências sísmicas adoptado. Costa (1989) - [2] - refere algumas dessas limitações, nomeadamente as que decorrem do facto de não se ter filtrado o catálogo de fenómenos premonitórios nem de réplicas, tendo-se utilizado apenas a informação do catálogo de ocorrências sísmicas correspondente ao período posterior a 1900.
- Uma segunda limitação diz respeito à natureza discreta do factor Espaço. (Costa, 1989) - [2] simplificou o carácter bidimensional do Espaço ao reduzir este factor a 21 zonas sísmicas na Península Ibérica.

Não podendo, pela natureza do factor Espaço, definir-se um conceito de "média", considerou-se no presente trabalho, a dicotomia "pertença/não pertença à região em estudo", sendo cada região estudada um agrupamento de duas das 21 zonas sísmicas.

- Considera-se ainda que o pressuposto de independência entre as características comparadas (num dado nível), subjacente à metodologia AHP, não é tido em conta em muitas aplicações práticas (veja-se o exemplo apresentado em Saaty, 1990 - [9], pág. 128).

A aplicação proposta no presente trabalho, pela própria natureza do fenómeno físico envolvido, não permite afirmar (nem rejeitar) linearmente o respeito pelo referido pressuposto.

- A abordagem seguida não permitiu uma diferenciação muito expressiva das componentes do vector de prioridades, pelo que, num estudo posterior à elaboração deste trabalho, e actualmente ainda em desenvolvimento, experimentou-se estudar cenários sísmicos com valores de condições iniciais gerados aleatoriamente, dentro de determinados intervalos, efectuando-se no final a média dos resultados obtidos. Os primeiros resultados desta nova abordagem permitem observar uma maior diferenciação das componentes do vector de prioridades.

5.3 Potencialidades da abordagem seguida

A metodologia AHP teve a sua origem ligada ao apoio multicritério à decisão. O presente trabalho apresenta uma aplicação da metodologia AHP num contexto bem diverso do original: comparação da influência de condições iniciais num modelo de simulação do processo de ocorrências sísmicas.

Crê-se que a metodologia apresentada pode encontrar várias extensões, nomeadamente na comparação da influência de condições iniciais em modelos de simulação.

5.4 Desenvolvimento

- Relativamente ao factor espaço poder-se-á introduzir o conceito de "região vizinha" pelo que, em lugar da dicotomia referida, se teriam três situações: "pertença à região em estudo", "pertença à região vizinha" e "não pertença à região em estudo ou à sua vizinhança". Esta abordagem permitiria uma análise mais fina (embora significativamente mais complexa) deste factor.
- Como se referiu anteriormente, presentemente leva-se a cabo uma extensão deste trabalho: procede-se à geração de condições iniciais pseudo-aleatórias e posterior comparação da sua influência no processo de ocorrências sísmicas na Península Ibérica.

Por exemplo, para o estudo do factor Espaço, em vez de se comparar uma condição inicial em que E_{i-1} se situava na zona em estudo com outra condição em que E_{i-1} se situava fora dela, no primeiro caso o valor de E_{i-1} era gerado aleatoriamente no conjunto dos números das zonas sísmicas que correspondem à zona em estudo, e no segundo caso, E_{i-1} era gerado aleatoriamente, podendo percorrer toda a Península Ibérica excepto as zonas sísmicas correspondentes à zona em estudo.

Os resultados já obtidos permitem observar uma maior diferenciação da relevância dos diferentes cenários sísmicos iniciais, face ao que se observou no presente trabalho.

Deve-se, no entanto, ter em conta que, em **situações reais** de comparação de cenários sísmicos, o processo indicado é o da afectação das condições iniciais com valores concretos e não gerados aleatoriamente.

References

[1] Abramowitz, Minton and Stegun, Irene A., *Handbook of Mathematical Functions*, Dover Publications, Inc., New York (1972).
[2] Costa, R.A., *Modelação do Processo Estocástico Sísmico na Península Ibérica*, Dissertação de Doutoramento, IST-UTL (1989)
[3] Costa, R.A and Oliveira, C.S., *Defining seismic zones in the Ibero Mogrebi Region*, Proc.9 Th. European Conference on Earthquake Engineering, Moscovo (1991) 279-288
[4] Grogono, Peter, *Programming in Pascal*, Addison-Wesley (1984).
[5] Hastings, N.A.J. and Peacock, J.B., *Statistical Distributions*, Buttherworth & Co Ltd. (1975).
[6] LNEC, *Compilação de Catálogos Sísmicos da Região Ibérica*, Departamento de Estruturas, Núcleo de Dinâmica Aplicada (1992).
[7] Naylor, Thomas H., Balintfy, Joseph L. and Burdick, Donald S., *Computer Simulation Techniques*, Wiley (1968).
[8] Pidd, Michael, *Computer Simulation in Management Science*, John Wiley & Sons Ltd. (1984).
[9] Saaty, T., *Multicriteria Decision Making: The Analytic Hierarchy Process*, Vol. 1 of AHP series, Expert Choice, Inc. (1990).

INVESTIGAÇÃO OPERACIONAL, vol. 17 (Junho 1997) 41-59

MODELAÇÃO DAS VENDAS DE COMBUSTÍVEIS LÍQUIDOS RECORRENDO A MODELOS GRAVITACIONAIS

Carla Fernandes
SPC - Cargas Terrestres, S.A.

Isabel Themido
CESUR - IST-UTL
Av. Rovisco Pais
1000 Lisboa - Portugal

Abstract

This paper describes the development of models for the sales of liquid fuels in service stations, which will inform decision making in the area of evaluation of sales of new service station locations. A gravitational model for a market area located in an urban peripherical district of Portugal was developed, parameter estimation being accomplished through a log-centered transformation. The implementation of the gravitational models lead to the development of the regression model of the total consumption of the market area. The final gravitational model with an absolute global error of 17% includes, as explanatory variables, the travel time, the capacity and the visibility of the service station. The models which have a smaller error are those where proximity was measured by the travel time between home and the service station rather than by distance in kilometres.

Resumo

Este artigo descreve o desenvolvimento de modelos de vendas de combustíveis líquidos, em postos de abastecimento de rodovia, utilizados na avaliação de potenciais localizações para a instalação de novas posições. Desenvolveram-se modelos gravitacionais para uma área piloto de mercado, situada num distrito urbano de periferia português, recorrendo-se a uma transformação log-centrada para a estimação dos parâmetros. A aplicação dos modelos gravitacionais exigiu o desenvolvimento de um modelo de regressão para estimar o consumo potencial da área de mercado. O modelo gravitacional final apresenta um erro absoluto global de 17% e tem como variáveis explicativas a duração do trajecto, a capacidade física do posto e a visibilidade do posto. Os modelos que apresentam menores erros médios são aqueles em que a proximidade é medida pela duração do trajecto entre a residência do consumidor e o posto em vez de ser medida pela distância (em km).

Keywords

Gasoline Sales, Service Stations, Gravitational Models, Store Location, Log-centered Transformation.

1. Introdução

O objecto deste trabalho foi o desenvolvimento de modelos gravitacionais para prever as vendas de combustíveis em postos de abastecimento situados em rodovia. As companhias petrolíferas, como qualquer empresa retalhista, têm necessidade de efectuar previsões de vendas para locais onde exista a possibilidade de instalar um novo ponto de venda.

Este problema enquadra-se na definição da estratégia de distribuição da organização e do grau de exposição no mercado. Concretamente a organização precisa de saber qual o número de posições (lojas, postos, balcões de atendimento, etc.) que deve possuir e qual a sua localização.

A sequência de decisões que os responsáveis da empresa enfrentam encontra-se representada na Figura 1. A organização tem que determinar, inicialmente, quais os mercados que devem ser seleccionados para localizar os pontos de venda. Numa segunda fase, deve determinar o número de pontos de venda a instalar, seguindo-se então a fase de selecção e avaliação das localizações alternativas. A decisão final, envolve a especificação da dimensão da loja e das suas características (configurações alternativas). Com este trabalho pretendeu-se responder aos problemas do nível 2 desenvolvendo-se modelos gravitacionais para a previsão de vendas de combustíveis líquidos em postos de abastecimento (vide Figura 1).

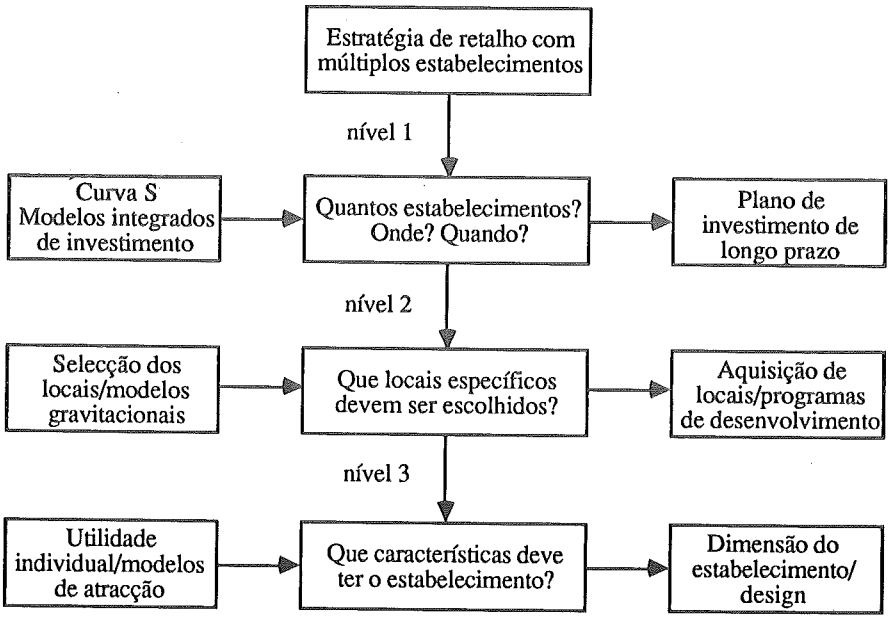

Figura 1 - Estruturação de alguns problemas de gestão de posições de retalho [Lilien e Kotler, 1983]

Este trabalho surge na sequência dos modelos de regressão desenvolvidos por Quintino (1994) para a modelação de Vendas de Combustíveis em Postos de Rodovia de uma Área Metropolitana os quais são actualmente utilizados por uma companhia petrolífera na previsão de vendas de novas posições a instalar naquela área [Themido et al, 1995].

2. Modelação de vendas de novas posições

Na literatura científica encontram-se, essencialmente, duas abordagens alternativas para a previsão de vendas de novas posições de retalho - modelos de regressão múltipla e modelos

gravitacionais. Os modelos referidos na literatura são bastante gerais e frequentemente a um nível conceptual, para proteger os interesses comerciais das organizações.

Nas últimas décadas a técnica mais utilizada para a construção de modelos de avaliação/selecção de localizações, para o retalho, tem sido a regressão múltipla. Os modelos de regressão múltipla tornam explícitas relações entre o desempenho dos pontos de venda (volume de vendas, quota de mercado, lucro, etc.) e os seus factores chave ou variáveis explicativas (características do ponto de venda, da localização, da população, da concorrência, etc.) podendo servir para prever as vendas de um novo local se as suas características forem conhecidas.

As principais vantagens dos modelos de regressão são o estabelecimento de relações de causalidade entre o desempenho da loja e as variáveis explicativas, a par da obtenção de intervalos de confiança para as estimativas. Contudo, os modelos de regressão têm um tempo de vida limitado, em consequência das mutações constante do mercado [Rogers, 1992] e exigem um grande número de observações. Estes modelos têm ainda dificuldade em representar o efeito da concorrência.

Os modelos gravitacionais são considerados por vários autores como o *estado da arte* em modelos de previsão de vendas para estações de serviço [Reinitz, 1968], supermercados [Stanley e Sewall, 1976; Penny e Broom, 1988], restaurantes, cadeias de *franchise* [Black et al, 1985; Kaufman e Rangan, 1990], investimentos imobiliários [Bottum, 1989; Carter, 1993], lojas de conveniência e outros retalhistas [Lilien e Kotler, 1983; Rogers, 1992].

A metodologia, inicialmente desenvolvida por Huff (1964), baseia-se na teoria de Luce sobre a escolha individual [Lilien e Kotler, 1983]. O modelo gravitacional trata o cliente, da área de mercado em estudo, como gravitando em torno do posto; aumentando a sua frequência de deslocação ao posto com a proximidade deste à residência do cliente e com a atractividade do posto (em qualquer dos casos relativamente à concorrência). Para além de fornecer estimativas sobre as vendas de novos postos, estes modelos permitem analisar rapidamente estratégias alternativas [Rogers, 1992], nomeadamente o impacto da abertura de novos postos, nas vendas dos já existentes, e o efeito da concorrência [Kaufman e Rangan, 1990].

Modelos gravitacionais mais complexos permitem também estabelecer a localização simultânea de vários pontos de vendas por forma a optimizar a rede [Achabal, et al 1982; Ghosh e Craig, 1983]. O modelo desenvolvido por Ghosh e Craig (1983) toma em consideração as acções dos concorrentes, considerando que a tomada de decisões estratégicas dos retalhistas, de um mesmo mercado, pode ser modelada como um processo de equilíbrio competitivo.

Os modelos gravitacionais mais gerais, para além da dimensão do posto e da distância deste ao cliente, consideram outras características (atributos) das lojas [Stanley e Sewall, 1976; Gautschi, 1981; Lilien e Kotler, 1983; Rogers, 1992], nomeadamente a imagem e variáveis de marketing.

Penny e Broom (1988) desenvolveram um modelo gravitacional, para prever o potencial de vendas de um dado local, para a cadeia de retalho Britânica TESCO. Este modelo considera que

o consumo num ponto de venda irá depender da procura total disponível e da atractividade do local quando comparado com os concorrentes. O erro absoluto médio das previsões obtidas com este modelo é da ordem dos 13% [Penny e Broom, 1988], sendo importante notar que este modelo é dos poucos modelos gravitacionais desenvolvidos e aplicados na Europa, de que há conhecimento.

O modelo gravitacional mais geral é usualmente designado na literatura como *Multiplicative Competitive Interaction Model* (MCI), sendo dado pela expressão seguinte:

$$\textbf{Eq 1} \qquad \pi_{ij} = \frac{\prod_{k=1}^{q} x_{kij}^{\beta_k}}{\left(\sum_{j=1}^{m} \prod_{k=1}^{q} x_{kij}^{\beta_k}\right)}$$

onde: π_{ij} é a probabilidade do consumidor do sector **i** (i=1,..., I) escolher a loja **j** (j=1,..., m); x_{kij} é a variável **k** que descreve a loja **j** na escolha do consumidor do sector **i**; β_k é o parâmetro que mede a sensibilidade de π_{ij} à variável **k**.

As primeiras aplicações dos modelos gravitacionais estavam de certo modo limitadas pela dificuldade de estimação dos parâmetros do modelo MCI, mas Nakanishi e Cooper (1974) demonstraram que o modelo MCI pode ser calibrado usando os procedimentos dos mínimos quadrados.

Os parâmetros β_k, do modelo gravitacional, podem ser estimados de dois modos [Nakanishi e Cooper, 1982][1]:

$$\textbf{Eq 2} \qquad \log\left(\frac{p_{ij}}{\bar{p}_i}\right) = \sum_{k=1}^{q} \beta_k \log\left(\frac{x_{kij}}{\bar{x}_{ki}}\right) + \varepsilon_{ij}$$

$$\textbf{Eq 3} \qquad \log p_{ij} = \sum_{i'=1}^{I} \alpha_{i'} D_{i'} + \sum_{k=1}^{q} \beta_k \log x_{kij} + \varepsilon_{ij},$$

onde: p_{ij} é a proporção de habitantes do sector **i** que abastecem habitualmente no posto **j**; $\bar{p}_j$ é média geométrica em **j** de p_{ij}; x_{kij} é a variável **k** para o habitante do sector **i** abastecer no posto **j**; $\bar{x}_{ki}$ é média geométrica em **j** de x_{kij}; $D_{i'}$ é uma variável binária igual a 1 quando i' = i e zero caso contrário; $\alpha_{i'}$ é o coeficiente da regressão para $D_{i'}$.

Para estimar os coeficientes dos modelos gravitacionais optou-se pela Eq 2 por ser mais intuitiva a sua utilização posterior, embora implique o cálculo de médias geométricas para cada variável.

O desenvolvimento de modelos gravitacionais exige um elevado volume de informação, nomeadamente de dados demográficos e socioeconómicos das populações estudadas.

---

[1] Nestes métodos apenas se considera o efeito do sector. Isto é as variáveis que carcaterizam a actividade de um posto são independentes do sector de origem do condutor. Cooper e Nakanishi (1993) para situações de estimação de quotas de mercado consideram simultaneamente os efeitos do sector e do posto.

3. Descrição do caso de estudo

Foi escolhido, como caso de estudo, um distrito do interior de Portugal, que passaremos a designar por **Distrito**. Ao nível do **Distrito** recolheu-se, no local, informação sobre as características internas de **todos** os postos de abastecimento[2] tais como: dimensão, layout, localização, serviços disponíveis, horário de funcionamento, etc. Paralelamente fizeram-se contagens de tráfego de curta duração que permitiram determinar o potencial local do posto [Fernandes, 1996].

Para a construção do modelo gravitacional foi necessário seleccionar dentro do Distrito uma área homogénea, estanque, e com elevado número de postos de abastecimento que designaremos por **Área de Mercado** (vide **Figura 2**).

Figura 2 - Áreas consideradas neste trabalho: Distrito e Área de Mercado.

A **Área de Mercado** foi posteriormente dividida em sectores, consoante as suas características socio-económica, barreiras geográficas e acessos rodoviários (vide **Figura 3**). Para dividir a **Área de Mercado** em **sectores** utilizou-se a divisão administrativa, *freguesia*, tendo estas sido associadas em **sectores** em função das vias rodoviárias de acesso ao centro urbano principal. Os sectores de 1 a 4 são constituídos por freguesias que pertencem a esse centro urbano.

Para determinar o potencial da área de mercado[3] foram recolhidos dados demográficos e socioeconómicos dos habitantes do distrito.

A estimação da proporção dos habitantes, de cada sector, que abastece em cada um dos postos de abastecimento, exigiu a realização de um inquérito telefónico da responsabilidade de uma empresa da especialidade.

Todos os dados recolhidos são referentes ao ano de 1994.

---

2 O universo de postos de abastecimento estudados foi de 125 postos.

3 Entende-se por Potencial da Área de Mercado a procura total gerada pelos habitantes dessa área.

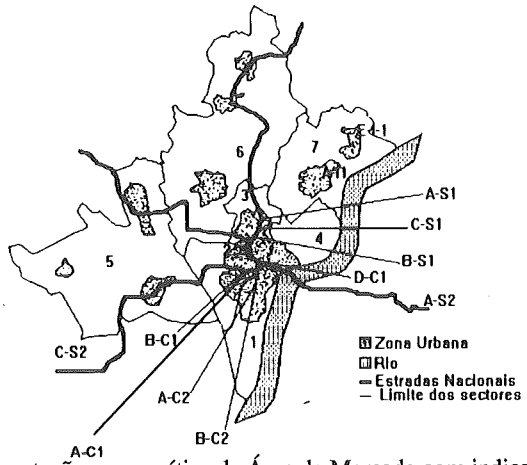

Figura 3 - Representação esquemática da Área de Mercado com indicação da localização aproximada dos postos de abastecimento existentes[4]

3.1 Levantamento das características dos postos de abastecimento

Fez-se um levantamento exaustivo das características dos postos de abastecimento situados em rodovia em todo o Distrito, utilizando para tal os critérios definidos anteriormente por Quintino (1994), apresentando-se nas **Tabela 1** e **Tabela 2**, as estatísticas das variáveis escalares e binomiais, respectivamente.

As variáveis que traduzem a dimensão do posto são a ***Área de Vendas do Posto*** e a ***Capacidade do Posto*** (Cap); esta última traduz o número máximo de veículos que podem abastecer simultaneamente no posto. As variáveis ***Acessibilidade***, ***Visibilidade*** e ***Imagem*** são variáveis de carácter qualitativo, tendo sido medidas numa escala ordinal de 1 a 5.

n = 66	Média	Mediana	Mínimo	Máximo	D.Padrão
Área de Vendas	277	188	100	800	146
Capacidade do Posto[5]	5.11	4.00	2	24	3.30
Imagem	2.15	2.00	1	5	1.13
Visibilidade	2.49	2.00	1	5	0.93
Acessibilidade	2.30	2.00	1	4	0.92
Nº de Serviços	1.96	2.00	0	9	1.36
Potencial Local	11.51	9.18	1.63	36.05	8.07
Nº de Veículos na Área de Mercado	10303	10021	624	19610	5499
Nº de Postos na Área de Mercado	9.61	10.00	1	18	4.51

Tabela 1 - Estatísticas das variáveis escalares

[4] A identificação dos postos está codificada, sendo que a primeira letra identifica a companhia, a segunda a localização do posto (C centro da área urbana, S à saída da área urbana e I posto isolado).

[5] No mercado Norte Americano o valor médio desta variável é de 5.455 e o desvio padrão igual a 0.867 [Ingene e Brown, 1987].

n = 66	Nº de casos com x = 1	% Total
Ilhas em Portagem	8	12%
Ilhas em Quadrado	1	2%
Ilhas em Espinha	2	3%
Ilhas em Linha	37	56%
Ilhas em Passeio	18	27%
Self-Service	2	3%
Pagamento Automático	53	80%
Cartão de Marca	28	42%
Loja	7	11%
Serviço de Restaurante/Cafetaria	10	15%
Lavagem Automática	2	3%
Estação de Serviço	7	11%
Serviços Assistência ao Veículo	7	11%
Posto Moderno	16	24%
Posto com Pala	26	39%
Via com Traço Contínuo	17	26%
Via com Traço Semi Contínuo[6]	28	42%
Via com Traço Descontínuo	20	31%
Posto à Entrada de uma Localidade	15	23%
Posto à Saída de uma Localidade	23	35%
Posto em Meio Urbano	56	85%
Posto de Estrada	10	15%

Tabela 2 - Estatísticas das variáveis do tipo *dummy*

A existência de serviços complementares tais como loja, meios de pagamento automático, serviço de restaurante/cafetaria, lavagem automática ou estação de serviço são representados por variáveis do tipo dummy. Estas variáveis encontram-se muito correlacionadas com as variáveis de dimensão, uma vez que os postos de reduzidas dimensões tendem a comercializar apenas combustíveis líquidos enquanto que os postos modernos, de maiores dimensões, tendem a possuir vários serviços complementares.

O ***Potencial Local*** é uma medida do tráfego médio diário anual junto ao posto[7]. Esta variável tem carácter estocástico, existindo variações de tráfego importantes com factores distintos de sazonalidade, o que exigiu um cuidado especial na sua estimação.

---

6 A variável Via com Traço Semi Contínuo indica que o posto está situado numa via onde é legalmente permitido transpor o eixo da via mas a manobra é feita com dificuldade.

7 O processo de estimação desta variável exige a realização de contagens de tráfego de curta duração. O processo de estimação do TMDA (Tráfego Médio Diário Anual) encontra-se descrito por Quintino [1994].

3.2 Inquérito telefónico

A aplicação dos modelos gravitacionais requer o conhecimento do padrão de deslocações dos consumidores aos postos de venda. Para estimar as distâncias entre a residência dos consumidores e os postos de abastecimento e identificar os postos preferidos pelos consumidores realizou-se um inquérito telefónico na Área de Mercado.

Os inquiridos tinham que indicar o posto em que abasteciam habitualmente, qual a frequência com que abasteciam nesse posto e a distância (em minutos) da sua casa ao posto. No caso do inquirido abastecer também noutros postos a pergunta era repetida.

- **Grau de Fidelidade ao Posto de Abastecimento**

A frequência com que os inquiridos abastecem no seu posto "principal" (habitual) indica o seu grau de fidelidade ao posto (e não à marca). Verificou-se que os condutores desta área de mercado abastecem no mesmo posto pelo menos 75% das ocasiões.

- **Probabilidade de Abastecer num Dado Posto de Abastecimento**

Na **Tabela 3** encontra-se a probabilidade de um condutor do sector **i** abastecer no posto **j** (P_{ij}), obtida, ponderando as preferências indicadas pelos consumidores com a frequência de visita a cada um dos postos referidos.

Os sectores 1, 2, 3 e 4, sendo freguesias do centro urbano principal, onde existem muitos postos, não permitem uma leitura fácil das preferências evidenciadas pelos condutores aí residentes. Os sectores periféricos são mais fáceis de analisar, por existir um pólo de atracção definido (centro urbano principal) e vias rodoviárias preferenciais de acesso a esse pólo.

