ISSN 0873-4275

INSTITUTO NACIONAL DE ESTATÍSTICA

PORTUGAL

REVISTA DE ESTATÍSTICA

STATISTICAL REVIEW

VOLUME III
3º QUADRIMESTRE 2001

REVISTA DE ESTATÍSTICA

STATISTICAL REVIEW

ÍNDICE

INDEX

ARTIGOS

REVISTA DE ESTATÍSTICA
7ª PÁGINA
3º QUADRIMESTRE DE 2001

Controlo de Qualidade em Registos de Cancro: Uma Nova Abordagem para o ROR-SUL

Autores:
Sónia Dória Nóbrega
e
Carlos Daniel Paulino

CONTROLO DE QUALIDADE EM REGISTOS DE CANCRO: UMA NOVA ABORDAGEM PARA O ROR-SUL

QUALITY CONTROL IN CANCER REGISTRIES: A NEW APPROACH FOR ROR-SUL

Autores: Sónia Dória Nóbrega
- Consultora Estatística do Registo Oncológico Regional Sul Instituto Português de Oncologia de Francisco Gentil
Carlos Daniel Paulino
- Professor Associado do Departamento de Matemática do Instituto Superior Técnico da Universidade Técnica de Lisboa

RESUMO:

- No presente estudo, criou-se um modelo de predição das observações com maior probabilidade de erro de exactidão dos dados e causadores de maior impacte na qualidade dos indicadores epidemiológicos produzidos pelo ROR-Sul. Utilizaram-se modelos de regressão logística e, para avaliar a sua reprodutibilidade, calculou-se a sensibilidade e especificidade. O modelo encontrado dá-nos a indicação de existirem variáveis capazes de predizer a existência de erros de registo; no entanto, a avaliação da reprodutibilidade do modelo leva-nos a concluir que caso não se encontrem meios que permitam alcançar resultados com melhores sensibilidades, este método de predição de erros de registo não é eficiente nem rendível.

PALAVRAS-CHAVE:

- *Registos de Cancro; Controlo de Qualidade; Exactidão; Validade; Fiabilidade.*

ABSTRACT:

- In this study we develop a prediction model to detect observations with high probability of data accuracy errors with impact on quality of epidemiological measures produced by ROR-Sul. Logistic regression was used to obtain the model and its reproducibility was evaluated by means of sensibility and specificity. The results identified predicting factors of registry errors; however, the model's reproducibility indicate that, if there is no way to reach better sensitivities, this method of predicting cancer errors is neither efficient nor profitable.

KEY-WORDS:

- *Cancer Registry; Quality Control; Accuracy; Data Validity; Data Reliability.*

1. INTRODUÇÃO

O registo de cancro de base populacional é um instrumento essencial de qualquer programa de controlo de cancro a nível europeu (Muir *et al.*, 1985). À sua actividade está associada a gestão de todos os registos de casos de cancro diagnosticados numa população definida, onde estão documentadas as características individuais dos doentes assim como os dados clínicos e anátomo-patológicos dos tumores, recolhidos de forma contínua e sistemática em diferentes fontes de informação. O registo analisa e interpreta estes dados, proporcionando anualmente informação sobre a incidência de todos os tumores malignos, através da sua distribuição geográfica, por sexo e grupo etário. Esta informação não serve apenas como ponto de partida para a investigação epidemiológica sobre os agentes que determinam o cancro, mas também para o planeamento dos cuidados a prestar, tanto ao nível da prevenção como do diagnóstico e tratamento da doença oncológica.

O registo de cancro é, acima de tudo, uma fonte de informação, e como tal a busca da excelência deve ser uma prioridade. Como tem sido alegado, uma informação que não merece confiança é pior do que a falta de informação (Skeet, 1991).

Os princípios básicos do controlo de qualidade em registos de cancro surgiram e evoluiram independentemente do seu desenvolvimento em outras disciplinas, nomeadamente na indústria. Para tal contribuíram programas de controlo de qualidade de prestação de cuidados de saúde a nível hospitalar e de outros sistemas de informação médica (Zippin *et el.*, 1985).

Existem cinco características reconhecidas como indicadores de qualidade de registos de cancro (Hilsenbeck, 1994, Goldberg *et al.*, 1980):

- a ***exaustividade do registo*** (*case completeness*) – definida como a proporção de todos os casos de cancro incidentes numa população definida que constam da base de dados do registo, relativamente a todos os casos de cancro incidentes nessa população (Parkin *et al.*, 1992);
- a ***exactidão*** (*accuracy*) – definida como a proporção de casos registados com uma dada característica (sexo, idade, diagnóstico) que realmente têm esse atributo. Na prática, a exactidão traduz-se na percentagem de concordância entre a informação que consta no registo de cancro e a informação de uma fonte independente (Hilsenbeck *et al.*, 1985);
- a ***disponibilização atempada da informação*** (*timeliness*) - a sua definição está associada ao cumprimento de um determinado nível de calendarização das fases de registo, que permita a tomada de acções decorrentes da informação produzida (Hilsenbeck, 1994);
- a ***estabilidade de codificação*** (*coding constancy*) - a sua definição está associada a práticas de codificação uniformizadas entre e intra instituições e, tanto quanto possível, constantes ao longo dos anos;

- e a ***estabilidade de publicação*** (*publication constancy*) - a sua definição está associada à padronização das publicações periódicas, quer ao nível do formato de apresentação quer ao do uso de metodologias estatísticas.