Os habitantes do sector 5 não têm um posto de abastecimento na sua zona de residência, sendo os postos mais convenientes **C-S1** e **B-S1**, situados no acesso ao centro urbano principal (vide **Figura 3**), e **C-S2**, situado numa das principais saídas da Área de Mercado. Surge assim natural que o posto com probabilidade mais elevada neste sector seja o posto **C-S1** (0.268). Contudo, o posto **B-S1**, situado na mesma via, tem uma probabilidade muito baixa (0.04) o que poderá ficar a dever-se a factores de imagem, dimensão do posto ou qualidade do atendimento.

Tal como o sector 5 também o sector 6 não tem postos de abastecimento sendo os postos mais convenientes **A-S1**, **C-S1** e **B-S1**. O inquérito revelou que o posto com maior probabilidade de ser utilizado pelos condutores do sector 6 é o posto **A-S1** (0.597).

No sector 7 existem dois postos de abastecimento, **A-I1** e **E-I1**, que foram os mais citados pelos inquiridos desse sector. Podemos verificar na Tabela 4 que esses postos são os mais próximos da residência dos condutores desse sector.

	Sector						
Posto	1	2	3	4	5	6	7
A-C1	0.030	0.108	0.178	0.246	0.020	0.020	0.089
A-C2	0.390	0.141	0.122	0.223	0.112	0.005	0.089
A-I1		0.005	0.020	0.019		0.031	0.335
A-S1	0.055	0.009	0.152	0.028	0.073	0.597	0.123
A-S2	0.020	0.005	0.020	0.019			
B-C1	0.105	0.413	0.208	0.057	0.107	0.061	0.025
B-C2	0.270	0.174	0.061	0.052	0.078	0.020	0.034
B-S1	0.005		0.015	0.014	0.039		0.005
C-S2	0.015	0.023	0.005	0.043	0.190		0.015
C-S1	0.095	0.042	0.188	0,128	0.268	0.189	0.079
D-C1	0.015	0.019	0.005	0.076		0.005	0.015
E-I1							0.194
Outros		0.029		0.024	0.083	0.066	0.010

Tabela 3 - Probabilidade dos condutores do sector **i** abastecerem no posto **j**

- **Duração Média do Trajecto Casa/Posto de Abastecimento**

Para cada posto referido pelo inquirido, este indicava a duração (em minutos) do trajecto de sua casa até ao posto. Na **Tabela 4** encontram-se os valores médios entre cada sector e posto, calculados com base nas respostas dos inquiridos.

	Sector						
Posto	1	2	3	4	5	6	7
D-C1	4.50	3.50	5.00	5.00		15.00	20.00
A-C1	9.33	5.00	6.75	8.04	15.00	18.75	19.58
B-C1	5.80	5.31	5.89	7.00	16.82	16.43	17.50
B-C2	4.70	6.00	7.20	8.29	15.29	16.67	20.00
A-C2	5.42	5.43	6.27	8.14	13.36	25.00	17.10
B-S1	5.00		7.00	7.00	8.75		15.00
C-S1	10.71	8.57	4.54	8.25	14.50	13.47	11.17
A-S1	13.00	12.50	7.00	8.40	12.57	10.15	12.54
A-I1		25.00	20.00	11.00		10.33	4.44
E-I1							4.00
C-S2	9.00	10.00	15.00	16.67	9.54	40.00	40.00
A-S2	10.00	20.00	15.00	15.00			

Tabela 4 - Durações médias dos trajectos domicílio/posto de abastecimento (em minutos)

Comparando as **Tabela 3** e **4** verifica-se, para cada sector, que os postos de abastecimento com maior probabilidade são aqueles cuja duração do trajecto domicílio/posto de abastecimento é menor. Este facto reflecte a tendência geral de o abastecimento dos veículos ocorrer próximo da residência do consumidor.

Apesar do factor distância ser crucial na escolha do posto de abastecimento, os factores imagem e dimensão do posto também influenciam a escolha dos condutores. O posto **D-C1** encontra-se próximo da residência dos inquiridos dos sectores centrais mas, como é um posto de reduzidas dimensões e degradado, é pouco citado.

4. Desenvolvimento do Modelo Gravitacional

4.1 Estratégia de modelação

Para obtenção de previsões de vendas baseadas num modelo gravitacional começou-se por construir uma base de dados com as variáveis independentes e a variável dependente (P_{ij}) transformadas segundo a **Eq 2**. Numa segunda fase determinaram-se os parâmetros dos Modelos Gravitacionais e subsequente estimação das porbabilidades P_{ij}. Por fim, estimaram-se as Vendas de cada posto.

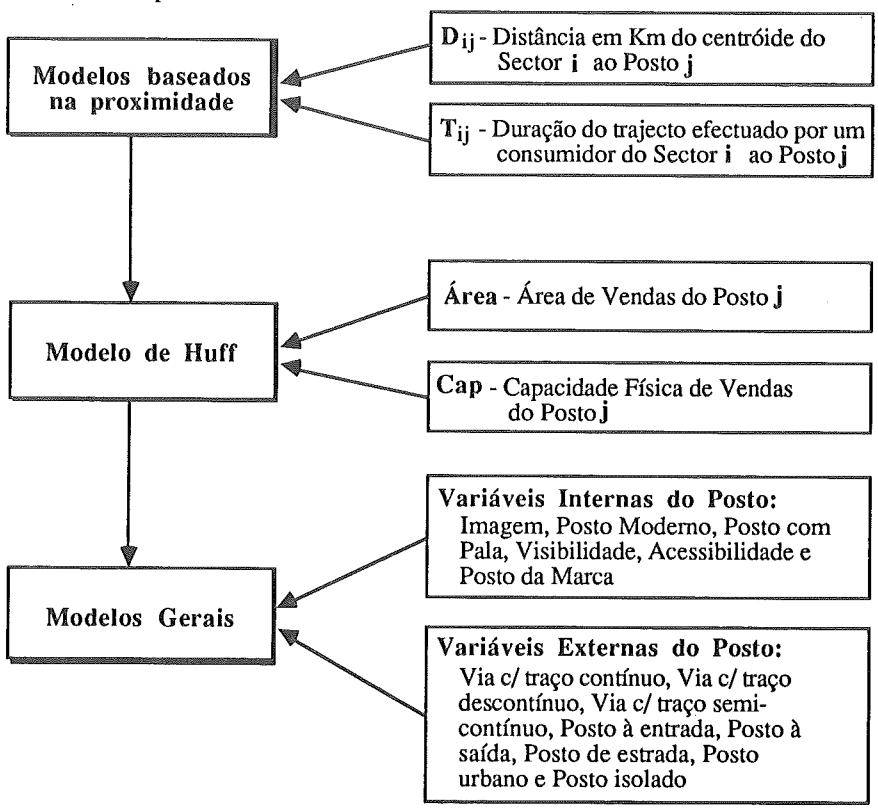

Figura 4 - Estratégia de modelação seguida para o desenvolvimento dos modelos gravitacionais

A estratégia adoptada para a obtenção dos modelos gravitacionais, segue os seguintes passos esquematicamente representados na **Figura 4**:

1. Determinação dos parâmetros de um modelo gravitacional baseado exclusivamente em variáveis de proximidade.
2. Determinação dos parâmetros para modelos do tipo de Huff, em que são consideradas variáveis de proximidade e de dimensão do posto [Huff, 1964].
3. Construção de modelos gravitacionais mais complexos com a inclusão de outro tipo de variáveis, nomeadamente variáveis de imagem, visibilidade, acessibilidade do posto e variáveis de localização.

Neste artigo iremos apenas apresentar os Modelos Gerais uma vez que apenas estes apresentam erros considerados aceitáveis (inferiores a 20%).

4.2 Modelação do consumo potencial de uma área

A utilização de modelos gravitacionais implica a construção de modelos para estimar o consumo potencial da área de mercado considerada. Existem essencialmente dois factores que afectam o consumo potencial de uma área - a oferta e o potencial da área (traduzido por variáveis demográficas e socio-económicas dos habitantes).

O inquérito telefónico realizado permitiu concluir que os consumidores de combustíveis líquidos abastecem preferencialmente na proximidade da sua zona de residência o que sugere que o aumento do número de posições de abastecimento tenderá a aumentar o consumo no sector.

A fórmula funcional adoptada[8] para explicar as vendas totais de combustíveis numa área de mercado (potencial do mercado) foi a seguinte:

$$\textbf{Eq 4} \qquad V = \alpha \cdot X_1^{\beta_1} \cdot X_2^{\beta_2}$$

onde V são as vendas totais anuais de combustíveis na Área de Mercado (somatório das vendas de todos os postos existentes nessa área); X_1 representa a oferta da Área de Mercado (somatório das áreas de venda de todos os postos); X_2 é o Potencial da Área de Mercado (representado por variáveis demográficas e/ou socio-económicas dessa área).

Os dados de consumo de combustíveis líquidos disponíveis em Portugal, são escassos e encontram-se agregados não permitindo uma análise por áreas geográficas da dimensão exigida neste estudo (concelhos)[9]. Uma vez que a escolha das variáveis demográficas estava à partida muito limitada optou-se por construir modelos fáceis de calibrar e de utilizar em que a variável do Potencial da Área de Mercado utilizada foi o número de habitantes. Na **Tabela 5** encontram-se os resultados do modelo do potencial, o qual foi estimado com base nas vendas dos vários

---

8 Esta fórmula funcional foi escolhida por comparação com modelos lineares. Estes apresentavam todos elevados desvios padrão além da qualidade do ajustamento ser sempre inferior à dos modelos multiplicativos.

9 Os dados de consumo de combustíveis líquidos pelas famílias portuguesas encontram-se disponíveis no INE agregados por regiões NUTE II.

concelhos que constituem o Distrito:

$$\text{Potencial}_{area} = \alpha \cdot \text{Hab}_{area}^{0.514} \cdot \text{Area Total Vendas}^{0.489}$$

Variáveis	Coef a,b,c,...	Desvio Padrão	t	Signif.
Constante	2.724	0.987	2.577	0.0130
Nº de Habitantes	0.514	0.115	4.449	0.0003
Área Total de Vendas	0.489	0.073	6.652	0.0000

R^2	0.949	D. Padrão	0.231	F da regressão	169.2
R^2 (ajust.)	0.944	Erro absoluto médio	11%	Graus de Liberdade	18

Tabela 5 - Resultados do Modelo do Consumo Potencial (variáveis logaritmizadas)

4.3 Modelo de previsão de vendas

Os modelos gravitacionais permitem estimar as probabilidades dos habitantes do sector **i** abastecerem no posto **j**. Os modelos estimados, baseados nas respostas de condutores residentes na área do inquérito, só permitem explicar a fracção das vendas originadas na área e não as vendas totais dos postos de abastecimento.

Para avaliar a importância das vendas a clientes residentes fora da Área de Mercado realizaram-se inquéritos aos condutores em dois postos de abastecimento: um posto urbano e um posto de estrada. Verificou-se que, em ambos os postos de abastecimento, cerca de 25% dos consumidores residiam fora da Área de Mercado. É, no entanto, importante salientar que cerca de 55% dos consumidores, não residentes na Área de Mercado, residem nas zonas limítrofes desta área. Logo, é admissível considerar que o consumo proveniente do tráfego de atravessamento é reduzido o que, na impossibilidade de o determinar em todas as posições, levou a ignorá-lo, admitindo-se que a totalidade das vendas são atribuíveis a residentes.

O potencial da Área de Mercado foi determinado pelo Modelo de Consumo Potencial (vide alínea anterior), substituindo a Área Total de Vendas e o número de habitantes pelos valores correspondentes:

Consumo Potencial da Área de Mercado = 16 620.6 m^3/ano

A previsão das vendas para cada posto será feita através da equação seguinte:

$$\textbf{Eq 5} \qquad \text{Vendas}_j = \sum_{i=1}^{I} (P_{ij} \cdot \text{Hab}_i) \cdot \frac{\text{Potencial}_{area}}{\text{Hab}_{area}}$$

onde: $\textbf{Vendas}_j$ são as vendas anuais de combustíveis (m^3) dos posto **j**; $\mathbf{P_{ij}}$ é a proporção de habitantes do sector **i** que abastecem habitualmente no posto **j**; $\textbf{Potencial}_{area}$ é o consumo potencial de toda a Área de Mercado; $\textbf{Hab}_i$ é o número de habitantes do sector **i**[10]; $\textbf{Hab}_{area}$ é o número total de habitantes da área de mercado ($\text{Hab}_{area} = \sum_{I} \text{Hab}_i$).

10 Fonte: Census de 1991, Instituto Nacional de Estatística.

4.4 Modelos gravitacionais generalizados

Os modelos gravitacionais, após a transformação log-centrada, podem ser tratados como sendo modelos de regressão comuns. Às variáveis binárias aplicou-se uma transformação exponencial sugerida por Mahajan et al (1978).

Os modelos apresentados na **Tabela 6** foram obtidos por regressão passo a passo com conjuntos diferentes de variáveis independentes. Estes modelos são aqueles cujos sinais dos coeficientes traduzem um sentido de variação que consubstancia o conhecimento teórico do fenómeno, são estatisticamente significativos para um nível de significância de 5% e simultaneamente as vendas apresentam um erro absoluto global inferior a 20%, para a área de mercado. O erro absoluto global de toda a Área ($erro_{area}$) é estimado do seguinte modo:

$$\textbf{Eq 6} \qquad erro_{area} = \frac{\sum_{j=1}^{m} |V_j^{obs} - V_j^{prev}|}{\sum_{j=1}^{m} V_j^{obs}}$$

onde: V_j^{obs} são as vendas anuais de combsutíveis no posto **j**; V_j^{prev} são as vendas anuais de combustíveis previstas para o posto **j**.

A variável representativa da proximidade é sempre a duração do percurso entre o sector **i** e o posto **j** (T_{ij}). Nos modelos em que a variável representativa da proximidade é a distância (D_{ij}), o erro absoluto global para a área de mercado é sempre superior a 20% o que inviabiliza a sua utilização na previsão de vendas dos postos de abastecimento.

Como variável representativa da dimensão dos postos de abastecimento figura sempre a Capacidade em vez da Área de Vendas, o que poderá ficar a dever-se ao facto da variável Capacidade estar mais correlacionada com as proporções $\mathbf{p}_{ij}$ do que a variável Área de Vendas.

A variável Potencial Local, uma medida do tráfego junto ao posto de abastecimento, não figura em nenhum modelo gravitacional, o que poderá ser explicado atendendo ao facto dos postos centrais (i.e. os mais próximos para a maioria dos consumidores) serem também aqueles onde o tráfego é mais intenso.

O valor R^2 (ajustado), uma medida da qualidade do ajustamento, é sempre superior a 54%. Comparando com os valores apresentados na literatura, embora para outros modelos gravitacionais e outros tipos de redes de retalho, eventualmente mais complexas, verifica-se que aquele é pelo menos da mesma ordem de grandeza: Stanley e Sewall (1976) apresentam um R^2 de 51.6%, Gautschi (1981) um R^2 de 35% e Black, et al (1985) apresentam um R^2 (ajustado) de 55%.

Para determinar as Vendas Anuais de Combustíveis (m^3) previstas para cada posto aplicou-se a Eq 5, obtendo-se $erros_{área}$ inferiores a 17%.

		MODELOS		
		I	II	III
VARIÁVEIS		$P_{ij} = \frac{T_{ij}^{\beta_1} I_j^{\beta_3} A_j^{\beta_4} PU_j^{\beta_5}}{\sum_{j=1}^{m} T_{ij}^{\beta_1} I_j^{\beta_3} A_j^{\beta_4} PU_j^{\beta_5}}$	$P_{ij} = \frac{T_{ij}^{\beta_1} Cap_j^{\beta_2} V_j^{\beta_6}}{\sum_{j=1}^{m} T_{ij}^{\beta_1} Cap_j^{\beta_2} V_j^{\beta_6}}$	$P_{ij} = \frac{T_{ij}^{\beta_1} I_j^{\beta_3} PC_j^{\beta_7}}{\sum_{j=1}^{m} T_{ij}^{\beta_1} I_j^{\beta_3} PC_j^{\beta_7}}$
Tempo (em min) - T_{ij}	β_1	-1.666 (0.000)	-1.831 (0.000)	-1.694 (0.000)
Capacidade - **Cap**	β_2		1.257 (0.000)	
Imagem - **I**	β_3	0.645 (0.009)		1.276 (0.000)
Acessibilidade - **A**	β_4	1.572 (0.002)		
Posto Urbano - **PU**	β_5	0.360 (0.022)		
Visibilidade - **V**	β_6		0.861 (0.001)	
Posto Comp. A - **PC**	β_7			0.481 (0.001)
R^2 (ajust)		0.597	0.544	0.600
D. Padrão		0.801	0.853	0.799
F		26.63	28.40	35.57
Durbin-Watson		1.91	1.78	1.89
$erro_{area}$		14%	17%	12%

Tabela 6 - Resumo dos modelos gravitacionais generalizados (número de observações = 69)

4.5 Escolha do modelo

A verificação da existência de heteroscedasticidade, é particularmente importante no caso dos modelos gravitacionais porque afecta o processo de estimação dos coeficientes. Nakanishi e Cooper (1974) demonstram que o termo do erro, nos modelos gravitacionais, não tem variância constante, quando os coeficientes são estimados pela **Eq 2** ou pela **Eq 3** através de modelos de regressão de mínimos quadrados (OLS). Para reduzir o erro das estimativas Nakanishi e Cooper (1974) sugerem a estimação dos coeficientes através da técnica GLS (Two-Stage Generalized Least-Squares).

Tendo em conta o erro de previsão, a não violação das hipóteses básicas dos modelos (linearidade, ausência de multicolinearidade, normalidade e homoscedasticidade dos resíduos) e

a robustez do modelo, seleccionou-se o modelo gravitacional II como o mais adequado:

$$\textbf{Eq 7} \qquad \text{Vendas}_j = \sum_{i=1}^{I} \left[\frac{T_{ij}^{-1.83} \cdot Cap_j^{1.25} \cdot V_j^{0.861}}{\sum_{j=1}^{m} T_{ij}^{-1.83} \cdot Cap_j^{1.25} \cdot V_j^{0.861}} \right] \cdot Hab_i \cdot \frac{Potencial_{area}}{Hab_{area}}$$

onde, $\mathbf{Vendas_j}$ são as vendas previstas de combustíveis líquidos (m^3/ano) para o posto **j** (j = 1, 2,..., m); $\mathbf{T_{ij}}$ é a duração média do trajecto (em minutos) entre o sector **i** (i = 1, 2,..., I) e o posto **j**, para o condutor do sector **i**; **Cap** é a capacidade do posto (em unidades de capacidade); **V** é a visibilidade do posto **j** medida numa escala de 1 a 5; $\mathbf{Hab_i}$ é o número de habitantes existentes no sector **i**; $\mathbf{Potencial_{area}}$ é o potencial de consumo da área do inquérito; $\mathbf{Hab_{area}}$ é o número total de habitantes da área de mercado. Refira-se que embora estimado pelo método dos mínimos quadrados (OLS), o modelo é homoscedástico.

Considerou-se que o modelo II, embora com um erro absoluto global ligeiramente superior aos outros dois, era preferível por incluir uma variável que mede a capacidade física do posto. Esta é uma variável estruturante que se julgou dever figurar até pelo risco de a sua não inclusão poder de futuro ser interpretado como sinal de que as vendas não dependem da capacidade do posto, o que não é a verdade. Acontece sim que, face à correlação existente entre a variável Capacidade e a variável Imagem (r = 0.86), esta substitui aquela nos modelos I e III.

Na **Figura 5** encontram-se os valores das Vendas Anuais de Combustíveis previstas e observadas para o modelo gravitacional II.

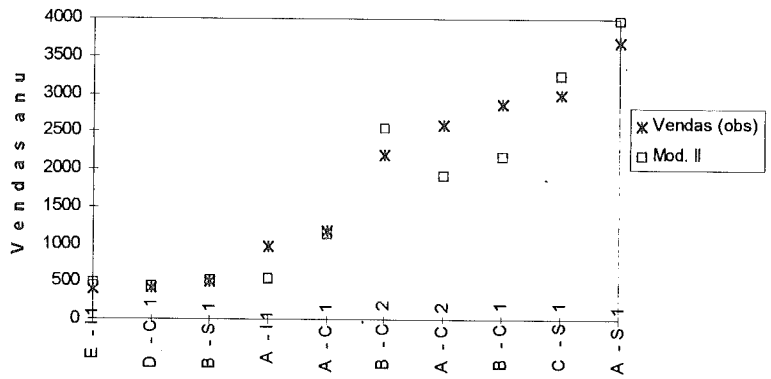

Figura 5 - Valores das Vendas Anuais (m^3) observadas e previstas pelo modelo gravitacional II, por posto de abastecimento

5. Aplicação do Modelo

Uma das vantagens dos modelos gravitacionais é permitirem análises sobre as vendas individuais dos postos de abastecimento da área se as características de um deles (ou vários) se alterarem. Num mercado em permanente alteração, a melhoria da imagem, condições de

atendimento ou dimensões de um posto implica modificações nas quotas de mercado que poderão afectar negativamente postos de abastecimento da mesma companhia.

Em todas as análises realizadas considerou-se o modelo II e os efeitos sobre a companhia A, sendo obviamente possível efectuar um estudo semelhante para outra companhia.

5.1 Introdução de um posto duplo

A introdução de um novo posto de abastecimento na área irá provocar a diminuição ou manutenção das quotas de vendas dos postos existentes, provocando sempre um aumento da quota global da companhia.

A introdução de um novo posto poderá traduzir-se pela duplicação de um posto já existente, do outro lado da via, transformando-o num posto duplo. O único posto da companhia A que poderia ser eventualmente transformado em posto duplo é o posto **A-S1**. Admitiu-se que o novo posto teria exactamente as mesmas características do posto já existente.

A transformação de **AS-1** em posto duplo implica um aumento da quota de mercado global da companhia de 43% para 50%, registando os postos da companhia A já existentes uma diminuição na quota de vendas de 6.7% (vide **Figura 6**). A quota do posto **AS-1** passa de 22% para 18%, mas no conjunto o posto **AS-1** e o novo posto passariam a representar 31% do mercado.

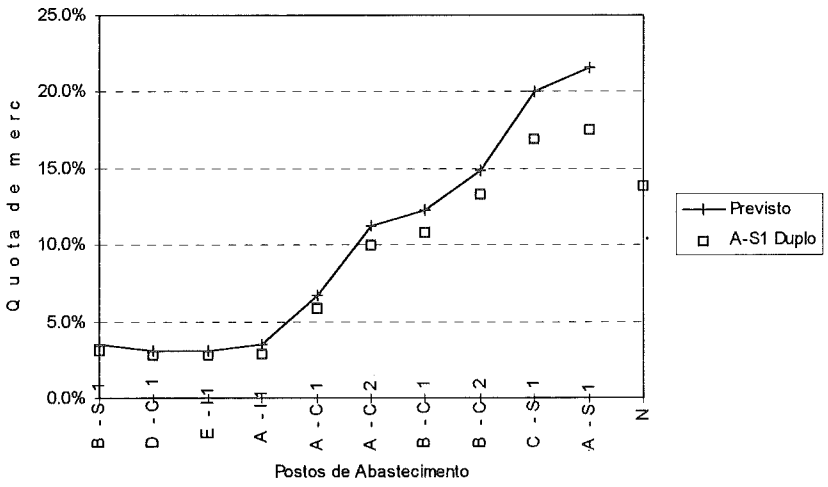

Figura 6 - Transformação do posto de abastecimento **A-S1** em posto duplo, designando-se o posto novo por **N**

A utilização deste resultado deve ser feita com algum cuidado, uma vez que na amostra não existem postos duplos.

5.2 Introdução de um posto num sector sem postos

A introdução de postos de abastecimento em sectores que actualmente não possuem nenhum posto de abastecimento pode ser uma alternativa interessante. Escolheram-se três localizações possíveis:

1. Posto no sector 4, na fronteira da área.
2. Posto no centro do sector 5.
3. Posto no centro do sector 6.

Considerou-se que todos os postos tinham as características médias dos postos existentes na área (Capacidade = 5; Visibilidade = 2). Na impossibiliade de fazer novo inquérito telefónico, a duração média do percurso entre os novos postos e os centróides dos sectores foi estimada através da relação entre a duração do percurso e a distância quilométrica obtida através dos dados recolhidos pelo inquérito.

Na **Figura 7** encontra-se representado o efeito que teria a introdução destes novos postos. Um posto novo no sector 4 não traria benefícios assinaláveis para a companhia **A**, passando a sua quota de mercado de 43% para 46%. A quota do novo posto seria de 4.7% e os postos já existentes da companhia perderiam 2%.

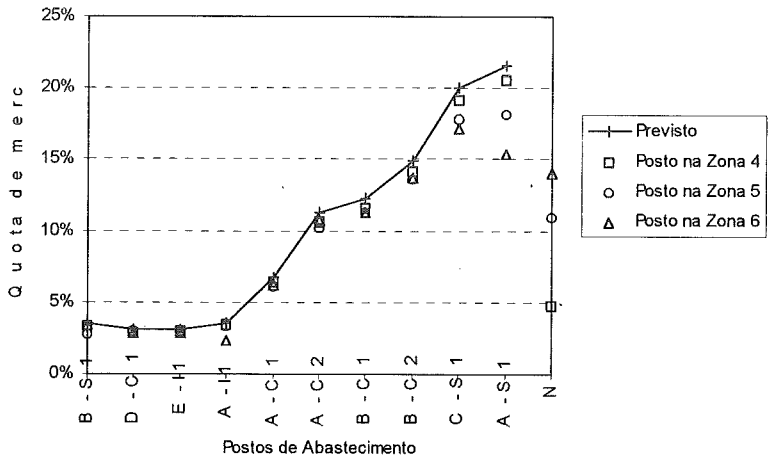

Figura 7 - Efeito da introdução de postos de abastecimento em sectores sem postos, designando-se o posto novo por N

A introdução de um posto no sector 5 já seria mais favorável; a quota global de mercado da companhia **A** passaria para 48.5%, registando os postos já existentes 5.4% de perdas sendo a quota do novo posto 11%. Note-se que o posto mais prejudicado seria o posto da concorrência **C-S1**.

A construção de um posto no sector 6 seria muito prejudicial para as vendas do posto de abastecimento **A-S1** que passaria de uma quota de 21% para 14%, não sendo os restantes

postos da companhia praticamente afectados. A quota global da companhia A passaria a ser de 48.6% contra os actuais 43%, sendo a quota do novo posto 14%.

Em conclusão pode afirmar-se que a construção de um posto no sector 4 praticamente não afecta os postos já existentes da companhia A, mas pouco melhora o desempenho global. A introdução de um posto no sector 6 é particularmente penalizador para o posto A-S1, sendo cerca de 60% das vendas do novo posto geradas a partir dos postos da companhia A já existentes. Assim, a localização mais favorável para o posto novo da companhia A seria no sector 5, embora acarretando apenas um aumento de 5.4% na quota de mercado dessa companhia.

6. Conclusões

Os resultados do inquérito telefónico, aos condtutores da Área de Mercado, mostraram que 75% dos condutores abastecem habitualmente no mesmo posto o qual se situa próximo da sua residência. Deve salientar-se que este resultado está de acordo com os outros estudos de mercado realizados pela empresa, a nível nacional, nos quais se conclui que 61% dos condutores abastece próximo da sua residência.

A utilização dos modelos gravitacionais implicou o desenvolvimento de um modelo auxiliar para estimar o consumo potencial da Área de Mercado. As variáveis explicativas desse modelo de regressão são o número de habitantes e a área total de vendas, da área de mercado, variável que traduz a oferta nessa área. A capacidade explicativa alcançada foi de $R^2 = 94.4\%$, sendo o erro absoluto médio igual a 11%.

Desenvolveram-se modelos gravitacionais generalizados em que se considerou um conjunto lato de variáveis que foram seleccionadas recorrendo à metodologia passo a passo, tendo-se obtido erros absolutos globais entre 12% e 17%.

A variável de proximidade que demonstrou maior poder explicativo foi a duração do percurso entre o consumidor **i** e o posto de abastecimento **j**, variável esta que traduz a percepção que os consumidores têm da distância da sua residência ao posto de abastecimento. Demonstrou-se ainda que os modelos gravitacionais permitiram estudar o efeito da alteração do número de postos na área.

A qualidade dos modelos gravitacionais depende, dos padrões de deslocação das populações. O tipo de produto estudado, combustíveis líquidos, é particularmente adaptado para a aplicação destes modelos, já que o padrão de deslocação é único - veículos automóveis, e o produto não é substituível. Estas características do mercado são certamente responsáveis pelos bons resultados obtidos, quando comparados com os citados na literatura para outros mercados.

References

[1] Achabal, Dale, D., Wilpen, L.G. and Vivaj M., *Multiloc: a multiple store location decision model*, Journal of Retailing 58 (1982) 5-25.

[2] Black, W.C., Ostlund, L.E. and Westbrook, R.A, *Spatial demand models in an intrabrand context*, Journal of Marketing 49 (1985) 106-113.

[3] Bottum, M.S., *Retail gravity model*, The Appraisal Journal (1989) 166-172.

[4] Carter, C.C, *Assumptions underlying the retail gravity model*, The Appraisal Journal (1993) 509-517.

[5] Cooper, L. and Nakanishi, M., *Market-share analysis - evaluating competitive marketing effectiveness*, Kluwer Academic Publishers, Boston (1993) 103-176.

[6] Fernandes, C., *Modelação das vendas de combustíveis líquidos recorrendo a modelos gravitacionais*, Tese de Mestrado, Instituto Superior Técnico (1996) Lisboa.

[7] Gautschi, D.A., *Specification of patronage models for retail center choice*, Journal of Marketing Research 18 (1981) 162-174.

[8] Ghosh, A. and Craig, C.S., *Formulating retail location strategy in a changing environment*, Journal of Marketing 47 (1983) 56-68.

[9] Huff, D.L., *Defining and estimating a trading area*, Journal of Marketing 28 (1964) 34-38.

[10] Ingene, C.A. and Brown, J.R., *The structure of gasoline retailing*, Journal Retailing 63 (1987) 365-392.

[11] Kaufmann, P.J. and Rangan, V.K., *A model for managing system conflict duting franchise expansion*, Journal of Retailing 66 (1990) 155-173.

[12] Lilien, G.L. and Kotler, P., *Marketing decision making - a model building approach*, Harper & Row Publishers, New York (1983) 445-464.

[13] Mahajan, V., Jain, A.R. and Ratchford, B.T., *Use of binary attributes in the multiplicative competitive interactive choice model*, Journal of Consumer Research 5 (1978) 210-215.

[14] Nakanishi, M. and Cooper, L.G., *Parameter estimation for a multiplicative competitive interaction model-least squares approach*, Journal of Marketing Research 11 (1974) 303-311.

[15] Nakanishi, M. and Cooper L.G., *Simplified estimation procedures for MCI models*, Marketing Science 1 (1982) 314-322.

[16] Penny, N.J. and Broom, D., *The tesco approach to store location*, in N. Wrigley (ed.), Store Choice Location and Market Analysis, London Routledge (1988).

[17] Quintino, A., *Modelação de vendas de combustíveis líuidos*, Tese de Mestrado, Instituto Superior Técnico, Lisboa (1994).

[18] Reinitz, E.C., *Sales forecasting model for gasoline service stations*, private correspondence, in Lilien, G.L., Kotler, P., Marketing Decision Making - A Model Building Approach, Cap. 13 (1983) 445-464, Harper & Row Publishers, New York.

[19] Rogers, D., *A Review of Sales Forecasting Models Most Commonly Applied in Retail Site Evaluation*, International Journal of Retail & Distribution Management 20 (1992) 3-11.

[20] Stanley, T.J. and Sewall, M.A., *Image Inputs to a Probabilistic Model: Predicting Retail Potencial*, Journal of Marketing 40 (1976) 48-53.

[21] Themido, I.H., Quintino, A. and Leitão, J., *Modelling the Retail Sales of Gasoline in a Portuguese Metropolitan Area*, contribuição nacional para a conferência internacional IFORS, 14 Julho 96, Canadá, a publicar na International Transactions in Operational Research (ITOR).

INVESTIGAÇÃO OPERACIONAL, vol. 17 (Junho 1997) 61-70

ARE ISO 9000 QUALITY SYSTEMS COMPATIBLE WITH TQM?

J.A. Sarsfield Cabral

DMEIM
Universidade do Porto
Rua dos Bragas
4099 Porto Codex - Portugal

Abstract

The implementation of the ISO 9000 standards is increasing significantly all over the world. However, many companies found that the ISO 9000 systems have not improved significantly their competitiveness or even the quality of their products or services. Some authors believe that the ISO 9000 standards are good foundations for a Total Quality system, but others claim that the ISO 9000 quality assurance models can cause difficulties when implementing a Total Quality programme. This question is analysed and the role of those formal quality systems is clarified.

Resumo

A utilização das normas ISO 9 000 tem crescido de forma significativa em todo o mundo. No entanto, muitas empresas verificaram que com os sistemas ISO 9 000 a competitividade não aumentou significativamente, nem mesmo a qualidade dos seus produtos ou serviços. Algures autores pensam que as normas ISO 9 000 constituem uma boa base de sustentação para um sistema de gestão pela Qualidade Total, enquanto que outros afirmam que tais modelos de sistemas de garantia da qualidade podem causar sérias dificuldades quando se pretende implementar um verdadeiro programa de Qualidade Total. Neste artigo, esta questão é analisada, clarificando-se o papel daqueles sistemas formais de garantia da qualidade.

Keywords

Quality assurance, quality standards, quality systems, TQM implementation, TQM philosophy.

1. Introduction

The world-wide implementation of the ISO 9000 standards increased significantly over the last years. According to the last Mobil Survey (ISO 9000 News, 1995), in January 1993 there were 27921 certified companies in the 48 countries that have adopted the ISO 9000 series as their national standards for quality systems. In June 1994 (i.e. one and a half year later), those figures rose to 70 544 companies over 76 countries, and at least 95 676 certificates of conformance to ISO 9000 series had been awarded in 86 countries up to the end of March 1995, (see Figure 1). In 1995 about 2500 new ISO 9000 certificates were issued each month. In some countries like New Zeland, Italy, USA and Japan, the growth rates were 1900%, 2300%, 4400% and 6400%, respectively. There is no doubt that the "ISO 9000 phenomenon represents an impact of standardisation unprecedented in modern history" (Durand et al 1993).

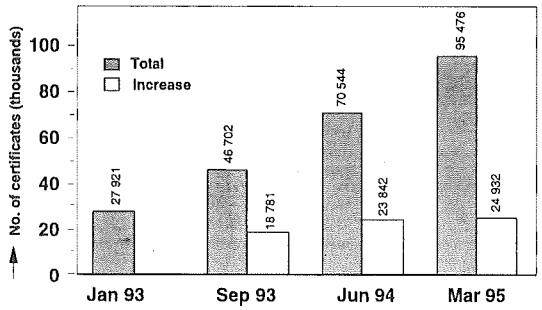

Figure 1 - ISO certificates awarded worldwide (Source: Mobil Survey, ISO 9000 News, 1995)

Although those figures are quite impressive, for the time being the ISO 9000 business can be regarded as an essentially European event: about 75% of the total certificates belong to European countries, the United Kingdom alone sharing 46% of the world total - 44107 certificates by March 1995 (see Figure 2). Moreover, between June 1994 and March 1995 the United Kingdom was still the country where the growth of certificates had the highest increase (7284), followed by the United States (5954) and by Germany (5875).

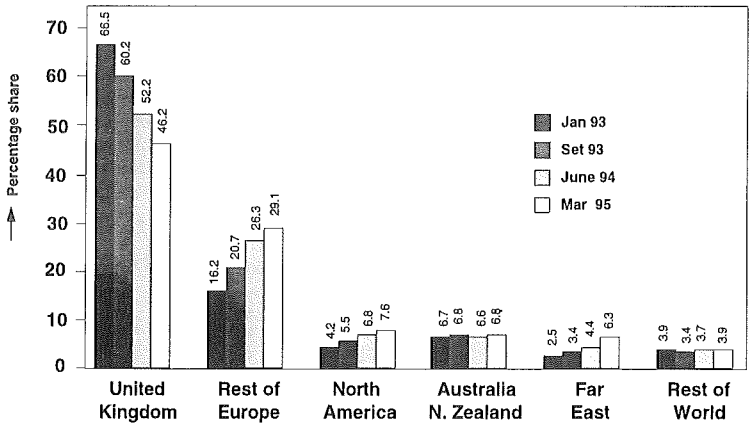

Figure 2-ISO certificates: regional comparison (Source: Mobil Survey, ISO 9000 News, 1995)

Many companies move to ISO 9000 because they expect that internal improvements will arise from the ISO 9000-certification process. According to the author's experience, the exercise is usually beneficial: at least, the certification process allows companies to 'clean' and to stabilise the functional process, and sustain the quality assurance procedures under control. "The main benefit came not from ownership of an ISO 9000 series certificate, but through the process of meeting the standard's requirements" (McTeer and Dale, 1995).

But the success of the ISO 9000 series derives not only from the increasing concern and commitment to Total Quality Management (TQM). It seems that the main force driving companies to pursue ISO 9000 registration is to fulfil a prerequisite for selling goods to client

companies or to obtain a "passport" for exporting. Quoting Blackham (1994), "ISO 9000 is becoming a fundamental contractual requirement for doing business with governments, trading blocs and withim many major economic sectors". Refering to 55% of the respondents of a quality managers survey, the same author states that "... their entire motivation in implementing ISO 9000 has been to get their costumers off their blacks!". Based on the opinions of 105 European experts, Chauvel (1994) says that "the reasons that lead companies to pursue certification are complementary. These reasons are new requirements by client companies and export demands". According to the International Secretary of the ISO/TC 176 (the ISO Quality Management and Quality Assurance Technical Committee), "for the global market place at the moment, it can be reasonably estimated that at least 85% of the use of these Standards is primarily for third party supplier registration, i.e., where the purpose is the affirmation of meeting minimum requirements in procurement or commercial contracts rather than a self imposed internal system for quality effectiveness and continuous improvement" (Ford 1994). More recent empirical evidence on the reasons for seeking ISO 900 registration goes in the same direction (see Weston, 1995 and Struebing, 1996).

Up to this moment the role of ISO 9000 series in quality or competitiveness improvement is not clearly established. For example, a survey based on a sample of 222 British companies from the mechanical engineering manufacturing sector reveals that the ISO 9000-registered companies are two times more profitable than the non-registered competitors (ISO 9000 News, 1996). Another survey conducted with ISO 9000-certified organisations in Belgium finds that many of their quality managers consider that ISO 9000 enabled them to deliver better service to customers (Vloeberghs and Bellens, 1996). On the other hand, a study ordered by the European Commission says that "more than two thirds of the companies stated that there is no significant difference between certified and non-certified suppliers with respect to the reliability of deliveries, to the quality of products and to the number of complaints", (European Commission, 1995). Dr. Tito Conti, former President of the European Organisation for Quality, argues that "certification in itself does not offer any long-term competitive advantages" (Conti, 1995). A number of articles in business newspapers and magazines supports the idea that those quality systems can even cause difficulties on the implementation of a 'true' Total Quality programme (see, for example, Juran, 1994). This debate is the central topic discussed in this paper.

2. The nature of the ISO quality systems

The ISO 9000 series were first published in 1987. After more than four years of international negotiations conducted by ISO/TC 176, several modifications were introduced to all the five original standards (ISO 9000, 9001, 9002, 9003 and 9004). The changing process evolved under the guidance of a strategy document called "Vision 2000" (see Marquard et al, 1991). New updated versions were published in 1993 and in 1994 which include, for example, guidelines for the application of the ISO 9001 to the development, supply and maintenance of software, guidelines for the quality management of services, and the revision of the three basic

ISO 9001, 9002 and 9003 standards. The new versions are basically extensions of the previous ISO 9000 series.

What is the nature of ISO 9000? The standard proposes three alternative management systems (ISO 9001, 9002 and 9003) for quality assurance, i.e., for the "prevention of quality problems through planned and systematic activities" (Oakland 1993, p.16). If a company wants to certify according to one of the three quality assurance models proposed in the ISO 9000 series, the system has not only to be properly designed and implemented but the quality procedures and activities must be written and documented. The same applies to the quality policy and the organisation structure (responsibility, authority, etc.). This involves a considerable amount of bureaucratic tasks.

The objectives of the quality systems are (see Corrigan 1994):

- To achieve and sustain the quality of the product or service.
- To give management confidence that quality is being met.
- To give customer confidence that consistency is being delivered in the product or service.

It is arguable that the ISO 9000 systems can assure that the quality of a product or service is achieved and sustained (although, they certainly increase the likelihood of that being the case). Fundamentally, the ISO 9000 systems are designed for controlling what is done, emphasising the prevention and the early detection of errors. The ISO 9000 series is not specific in relation to the product or service intrinsic quality (i.e., the degree of compliance between market requirements and product or service characteristics). What they prescribe is a set of quality activities that, once implemented, assures that the tangible characteristics of a product or service are consistently sustained at some level. In other words, no matter which quality assurance system is adopted, the quality of a product or service is defined independently.

In practice, with the ISO systems "there is too much emphasis on conformance rather than on adequacy and/or effectiveness; meeting the requirements is the principal concern" (Stephens 1994). Either those requirements are defined by the client, or by specific regulations or even internally by the producer, in general they correspond to what Kano et al. (1984) labelled as the "must-be" and "one-dimensional" quality attributes. Kano's model considers the relationship between two aspects of quality: an objective aspect involving the fulfilment of a quality attribute, and a subjective aspect involving the user's sense of satisfaction or dissatisfaction. The "must-be" quality attributes are expected by the customer and their presence is accepted without creating satisfaction (for example, the existence of towels in a hotel bathroom). In turn their absence causes dissatisfaction. The "one-dimensional" quality attributes are those whose presence gives satisfaction and absence causes dissatisfaction (for example, the cleanness of the bathroom towels).

It is obvious that the "must-be" and "one-dimensional" quality attributes constitute the minimum and basic characteristics that must be assured to have access and to stay in the market.

From this angle the ISO 9000 standards are of great value, as far as they provide a sound and world-wide recognised management structure for assuring that those attributes are consistently met.

That structure comprises several quality management aspects: the organisational framework, document control and process control, quality audits, etc. Design control is only included in the more demanding model out of the three proposed by the ISO 9000 series - the ISO 9001. This standard institutes the obligation of establishing and maintaining procedures to control and verify the design, in order to ensure that the *specified requirements* are met. According to the meaning of the term "quality", those requirements should be the customer needs and expectations or, in short, the market requirements. Although the new version of the ISO 9001 includes a clause that aims at demonstrating that the final design conforms to user's needs, it is not considered in the standard the problem of how to grasp the market requirements and how those requirements are monitored and translated into production and service manageable characteristics. Referring again Kano's model, the "attractive quality elements" - unexpected attributes whose presence gives satisfaction, but whose absence is accepted - are typically kept aside the certification process (remember that, according to Ford (1994), at least in 85% of the situations the purpose for using ISO 9000 is the affirmation of meeting minimum requirements in procurement or commercial contracts).

Another point deserving some consideration is quality improvement. As mentioned, the quality systems proposed by ISO 9000 series are specially suited for quality assurance. Since they emphasise the need to do things right, not questioning if those things are right or not, the standards are not designed to convey a quality improvement attitude. When the standard is properly used and applied, some improvements should certainly result from internal quality audits, preventive and corrective actions, handling of customer complaints, etc. But, essentially, ISO 9000 standards provide the discipline, the accuracy, and the evidence to assure that a particular quality level is being met.

Although there is some debate about the meaning and the value of an ISO 9000 certificate (see, for example, Day, 1994), it is now generally accepted that the ISO 9000 series are not market oriented nor specially suited to cope with the dynamics of the ever changing customer requirements. Quoting Stephens (1994), "ISO 9000 series is intended as a set of standards on quality assurance systems, not for total quality systems!".

3. From ISO 9000 Registration to TQM?

In 1992, the chairmen and chief executive officers of nine major U.S. corporations, deans and professors of majors universities, and eminent consultants in TQM methods an principles approved a definition of TQM, clarifying the modern meaning of that philosophy (see Baker et al, 1994). The following are some important topics addressed in that definition:

- TQM is a people-focused management system that aims at continual increase in customer satisfaction at continual lower real cost.

- TQM is a total system approach (not a separate area or program) and an integral part of high-level strategy.
- TQM works horizontally across functions and departments, involves all employees, top to bottom, and extends backward and forward to include the supply chain and the customer chain.
- TQM stresses learning and adaptation to continual change as key to organisational success.

It is very important to notice that increasing customer satisfaction is the ultimate objective of TQM. Customer preferences and requirements are volatile and evolve under uncontrollable factors. In practice, a company must have a strong market-oriented attitude and must adopt a continuous improvement attitude to pursue that objective.

Empowerment is another important characteristic of the TQM philosophy. According to Becker et al (1994), "that means assigning power and responsability for producing quality output, and for improving the process by which the output is produced, to those involved in the production of that output". Following again Becker et al (1994), TQM should support problem solving teams empowered to identify root causes of problems, to collect data, to verify hypotheses, to propose and test solutions, and to implement and verify the efficacy of solutions.

Based on empirical evidence, the same authors concluded that organisations will probably be flatter after adopting TQM, that is, have fewer levels of supervision and have larger spans of control. Apparently, that structure eases the company adaptation to continual change, a strong element of the TQM philosophy. In short, TQM demands a non-hierarchical organisational structure.

Finally, another significant dimension of the TQM culture is leadership. According to Fuchs (1993), traditional forms of leadership honour stability, systems control and procedures. Where traditional executives are managers, TQM require leaders. Managers are concerned with control of resources, with making things happen by giving orders and issuing policies. Leaders establish credibility with their passion and their personal deeds.

Having reviewed the basic characteristics of the TQM philosophy, one can question whether they are compatible with the formal nature of the quality assurance systems proposed by the ISO 9000 series. As it will be seen, this is a debatable matter.

Several authors can be found in the literature for whom ISO 9000 is not only compatible with TQM but also beneficial. For example, Askey and Dale (1994) say that "ISO 9000 series provide a sound basis from which to progress toward TQM". But these authors also warn that "if registration is to fit into the TQM approach of the organisation, it must be developed as a pragmatic, non-bureaucratic system". To Corrigan (1994), "ISO 9000 might not be *the* path to TQM, but it could be *a* path to TQM". According to this author, even if ISO 9000 is not the path to TQM, it can't hurt. Another example comes from Craig (1994), when referring to the ISO

9000 benefits: "These benefits will be proportional to the effort put into the quality system and the registration effort. If minimal effort is expended simply to satisfy the requirements, the payoff also be minimal. However, much greater benefits can be acheived from an approach in which registration is viewed as part of a larger business strategy for building toward a TQM philosophy". Dissent opinions are also easy to find. Tannock (1991), for instance, says that the ISO 9000 models "inculcate a bureaucratic 'Theory X' to quality management. This is in fundamental conflict with most recent thinking in quality, which is concerned with TQM". Citing distinguished business consultants, Shipman (1993) says that "TQM is more than just an ISO certificate. The fear is that ISO 9000 can tie a company to inferior quality standards by ignoring technical change, economy efficiency and customer feedback. It can mean you institutionalise non-value-added processes". He also add that "neglect of continuous improvement could fossilise outdate best practices".

These last views are more in line with the author's experience in helping companies to obtain the ISO 9000 certificate. In fact, it takes no less than twelve months to implement an ISO quality system, including a large period devoted to paper work. "As certification moves along, most work is narrowly focused on the development of procedures that meet the standards and many companies find the approach very bureaucratic: the processes are driven by systems on paper. This is particularly true if certification becomes an end in itself: people are motivated during the pre-certification time; but as soon as the organisation is certified, motivation decreases and the temporary attitude of improvement fades away" (Simon and Hoes, 1994).

The extraordinary expansion of the standards is putting companies under a great pressure. ISO registration is becoming an end in itself, not forcing the change of attitudes. In this situation (which is likely to be the more common), "the certification exercise will certainly result in processes being in control, a certificate, but no changes over the long term and no improvement in the company's competitive positioning" (Simon and Hoes, 1994). Moreover, in this situation the certification process will probably impairs any further TQM efforts.

More light was brought into this problem by a survey on medium-sized Norwegian engineering companies (see Bredrup, 1994). The findings from the survey reveal that not only there are no indicators showing that companies devoted to ISO 9000 have, on average, achieved better results than others, but also that ISO 9000 companies are less customer oriented - an astonishing result! From the study emerges the evidence of a division between marketing and quality work. According to the author, this could be summed up in the statement "ISO 9000 companies are more concerned about following the rules than satisfying the customer". Indeed, it is not unusual hearing from those who are engaged in a certification process that the company starts working as if the third party auditor suddenly becomes the customer. Bredrup (1994) concludes that the findings may also be an early warning for ISO-certified companies to revise their quality strategy because "trust in ISO 9000 could inhibit the adoption of quality

improvement techniqes" All this is, obviously, in absolute disagreement with the TQM philosophy.

It is then possible to conclude that ISO 9000 quality systems are not compatible with TQM? The answer can be yes or no, depending on how the standards are understood and implemented. In fact, a true Total Quality System has two main objectives:

- to assure that quality is attained at a minimum cost, and
- to foster continuous quality improvement, i.e., to increase customer satisfaction.

Any company that wants to stay in the market must assure that their processes are under control and no defects are being delivered to the clients, requiring a sound quality assurance system, stability, procedures and discipline: a product-oriented and process-oriented attitude. This is the domain for the ISO 9000.

The second aspect demands leadership, empowerment, flexibility: a continuous change towards increasing customer satisfaction. The author believes that the contribution of the ISO 9000 standards for the achievement of this objective is only marginal.