Os dois primeiros indicadores são os mais conhecidos e usados, concentrando-se neles a maior parte do desenvolvimento de metodologias para a sua avaliação (Goldberg *et al.*, 1980).

De acordo com os resultados demonstrados num estudo desenvolvido pela *North American Association of Central Cancer Registries* (Roffers, 1996), o autor refere que, na utilização dos dados para o controlo do cancro, a exaustividade do registo é menos importante do que a exactidão dos dados registados.

A exactidão dos dados é, assim, uma componente essencial no controlo de qualidade de qualquer registo de cancro. Depende da fiabilidade da informação contida nos documentos que dão origem ao registo (e.g. processo clínico, relatórios anátomo-patológicos), do nível de destreza na recolha dos dados, da codificação e da introdução dessa informação na base de dados.

Com base nesta perspectiva, a evolução dos procedimentos de trabalho do Registo Oncológico Regional Sul (ROR-Sul) nos últimos anos resultou na passagem de um registo feito em suporte de papel para um sistema de registo via *Internet*.

O ROR-Sul é responsável pelo registo de cancro da região Sul de Portugal. A sua área de actividade inclui o distrito de Santarém, Lisboa, Setúbal, Portalegre, Évora, Beja e Faro, cobrindo cerca de metade da área e população de Portugal (aproximadamente 4.200.000 habitantes). Funciona como registo de cancro de base-populacional desde 1989.

É constituído por uma estrutura central, sediada nas instalações do Centro Regional de Lisboa do IPOFG, e por núcleos locais sediados nos hospitais centrais e distritais, nas Sub-Regiões de Saúde e, ainda, em infraestruturas de saúde privadas, nomeadamente nos laboratórios de Anatomia Patológica e instituições hospitalares privadas.

Recebe, por ano, dados de cerca de 12.000 novos casos de cancro e cerca de 7000 óbitos por tumor maligno (Miranda, 1998).

As vantagens do sistema de transmissão de dados que está implementado são:

- permitir que a maioria das validações de edição dos dados sejam feitas no momento da introdução dos dados;
- reduzir substancialmente o tempo que decorre entre a identificação dos casos, em cada instituição, e a aceitação e integração numa base de dados central de toda a informação respeitante aos novos casos de cancro.

Este sistema depara-se agora com uma dificuldade; embora exista um programa de formação contínua que visa a manutenção da padronização dos conceitos e procedimentos de registo, dada a diversidade de pessoas que contribuem para o sistema, o ROR-Sul necessita de encontrar mecanismos capazes de detectar falhas

em todo o sistema de informação, por forma a garantir uma maior qualidade dos dados.

Uma das deficiências apontadas aos estudos de avaliação da qualidade realizados até à data pelo ROR-Sul, é o facto de estes terem um carácter pontual, ou seja, realizam-se anualmente apenas em alguns dos núcleos locais. Por outro lado, com o arranque da rede de transmissão de dados só agora se torna viável ao ROR-Sul centralizar toda a informação dos diferentes núcleos locais com o espaço temporal suficiente para poder trabalhar a base de dados da incidência, previamente à publicação dos seus indicadores.

Os estudos de avaliação da qualidade dos dados apontam para a existência de uma considerável proporção de erros na exactidão dos dados, produzidos de uma forma sistemática, ou seja, associados a determinadas características dos registos (por exemplo, lateralidade do tumor associada à topografia do tumor, datas de incidência e instituição que regista o caso, etc.).

Por esta razão, pretende-se com o presente trabalho testar uma metodologia que vise criar um modelo de predição das observações com maior probabilidade de erro de exactidão dos dados e causadoras de maior impacte na qualidade dos indicadores epidemiológicos produzidos pelo ROR-Sul (e.g. taxas de incidência). Pretende-se ainda avaliar a reprodutibilidade dos resultados obtidos pelo modelo.

Por razões de limitação de tempo, decidiu-se restringir a população-alvo a duas fontes de informação do ROR-Sul, designadas, por razões de confidencialidade, por *Instituição A* e *Instituição B*.

2. MATERIAL E MÉTODOS

2.1. TIPO DE ESTUDO

Trata-se de um estudo transversal dirigido a todos os casos de tumores malignos incidentes (novos casos) num período de um ano e registados pela *Instituição A* e *Instituição B*.

2.2. POPULAÇÃO-ALVO

A população estudada compreende todos os novos casos de tumores malignos diagnosticados em 1996 que acorreram à *Instituição A* (634 casos), e diagnosticados em 1997 que acorreram à *Instituição B* (3509 casos).

2.3. AMOSTRA

Foram recolhidas duas amostras aleatórias simples, por instituição, uma para se proceder à aplicação do modelo de predição e outra para avaliação da reprodutibilidade dos resultados obtidos.