To a certain extent, those two objectives are conflicting, and it is very difficult to obtain high levels of performance regarding both criteria. Many consultants and managers corroborate this view. For example, when mentioning the implementation of a new quality policy in the Rover group, Cullen (1995) says that "the failure of the quality assurance system to deliver product and service quality on its own led to the total quality improvement programme (TQI) in 1987. Up until the mid 1980s, Austin Rover had been amongst the companies which had 'a system and no passion' as Tom Peters describes them. After introducing the TQI programme, it was crucial not to fall into the opposite trap of having 'passion but no system'. The quality policy provides quality assurance in a total quality culture".

The success or failure of a TQM program is related to the quality system approach. The structure organisation of a company depends on its complexity, but should also be suited to its quality philosophy. For example, for a small business with a quality assurance philosophy, it is probably best to have a strong quality assurance department which performs all inspection and test. Alternatively, if the small business has a quality improvement philosophy, it is best for inspection and test to be performed by the manufacturing department, and the quality department should then be mainly responsible for obtaining, analysing, and publishing quality data. In this situation, there should be a multifunctional quality improvement board, chaired by a senior department head, which should not be the quality manager (Groocock, 1994). When the certificate is the main objective on the ISO 9000 process a quality assurance structure will be probably adopted. Further efforts towards an improvement approach will suffer from an inappropriate structure, causing strong resistance to change. When this occurs, the implementation of the ISO 9000 models can indeed be harmful to TQM.

On the other hand, "companies can benefit from the process of ISO certification if they approach the challenge with a focus on continuous improvement" (Simon and Hoes, 1994).

But, unfortunately, this is an unusual situation: all the evidence shows that the priority is to get the ISO certificate when answering to procurement or commercial requirements (see Ford, 1994).

Conclusions

There has been a growing interest on the ISO 9000 standards. Thousands of companies all over the world have been certified against these quality assurance standards. This phenomenon is basically motivated by the need of complying with commercial or contractual requirements rather than a self imposed internal system for quality effectiveness and continuous improvement.

ISO 9000 carries the risk of diverting companies from adopting the TQM philosohpy. In fact, some empirical evidence shows that, on average, ISO 9000-certified companies are more distant from their customers than others companies that are not devoted to the standards. Does this means that ISO 9000 standards are not compatible with TQM? Unless the implementation of the standard is integrated on a TQM approach, it seems that ISO 9000 models can impar a long term attitude towards quality improvement. This is in agreement with the opinions of senior European Union officials which recently voiced concern that the current emphasis on earning ISO 9000 certificates is not the path to quality assurance they want to promote. Accordingly, the European Council will back a proposal to desemphasize ISO 9000 registration and promote the European Quality Award (Zuckerman, 1996).

Finally, it can be concluded that "ISO 900 deals with vertical (or functional) processes and stand-alone certification can easily be turned into bureaucratic exercise. But as the first step toward TQM, certification can be used to 'clean' the functional processes, allowing organisations to define their quality mission and policy, build a quality assurance system and bring their procedures under controlled conditions, before launching systematic improvement efforts directed at their horizontal processes. Building TQM must drive the approach to certification" (Simon and Hoes 1994).

References

[1] Askey, J.M. and Dale, B.G., *From ISO 9000 Series Registration to Total Quality Management: An Examination*, Quality Management Journal 1 (1994) July.
[2] Becker, S.W., Golomski, W.A. and Lory, D.C., *TQM and Organisation of the Firm: Theoretical and Empirical Perspectives*, Quality Management Journal 1 (1994) Januray.
[3] Blackam, L.A., *What's Gone Wrong with ISO 9000?*, ISO 9000 News 3 (1994) May/June.
[4] Bredrup, H., *Standard Illusions*, European Quality 1 (1994) Sptember/October.
[5] Chauvel, A.M., *Quality in Europe: Toward the Year 2000*, Total Quality Management 5 (1994).
[6] Conti, T., *Critical Time for Certification*, European Quality 2 (1995) March/April.
[7] Corrigan, P.J., *Is ISO 9000 the Path to TQM?*, Quality Progress 27 (1994) May.
[8] Craig, R.J., *The No-Nonsense Guide to Achieving ISO 9000 Registration*, ASME Press (1994).
[9] Cullen, J., *Visualizing Improvement*, Automotive Technology International 95, Stirling Publications Limited (1995).
[10] Day, I.J., *The Development of Quality System Certification in Europe: What Have we Learned? What Are we Going?*, Proceedings of the 38th EOQ Annual Congress 1 (1994) June, Lisbon.
[11] Duran, I.G., Marquardt, D.W., Peach, R.W. and Pyle, J.C., *Updating the ISO 9000 Quality Standards: Responding to Marketplace Needs*, Quality Progress 26 (1994) July.

[12] European Commission, *Working document on A European Quality Promotion Policy or The European way towards Excellence*, Directorate-General III, Industry (1995) February, Brussels.
[13] Ford, K.C., *Perspectives for Evolution of the ISO 9000 Series of Standards*, 38th EOQ Annual Congress (1994) June, Lisbon.
[14] Fuchs, E., *Total Quality Management from the Future: Practices and Paradigms*, Quality Management Journal 1 (1993) October.
[15] Groocock, J.M., *Organising for Quality - Including a Study of Corporate-Level Quality Management in Large UK - Owned Companies*, Quality Management Journal 1 (1994) January.
[16] ISO 9000 News, *The Mobil Survey*, 4 (1995).
[17] ISO 9000 News, *ISO 9000-registered Companies are Twice as Profitable, Says Survey*, 5 (1996).
[18] Juran, J., *Juran's Message for Europe*, European Quality 1 (1994).
[19] Kano, N., Sheraku, N., Takahasi, F. and Tsuji, F., *Attractive Quality and Must-Be Quality*, Journal of the Japanese Society for Quality Control 14 (1984) 39-48.
[20] Markardt, D., Chove, J., Kensen, K.E., Petrick, K., Pyle, J. and Strahle, D., *Vision 2000: The Stategy for the ISO 9000 Series Standards in the '90s*, Quality Progress 24 (1991) May.
[21] McTeer, M.M. and Dale, B.G., *How to Achieve ISO 9000 Series Registration: a Model for Small Companies*, Quality Management Journal 3 (1995) Fall.
[22] Oakland, J.S, *Total Quality Management*, Second Edition, Butterworth-Heinmann (1993) Oxford.
[23] Shipman, A., *Quality Defects*, International Management (1993) May.
[24] Simon, J.C. and Hoes, F., *The Use of ISO 9000 Standards and Total Quality*, Proceedings of the 38th EOQ Annual Congress 1 (1994) June, Lisbon.
[25] Stephens, K.S., *ISO 9000 and Total Quality*, Quality Management Journal 2 (1994) Fall.
[26] Struebing, L., *9000 Standards?*, Quality Progress 29 (1996) January.
[27] Tannok, J.D, *Industrial Quality Standards and Total Quality Management in Higher Education*, European Journal of Engineering Education 15 (1991).
[28] Weston Jr., F.C., *What Do Managers Really Think of the ISO 9000 Registration Process?*, Quality Progress 28 (1995) October.
[29] Vloeberghs, D. abd Bellens, J., *Implementing ISO 9000 Standards in Belgium*, Quality Progress 28 (1995) October.
[30] Zuckerman, A., *European Standards Officials Push Reform of ISO 9000 and QS-9000 Registration*, Quality Progress 29 (1996) September.

INVESTIGAÇÃO OPERACIONAL, vol. 17 (Junho 1997) 71-86

A STANDARD GENETIC ALGORITHM FOR CLUSTERING WITH PRECEDENCE CONSTRAINTS

Margarida Vaz Pato*

Instituto Superior de Economia e Gestão
Universidade Técnica de Lisboa
Rua Miguel Lupi, 20
1200 Lisboa - Portugal

Lídia Lampreia Lourenço

Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
Quinta da Torre
2825 Monte da Caparica - Portugal

Resumo

O problema tratado neste artigo refere-se à classificação de N elementos num número máximo de M grupos disjuntos, satisfazendo as restrições de capacidade destes e as de precedência no agrupamento dos elementos. Como critério de agregação usa-se a minimização da dissemelhança total entre elementos colocados no mesmo grupo. Este problema de classificação pode ser aplicado, por exemplo, ao desenho de software.

É apresentada uma heurística genética com base numa codificação dos agrupamentos, caracterizada pela identificação do índice do grupo em que cada elemento é colocado. Os resultados da experiência computacional envolvendo a comparação da heurística genética com uma heurística melhorativa, e uma híbrida, indicam um melhor comportamento da heurística genética para problemas de pequena dimensão e para os problemas sem restrições de capacidade. Em relação ao tempo computacional a genética monstrou-se mais desfavorável do que a melhorativa.

Abstract

Our paper reports on the clustering of N items into a maximum of M non-overlapping groups subject to capacity and precedence constraints when grouping the items. The clustering criterion employed is that of total dissimilarity of items grouped together. This classification problem can, for instance, be applied to the clustering of tasks in software production projects.

The authors developed a genetic heuristic, based on a specific encoding to identify the group in which each element is inserted. Results of the computational experiments, involving comparison of the genetic heuristic with another improvement heuristic and a hybrid heuristic, indicate a favourable behaviour of the basic genetic for the smaller problems, as well as for the uncapacitated problems, in terms of the quality of the solution. However, for problems with a larger number of items, the genetic and the hybrid heuristics did not perform so well as the standard improvement heuristic. Although, in terms of computing time, the genetic heuristic is more expensive compared with the standard improvement heuristic, these experiements will encourage us to redefine the genetic procedure.

Keywords

clustering, process organization, precedence constraints, genetic heuristics.

---

* This research has been partly supported by the Centro de Investigação Operacional (FC-Universidade de Lisboa) within the PRAXIS XXI (NJICT) Project nr 2/2.1/MAT/139/94 and by the Centro de Matemática Aplicada (FCT-Universidade Nova de Lisboa).

1. Introduction

When dealing with a huge number of items identified by specific characteristics, it is often necessary to cluster them in order to study each group separately, or even to take each group as a single entity. Determination of the number of groups required for clustering, together with selection of the criterion used for the grouping are the first steps of cluster analysis. This is followed by the task of clustering the items on the basis of the criterion chosen and according to the specific nature of the situation.

Cluster analysis problems arise in many scientific areas such as biology, medicine, psychology, economics, computer science and even literature. A survey of methods and applications may be found, for instance, in Duran and Odell[5] or in Gnanadesikan and Kettenring[8].

In this paper we study a specific cluster analysis problem - the Clustering with Precedence Constraints Problem - where a maximum number of groups is pre-defined, as well as subsidiary constraints on group capacities and precedence constraints in the grouping of items. The clustering criterion calls for total dissimilarities among the items clustered in the same groups. The major applications for this problem are related to the assignment of tasks to work groups in large software production projects.

Some authors have formulated the problem as a mixed integer linear problem, and have used heuristics and branch-and-bound methods to tackle it. The problems solved are of small and medium size, and one may speculate that if the number of items ranges from some tens to hundreds, the methods, as they stand, would not be applicable, and enhanced heuristic versions should be developed. As constructive heuristics and standard exchange scheme heuristics could not, apparently, be significantly perfected, we decided to explore the evolutionary approach. Though the problem is highly constrained, the genetic algorithm may be used if the encoding and genetic operators incorporate all the problem constraints reasonably well.

The Clustering with Precedence Constraints Problem is described in Section 2. Section 3 presents a genetic algorithm for that problem, while Section 4 gives the results of the computational experiments for small and medium-sized instances, partly taken from literature. Section 5, the final section, points to modifications that could lead us to handle larger problems through genetic search.

2. The Clustering with Precedence Constraints Problem

Let us consider N items and real numbers d_{ij} (i=1,..., N-1; j=i+1,..., N), which stand for the dissimilarities between each pair of items. Known data include the maximum number of groups, M, the weight of each item (p_i, i=1,..., N), the upper bound on the number of items per group (B_k, k=1,..., M), and on the weight of each group (C_k, k=1,..., M), besides several precedence constraints that prevent items from being clustered in a group if their preceding items have not, as yet, been placed in the current group, or in any previous group.

The Clustering with Precedence Constraints Problem (CPCP) addresses the partition of the set of items into a maximum of M disjoint subsets, according to the criterion of minimization of the total amount of the dissimilarities among items placed together, while satisfying the capacity and precedence constraints explained earlier.

This problem has been proposed by Karimi[13] and Klein, Beck and Konsynski[14] to process organization in an information system. The processes of a system and their connections may reach a high level of complexity and in that case clustering into modules with minimum interconnections is a way of organizing the overall system.

Further possible applications suggested concern support for CAD/CAM software production, production of sofware for industrial process control and robotics. The problem of assigning jobs to independent processors of a computer, referred to in Sofianopoulou[18], with a slightly different clustering criterion (minimizing costs of communication between jobs, as well as the execution costs of the jobs) may share the same kind of constraints as the CPCP, besides a similar objective function.

Aronson and Klein[1] formulated the CPCP as an extension of another formulation[14] for a CPCP without capacity constraints. In the abovementioned CPCP formulation[1], the following parameters are required:

M - maximum number of groups, $M \geq 2$ and integer;

N - number of items, $N > M$ and integer;

d_{ij} - dissimilarity between item i and item j, $d_{ij} > 0$ for $i = 1,\ldots, N-1$ and $j=i+1,\ldots, N$;

B_k - maximum number of items for group k, $B_k \geq 0$ and integer for $k=1,\ldots, M$;

C_k - maximum weight for group k, $C_k \geq 0$ for $k = 1,\ldots, M$;

p_i - weight of item i, $p_i \geq 0$ for $i = 1,\ldots, N$.

Moreover, the precedence relations for some pairs of items, say (i,j), require that item j may only be grouped if item i has already been placed in a previous group. This assumes that the groups are ranked according to a specific order.

Let us now present the objective function and the model constraints:

$$\text{minimize} \qquad z = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}\, y_{ij} \qquad (0)$$

subject to

$$x_{ik} + x_{jk} - y_{ij} \leq 1 \qquad (i=1,\ldots, N-1;\ j=i+1,\ldots, N;\ k=1,\ldots, M) \qquad (1)$$

$$\sum_{k=1}^{M} x_{ik} = 1 \qquad (i=1,\ldots, N) \qquad (2)$$

$$\sum_{i=1}^{N} x_{ik} \leq B_k \qquad (k=1,\ldots, M) \qquad (3)$$

$$\sum_{i=1}^{N} p_i x_{ik} \leq C_k \qquad (k=1,\ldots, M) \qquad (4)$$

$$\sum_{k=1}^{M} k\, x_{ik} \leq \sum_{k=1}^{M} k\, x_{jk} \qquad \text{(for each pair (i,j) such that i precedes j)} \qquad (5)$$

$$y_{ij} \geq 0 \qquad (i=1,\ldots, N-1;\ j=i+1,\ldots, N) \qquad (6)$$

$$x_{ik} = 0, 1 \qquad (i=1,\ldots, N;\ k=1,\ldots, M) \qquad (7)$$

The variable x_{ik} (for all i and k) takes the value 1 if item i is clustered into group k, or 0 otherwise. As for the variable y_{ij} (also defined for all i and all j greater than i), when it is equal to 1 it means that items i and j belong to the same group, being 0 otherwise. Constraints (6) on the variables y_{ij} are simply of the non-negativity type, because minimization of the objective function (0) and constraints (1) and (7) force such variables to be binary in the optimum, without the need to impose the binary conditions.

As for the objective function (0), it represents the total sum of the dissimilarities for the items grouped together.

The first set of constraints (1) establishes the relation between the x_{ik} and the y_{ij} variables. The following set (2) enforces each item to be assigned to a unique group, whilst constraints (3) and (4) are defined to limit the number of elements and the total weight of each group. Finally, the precedence constraints are expressed in terms of inequalities (5).

In this formulation the size of the instances tends to be very high, even for small-sized problems. For instance, if the CPCP is used to classify 13 items into 8 groups with 12 precedence relations, the formulation has 182 variables and 685 constraints, without considering the constraints for definition of the variables ((6) and (7)).

Moreover, the CPCP is a generalization of a clustering problem classified as NP-hard in Garey and Johnson[7], thus placing the CPCP itself in the NP-hard class.

Therefore LP based techniques are not advisable and of course a complete enumeration of solutions is impracticable. Hence heuristics arise as natural approaches, at least for the larger-sized problems that may not be solved by exact methods within a reasonable amount of computing time.

Let us now refer to other similar clustering problems that could lead to different formulations and solution-finding methods.

It should be remembered that in formulation (0) to (7) for the CPCP, the y_{ij} variables are used to establish the objective criterion, whereas the other variables, the x_{ik}s, are used to impose the clustering restrictions.

If neither capacity nor precedence restrictions were imposed on the CPCP, the problem could be viewed as a quadratic semi-assignment problem, as shown in Gondran and Minoux[10]

(page 467). In their formulation, the unique set of (binary) variables translates the objective criterion and determines where each element should be placed. But if we consider precedence constraints, this quadratic model cannot be adapted to the situation. However, even if it were possible to model the CPCP as a binary quadratic problem, it would not be advisable to do so in view of the computing difficulties known for that kind of problem.

On the other hand, Jensen[12] presented a dynamic programming formulation for clustering N items into M groups without constraints, where the dissimilarities among the items placed in a group are penalized through the inverse of the cardinality of that group. Such a dynamic model may be applied, with minor modifications, to our CPCP, as the capacity and precedence constraints could easily be imposed on the model. But unfortunately dynamic programming calculations remain very hard to handle. Even in the case of small-sized problems, it would be difficult to solve this problem within a reasonable amount of computing time.

3. The Genetic Heuristic

Genetic heuristics are search procedures that rely to a large extent on basic biological rules for genetics. Holland[11] is considered to be the father of genetic algorithms as applied within the field of combinatorial optimization. Goldberg[9], Davis[4] and Michalewics[16] present the main features of genetic heuristics and the corresponding algorithms, besides many bibliographical references concerning their applications. Although theoretical support for genetics is, to date, mainly restricted to conventional binary codes, which should be applied only to particular problems (see for instance Reeves[17]), the genetic-based methodology has proved to be successful in practice, even for highly constrained combinatorial problems.

Literature has already referred to several uses of genetic algorithms in clustering problems (Michalewics[16] and Faulkenauer[6]), though they are simpler than our Clustering with Precedence Constraints Problem, as they do not embed precedence constraints.

On the other hand, Reeves[17] refers to a successful application of such algorithms to a problem with precedence constraints which, broadly speaking, is also a clustering problem - the instruction scheduling problem (Beaty[2]).

We shall now begin by explaining the genetic heuristic developed for the CPCP (Lourenço[15]). The algorithm is given in Subsection 3.1, while the four subsections that follow are devoted to the chromosome encoding and to the selection, crossover and mutation operators.

3.1 The Algorithm

The genetic algorithm, called GENET, is summarized in Figure 1.

Each individual of the population is identified by a chromosome encoding a feasible solution for the CPCP. The initial population has P individuals (here, the dimension of the population P was set to 20). Three of these 20 feasible solutions are created by a constructive

heuristic[15] briefly presented in Subsection 4.2, whereas the remaining 17 feasible solutions are randomly generated, following a procedure also given in that work[15].

```
procedure GENET
generate the population with P individuals
compute the fitness value for each individual
identify the fittest individual
gener = 1
while gener ≤ maxgener do
        act with selection operator
        act with crossover operator and update the fittest individual
        act with mutation operator and update the fittest individual
        gener = gener + 1
end GENET
```

Figure 1 - The Genetic Algorithm GENET

In each iteration of the algorithm or generation a new population is created. On 20 occasions, in each generation, the selection operator chooses one chromosome from the population, based on its fitness. Then the crossover, as well as the mutation operator, acts on the 20 chromosomes selected, thus creating the population for the subsequent generation.

When the maximum number of *maxgener* (here, equal to 100) generations is reached, the algorithm stops and produces the best individual (chromosome) found so far and the corresponding solution.

The features used in the algorithm were suggested by the abovementioned literature, as well as our own computing experience. Other versions were tried out but abandoned in view of their poorer behaviour, when compared to GENET, such as those below:

- a direct binary encoding scheme;
- the selection operator, acting only once at the initialization step;
- a crossover, where offsprings always replace parents if feasible.

3.2 The Encoding

Each individual is associated with one chromosome representing a feasible solution for the problem. The chromosome consists of a string of N genes, each standing for an item and indicating, through its allele, the group in which the item is placed (Figure 2).

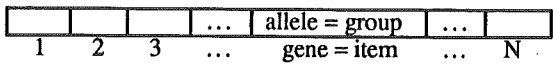

Figure 2 - A Chromosome

Let us consider a small instance of the CPCP, where 13 items are to be clustered into 8 groups. Other data for this case are:

$[d_{ij}]$ i=1,..., N-1; j=2,..., N =

$$\begin{bmatrix} 0.3 & 0.4 & 0.8 & 0.6 & 0.2 & 0.1 & 0.5 & 0.9 & 0.6 & 0.4 & 0.1 & 0.6 \\ - & 0.2 & 0.7 & 0.8 & 0.3 & 0.7 & 0.6 & 0.6 & 0.3 & 0.5 & 0.9 & 0.5 \\ - & - & 0.8 & 0.1 & 0.7 & 0.6 & 0.5 & 0.5 & 0.8 & 0.1 & 0.4 & 0.5 \\ - & - & - & 0.9 & 0.2 & 0.6 & 0.9 & 0.5 & 0.6 & 0.3 & 0.3 & 0.4 \\ - & - & - & - & 0.4 & 0.5 & 0.4 & 0.2 & 0.7 & 0.4 & 0.7 & 0.7 \\ - & - & - & - & - & 0.6 & 0.9 & 0.3 & 0.1 & 0.8 & 0.3 & 0.2 \\ - & - & - & - & - & - & 0.1 & 0.4 & 0.7 & 0.5 & 0.5 & 0.8 \\ - & - & - & - & - & - & - & 0.8 & 0.3 & 0.6 & 0.6 & 0.8 \\ - & - & - & - & - & - & - & - & 0.8 & 0.5 & 0.5 & 0.6 \\ - & - & - & - & - & - & - & - & - & 0.4 & 0.4 & 0.1 \\ - & - & - & - & - & - & - & - & - & - & 0.9 & 0.7 \\ - & - & - & - & - & - & - & - & - & - & - & 0.8 \end{bmatrix}$$

$[B_k]$ k=1,..., M = [5 6 5 4 3 4 8 8]

$[C_k]$ k=1,..., M = [9.0 15.0 14.0 4.0 9.0 3.0 10.0 12.0]

$[p_i]$ i=1,..., N = [1.4 2.6 3.1 2.8 2.0 1.2 2.2 2.2 1.9 2.4 2.2 1.0 1.8]

Finally, precedence constraints for this instance are given in Figure 3.

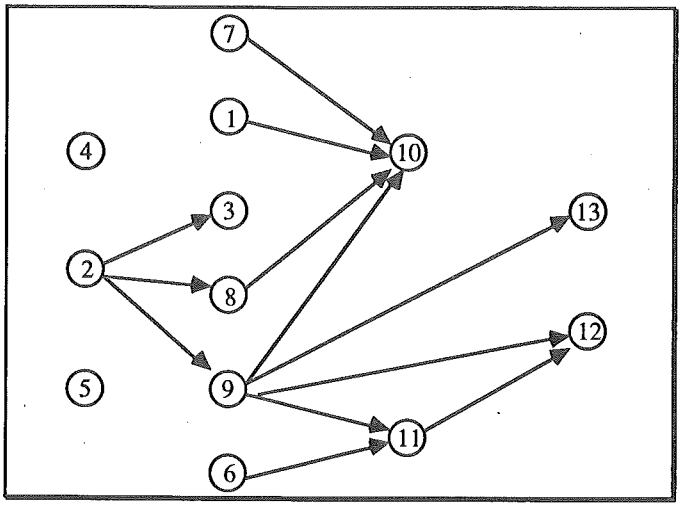

Figure 3 - Precedence Constraints for the Example

A clustering that fulfils all constraints for this case is illustrated in Figure 4. Here groups have the following composition: g_1={2}, g_2={5,9}, g_3={13}, g_4={4}, g_5={1,6}, g_6={11}, g_7={3,12} and g_8={7,8,10}.

The given encoding represents the specific clustering for the CPCP instance by the chromosome:

5	1	7	4	2	5	8	8	2	8	6	7	**3**

Here, item 13 is clustered in group 3 (the last gene with its allele in bold print) and both items 5 and 9 fall into group 2 (the fifth and ninth genes, respectively).

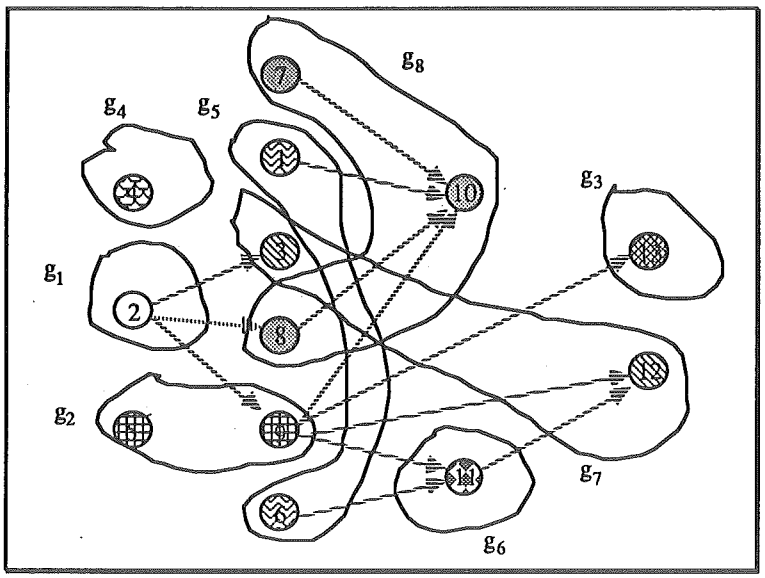

Figure 4 - A Clustering for the Example

We now have the corresponding solution for the CPCP formulation of Section 2: $x_{15} = x_{21} = x_{37} = x_{44} = x_{52} = x_{65} = x_{78} = x_{88} = x_{92} = x_{10,8} = x_{11,6} = x_{12,7} = x_{13,3} = 1$; $y_{59} = y_{3,12} = y_{16} = y_{78} = y_{7,10} = y_{8,10} = 1$; and all others equal to 0.

3.3 The Selection Operator

The selection of chromosomes to remain in the population for the current generation is based on their fitness values, that is, the probability of selection of an individual chromosome is proportional to its fitness. The fitness of each chromosome is supposed to reflect the objective value of our problem, which represents total dissimilarity of the corresponding feasible solution. Thus, the fitness value must be better for the lower dissimilarity chromosomes and always positive.

Considering that the minimization of the objective function of the CPCP given by (0) is equivalent to

$$\text{maximize} \quad z' = -\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}\, y_{ij} \tag{8}$$

and also to

$$\text{maximize} \quad z'' = -\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}\, y_{ij} + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij} \tag{9}$$

the fitness value for each chromosome will be expressed by the value of this last objective function (9) calculated in the respective feasible solution. Therefore, the fitness values are positive, as required, and better for the lower dissimilarity chromosomes.

The selection operator performs like a roulette wheel, where the fittest chromosome is associated with the largest sector of the wheel, thus having the greatest probability of being chosen.

Let us show how the roulette works in a case taken from the example. Take 5 chromosomes from the population and suppose that the corresponding feasible solutions have the following total dissimilarities or values of the objective function (0): 14.5, 3.6, 10.0, 10.6 and 7.0. These dissimilarities amount to 40.3 and, bearing in mind the reformulation of the objective given by (9), we find the following fitness values for the 5 chromosomes to be: 25.8, 36.7, 30.3, 29.7 and 33.3. Figure 5 displays the roulette for this selection process.

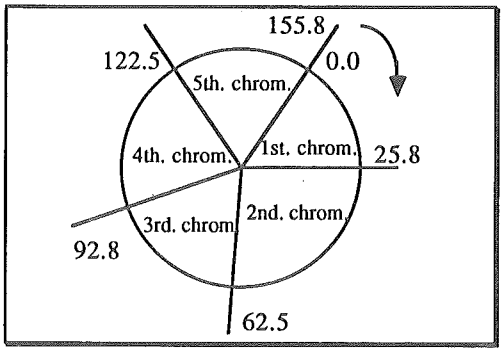

Figure 5 - Roulette for an Illustrative Case

Let us assume that the (uniformly) random generation of a real number between 0.0 and 155.8 produces number 65.0. The chromosome selected is therefore the third one, whose fitness value is 30.3. In fact, the addition of the first two fitness values gives 62.5 (below 65.0) and the addition of the first three values gives 92.8 (above 65.0).

The selection operator is repeated a certain number of times - in GENET on 20 occasions - and the resulting chromosomes build the population for the current generation. Some of the chromosomes may ultimately be identical.

3.4 The Crossover Operator

The number of crossovers in a generation is randomly generated between 1 and 5. These figures were set in accordance with the parameter for the dimension of the population ($P = 20$).

For each crossover a pair of chromosomes (the parents) is chosen by the roulette, thus giving priority to the fittest chromosomes. Then, in order to create the child-chromosomes, approximately 1/3 of the N genes are randomly chosen and their alleles are swapped from one chromosome to the other of the same pair.

As an illustration, let us take two parent-chromosomes of the instance given above:

7	1	7	5	4	5	8	8	2	8	6	7	3

1	2	3	3	1	1	1	2	2	3	2	2	2

Here [13×1/3] = 4, and so the offsprings will have 4 genes whose values result from swapping the corresponding alleles of their parents (all in bold print in the parents):

7	1	3	5	1	5	1	8	2	8	6	7	3

1	2	7	3	4	1	8	2	2	3	2	2	2

These correspond to the following two groupings respectively, and in this case happen to be feasible:

- $g_1=\{2,5,7\}$, $g_2=\{9\}$, $g_3=\{3,13\}$, $g_4=\{\ \}$, $g_5=\{4,6\}$, $g_6=\{11\}$, $g_7=\{1,12\}$ and $g_8=\{8,10\}$;
- $g_1=\{1,6\}$, $g_2=\{2,8,9,11,12,13\}$, $g_3=\{4\}$, $g_4=\{5\}$, $g_5=\{\ \}$, $g_6=\{\ \}$, $g_7=\{3\}$ and $g_8=\{7\}$.

This crossover operator was designed with the single purpose of creating solutions that are partly similar to the original ones, without considering the feasibility.

The two child-chromosomes, if feasible, substitute their parents, but only if they are better. If only one of the descendants is feasible, this feasible chromosome replaces the most ill-fitted parent. The parent-chromosomes may be selected for crossover twice or more often in the same generation.

As the population dimension is 20, once the crossovers occur, more than half of the population is kept identical to the previous generation and the remainder does at least retain some characteristics of their parents.

3.5 The Mutation Operator

The mutation operator may act on up to 1/5 of the population's chromosomes. In other words, the number of mutants can be 0, 1, 2, 3 or 4, in the case of the present implementation.

Each chromosome in which mutation will occur is randomly chosen among the population. This method also applies in the selection of two of its genes. For each of these two genes (items) the corresponding allele (the group wherein the item is clustered) is compared with the central index group (in the case of an even number of items, the central group is considered as the average of the two central index groups):

- if the allele is below the central group, one unit is added to it, i. e. the item shifts to the subsequent group;
- if the allele is above the central group, one unit is subtracted, and the item shifts to the previous group. This mutated chromosome is then tested against all the constraints, and only if it is feasible, does it replace the original.

To illustrate this operator, let us consider a chromosome for the same example problem:

7	1	3	5	1	5	1	8	2	8	6	7	3
1	2	3	4	5	6	7	8	9	10	11	12	13

central genes

which, after random determination of the two genes - the fourth and the ninth - mutates into:

7	1	3	4	1	5	1	8	3	8	6	7	3

The second chromosome still represents a feasible solution: $g_1=\{2,5,7\}$, $g_2=\{\ \}$, $g_3=\{3,9,13\}$, $g_4=\{4\}$, $g_5=\{6\}$, $g_6=\{11\}$, $g_7=\{1,12\}$ and $g_8=\{8,10\}$.

This chromosome-mutating operator was designed with two main aims. One is to concentrate the items in the neighbourhood of the central group. Such chromosomes will differ from those created by the constructive and improvement heuristics (see Subsection 4.2) because these procedures concentrate the items in the first and last groups, respectively. The other aim is to ensure a frequent verification of precedence constraints. However, as a lot of mutants are not feasible (often they do not respect group capacities) and are therefore omitted, we mutate frequently (up to 4 times in 20 chromosomes) to bring about some diversification among individuals in the population. In fact this operation is not rare in genetic heuristics, as opposed to natural biological evolution.

4. Computational Experiment

Computational experiments were carried out to test the performance of the genetic heuristic GENET following the algorithm of Figure 1 and two other heuristics: an improvement heuristic based on exchanges of items and a hybrid heuristic, produced by combining the improvement with the genetic procedure.

All the algorithms were coded in PASCAL and ran in a HP 486/33 VL personal computer.

We shall begin by describing the test instances in Subsection 4.1. This is followed by a short presentation of the comparison heuristics, in Subsection 4.2. Results are finally given in Subsection 4.3 and commented upon in Subsection 4.4.

4.1 Instances

The test instances came from a semi-random generation process, following the rules and parameters given in Aronson and Klein[1] (who took partial data for some problems from Tonge[19]), as no complete set of data for this kind of clustering problem - the CPCP - was found. The only exception is an instance to be referred to in Subsection 4.4, belonging to Type C problems.

Parameters of test problems were randomly generated within the bounds stated in Table 1: the number of groups, the maximum number of items, the maximum weight for groups and the weights of the items.

Generation of the dissimilarity matrix was carried out following two different rules:

- rule 1 - each element of the matrix is randomly drawn from the interval [0.1,0.9];
- rule 2 - two spatial coordinates are randomly generated between 1 and 100 for each item. Euclidean distances for each pair of items are then calculated, which leads to the dissimilarity matrix.

The precedence constraints were also set semi-randomly for some instances[15], whilst others were taken from literature - Class 2 problems.

Four different classes of instances were considered - Class 1 to Class 4 - and, for each one, 25 problems were generated by changing one of the parameters within the bounds.

For Class 1 two sets of 13 and 12 problems each were considered: problems Type A, with 2 to 5 groups and problems Type B, with 6 to 8 groups. Similar differences exhibit the following pairs of problem Types created respectively for Classes 2, 3 and 4:

- Type C (12 problems) and Type D (13 problems);
- Type E (7 problems) and Type F (18 problems);
- Type G (7 problems) and Type H (18 problems).

Some problems from Types C and D have a dissimilarity matrix copied from Aronson and Klein[1].

	Class 1	Class 2	Class 3	Class 4
	Types A and B	Types C and D	Types E and F	Types G and H
N	13	23	32	70
M	2 - 8	3 - 8	4 - 9	3 - 30
nº of precedence constraints	12	25	25	86
maximum nº of items per group	3 - 8	6 - 14	6 - 14	3 - 30
maximum weight per group	3.0 - 15.0	8.6 - 25.0	8.6 - 28.0	320.0 -1400.0
weight of items	0.1 - 3.1	0.1 - 6.0	0.1 - 6.0	1.0 - 319.0
dissimilarities	0.1 - 0.9 (rule 1)	0.1 - 0.9 (rule 1)	0.1 - 0.9 (rule 1)	(rule 2)

Table 1 - Bounds for Data of Test Problems

4.2 Other Heuristics

An exchange-based improvement heuristic and a hybrid genetic heuristic were used to assess the behaviour of the genetic heuristic.

Let us begin by briefly presenting the improvement heuristic IMPROVE, suggested in reference[1] and extensively described in the dissertation[15].

It starts from any feasible solution of the CPCP, in particular from one built through a constructive procedure that clusters the items according to increasing weights, filling the groups ranked by increasing products of the parameters B_k and C_k. These two building criteria - increasing weights of items and increasing products of group capacities - are used to cluster the greatest quantity of items in the less favourable groups first. This constructive procedure is also used to build 3 of the individuals of the initial population of GENET, as mentioned earlier.

As such a classification of items tends to leave the last groups empty, the improvement heuristic runs through the groups, starting by taking the last and continuing in decreasing group index order, displacing to the current group all items with their successors in the current group or in the groups analysed before the current group (that is, the groups with a higher index).

The criterion for choosing items to be displaced at each stage is determined by taking into account the maximum reduction that may be achieved in the total dissimilarity.

The hybrid heuristic, HYBRID, is basically a genetic search, characterized by the fact that, at the end of each generation, it calls the above improvement procedure from the chromosomes (feasible solutions) of the population.

4.3 Results

The measures used to evaluate the heuristics were the solution quality and the computing time in seconds. For each problem Type, average values were calculated, as seen in Table 2.

To assess the quality of the solution, we used the percentual gap between the total dissimilarity of the solution obtained by the genetic-based heuristics (GENET or HYBRID) and the value given by IMPROVE, that is, $100\times(z^{IMP}-z^{genetic})/z^{IMP}$. These values are listed in columns (5) and (8).

For the genetic heuristic GENET the percentage of non-discarded or accepted chromosomes over created chromosomes, in all the 100 generations, is also given. Columns (6) and (7) of Table 2 present these figures for the crossover and for the mutation respectively.

Table 3 presents similar results to those of Table 2 but refers only to the 4 problems with no capacity constraints that were used in our tests.

As for computing times, IMPROVE took up to 2 seconds for each of these 100 instances, whereas GENET on average spent about 14 seconds per instance and HYBRID, on average, took 39 seconds per instance.

problem Types	N	M	nº of prec. constr.	GENET			HYBRID
				diss.gap (%)	acc.cro. (%)	acc.mut. (%)	diss.gap (%)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A	13	2 - 5	12	10	64	32	8
B	13	6 - 8	12	12	86	40	5
C	23	3 - 5	25	2	56	28	3
D	23	6 - 8	25	11	73	48	10
E	32	4 - 6	25	-15	53	29	-9
F	32	7 - 9	25	-9	79	35	-4
G	70	3 - 10	86	-1	63	15	-4
H	70	11 - 30	86	-12	70	27	1

Table 2 - Computational Results - Average Values

problem (Type)	N	M	nº of prec. constr.	GENET diss.gap (%)	GENET acc.cro. (%)	GENET acc.mut. (%)	HYBRID diss.gap (%)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
problem 10 (A)	13	5	12	28	82	49	23
problem 35 (C)	23	3	12	28	47	5	14
problem 60 (F)	32	8	25	21	82	21	26
problem 85 (G)	70	9	86	10	79	27	10

Table 3 - Computational Results for Uncapacitated Problems

4.4 Comments

As Table 2 shows, heuristic GENET performed better, in terms of solution quality, over the shorter instances (problem Types A, B, C and D), whereas the IMPROVE outperformed it in the larger-sized instances (problem Types E, F, G and H). The hybrid process (column (8)) stands between the other two single heuristics - IMPROVE and GENET.

The occurrence of these instances with a poorer behaviour is due to the small number of individuals in the population - fixed and equal to 20 - with respect to the problem size, a direct consequence of the number of items which, for these instances, is N = 32 or 70. Such results are to be expected, as literature suggests higher cardinality populations for larger problems. The enlargement of the dimension of the population has not yet been tried for all the instances of CPCP, as the available computing resources were insufficient. However, in some cases we simulated that situation (enlargement of the population) and found a much better performance of both genetic-based heuristics.

As can be seen in Table 2, columns (6) and (7), the number of discarded chromosomes is very high as compared to the chromosomes included in the population (corresponding to feasible clusterings). We also observed through our experiment - though not noticeable in Tables 2 and 3 - that, in instances with a lower percentage of feasible chromosomes the results were worst.

This leads us to think that we could develop operators to attain feasibility more frequently or, alternatively, procedures to restore feasibility after performing the crossover and mutation operators.

The computing time is higher for the genetic-based heuristics but this measurement of heuristic behaviour is not significant if we assume that the CPCP is a problem whose solution is not called frequently.

It should be stressed that GENET obtained much better solutions than IMPROVE for the problems with no capacity constraints - column (5) of Table 3.

Comparison of the above figures with results found in literature[1,14] is only partially possible, as authors give neither complete problem data nor similar solution quality results.

Moreover, computing times are not comparable, in view of the considerable differences in computing resources.

However, Aronson and Klein[1] do provide some information. Where solution quality is concerned, they present the optimum value (19.7) for a single problem only (which is precisely one of ours belonging to the Type C problems). They also add that their constructive heuristic produced a solution with total dissimilarity equal to approximately seven times the optimum value, and that their improvement scheme (similar to the one described in Subsection 4.2) reduced the value to 25.4. Our results for this problem are: 25.4, 21.9 and 22.3 respectively for the three heuristics IMPROVE, GENET and HYBRID.

According to the authors quoted, the total of 29 problems tested (25 with parameter ranges similar to ours, and 1 with 140 items and 3 groups), the constructive heuristic failed to build up a feasible solution in 4 cases, whereas their specialized branch-and-bound method, embedding the constructive and improvement heuristics, produced the optimal solution in all other cases - though high computing expenses were incurred in some cases. The branch-and-bound is a generalization of Balas's method to embed multi-branches representing the groups, whilst the levels represent the items. Such a method is not suitable for larger problems.

The results of Klein, Beck and Konsynski[14] are even more difficult to compare with ours, as their test problems are CPCPs without capacity constraints, involving 10 to 20 items and 3 groups, and they used a standard ILP code to optimally solve the problems.

Though problems of Table 3 are also uncapacitated CPCPs, they cannot be considered for comparison with the above work[14] as the size and data generation processes are different.

On the strength of the above comparisons and computational results, one may conclude that the genetic-based heuristics we developed for the CPCP behaved well in the set of uncapacitated test instances, and poorly at capacitated medium sized test instances.

5. Conclusions and Future Work

This paper presented a clustering problem, where clusters are built up by considering both maximum capacities and precedence constraints when grouping the items, using a total dissimilarity clustering criterion. This is a typical highly constrained problem and the plain version of the genetic heuristic developed in this case performed fairly well for the set of small and medium sized test instances - at least when compared with results obtained from an improvement scheme, also custom-made for the CPCP.

Computing times are higher for the genetic and the hybrid heuristics, though this is not significant in view of the strategic nature of the applications for the CPCP mentioned earlier.

The computational experiments performed point to the interest in trying different genetic heuristic versions, characterized by imposing the feasibility over solutions created by crossover or mutation, or alternatively by an encoding process that naturally conveys most of the problem constraints through the operation of crossover and mutation (Caldas[3]). This should help to

reduce the non-productive computing time resulting from discarding non-feasible chromosomes.

We also suspect that, for the larger problems, an increase in the number of individuals in the population would produce much better results.

We would like to study lower bounds on the optimal total dissimilarity, so as to evaluate the quality of heuristic solution in absolute terms. We consider this to be one of the next most important steps to take in the study of heuristics for the CPCP. Lower bounds could be obtained through exploitation of the dynamic quality of the CPCP. This may involve the study of state space relaxations for dynamic programming formulations similar to or even different from the dynamic programming model referred to in Section 2.

Finally, we look forward to finding real situations to which the CPCP may be applied and therein experiment methodology developed to date. It is commonly known that huge software companies spend a large amount of their budgets on overtime and extra-fees due to delays. If tasks were better organized, resources would be more profitably used. We believe that these would be successful fields of application for the Clustering with Precedence Constraints Problem.

References

[1] J. E. Aronson and G. Klein, "A Clustering Algorithm for Computer-Assisted Process Organization", *Decision Sciences*, vol. 20 (1989), pp. 730-745.
[2] S.Beaty,*Instruction Scheduling Using Genetic Algorithms*,PHD thesis,Colorado State Univ.,USA (1991).
[3] J. C. Caldas, personal communication (1995).
[4] I. Davis, *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New-York, USA (1991).
[5] B. S. Duran and P. L. Odell, *Cluster Analysis-A Survey*, Springer-Verlag, Berlin, Rep. of Germ. (1974).
[6] E. Faulkenauer, "The Grouping Genetic Algorithms: Widing the Scope of the GAs", *Belgian Journal of Operations Research, Statistics and Computer Science*, vol. 33 (1993) pp. 79-102.
[7] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman and Company, San Francisco, USA (1979).
[8] R. Gnanadesikan and J. R. Kettenring, "Discriminant Analysis and Clustering", *Statistical Science*, vol. 4 (1989) pp. 34-69.
[9] D. E. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*, Addison-Wesley Publishing Company, Reading, USA (1989).
[10] M. Gondran and M. Minoux, *Graphs and Algorithms*, translated by S. Vajda, John Wiley and Sons, Chichester, USA (1986).
[11] J. Holland,*Adaption in Natural and Artificial Systems*, Univ.of Michigan Press, Ann Arbor, USA (1975).
[12] R. J. Jensen, "A Dynamic Programming Algorithm for Cluster Analysis", *Operations Research*, vol.17 (1969) pp. 1034-1057.
[13] J. Karimi, "An Automated Software Design Methodology Using CAPO", *Journal of Management Information Systems*, vol. 3 (1986-87) pp. 71-100.
[14] G. Klein, P. O. Beck and B. Konsynski, "Computer-Aided Process Structuring Via Mixed Integer Programming", *Decision Sciences*, vol. 19 (1988) pp. 750-761.
[15] L. L. Lourenço, *Contributos da Optimização Discreta para a Análise Classificatória. Aplicação de Heurísticas Genéticas a uma Classificação com Precedências*, Master dissertation, Instituto Superior de Economia e Gestão, Universidade Técnica de Lisboa, Portugal (1995).
[16] Z. Michalewics, *Genetic Algorithms + Data Structures = Evolution Programs*, Springer-Verlag, Berlin, Republic of Germany (1992).
[17] C. R. Reeves, *Modern Heuristic Techniques for Combinatorial Problems*, editor, Backwell Scientific Publications, Oxford, England (1993).
[18] S. Sofianopoulou, "The Process Allocation Problem: A Survey of the Application of Graph-Theoretic and Integer Programming Approaches",*Journal of the Operational Research Society*, vol.1 (1992) pp.317-336.
[19] F. M. Tonge, "Assembly Line Balancing Using Probabilistic Combinations of Heuristics", *Management Science*, vol. 11 (1961) pp. 726-735.

INVESTIGAÇÃO OPERACIONAL, vol. 17 (Junho 1997) 87-100

MÉTODO DE SUBSTITUIÇÃO DO VECTOR DOS MULTIPLICADORES BASEADO EM PSEUDO-DERIVADAS

M.T. Monteiro

Edite M.G.P.Fernandes

Departamento de Produção e Sistemas
Universidade do Minho
4700 Braga - Portugal

Abstract

An exact penalty technique based on an augmented Lagrangian function is used for solving equality constrained nonlinear optimization problems. For the Lagrange multiplier vector a substitution philosophy is used. To compute the stationary points of the augmented Lagrangian function, different versions are implemented. The first uses the exact Jacobian matrix of the constraints. The second replaces analytic Jacobian with a finite-difference approximation and the third defines a suitable pseudo-derivative approximation to the Jacobian. Numerical results show that two of them have similar performance. One can prove convergence to a stationary point of the augmented Lagrangian function regardless of the Jacobian approximation. We also prove that the pseudo-derivative problem formulation is equivalent to the penalty technique in sequential quadratic programming.

Resumo

Uma técnica de penalização exacta baseada numa função Lagrangiana aumentada é usada na resolução de problemas de optimização não lineares com restrições de igualdade. É implementada uma filosofia de substituição do vector dos multiplicadores. No cálculo dos pontos estacionários da função Lagrangiana aumentada, são implementadas três versões diferentes. A primeira utiliza o Jacobiano do vector função dos multiplicadores exacto. A segunda aproxima o Jacobiano por diferenças finitas e a terceira utiliza uma aproximação ao Jacobiano designada por pseudo-derivada. Os resultados experimentais mostram que duas das implementações têm desempenhos similares. Pode-se mostrar que a convergência para um ponto estacionário da função Lagrangiana aumentada não depende da aproximação à matriz do Jacobiano. Também se mostra que o problema resultante da introdução das pseudo-derivadas é equivalente à implementação de uma técnica de penalização em programação quadrática sequencial.

Keywords

Penalty technique, multiplier vector substitution, pseudo-derivative.

1. Introdução

1.1 Definição do problema

Considere-se o problema de optimização não linear com restrições do tipo igualdade

$$\min_{x \in \mathbb{R}^n} f(x) \qquad \text{(P1)}$$

$$\text{sujeita a} \quad c(x) = 0,$$

em que f(x) e c(x) são funções não lineares do tipo f: $\mathbb{R}^n \rightarrow \mathbb{R}$ e c: $\mathbb{R}^n \rightarrow \mathbb{R}^m$ com $m \leq n$. Designe-se o vector gradiente de f(x) por $\nabla f(x)$ e a matriz Hessiana por $\nabla^2 f(x)$. A matriz do Jacobiano das restrições é designada por $\nabla c(x)$ e contém os vectores gradientes das funções $c_i(x)$, $i = 1,\ldots, m$, ao longo das linhas. Considerem-se as seguintes hipóteses:

Hipótese 1: As funções f e c_i, $i = 1,\ldots, m$ são continuamente diferenciáveis até à segunda ordem.

Hipótese 2: A matriz do Jacobiano de c(x) tem m linhas linearmente independentes.

Seja $L(x, \lambda)$ a função Lagrangiana associada ao problema P1 definida por

$$L(x, \lambda) = f(x) - c(x)^T\lambda,$$

se o par (x^*, λ^*) verifica

$$\nabla_x L(x^*, \lambda^*) = \nabla f(x^*) - \nabla c(x^*)^T\lambda^* = 0$$

e

$$\nabla_\lambda L(x^*, \lambda^*) = c(x^*) = 0$$

então chama-se ponto Karush-Kuhn-Tucker (KKT) do problema P1; se, além disso,

$$u^T\nabla^2_{xx}L(x^*, \lambda^*)u > 0, \forall u \in \mathbb{R}^n$$

tal que $\nabla c(x^*)u = 0$, então o ponto (x^*, λ^*) é solução do problema P1.

Este artigo está estruturado da seguinte maneira; na subsecção 1.2, é feita uma abordagem sucinta à técnica de penalização exacta para a resolução do problema P1. Na secção 2 introduz-se uma fórmula de substituição para aproximar o vector dos multiplicadores de Lagrange e são definidas as pseudo-derivadas. Na secção 3 mostra-se a forma final das equações do tipo Newton para o cálculo da solução do problema P1 e introduzem-se os métodos para o cálculo da direcção de procura e do parâmetro de penalização. A secção 4 mostra a equivalência entre as equações do tipo Newton baseadas nas pseudo-derivadas e a técnica de penalização em programação quadrática sequencial, a secção 5 apresenta alguns resultados numéricos e na secção 6 são apresentadas as conclusões.

1.2 Técnica de penalização exacta

As técnicas de penalização transformam o problema de optimização não linear com restrições num problema sem restrições, cuja função objectivo se obtém adicionando à função objectivo de P1 termos que penalizam a violação das restrições. Veja-se por exemplo [7]. Se as restrições forem do tipo igualdade (como no problema P1), esse termo de penalização tem a forma

$$\frac{1}{2} r\, c(x)^T A\, c(x)$$

em que r é um escalar positivo designado por parâmetro de penalização e A é uma matriz definida positiva. O termo mais simples, deste tipo de penalização, considera A = I. Se o termo de penalização for adicionado à função Lagrangiana associada ao problema, em vez de adicionado a f(x), obtém-se uma função

$$\Phi(x, \lambda, r) = f(x) - c(x)^T\lambda + \frac{1}{2} r\, c(x)^T c(x) \qquad (1)$$

que se chama função Lagrangiana aumentada ([6], [8], [10] e [11]). Certas funções de penalização definem técnicas de penalização sequencial, uma vez que o mínimo da função de penalização sem restrições converge para a solução do problema P1, x^*, apenas quando $r \to \infty$. A técnica de penalização exacta é caracterizada pelo facto de, para um valor finito de r, $r > \bar{r}$, se ter o mínimo da função de penalização Φ, $x^*(r)$, igual a x^*, [1].

Assim, dado um valor para r, o objectivo consiste em calcular

$$\min_{\substack{x \in \mathbb{R}^n \\ \lambda \in \mathbb{R}^m}} \Phi(x, \lambda, r). \qquad \text{(P2)}$$

2. Método de substituição baseado em pseudo-derivadas

Se for possível conhecer o valor do vector dos multiplicadores na solução, λ^*, então o problema P2 reduz-se a um problema de minimização a n dimensões,

$$\min_{x \in \mathbb{R}^n} \Phi(x, \lambda^*, r)$$

2.1 Fórmula de substituição

Sem conhecer λ^*, é possível estimar este valor usando fórmulas de aproximação ([11]).

Uma das fórmulas de substituição do vector dos multiplicadores pode ser deduzida a partir de uma das condições de primeira ordem do problema P1.

Lema 1: Suponha que as Hipóteses 1 e 2, enunciadas em 1.1, são verdadeiras e que as condições de primeira ordem no ponto (x^*, λ^*) se verificam

$$\nabla_x L(x^*, \lambda^*) = 0 \quad \text{e} \quad \nabla_\lambda L(x^*, \lambda^*) = 0.$$

Então existe uma função $\lambda(x)$: $\mathbb{R}^n \to \mathbb{R}^m$, que determina a relação entre λ e o vector x, definida por

$$\lambda(x) = (\nabla c(x) \nabla c(x)^T)^{-1} \nabla c(x) \nabla f(x) \qquad (2)$$

de tal modo que quando $x = x^*$, $\lambda^* = \lambda(x^*)$.

Prova: Da condição $\nabla_x L(x^*, \lambda^*) = 0$ tira-se que $\nabla c(x^*)^T \lambda^* = \nabla f(x^*)$. A solução deste sistema de n equações com m incógnitas deve ser calculada pelo método dos mínimos quadrados. Assim, $\nabla c(x^*) \nabla c(x^*)^T \lambda^* = \nabla c(x^*) \nabla f(x^*)$ ou seja, $\lambda^* = \lambda(x^*) = (\nabla c(x^*)^+)^T \nabla f(x^*)$ em que $\nabla c(x)^+$ é a inversa generalizada à direita da matriz $\nabla c(x)$, definida por

$$\nabla c(x)^+ = \nabla c(x)^T (\nabla c(x) \nabla c(x)^T)^{-1}. \quad \blacklozenge$$

Se na formulação P2 substituirmos λ por $\lambda(x)$, obtém-se a função

$$\Phi(x, r) = f(x) - c(x)^T \lambda(x) + \frac{1}{2} r\, c(x)^T c(x), \qquad (3)$$

função apenas de $x \in \mathbb{R}^n$ e r, e o problema de minimização sem restrições passa a ser

$$\min_{x \in \mathbb{R}^n} \Phi(x, r). \qquad \text{(P3)}$$

Lema 2: Suponha que as Hipóteses 1 e 2, enunciadas em 1.1 são verdadeiras. No cálculo da solução do problema P3 considere a função de substituição $\lambda(x)$ definida pelo lema 1. Então a função $\lambda(x)$ é bem definida, diferenciável e tem derivada dada por

$$\nabla\lambda(x) = (\nabla c(x)^{+})^{T}\,[\nabla^2 f(x) - \sum_{i=1}^{m} \lambda_i(x)\nabla^2 c_i(x)] + \\ + (\nabla c(x)\nabla c(x)^{T})^{-1}\,(\sum_{i=1}^{m} \nabla^2 c_i(x)v(x)e_i^{T})^{T} \qquad (4)$$

em que $v(x) = \nabla f(x) - \nabla c(x)^{T}\lambda(x)$, e_i é a coluna i da matriz Identidade de ordem m, $\nabla^2 f(x)$ é a matriz Hessiana de f(x) e $\nabla^2 c_i(x)$ é a Hessiana, de ordem n, da função de restrição $c_i(x)$, i = 1,..., m.

Prova: De acordo com as Hipótese 1 e 2 e a equação (4), a matriz $\nabla c(x)\nabla c(x)^{T}$ é não singular, $\nabla c(x)^{+}$ existe e $\nabla\lambda(x)$ é bem definido e diferenciável.

Por sua vez, a equação $\nabla f(x) - \nabla c(x)^{T}\lambda(x) = 0$ é equivalente à seguinte conjunção

$$\begin{cases} v(x) = \nabla f(x) - \nabla c(x)^{T}\lambda(x) \\ \nabla c(x)v(x) = 0 \end{cases}$$

para algum vector $v(x) \in \mathbb{R}^n$, desde que a característica de $\nabla c(x)$ seja m. Derivando estas equações em ordem a x obtém-se

$$\begin{cases} \nabla v(x) = \nabla^2 f(x) - \sum_{i=1}^{m} \lambda_i(x)\nabla^2 c_i(x) - \nabla c(x)^{T}\nabla\lambda(x) \\ (\sum_{i=1}^{m} \nabla^2 c_i(x)v(x)e_i^{T})^{T} + \nabla c(x)\nabla v(x) = 0. \end{cases}$$

Multiplicando a primeira expressão à esquerda por $\nabla c(x)$ e substituindo na segunda chega-se a

$$\begin{cases} \nabla c(x)\nabla v(x) = \nabla c(x)\nabla^2 f(x) - \nabla c(x)\sum_{i=1}^{m} \lambda_i(x)\nabla^2 c_i(x) - \\ \qquad - \nabla c(x)\nabla c(x)^{T}\nabla\lambda(x) \\ (\sum_{i=1}^{m} \nabla^2 c_i(x)v(x)e_i^{T})^{T} + \nabla c(x)\nabla^2 f(x) - \nabla c(x)\sum_{i=1}^{m} \lambda_i(x)\nabla^2 c_i(x) - \\ \qquad - \nabla c(x)\nabla c(x)^{T}\nabla\lambda(x) = 0. \end{cases}$$

Explicitando em função do último termo da segunda equação vem

$$\nabla c(x)\nabla c(x)^{T}\nabla\lambda(x) = \nabla c(x)\nabla^2 f(x) - \nabla c(x)\sum_{i=1}^{m} \lambda_i(x)\nabla^2 c_i(x) + \\ + (\sum_{i=1}^{m} \nabla^2 c_i(x)v(x)e_i^{T})^{T}$$

ou, como $\nabla c(x)\nabla c(x)^{T}$ é uma matriz não singular, tem-se

$$\nabla\lambda(x) = (\nabla c(x)\nabla c(x)^{T})^{-1}\,\nabla c(x)\nabla^2 f(x) - \\ - (\nabla c(x)\nabla c(x)^{T})^{-1}\,\nabla c(x)\sum_{i=1}^{m} \lambda_i(x)\nabla^2 c_i(x) +$$

$$+ (\nabla c(x)\nabla c(x)^T)^{-1} (\sum_{i=1}^{m} \nabla^2 c_i(x) v(x) e_i^T)^T. \quad \blacklozenge$$

Proposição: Suponha que as Hipóteses 1 e 2, enunciadas em 1.1, são verdadeiras e que a função objectivo do problema P3 é a apresentada em (3). Então $\Phi(x, r)$ é diferenciável e tem primeira derivada, em ordem a x, dada por

$$\nabla_x \Phi(x, r) = \nabla f(x) - \nabla c(x)^T \lambda(x) - \nabla \lambda(x)^T c(x) + r \nabla c(x)^T c(x). \qquad (5)$$

A matriz Hessiana das segundas derivadas é definida por

$$\begin{aligned} \nabla^2_{xx} \Phi(x, r) = {} & \nabla^2 f(x) - \sum_{i=1}^{m} \lambda_i(x) \nabla^2 c_i(x) - \nabla c(x)^T \nabla \lambda(x) - \\ & - [\nabla \lambda(x)^T - r \nabla c(x)^T] \nabla c(x) - \\ & - \sum_{i=1}^{m} c_i(x) \nabla^2 \lambda_i(x) + r \sum_{i=1}^{m} c_i(x) \nabla^2 c_i(x) \end{aligned} \qquad (6)$$

em que $\lambda(x)$ e $\nabla\lambda(x)$ foram definidos respectivamente nos lemas 1 e 2, $\nabla^2 f(x)$ é a matriz Hessiana de f(x), $\nabla^2 c_i(x)$ é a Hessiana da restrição $c_i(x)$, i = 1,..., m e $\nabla^2 \lambda_i(x)$ é também a Hessiana do elemento $\lambda_i(x)$ do vector função $\lambda(x)$.

A solução do problema P3 é obtida calculando um ponto estacionário de Φ, $\nabla_x \Phi(x, r) = 0$. O método de Newton para a resolução de $\nabla_x \Phi(x, r) = 0$, origina o seguinte processo iterativo

$$x^{k+1} = x^k + d^k, \; k = 0, 1, 2, \ldots$$

em que a direcção d^k é calculada por

$$\nabla^2_{xx} \Phi(x^k, r)\, d^k = -\nabla_x \Phi(x^k, r).$$

O vector $\nabla_x \Phi(x, r)$ e a matriz $\nabla^2_{xx} \Phi(x, r)$ são os definidos na proposição.

2.2 Pseudo-derivadas

A hipótese seguinte vai ser necessária para a definição das pseudo-derivadas.

Hipótese 3: O vector c(x) das funções de restrição é aproximado por um modelo linear

$$l(x) = l(\bar{x} + d) = c(\bar{x}) + \nabla c(\bar{x}) d$$

numa vizinhança d de $\bar{x}$.

Definição 1: Suponha que as Hipóteses 1, 2 (enunciadas em 1.1.) e 3 são verdadeiras. Considere a função substituição definida em (2), então a pseudo-derivada da função $\lambda(x)$ é definida por

$$D\lambda(x) = (\nabla c(x)^+)^T \nabla^2 f(x). \qquad (7)$$

Com a aproximação l(x), a expressão para $\nabla\lambda(x)$, (4), fica com a forma

$$\begin{aligned} \nabla \lambda(x) = {} & (\nabla l(x)\, \nabla l(x)^T)^{-1} \nabla l(x) \, [\nabla^2 f(x) - \sum_{i=1}^{m} \lambda_i(x) \nabla^2 l_i(x)] + \\ & + (\nabla l(x)\, \nabla l(x)^T)^{-1} (\sum_{i=1}^{m} \nabla^2 l_i(x) v(x) e_i^T)^T. \end{aligned}$$

Para qualquer x da vizinhança de $\bar{x}$, o Jacobiano da aproximação $l(\bar{x} + d)$ é

$$\nabla l(x) = \nabla c(\bar{x})$$

uma vez que $d = x - \bar{x}$. Também $\nabla^2 l_i(x) = 0$, $i = 1,\ldots, m$ e

$$D\lambda(x) = (\nabla c(x)\, \nabla c(x)^T)^{-1}\, \nabla c(x) \nabla^2 f(x).$$

Definição 2: Suponha que as Hipóteses 1, 2 (enunciadas em 1.1) e 3 são verdadeiras, que o vector função de substituição dos multiplicadores é dado por (2) e que a pseudo-derivada de $\lambda(x)$ é dada por (7).

Então o vector das primeiras pseudo-derivadas de $\Phi(x, r)$ é definido por

$$D\Phi(x, r) = \nabla f(x) - \nabla c(x)^T (\nabla c(x)\nabla c(x)^T)^{-1} \nabla c(x)\nabla f(x) - \\ - \left((\nabla c(x)\nabla c(x)^T)^{-1} \nabla c(x)\nabla^2 f(x)\right)^T c(x) + r\, \nabla c(x)^T c(x) \qquad (8)$$

e se, além disso, x^* é um ponto KKT, a matriz das segundas pseudo-derivadas é definida por

$$D^2\Phi(x, r) = \nabla^2 f(x) - \nabla c(x)^T \left((\nabla c(x)\nabla c(x)^T)^{-1} \nabla c(x)\nabla^2 f(x)\right) - \\ - \left(\left((\nabla c(x)\nabla c(x)^T)^{-1} \nabla c(x)\nabla^2 f(x)\right)^T - r\, \nabla c(x)^T\right) \nabla c(x). \qquad (9)$$

A fórmula (8) obtém-se usando $D\lambda(x)$, (7), na fórmula (5), no lugar de $\nabla\lambda(x)$.

Da expressão (6) para $\nabla^2_{xx}\Phi(x, r)$, substituindo $\nabla\lambda(x)$ por $D\lambda(x)$, usando a Hipótese 3 e desprezando o penúltimo termo da definição de $\nabla^2_{xx}\Phi(x, r)$, pois tende para zero quando $x \to x^*$, obtém-se a expressão (9) para $D^2\Phi(x, r)$.

3. Forma final das equações Newton

3.1 Raízes de $D\Phi = 0$

Teorema: Suponha que as Hipóteses 1, 2 (enunciadas em 1.1) e 3 são verdadeiras. Se o par (x^*, λ^*) for um ponto KKT do problema P1, se $\lambda^* = \lambda(x^*) = (\nabla c(x^*)^+)^T \nabla f(x^*)$, então x^* é ponto estacionário de $\Phi(x, r)$, função objectivo do problema P3 e raiz de $D\Phi(x, r) = 0$. Por sua vez, qualquer raiz de $D\Phi(x, r) = 0$ que verifique as restrições, $c(x) = 0$, é ponto estacionário de $\Phi(x, r)$.

Prova: Da equação (5) tem-se, no ponto (x^*, λ^*),

$$\nabla\Phi(x^*, r) = \nabla f(x^*) - \nabla c(x^*)^T\lambda(x^*) - \nabla\lambda(x^*)^T c(x^*) + r\nabla c(x^*)^T c(x^*) \\ = \nabla f(x^*) - \nabla c(x^*)^T\lambda(x^*) = \nabla f(x^*) - \nabla c(x^*)^T\lambda^* = 0$$

de acordo com as condições de primeira ordem do problema P1 e x^* é ponto estacionário de $\Phi(x, r)$.

Como $D\Phi(x, r)$ calculada no ponto (x^*, λ^*) origina

$$D\Phi(x^*, r) = \nabla f(x^*) - \nabla c(x^*)^T\lambda(x^*) - D\lambda(x^*)^T c(x^*) + r\nabla c(x^*)^T c(x^*) \\ = \nabla f(x^*) - \nabla c(x^*)^T\lambda(x^*) = 0$$

então x^* é também raiz de $D\Phi(x, r) = 0$.

Por sua vez, se x^* é uma raiz de $D\Phi(x, r) = 0$, então

$$D\Phi(x^*, r) = \nabla f(x^*) - \nabla c(x^*)^T\lambda(x^*) - D\lambda(x^*)^T c(x^*) + r\, \nabla c(x^*)^T c(x^*) = 0.$$

Como $\nabla\Phi(x, r) = D\Phi(x, r) + D\lambda(x)^T c(x) - \nabla\lambda(x)^T c(x)$ tem-se para qualquer raiz x^* de $D\Phi(x, r) = 0$, que verifique $c(x^*) = 0$, $\nabla\Phi(x^*, r) = 0$. ♦

No seguimento deste teorema, a solução do problema P3 passa a ser obtida pela resolução de $D\Phi(x, r) = 0$. A forma final das equações pseudo-Newton é agora

$$D^2\Phi(x^k, r)\, d_N^k = -D\Phi(x^k, r) \qquad k = 0, 1, 2,... \qquad (10)$$

com $x^{k+1} = x^k + d_N^k$.

3.2 Utilização das diferenças finitas

A aproximação à matriz Jacobiano $\nabla\lambda(x)$ baseada em diferenças finitas designada por $DF\lambda(x)$ é representada por

$$(DF\lambda(x))_j = \left[\frac{\lambda(x+h_j e_j) - \lambda(x)}{h_j}\right]_j, \qquad j = 1,..., n \qquad (11)$$

em que $(DF\lambda(x))_j$ define a coluna j da matriz aproximação, $h_j = \sqrt{\eta}x_j$ é o espaçamento entre o ponto x e o novo ponto $x + h_j e_j$ ao longo do eixo e_j, coluna j da matriz Identidade de ordem n, x_j é o elemento j do vector x e η é uma constante positiva próxima de zero que define a precisão da máquina.

A utilização da aproximação $DF\lambda(x)$ a $\nabla\lambda(x)$, baseada em diferenças finitas, origina uma simplificação razoável das formas $D\Phi(x, r)$ e $D^2\Phi(x, r)$, passando a ser designadas respectivamente por

$$DF\Phi(x, r) = \nabla f(x) - \nabla c(x)^T\lambda(x) - DF\lambda(x)^T c(x) + r\nabla c(x)^T c(x)$$

e

$$D^2F\Phi(x, r) = \nabla^2 f(x) - \nabla c(x)^T DF\lambda(x) - [DF\lambda(x)^T - r\nabla c(x)^T]\nabla c(x)$$

com as colunas da matriz $DF\lambda(x)$ definidas por (11).

3.3 Cálculo da direcção de procura

A solução adoptada para o cálculo da direcção de procura baseia-se no artigo de Fernandes [4] e consiste em definir uma nova direcção, como combinação linear de duas direcções, a $-D\Phi(x, r)$ e a direcção definida em (10). A expressão desta nova direcção é

$$d = -W(x)D\Phi(x, r) + (1 - W(x))d_N \qquad (12)$$

em que d_N é a direcção (10). Para que esta nova direcção possa ser de descida, a função W(x) é definida por

$$W(x) > \frac{D\Phi(x,r)^T d_N}{\|D\Phi(x,r)\|_2 + D\Phi(x,r)^T d_N}.$$

Se a matriz $D^2\Phi(x, r)$ do sistema pseudo-Newton for singular a direcção escolhida é $-D\Phi(x, r)$, se a matriz $D^2\Phi(x, r)$ é definida positiva então a direcção é a resultante do sistema (10), ou seja, d_N. Para os restantes casos, usa-se a direcção definida por (12).

A globalização do método de Newton foi feita recorrendo a uma técnica de procura unidimensional na definição do passo α, em que, $x^{k+1} = x^k + \alpha\, d^k$, $k = 0, 1, 2,...$

O critério para escolha do escalar α é o de Curry-Altman-Armijo que pode ser consultado em [5].

3.4 Parâmetro de penalização

Nas técnicas de penalização exacta a escolha do valor para o parâmetro r é crucial. Se o valor atribuído não for adequado é possível ajustá-lo ao longo do processo iterativo. O esquema de ajuste do valor de r, usado neste trabalho, é o que a seguir se descreve.

Durante o processo iterativo, se a aproximação à solução estiver muito afastada da região admissível, o termo de penalização deve dominar e r = 10; por sua vez, numa vizinhança δ da região admissível, a penalização da função Lagrangiana não é tão necessária e r toma o valor unitário. Nas outras situações, r deve variar na proporção inversa da proximidade do ponto em relação à região admissível.

Detalhes deste esquema, bem como alguns estudos comparativos com outros esquemas podem ser vistos em [3] e [4].

4. Penalização em programação quadrática sequencial

O modelo quadrático de aproximação a f(x) é

$$q(d) = f(x) + \nabla f(x)^T d + \frac{1}{2}\, d^T \nabla^2 f(x)\, d \qquad (13)$$

e a técnica de programação quadrática sequencial calcula a solução do problema P1, resolvendo uma sequência de problemas de programação quadrática com a forma

$$\begin{aligned} &\min_{d \in \mathbb{R}^n} \; q(d) \\ \text{sujeita a} \quad &l(d) = c(x) + \nabla c(x)\, d = 0 \end{aligned} \qquad (P4)$$

em que l(d) é a aproximação linear a c(x).

A resolução deste problema quadrático pode ser baseada nas condições de primeira ordem que envolvem o vector gradiente da função Lagrangiana associada a P4,

$$L_q(d, \lambda) = f(x) + \nabla f(x)^T d + \frac{1}{2}\, d^T \nabla^2 f(x)\, d - (c(x) + \nabla c(x)\, d)^T \lambda.$$

O par (d, λ) que verifica as equações

$$\begin{cases} \nabla_d L_q(d, \lambda) = 0 \Leftrightarrow \nabla^2 f(x) d - \nabla c(x)^T \lambda = -\nabla f(x) \\ \nabla_\lambda L_q(d, \lambda) = 0 \Leftrightarrow \nabla c(x) d = -c(x) \end{cases} \qquad (14)$$

é solução de P4 e define uma aproximação (x + d, λ) à solução (x^*, λ^*) do problema P1. O sistema (14) é de dimensão n + m.

A resolução do problema P4 pode ser feita recorrendo-se às técnicas de penalização exacta. Assim, a função Lagrangiana aumentada, que é quadrática na variável d, apresenta a seguinte forma

$$\begin{aligned} \Phi_q(d, \lambda, r) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x) d - (c(x) + \nabla c(x) d)^T \lambda + \\ + \frac{1}{2} r\, (c(x) + \nabla c(x) d)^T (c(x) + \nabla c(x) d). \end{aligned} \qquad (15)$$

Da condição de primeira ordem do problema P4, $\nabla_d L_q(d, \lambda) = 0$, resolvendo em ordem a λ, define-se uma função $\lambda(d)$: $\mathbb{R}^n \rightarrow \mathbb{R}^m$. Assim, de

$$\nabla f(x) + \nabla^2 f(x) d - \nabla c(x)^T \lambda = 0$$

tira-se que a solução dos mínimos quadrados é

$$\lambda(d) = (\nabla c(x)\nabla c(x)^T)^{-1}\nabla c(x)\,[\nabla f(x) + \nabla^2 f(x)d] = (\nabla c(x)^+)^T \nabla q(d) \qquad (16)$$

supondo a Hipótese 2 verdadeira. $\nabla q(d)$ é o gradiente do modelo quadrático q(d) em (13).

Substituindo $\lambda(d)$ na expressão da função Lagrangiana aumentada (15), resulta a função das duas variáveis independentes d e r:

$$\begin{aligned}\Phi_q(d, r) = f(x) - \nabla f(x)^T d + \tfrac{1}{2} d^T \nabla^2 f(x) d - \\ - [c(x) + \nabla c(x)d]^T (\nabla c(x)^+)^T \nabla q(d) + \\ + \tfrac{1}{2} r\,[c(x) + \nabla c(x)d]^T [c(x) + \nabla c(x)d].\end{aligned} \qquad (17)$$

Se $\nabla q(d)$ for substituído por $\nabla f(x) + \nabla^2 f(x)d$ obtém-se a função Lagrangiana aumentada

$$\begin{aligned}\Phi_q(d, r) = f(x) - c(x)^T(\nabla c(x)^+)^T \nabla f(x) + \tfrac{1}{2} r c(x)^T c(x) + \\ + \{\nabla f(x)^T - c(x)^T(\nabla c(x)^+)^T \nabla^2 f(x) + \tfrac{1}{2} r c(x)^T \nabla c(x)\}\, d + \\ + d^T \{-\nabla c(x)^T(\nabla c(x)^+)^T \nabla f(x) + \tfrac{1}{2} r \nabla c(x)^T c(x)\} + \\ + d^T \{\tfrac{1}{2}\nabla^2 f(x) - \nabla c(x)^T(\nabla c(x)^+)\nabla^2 f(x) + \tfrac{1}{2} r \nabla c(x)^T \nabla c(x)\}\, d\end{aligned}$$

quadrática em d.

O problema P4 é agora reformulado como

$$\min_{d \in \mathbb{R}^n} \Phi_q(d, r) \qquad (P4')$$

em que $\Phi_q(d, r)$ é uma função apenas de d e r.

Usando a condição de primeira ordem, $\nabla_d \Phi_q(d, r) = 0$, no cálculo do mínimo de P4', tem-se

$$\begin{aligned}\nabla f(x) - \nabla^2 f(x)\nabla c(x)^+ c(x) + \tfrac{1}{2} r \nabla c(x)^T c(x) - \\ - \nabla c(x)^T(\nabla c(x)^+)^T \nabla f(x) + \tfrac{1}{2} r \nabla c(x)^T c(x) + \\ + \{\nabla^2 f(x) - 2\nabla c(x)^T(\nabla c(x)^+)^T \nabla^2 f(x) + r\nabla c(x)^T \nabla c(x)\}\, d = 0.\end{aligned}$$

Esta expressão pode ser rearranjada tomando a forma

$$\begin{aligned}\left\{ [I - 2\nabla c(x)^T(\nabla c(x)^+)^T]\nabla^2 f(x) + r\nabla c(x)^T \nabla c(x) \right\} d = \\ = [\nabla c(x)^T(\nabla c(x)^+)^T - I]\nabla f(x) + \nabla^2 f(x)\nabla c(x)^+ c(x) - r\nabla c(x)^T c(x)\end{aligned}$$

de um sistema de n equações lineares para o cálculo do vector d.

O vector do segundo membro deste sistema é $-D\Phi(x, r)$ (veja-se em (8)). Como $\nabla^2 f(x)$ e $\nabla c(x)^T(\nabla c(x)^+)^T$ são simétricas, a matriz deste sistema coincide com $D^2\Phi(x, r)$ (em (9)) se aquelas matrizes comutarem. Assim, nas condições enunciadas as equações pseudo-Newton (veja-se em (10)) definem um processo iterativo equivalente à técnica de penalização em programação quadrática sequencial.

5. Experiências computacionais

As três versões para o cálculo do Jacobiano de $\lambda(x)$, a versão analítica, $\nabla\lambda(x)$, a baseada na pseudo-derivada, $D\lambda(x)$ e a baseada nas diferenças finitas, $DF\lambda(x)$, foram testadas com um conjunto de 25 problemas teste que se encontram definidos no Apêndice. A aproximação inicial do processo iterativo bem como a classificação de cada problema, estão de acordo com as indicações de [2] e [9]. Os algoritmos foram implementados em Fortran77 num computador pessoal em aritmética de precisão dupla, usando o critério baseado nas condições

$$|\alpha|\,\|d^k\|_2 \le \varepsilon\,\|x^k\|_2,\ |\Phi(x^{k+1}, r) - \Phi(x^k, r)| \le \varepsilon^2\,|\Phi(x^k, r)|$$

$$\text{e}\quad \|D\Phi(x^{k+1}, r)\|_2 \le \sqrt[3]{\varepsilon},$$

para a paragem do processo iterativo.

$P_{n/m}$	$D\lambda(x)$ Nit/Nf	$\|\nabla L\|_2$	$DF\lambda(x)$ Nit/Nf	$\|\nabla L\|_2$	$\nabla\lambda(x)$ Nit/Nf	$\|\nabla L\|_2$
$1_{3/1}$	2/5	0.0	a)		2/5	0.0
$2_{5/2}$	2/5	0.0	a)		2/5	0.0
$3_{5/3}$	2/5	0.0	a)		2/5	0.0
$4_{5/3}$	>100		a)		>100	
$5_{5/2}$	28/57	D-14	a)		28/57	D-14
$6_{5/3}$	9/18	0.0	a)		9/18	0.0
$7_{2/1}$	103/207	D-10	>100		>100	
$8_{4/2}$	13/27	D-06	>100		13/27	D-06
$9_{3/2}$	7/15	D-06	8/17	D-06	7/15	D-06
$10_{3/1}$	34/69	D-06	>100		34/69	D-06
$11_{3/1}$	10/21	D-16	>100		10/21	D-10
$12_{4/3}$	7/15	D-06	17/35	D-06	7/15	D-06
$13_{5/3}$	24/49	D-14	34/179	D-13	24/49	D-13
$14_{5/3}$	12/25	D-06	16/33	D-06	12/25	D-06
$15_{5/3}$	15/31	D-07	87/251	D-06	15/31	D-07
$16_{5/2}$	39/79	D-15	31/63	D-15	39/79	D-15
$17_{7/4}$	>100		>100		>100	
$18_{5/2}$	11/23	D-06	11/23	D-07	11/23	D-06
$19_{2/1}$	18/37	D-06	a)		20/41	D-07
$20_{5/3}$	2/5	D-15	a)		2/5	D-15
$21_{3/1}$	6/13	D-07	6/13	D-07	6/13	D-07
$22_{5/3}$	7/15	D-06	>100		7/15	D-06
$23_{5/2}$	10/21	D-06	10/21	D-06	10/21	D-06
$24_{5/3}$	15/31	D-07	23/47	D-07	15/31	D-06
$25_{2/1}$	21/43	D-07	24/49	D-07	21/43	D-07

Tabela - Resultados numéricos das três versões

Porque nem sempre foi possível atingir a precisão indicada, o processo iterativo também foi terminado quando o número de iterações excedeu o valor 100. Na tabela apresentam-se os resultados relativos às três implementações, usando para cálculo do passo α o critério de Curry-Altman-Armijo. Nit designa o número de iterações necessárias até convergir, Nf designa o número de cálculos de valores da função $\Phi(x, r)$ e a), na versão $DF\lambda(x)$, refere-se aos problemas com restrições lineares em que $\nabla c(x)$ é uma matriz constante e os resultados coincidem com os obtidos com $D\lambda(x)$. $\|\nabla L\|_2$ designa a norma 2 do gradiente da função Lagrangiana. Para as constantes ε do critério de paragem e δ de definição de vizinhança da região admissível, no cálculo de um valor aceitável para r, foram usados respectivamente 10^{-6} e 10^{-2}.

Conclusões

Da análise dos resultados pode-se concluir que a versão do método de substituição baseada em pseudo-derivadas assegura resultados tão bons quanto a versão analítica e não recorre ao cálculo das matrizes Hessianas das restrições e do vector substituição dos multiplicadores. Apenas em dois exemplos, foram verificados comportamentos ligeiramente distintos ($P7_{4/2}$ e $P19_{5/3}$). A versão baseada em diferenças finitas levou, em geral, mais iterações do que a versão pseudo e em 35% dos 17 problemas testados o processo não parou. Nos casos em que se assinala > 100 (na tabela), verificou-se convergência mas a implementação não consegue atingir a exigente precisão pretendida ($\varepsilon = 10^{-6}$).

As equações Newton baseadas nas pseudo-derivadas da função Lagrangiana aumentada são obtidas a partir de uma aproximação linear às funções de restrição, têm em consideração o facto de $c(x^*) \rightarrow 0$, quando $x \rightarrow x^*$ e originam implementações eficientes, económicas e razoavelmente simples.

Referências

[1] Bazaraa, Mokhtar S. e Shetty, C.M., *Nonlinear Programming*, John Wiley & Sons (1979).

[2] Celis, M.R., Dennis, J.E. e Tapia, R.A., *A Trust-Region Strategy for Nonlinear Equality Constrained Optimization*, em P.T. Boggs, R.H. Byrd e R.B. Schnabel (Eds.), Numerical Optimization 1984, SIAM Philadelphia (1985) 71-82.

[3] Fernandes, E.M.G.P., *An Exact Penalty Approach with Newton Based Descent Directions for Nonlinear Optimization*, (Relatório interno) 5th Stockholm Optimization Days Conference, Stockholm (1994).

[4] Fernandes, E.M.G.P., *Newton Based Exact Penalty Techniques for Nonlinear Optimization with Constraints*, Operations Research Proceedings 1994, Springer-Verlag, Berlin (1995) 39-44.

[5] Fernandes, E.M.G.P., *A Curry-Altman-Armijo Line Search Algorithm Which Uses Newton Based Descent Directions*, Proceedings of Applied Mathematical Programming and Modelling (1995) 27-30.

[6] Fletcher, R., *A Class of Methods for Nonlinear Programming with Termination and Convergence Properties*, Integer and Nonlinear Programming (1970) 157-175.

[7] Gill, P.E., Murray, W. e Wright, M.H., *Practical Optimization*, Academic Press (1981).

[8] Glad, T. e Polak, E., *A Multiplier Method with Automatic Limitation of Penalty Growth*, Mathematical Programming 17 (1979) 140-155.

[9] Hock, W. e Schittkowski, K., *Test Examples for Nonlinear Programming Codes*, Springer-Verlag (1981).

[10] Powell, M.J.D. e Yuan, Y., *A Recursive Quadratic Programming Algorithm that Uses Differentiable Exact Penalty Functions*, Mathematical Programming 35 (1986) 265-278.

[11] Tapia, R.A., *Diagonalized Multiplier Methods and Quasi-Newton Methods for Constrained Optimization*, Journal of Optimization Theory and Applications 22 (1977) 135-194.

Apêndice - Problemas Teste

1, QLRT, 3/1, $x^0 = (-4,1,1)^T$

$f(x) = (x_1 + x_2)^2 + (x_2 + x_3)^2$

$c_1(x) = x_1 + 2x_2 + 3x_3 - 1$

2, QLRT, 5/2, $x^0 = (3,5,-3,2,-2)^T$

$f(x) = (x_1 - 1)^2 + (x_2 - x_3)^2 + (x_4 - x_5)^2$

$c_1(x) = x_1 + 2x_2 + 3x_3 - 1$

$c_2(x) = x_3 - 2(x_4 + x_5) + 3$

3, QLRT, 5/3, $x^0 = (2.5,0.5,2,-1,0.5)^T$

$f(x) = (x_1 - x_2)^2 + (x_2 + x_3 - 2)^2 + (x_4 - 1)^2 + (x_5 - 1)^2$

$c_1(x) = x_1 + 3x_2 - 4$

$c_2(x) = x_3 + x_4 - 2x_5$

$c_3(x) = x_2 - x_5$

4, QLRT, 5/3, $x^0 = (2,2,2,2,2)^T$

$f(x) = (4x_1 - x_2)^2 + (x_2 + x_3 - 2)^2 + (x_4 - 1)^2 + (x_5 - 1)^2$

$c_1(x) = x_1 + 3x_2$

$c_2(x) = x_3 + x_4 - 2x_5$

$c_3(x) = x_2 - x_5$

5, PLRT, 5/2, $x^0 = (10,7,2,-3,0.8)^T$

$f(x) = (x_1 - x_2)^2 + (x_3 - 1)^2 + (x_4 - 1)^4 + (x_5 - 1)^6$

$c_1(x) = x_1 + x_2 + x_3 + 4x_4 - 7$

$c_2(x) = x_3 + 5x_5 - 6$

6, PLRT, 5/3, $x^0 = (35,-31,11,5,-5)^T$

$f(x) = (x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_3 - x_4)^4 + (x_4 - x_5)^2$

$c_1(x) = x_1 + 2x_2 + 3x_3 - 6$

$c_2(x) = x_2 + 2x_3 + 3x_4 - 6$

$c_3(x) = x_3 + 2x_4 + 3x_5 - 6$

7, QQRT, 2/1, $x^0 = (-1.2,1)^T$

$f(x) = (1 - x_1)^2$

$c_1(x) = 10(x_2 - x_1^2)$

8, QQRT, 4/2, $x^0 = (1,1,1,1)^T$

$f(x) = (x_1 - 2)^2 + (x_2 - 2)^2 + (x_3 - 3)^2 + (x_4 - 4)^2$

$c_1(x) = x_1 - 2$

$c_2(x) = x_3^2 + x_4^2 - 2$

9, QQRT, 3/2, $x^0 = (4,-3,4)^T$

$f(x) = 4x_1^2 + 2x_2^2 + 2x_3^2 - 33x_1 + 16x_2 - 24x_3$

$c_1(x) = 3x_1 - 2x_2^2 - 7$

$c_2(x) = 4x_1 + x_3^2 - 11$

10, PQRT, 3/1, $x^0 = (2,2,2)^T$

$f(x) = 0.01(x_1 - 1)^2 + (x_2 - x_1^2)^2$

$c_1(x) = x_1 + x_3^2 + 1$

11, PPRT, 3/1, $x^0 = (-2.6,2,2)^T$

$f(x) = (x_1 - x_2)^2 + (x_2 - x_3)^4$

$c_1(x) = (1 + x_2^2)\, x_1 + x_3^4 - 3$

12, PPRT, 4/3, $x^0 = (0.8,0.8,0.8,0.8)^T$

$f(x) = -x_1x_2x_3x_4$

$c_1(x) = x_1^3 + x_2^2 - 1$

$c_2(x) = x_1^2x_4 - x_3$

$c_3(x) = x_4^2 - x_2$

13, PPRT, 5/3, $x^0 = (2,1.41421,-1,0.58578,0.5)^T$

$f(x) = (x_1 - x_2)^2 + (x_2 - x_3)^3 + (x_3 - x_4)^4 + (x_4 - x_5)^4$

$c_1(x) = x_1 + x_2^2 + x_3^3 - 3$

$c_2(x) = x_2 - x_3^2 + x_4 - 1$

$c_3(x) = x_1x_5 - 5$

14, PPRP, 5/3, $x^0 = (-2,2,2,2,2)^T$

$f(x) = x_1x_2x_3x_4x_5$

$c_1(x) = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 - 10$

$c_2(x) = x_2x_3 - 5x_4x_5$

$c_3(x) = x_1^3 + x_2^3 + 1$

15, PPRP, 5/3, $x^0 = (2,2,2,2,2)^T$

$f(x) = (x_1 - 1)^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_3 - x_4)^4 + (x_4 - x_5)^4$

$c_1(x) = x_1 + x_2^2 + x_3^3 - 2 - 3\sqrt{2}$

$c_2(x) = x_2 - x_3^2 + x_4 + 2 - 2\sqrt{2}$

$c_3(x) = x_1x_5 - 2$

16, PGRT, 5/2, $x^0 = (0.707107,1.75,0.5,2,2)^T$

$f(x) = (x_1 - x_2)^2 + (x_3 - 1)^2 + (x_4 - 1)^4 + (x_5 - 1)^6$

$c_1(x) = x_1^2x_4 + \sin(x_4 - x_5) - 1$

$c_2(x) = x_2 + x_3^4x_4^2 - 2$

17, PGRT, 7/4, $x^0 = (1,1,1,0.509,0.509,0.509,0.98)^T$

$f(x) = -x_1x_2x_3$

$c_1(x) = x_1 - 4.2 \sin^2(x_4)$

$c_2(x) = x_2 - 4.2 \sin^2(x_5)$

$c_3(x) = x_2 - 4.2 \sin^2(x_6)$

$c_4(x) = x_1 + 2x_2 + 2x_3 - 7.2 \sin^2(x_7)$

18, PGRP, 5/2, $x^0 = (2,2,2,2,2)^T$

$f(x) = (x_1 - 1)^2 + (x_1 - x_2)^2 + (x_3 - 1)^2 + (x_4 - 1)^4 + (x_5 - 1)^6$

$c_1(x) = x_1^2 x_4 + \sin(x_4 - x_5) - 2\sqrt{2}$

$c_2(x) = x_2 + x_3^4 x_4^2 - 8 - \sqrt{2}$

19, GLRT, 2/1, $x^0 = (-1,-1)^T$

$f(x) = \sin(\pi \frac{x_1}{12}) \cos(\pi \frac{x_2}{16})$

$c_1(x) = 4x_1 - 3x_2$

20, QLRT, 5/3, $x^0 = (2,2,2,2,2)^T$

$f(x) = (x_1 - x_2)^2 + (x_2 + x_3 - 2)^2 + (x_4 - 1)^2 + (x_5 - 1)^2$

$c_1(x) = x_1 + 3x_2$

$c_2(x) = x_3 + x_4 - 2x_5$

$c_2(x) = x_2 - x_5$

21, PPRT, 3/1, $x^0 = (1.5,1.5,1.5)^T$

$f(x) = (x_1 - 1)^2 + (x_1 - x_2)^2 + (x_2 - x_3)^4$

$c_1(x) = x_1(1 + x_2^2) + x_3^4 - 4 - 3\sqrt{2}$

22, PPRT, 5/3, $x^0 = (2,2,2,2,2)^T$

$f(x) = (x_1 - 1)^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_4 - 1)^4 + (x_5 - 1)^4$

$c_1(x) = x_1 + x_2^2 + x_3^3 - 2 - 3\sqrt{2}$

$c_2(x) = x_2 + x_3^2 + x_4 + 2 - 2\sqrt{2}$

$c_3(x) = x_1 x_5 - 2$

23, PGRT, 5/2, $x^0 = (2,2,2,2,2)^T$

$f(x) = (x_1 - 1)^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_4 - 1)^4 + (x_5 - 1)^6$

$c_1(x) = x_1^2 x_4 + \sin(x_4 - x_5) - 2\sqrt{2}$

$c_2(x) = x_2 + x_3^4 x_4^2 - 8 - \sqrt{2}$

24, GPRP, 5/3, $x^0 = (2,2,2,-1,-1)^T$

$f(x) = \exp(x_1 x_2 x_3 x_4 x_5)$

$c_1(x) = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 - 10$

$c_2(x) = x_2 x_3 - 5x_4 x_5$

$c_3(x) = x_1^3 + x_2^3 + 1$

25, GPRT, 2/1, $x^0 = (2,2)^T$

$f(x) = \ln(1 + x_1^2) - x_2$

$c_1(x) = (1 + x_1^2)^2 + x_2^2 - 4$

INVESTIGAÇÃO OPERACIONAL, vol. 17 (Junho 1997) 101-111

UM MODELO PARA A SELECÇÃO DA MELHOR CARTEIRA OBRIGACIONISTA

João J. Corado

Direcção Financeira
Banco Português do Atlântico

Joaquim J. Júdice

Departamento de Matemática
Universidade de Coimbra
3000 Coimbra

Abstract

The portuguese financial market has been putting tremendous effort in creating a strong technological basis to face the natural evolution of the major international financial markets. This paper describes an optimization portfolio selection model based on yield and risk indeces to be applied in bond markets. Assuming a certain scenario, the model gives a best portfolio in real time that provides the bonds to be kept and sold. The mathematical formulation leads into a bilinear fractional programming problem that reduces to linear program after some simple transformations. Finally the validation of the model is discussed by simulation in a real business environment.

Resumo

A disponibilidade de informação em tempo real é, na actualidade, não um luxo mas uma necessidade. O mercado financeiro português, através dos agentes que nele operam, tem nos últimos anos efectuado um grande esforço procurando criar uma base tecnológica sólida de forma a estar apto a enfrentar a permanente evolução inerente aos maiores mercados financeiros internacionais. Neste artigo é investigado um modelo de optimização direccionado para o mercado obrigacionista e baseado em índices de rentabilidade e de risco (lineares, bilineares e fraccionários). Este modelo procura apresentar, num dado momento e mediante uma determinada conjuntura de mercado que lhe é fornecida, a solução óptima ("melhor" carteira) em tempo real. Pretende-se, pois, determinar a carteira de obrigações óptima tendo em consideração a carteira existente, a compra de unidades no mercado e a venda de unidades em carteira. A formulação matemática do modelo consiste em um problema de programação bilinear fraccionária, que é reduzido a um programa linear através de transformações simples. A validade do modelo foi testado por simulação num ambiente de negociação real e as conclusões desse estudo são apresentadas no final do artigo.

Keywords

portfolio selection model, bilinear fractional programming, linear programming.

1. Introdução

O mercado obrigacionista nacional tem registado nos últimos anos uma evolução muito rápida a que não é alheio o facto do panorama financeiro também ter sofrido profunda alterações no decorrer da última década. Uma maior competitividade entre as instituições financeiras, um maior interesse e participação do pequeno e médio investidor e uma conjuntura

politico/económica que tem favorecido o desenvolvimento dos mercados financeiros são alguns dos factores que contribuem para essa evolução. No entanto, os sistemas de apoio à decisão existentes nesta área não têm acompanhado esta tendência, não considerando toda a informação disponível, nomeadamente a utilidade dos modelos de optimização. Estes modelos revestem-se de grande importância para o desenvolvimento dos sistemas de informação existentes, porque assentam o seu funcionamento numa base científica sólida e podem resolver problemas de elevada complexidade em tempo útil, o que é fundamental para a sua integração num ambiente em que as decisões são medidas ao segundo.

Como exemplo da importância que estes modelos podem assumir, algumas instituições financeiras consideradas "market-makers" nas principais praças mundiais têm gabinetes de investigação dedicados ao desenvolvimento de sistemas de apoio à decisão nas mais diversas áreas [5,10,12]. A Goldman-Sachs implementou um modelo de programação linear de razoável complexidade para a constituição e gestão de uma carteira obrigacionista multidivisas. O JP Morgan desenvolveu e implementou um sistema próprio de gestão de risco utilizando conceitos de Estatística, e obteve um sucesso tão grande que os principais bancos mundiais procuram implementá-lo ou seguir a sua lógica de funcionamento nos seus sistemas internos. Alguns investigadores universitários desenvolvem em colaboração com instituições financeiras modelos de igual importância [7], sendo exemplo disso, H. Konno, que desenvolveu e implementou dois modelos de optimização para a gestão de carteiras obrigacionistas [8,9]. Um dos modelos designado por modelo de optimização total procura optimizar um determinado índice associado à carteira; o outro, denominado modelo de optimização parcial, optimiza um determinado índice associado apenas ao universo dos títulos transaccionados num dado instante. Estes modelos foram estudados e adaptados ao caso português num trabalho de dissertação de mestrado [13].

Como é discutido em [13], o modelo de optimização total prefigura-se como sendo o que melhor poderá desempenhar o papel de conselheiro e também de pioneiro, junto dos investidores nacionais. A adaptação deste modelo ao mercado obrigacionista português poderá ser feito, como adiante se verá, de forma pacífica e com benefícios comprovados, suscitando, deste modo, também o interesse de potenciais utilizadores relativamente a este tipo de modelos. Nesse trabalho, houve a preocupação de construir um modelo que sem perda de generalidade fosse eficaz na gestão de uma carteira obrigacionista, mas também reflectisse a necessidade de tratamento da informação em tempo real de forma a ser possível tomar uma decisão rápida e fundamentada. Assim se obteve um modelo de optimização constituído por uma função objectivo que maximiza um índice de rentabilidade, no caso a "yield to maturity", com uma restrição relativa ao risco associado à carteira dado pela "modified duration", limites de variação para os preços de compra e venda dos títulos e limites relativos às quantidades disponíveis para compra e venda desses mesmos títulos. A sua formulação matemática conduz a um problema de programação bilinear fraccionária, de resolução bastante complexa. Contudo, através de

algumas transformações de variáveis e utilizando a técnica de Charnes e Cooper [3], é possível obter um problema de programação linear que não apresenta dificuldades computacionais.

Este modelo foi avaliado através da simulação de um ambiente de negociação real que tomou em consideração os dois pontos acima referidos, isto é, a fiabilidade das soluções encontradas e a rapidez com que as mesmas são conseguidas.

Na organização deste artigo são apresentadas na secção 2 a descrição do modelo de optimização total e a respectiva formulação matemática. Na secção 3 é efectuada a resolução do problema através da transformação de um problema de programação bilinear fraccionária num problema de programação linear. A validação do modelo é descrita na secção 4 e, por último, na secção 5 são apresentadas as conclusões do trabalho realizado.

2. Modelo de Optimização Total

O modelo de optimização total pretende optimizar um determinado índice relativo à carteira resultante da transacção (compra e/ou venda) efectuada sujeito a um conjunto de restrições baseadas nos restantes índices. Desta forma podemos considerar três conjuntos de índices associados à rentabilidade, aos risco e ao lucro/prejuízo decorrente da especulação.

Os índices de rentabilidade são construídos com base nos conceitos de "direct yield", "yield to maturity" e "effective yield". O risco é apresentado através da "duration de Macaulay", da "modified duration", da variação do preço da obrigação resultante da alteração de um *basis point* na taxa de juro e da convexidade [5]. Recorde-se que a "duration" pode ser interpretada como o período médio de recuperação do investimento feito e a "modified duration" como uma medida da sensibilidade do preço de uma obrigação em relação a uma alteração na "yield to maturity" ("effective yield"). Uma visão mais completa do risco associado a uma carteira pode ser obtida incorporando no modelo um índice que reflicta o impacto no valor da carteira resultante de uma determinada variação das taxas de juro. Esta alteração da estrutura temporal das taxas de juros pode ocorrer no presente (gap analysis) ou algures no futuro (horizon analysis). É importante referir que o risco cambial deve ser considerado, visto ser possível a constituição de carteiras com títulos em moeda estrangeira. Indicadores de lucro/prejuízo reflectem as mais ou menos valias decorrentes de cada transacção efectuada, podendo ser incluídos neste conjunto índices de variação cambial.

A selecção de uma função objectivo é feita nos conjuntos de índices de rentabilidade e risco. Os índices de lucro/prejuízo não são considerados como objectivos primários, pois no modelo em questão pretende-se construir uma carteira obrigacionista com determinadas características de rentabilidade e/ou risco inerentes a uma perspectiva de investimento e/ou negociação, excluindo a componente negociação com objectivos especulativos.

O modelo de optimização total assenta nas considerações apresentadas. Pretende-se que a estrutura do problema seja simples embora deva incluir os principais tipos de restrições, para que tipifique o maior número de possibilidades. Assim qualquer outro problema construído

com base no conjunto de índices descritos deve seguir os mesmos passos que são dados para resolver este problema-tipo.

O principal objectivo do modelo consiste em determinar as quantidades a comprar dos títulos disponíveis no mercado e/ou as quantidades a vender dos títulos disponíveis em carteira de forma a maximizar um dos índices relativo à rentabilidade (no caso a "yield to maturity").

De forma a limitar o risco associado à carteira a determinar incluímos uma restrição constituída pela média da "duration" ponderada pelas quantidades. É um indicador simples que pode reflectir o conceito de prazo médio do investimento realizado, fornecendo informação ao investidor relativa ao período temporal associado à sua carteira e necessário para recuperar o investimento.

É possível ainda, introduzir nos modelos outras variáveis baseadas no preço de compra e/ou venda de cada título, criando situações que induzem pequenas variações nos preços associados a uma determinada transacção. É claro que, estes casos dependem dos mercados financeiros em que se está a actuar, nomeadamente do tipo de legislação que está em vigor, da relação entre os intervenientes numa transacção e dos condicionalismos a nível de impostos que existam no momento.

O caso típico de uma situação deste género ocorre quando um "trader" compra e vende títulos, simultaneamente, a uma mesma entidade, por exemplo, a um corretor. Então, e tendo em consideração os factores acima descritos, pode ser possível escolher o preço de cada título, desde que respeite um determinado intervalo de variação previamente acordado com o corretor. Pode-se verificar uma transacção, em que se vende um determinado número de títulos a preços inferiores aos de mercado com o objectivo de reduzir o lucro dessa transacção e dessa forma diminuir o montante sobre o qual incidem os impostos a pagar.

Qualquer outra restrição com base nos índices descritos anteriormente, tem um tratamento em termos matemáticos idêntico à restrição seleccionada.

A variabilidade dentro de certos limites da componente preço vem introduzir uma maior complexidade no modelo, reduzindo de forma significativa o universo de algoritmos que podem resolver este tipo de problemas satisfatoriamente. Embora esta característica não seja sempre válida, deve ser tomada em consideração, pois flexibiliza o modelo aproximando-o da realidade que é suposto representar. Da mesma forma, não é obrigatório que numa mesma transacção se verifiquem simultaneamente compras e vendas, sendo perfeitamente admissível ocorrerem situações em que se pretende avaliar se os títulos disponíveis no mercado representam uma melhoria em relação à carteira actual. Portanto esta agilidade do modelo acarreta maiores responsabilidades para o investidor na medida em que a respectiva parametrização deve ser correcta e efectuada rapidamente de forma a que eventuais oportunidades de negócio não sejam desperdiçadas. No entanto, este último aspecto dependerá bastante da ligação e interacção entre o modelo e os sistemas de informação em que estiver integrado.

No sentido de construir a formulação matemática, iremos a seguir apresentar as variáveis, os dados e os parâmetros com que iremos trabalhar. Para aliviar um pouco a escrita das definições que se seguem, assumimos que as variáveis j e k tomam os seguintes valores:

$j = 1,\ldots, n_1, \quad k = 1,\ldots, n_2, \quad$ para $\quad n_1 \leq N, n_2 \leq N$

As variáveis do problema são:

x_j número de unidades da obrigação B_j a vender;

x_k número de unidades da obrigação B_k' a comprar;

y_j preço de venda por unidade da obrigação B_j;

y_k' preço de compra por unidade da obrigação B_k';

Os dados do problema são:

μ_j rentabilidade por unidade da obrigação B_j, $j = 1,\ldots, N$;

μ_k' rentabilidade por unidade da obrigação B_k';

t_j número de anos até ao vencimento da obrigação B_j, $j = 1,\ldots, N$;

t_k' número de anos até ao vencimento da obrigação B_k';

p_j preço de mercado para venda de uma unidade da obrigação B_j, $j = 1,\ldots, N$;

p_k' preço de mercado para compra de uma unidade da obrigação B_k';

u_j número de unidades, da obrigação B_j, existentes em carteira;

π_j prazo médio de retorno por unidade da obrigação B_j, $j = 1,\ldots, N$;

π_k' prazo médio de retorno por unidade da obrigação B_k';

Os parâmetros do problema são:

λ_j percentagem de variação admitida sobre o preço de venda da obrigação B_j;

λ_k' percentagem de variação admitida sobre o preço de compra da obrigação B_k';

Prazo máximo temporal de retorno admissível para a carteira;

a_j número mínimo de unidades da obrigação B_j disponíveis para venda;

a_k' número mínimo de unidades da obrigação B_k' disponíveis para compra;

b_j número máximo de unidades da obrigação B_j disponíveis para venda;

b_k' número máximo de unidades da obrigação B_k' disponíveis para compra;

Por último, é definido um parâmetro P_j a partir de

$P_j = (1+\lambda_j)p_j, \quad j = 1,\ldots, N$

A sua justificação será apresentada mais adiante. Tendo em conta o objectivo e as restrições referidas anteriormente e o significado das variáveis, dados e parâmetros também enunciados, obtém-se a seguinte formulação matemática para o modelo em causa

$$\text{Maximizar} \quad f = \frac{\sum_{j=1}^{N} \mu_j t_j P_j u_j - \sum_{j=1}^{n_1} \mu_j t_j x_j y_j + \sum_{k=1}^{n_2} \mu_k' t_k' x_k' y_k'}{\sum_{j=1}^{N} t_j P_j u_j - \sum_{j=1}^{n_1} t_j x_j y_j + \sum_{k=1}^{n_2} t_k' x_k' y_k'}$$

$$\text{sujeito a} \quad \frac{\sum_{j=1}^{N} \pi_j u_j - \sum_{j=1}^{n_1} \pi_j x_j + \sum_{k=1}^{n_2} \pi'_k x'_k}{\sum_{j=1}^{N} u_j - \sum_{j=1}^{n_1} x_j + \sum_{k=1}^{n_2} x'_k} \leq \text{Prazo} \qquad (1)$$

$$(1-\lambda_j)p_j \leq y_j \leq (1+\lambda_j)p_j, \qquad j = 1,\ldots, n_1$$

$$(1-\lambda'_k)p'_k \leq y'_k \leq (1+\lambda'_k)p'_k, \qquad k = 1,\ldots, n_2$$

$$a_j \leq x_j \leq b_j, \qquad j = 1,\ldots, n_1$$

$$a'_k \leq x'_k \leq b'_k, \qquad k = 1,\ldots, n_2$$

3. Resolução do Problema de Optimização

A formulação matemática apresentada na secção anterior constitui um problema de programação bilinear fraccionária de complexidade elevada, para o qual não existem algoritmos para o resolver em tempo real [6]. No entanto, dada a relação entre programas fraccionários e lineares [1,3] e explorando convenientemente as estruturas das restrições e a definição da função objectivo, podemos, de modo semelhante ao apresentado em [4], obter uma formulação linear equivalente.

Nesse sentido começaremos por efectuar as seguintes substituições:

$$z_j = x_j y_j, \qquad j = 1,\ldots, n_1$$

$$z'_k = x'_k y'_k \qquad k = 1,\ldots, n_2$$

Deste modo eliminamos a existência de multiplicações entre variáveis. Como ambas as variáveis, x_t e y_t, são não negativas também a variável de substituição, z_t, é não negativa e o problema toma a seguinte forma

$$\text{Maximizar} \quad f = \frac{\sum_{j=1}^{N} \mu_j t_j P_j u_j - \sum_{j=1}^{n_1} \mu_j t_j z_j + \sum_{k=1}^{n_2} \mu'_k t'_k z'_k}{\sum_{j=1}^{N} t_j P_j u_j - \sum_{j=1}^{n_1} t_j z_j + \sum_{k=1}^{n_2} t'_k z'_k}$$

$$\text{sujeito a} \quad \frac{\sum_{j=1}^{N} \pi_j u_j - \sum_{j=1}^{n_1} \pi_j x_j + \sum_{k=1}^{n_2} \pi'_k x'_k}{\sum_{j=1}^{N} u_j - \sum_{j=1}^{n_1} x_j + \sum_{k=1}^{n_2} x'_k} \leq \text{Prazo} \qquad (2)$$

$$(1-\lambda_j)p_j \leq y_j \leq (1+\lambda_j)p_j, \qquad j = 1,\ldots, n_1$$

$$(1-\lambda'_k)p'_k \leq y'_k \leq (1+\lambda'_k)p'_k, \qquad k = 1,\ldots, n_2$$

$$a_j \leq x_j \leq b_j, \qquad j = 1,\ldots, n_1$$

$$a'_k \leq x'_k \leq b'_k, \qquad k = 1,\ldots, n_2$$

Como referimos anteriormente, após termos concluido da qualidade das soluções fornecidas pelo modelo, é necessário verificar que o processo consegue disponibilizar a informação em tempo real. Nesse sentido, foram realizadas algumas experiências com três conjuntos de títulos de dimensões diferentes, 50, 100 e 150. O período em que foram recolhidos os dados utilizados é relativo à primeira quinzena de Novembro de 1995. Para cada conjunto foi iniciada uma nova carteira constituída por 50% dos títulos em teste de forma a que o número de eventuais transações fosse elevado. Este procedimento foi repetido 5 vezes para evitar oscilações momentâneas na capacidade de processamento da máquina onde decorreram os testes.

No final dos testes respeitantes a cada conjunto de títulos foram retirados os tempos de processamento e agrupados em oito intervalos disjuntos, como se pode verificar no Gráfico III. Os resultados apresentados demonstram que, mesmo para um problema de grande dimensão com 150 unidades, é possível extrair soluções em tempo real (0.25 segundos).

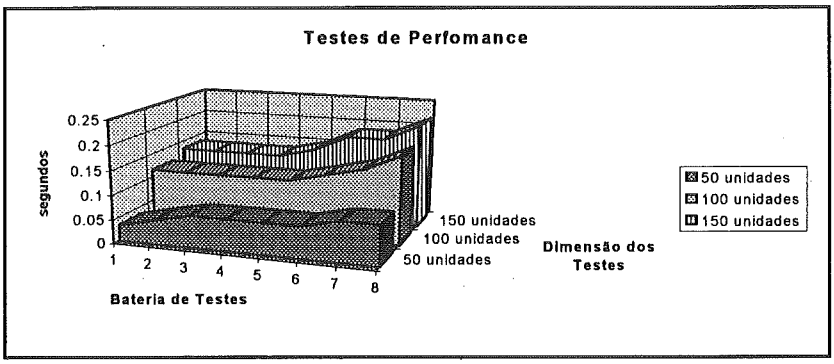

Gráfico III

5. Conclusões

A resposta em tempo real do modelo de optimização total fornece ao investidor uma leitura imediata das condições de mercado e o respectivo impacto na sua carteira. A consequente tomada de decisão incorpora informação única, tornando todo este processo mais transparente e racional. A qualidade das soluções apresentadas, aliada à rapidez do modelo, liberta o investidor para questões exclusivamente do âmbito negocial, mais do seu agrado, melhorando significativamente o seu desempenho.

Este modelo, apesar de simples, constitui uma ferramenta excelente que poderá ser utilizada nos ambientes mais diversos como sejam, gabinetes de Investigação e Consultoria Financeira, onde a qualidade das soluções é preferida à velocidade, em Salas de Mercados em que a rapidez e qualidade das soluções são medidas ao segundo, ou até mesmo pelo pequeno e médio investidor que poderá avaliar, de forma fácil, a sua carteira de títulos.

Por último aparaz-nos registar que a colaboração obtida para este trabalho por parte dos operadores de mercado e o entusiasmo que os resultados obtidos produziram, constituem fortes incentivos para o desenvolvimento deste tipo de modelos em trabalhos futuros.

Nas experiências efectuadas, as Obrigações do Tesouro a taxa fixa constituiram o universo de títulos disponíveis para a composição de uma carteira. Estes títulos são normalmente designados pelas siglas OT, e pelos ano e mês de vencimento (por exemplo: OT-2005/02). Os preços foram obtidos a partir de agências noticiosas financeiras e de corretores. O período sobre o qual incidiu o estudo é referente ao mês de Janeiro de 1996. Foi ainda considerada a existência de uma carteira inicial.

O gráfico I indica a evolução da rentabilidade da carteira ao longo do mês experimental, sugerida pelos resultados obtidos, e mostra a estabilidade da mesma entre os níveis 9.5% e 10%. Este facto é explicado pelo elevado número de unidades em carteira do título OT-1998/02 nos primeiros dez dias, e do título OT-2005/02 na parte final do mês.

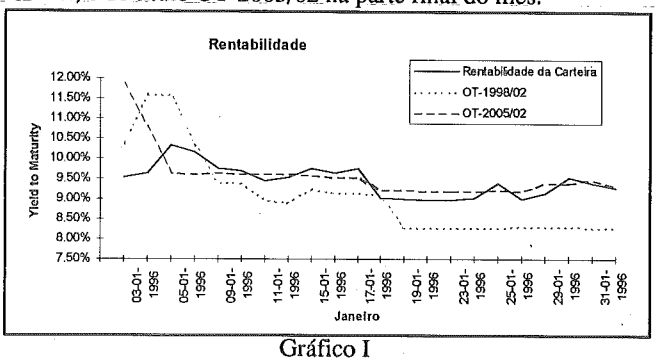

Gráfico I

Ao longo do mês a carteira seleccionada foi composta por 4 títulos, sempre em quantidades significativas. Estes resultados demonstram a forma eficaz como o modelo efectua a selecção dos títulos, evitando quebras de rentabilidade repentinas (exemplo: OT-1998/02) sem prejuízo das restrições impostas. Conclui-se assim, que a rentabilidade apurada evidencia a excelente capacidade do modelo em fazer representar na carteira apenas os títulos que melhor possam assegurar o cumprimento dos objectivos propostos. Além disso é muito importante referir a eliminação de todas as transacções consideradas não significativas, evitando assim os elevados custos que lhes estão sempre associados. Este último facto pode ser comprovado pelo Gráfico II, que mostra que apenas seis títulos foram negociados num total de onze.

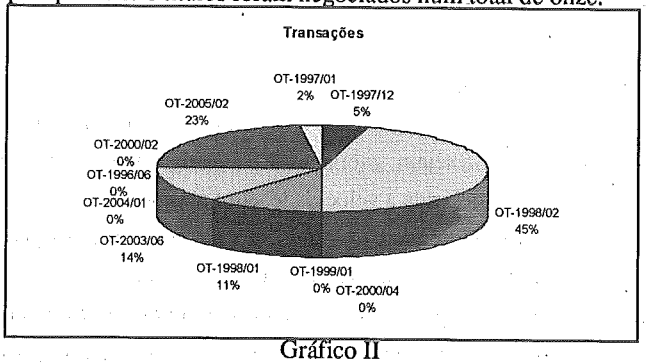

Gráfico II

Apesar de bilinear, este problema tem uma forma muito mais simples e poderá ser transformado num problema de programação linear por eliminação dos factores que consistem em multiplicações resultantes da transformação de Charnes e Cooper. Assim, fazendo as substituições

$$S_j = x_j t, \quad j = 1,\ldots, n_1$$
$$S'_k = x'_k t, \quad k = 1,\ldots, n_2$$

obtemos finalmente o problema de optimização com variáveis R_j, R'_k, S_j, S'_k e t da forma

$$\text{Maximizar} \quad f = \sum_{j=1}^{N} \mu_j t_j P_j u_j t - \sum_{j=1}^{n_1} \mu_j t_j R_j + \sum_{k=1}^{n_2} \mu'_k t'_k R'_k$$

$$\text{sujeito a} \quad \sum_{j=1}^{N} t_j P_j u_j t - \sum_{j=1}^{n_1} t_j R_j + \sum_{k=1}^{n_2} t'_k R'_k = 1$$

$$\sum_{j=1}^{N} (\pi_j - \text{Prazo}) u_j t - \sum_{j=1}^{n_1} (\pi_j - \text{Prazo}) S_j + \sum_{k=1}^{n_2} (\pi'_k - \text{Prazo}) S'_k \leq 0$$

$$(1-\lambda_j) p_j S_j \leq R_j \leq (1+\lambda_j) p_j S_j, \qquad j = 1,\ldots, n_1$$

$$(1-\lambda'_k) p'_k S'_k \leq R'_k \leq (1+\lambda'_k) p'_k S'_k, \qquad k = 1,\ldots, n_2$$

$$a_j t \leq S_j \leq b_j t, \qquad j = 1,\ldots, n_1$$

$$a'_k t \leq S'_k \leq b'_k t, \qquad k = 1,\ldots, n_2$$

É fácil de concluir que a função objectivo e as restrições deste último problema de optimização são lineares nas variáveis referidas anteriormente. Assim obtivemos um problema de programação linear, para o qual existem algoritmos muito eficientes para o resolver em tempo real [2,11].

4. Validação do Modelo

Na primeira fase para a validação do modelo procurámos comprovar, que a sua implementação constitui uma melhoria significativa nos métodos de trabalho e na qualidade dos resultados que os investidores nacionais têm ao seu dispôr. Seguidamente, estudámos a eficiência que o modelo demonstra na resolução de problemas de grande dimensão.

A possibilidade de parametrizar um conjunto de informação de forma a que em determinado momento se conheçam quais as transacções a realizar, representa um grande passo em frente relativamente à forma como este tipo de negócio é efectuado actualmente. No entanto, dois factores são fundamentais para que isso seja conseguido. Assim, a parametrização deve ser suficientemente flexível para que exista correspondência entre o modelo e a realidade que é suposto representar. Além disso, o processo deve ser muito rápido a sugerir um eventual cenário transaccional ao investidor.

A restrição relativa ao prazo deve ser simplificada de forma a ser retirada a respectiva componente fraccionária. O denominador desta restrição é sempre positivo, de forma que é possível multiplicar o segundo membro na inequação pelo denominador do primeiro membro sem o sinal da mesma ser alterado. Assim se obtém um programa linear fraccionário. Tal como é referido em [4], iremos seguidamente retirar a componente fraccionária da função objectivo. Entre as várias formas de abordar a questão foi escolhida a transformação de Charnes e Cooper [1,3], pela sua simplicidade e facilidade de implementação.

Antes de efectuarmos esse tipo de abordagem, notemos que da definição da variável x_j e do significado de u_j facilmente se conclui que

$$x_j \le u_j, \qquad j = 1,\ldots, n_1$$

Por outro lado $P_j = (1+\lambda_j)p_j$ e portanto

$$P_j u_j \ge x_j y_j = z_j$$

Donde

$$\sum_{j=1}^{N} t_j P_j u_j - \sum_{j=1}^{n_1} t_j z_j \ge 0, \quad j = 1,\ldots, n_1$$

e o denominador da função objectivo é positivo. Tal como é descrito em [4], consideramos agora a variável t definida por

$$t = \frac{1}{\sum_{j=1}^{N} t_j P_j u_j - \sum_{j=1}^{n_1} t_j z_j + \sum_{k=1}^{n_2} t'_k z'_k}$$

Fazendo agora as substituições

$$R_j = z_j t, \; R'_k = z'_k t, \; j = 1,\ldots, n_1; \; k = 1,\ldots, n_2$$

obtemos o seguinte problema equivalente a (2)

$$\text{Maximizar} \quad f = \sum_{j=1}^{N} \mu_j t_j P_j u_j t - \sum_{j=1}^{n_1} \mu_j t_j R_j + \sum_{k=1}^{n_2} \mu'_k t'_k R'_k$$

$$\text{sujeito a} \quad \sum_{j=1}^{N} t_j P_j u_j t - \sum_{j=1}^{n_1} t_j R_j + \sum_{k=1}^{n_2} t'_k R'_k = 1$$

$$\sum_{j=1}^{N} (\pi_j - \text{Prazo}) u_j - \sum_{j=1}^{n_1} (\pi_j - \text{Prazo}) x_j + \sum_{k=1}^{n_2} (\pi'_k - \text{Prazo}) x'_k \le 0 \qquad (3)$$

$$(1-\lambda_j)p_j x_j t \le R_j \le (1+\lambda_j)p_j x_j t, \qquad j = 1,\ldots, n_1$$

$$(1-\lambda'_k)p'_k x'_k t \le R'_k \le (1+\lambda'_k)p'_k x'_k t, \qquad k = 1,\ldots, n_2$$

$$a_j t \le x_j t \le b_j t, \qquad j = 1,\ldots, n_1$$

$$a'_k t \le x'_k t \le b'_k t, \qquad k = 1,\ldots, n_2$$

Referências

[1] Bazaraa, M.S., Sherali, H.D e Shetty, C.M., *Nonlinear Programming*, John Wiley and Sons, New York (1993).

[2] Brooke, A., Kendrick, D. e Meeraus, A.,*GAMS User's Guide*, The Scientific Press, San Francisco (1992).

[3] Charnes, A., Cooper, W.W., *Programming with Linear Fractional Functions*, Naval Research Logistics Quarterly 9 (1962) 181-186.

[4] Floudas, C.A., Hansen, P. e Jaumard, B., *Reformulation of Two Bond Portfolio Optimization Models*, Les cahiers du Gerard (1991).

[5] Gruber, E.J. e Gruber, N.J., *Modern Portfolio Theory and Investment Analysis*, John Wiley and Sons, New York (1991).

[6] Horst, R. e Tuy, H., *Global Optimization*, Springer Verlag (1991).

[7] Khan, R.N. e Rudd, A., *Does Historical Performance Predict Future Performance?*, Barra Newsletter (1995).

[8] Konno, H. e Inori, M., *Bond Portfolio by Bilinear Fractional Programming*, Institute of Human and Social Sciences, Tokyo Institute of Technology (1989).

[9] Konno, H. e Watanabe, H., *Nonconvex Bond Portfolio Optimization Problems and Their Applications to Index Tracking*, Institute of Human and Social Sciences, Tokyo Institute of Technology (1994).

[10] Markowitz, H., *Portfolio Selection: Efficient Diversification of Investments*, John Wiley and Sons, New York (1959).

[11] Murty, K., *Linear Programming*, John Wiley and Sons, New York (1983).

[12] Sharpe, W., *Portfolio Theory and Capital Markets*, McGraw-Hill, New York (1970).

[13] Corado, J., *Optimização de Portfolios Obrigacionistas*, Tese de Mestrado, Instituto Superior Técnico, Lisboa (1996).

INSTRUÇÕES AOS AUTORES

Os autores que desejem submeter um artigo à Investigação Operacional devem enviar três cópias desse trabalho para:

Prof. Joaquim J. Júdice
Departamento de Matemática
Universidade de Coimbra
3000 Coimbra, Portugal

Os artigos devem ser escritos em Português ou Inglês. A primeira página deve conter a seguinte informação:

- Título do artigo
- Autor(es) e instituição(ões) a que pertence(em)
- Abstract (em inglês)
- Resumo
- Keywords (em inglês)
- Título abreviado

As figuras devem aparecer em separado de modo a poderem ser reduzidas e fotocopiadas. As referências devem ser numeradas consecutivamente e aparecer por ordem alfabética de acordo com os seguintes formatos:

Artigos: autor(es), título, título e número da revista (livro com indicação dos editores), ano, páginas.

Livros: autor(es), título, editorial, local de edição, ano.

Fotografia, Montagem
Impressão e Acabamentos
Tip.Nocamil
COIMBRA

ÍNDICE

Associação Portuguesa para o Desenvolvimento da Investigação Operacional

CESUR - Instituto Superior Técnico - Avenida Rovisco Pais 1000 Lisboa - Telef. 80 74 55