Não foi possível determinar a dimensão da amostra de acordo com os objectivos do estudo, uma vez que se desconhece a existência de qualquer tipo de estudo idêntico ao presente que fornecesse alguma informação *a priori*. Como tal, optou-se por recolher uma amostra que conseguisse garantir a ocorrência de uma determinada prevalência de erros, conforme demonstravam estudos anteriores de avaliação da qualidade dos dados do ROR-Sul.

Considerando uma situação de amostragem sem reposição, o cálculo da dimensão das amostras de cada instituição fez-se com base na fórmula derivada do intervalo de confiança para uma proporção numa população finita

$$n \geq \frac{p(1-p)\frac{N}{N-1}}{\frac{\Delta^2}{z^2_{1-\frac{\alpha}{2}}} + \frac{p(1-p)}{N-1}}$$

onde se assumiram os seguintes valores:

- uma estimativa da proporção p de concordância dos dados (i.e. de não haver neles erros de registo considerados graves) de 80%;
- uma margem de precisão de Δ=5%;
- um nível de confiança 1-α=95%, pelo que $z_{1-\frac{\alpha}{2}} = 1{,}96$;
- N é a dimensão da população da instituição em estudo.

Obteve-se, assim, uma amostra de 177 casos para a *Instituição A* e 230 casos para a *Instituição B*.

Para o estudo da reprodutibilidade do modelo, estipulou-se, por questões práticas, que a dimensão da amostra fosse 50% da amostra inicial, de cada instituição.

Assim, o total de casos seleccionados, para a *Instituição A*, foi de 266 (177+89) e de 345 para a *Instituição B* (230+115).

2.4. COLHEITA DA INFORMAÇÃO

Depois de seleccionadas as amostras de cada instituição, procedeu-se à revisão dos dados dos casos identificados, consultando, para tal, os processos clínicos em cada uma das instituições.

A informação recolhida é aquela que consta do instrumento de "Entrada para Registo" do ROR-Sul, contendo variáveis referentes ao doente tais como nome, sexo, data de nascimento, naturalidade, residência, profissão e situação do doente, ou variáveis clínicas referentes ao diagnóstico e tratamento do tumor, nomeadamente, meios de diagnóstico utilizados, localização do tumor, tipo histológico, datas de início dos sintomas/diagnóstico/tratamento, estádio da doença, tipo de tratamento, etc.

2.5. MODELO DE PREDIÇÃO DOS ERROS DE REGISTO

Antes de iniciar uma breve abordagem ao modelo de regressão logística, é importante realçar que o objectivo da análise, por este método, é igual ao objectivo postulado em qualquer técnica de construção de um modelo de regressão: encontrar o modelo que melhor se ajusta e mais parcimonioso, e que seja biologicamente plausível na descrição da variabilidade de uma resposta (variável dependente) em função de um conjunto de variáveis independentes.

O que distingue o modelo de regressão logística de um modelo de regressão linear é o facto de a variável resposta, no primeiro, ser dicotómica (ou, pelo menos, categorizada). Esta diferença reflecte-se nos pressupostos e na escolha do modelo paramétrico subjacente aos dados.

Relembremos que o presente estudo tem como objectivo criar um modelo de predição das observações com maior probabilidade de erro nos dados e causadoras de maior impacte na qualidade dos indicadores epidemiológicos produzidos pelo ROR-Sul (e.g. taxas de incidência);

Para cada observação o nível de desacordo é definido em função da repercussão dos erros nos indicadores epidemiológicos. Assim, um erro que altere os dados referentes ao cálculo das taxas de incidência (alteração do número de novos casos) é considerado um desacordo máximo (e.g. alteração do sexo ou do grupo etário); ao invés, uma alteração da informação registada que não produza mudanças nos dados para o cálculo das taxas de incidência (e.g. alteração da idade mas manutenção do mesmo grupo etário) é considerada uma discordância mínima. Com base nestes níveis de desacordo, definem-se os seguintes níveis de erro para cada observação da amostra:

Nível 1 – quando ocorreu um erro com repercussões no cálculo da incidência, ou seja, desacordo máximo em pelo menos uma das variáveis sexo, idade, distrito de residência, data do diagnóstico, topografia, comportamento;

Nível 0 – quando não ocorreu nenhum erro de registo ou o erro é mínimo, não tendo qualquer influência nas variáveis que interferem com o cálculo da incidência.

Uma vez revistos todos os casos, cada variável em estudo foi classificada segundo o nível de discordância encontrado, de acordo com os critérios propostos por Hilsenbeck (Hilsenbeck, 1994 e Hilsenbeck *et al.*, 1985). Posteriormente, cada caso foi classificado de acordo com o nível de erro 0 ou 1, definido anteriormente. Esta será a variável resposta do modelo de predição de erros de registo.

As variáveis independentes, candidatas a este modelo, são as que têm os valores que constam do registo original. Nalguns casos decidiu-se definir alguns agrupamentos por forma a favorecer a discriminação de uma observação com maior probabilidade de erro nos dados.

Assim, foram seleccionadas as seguintes variáveis:

- Instituição
  - A
  - B
- Idade (anos) ou, alternativamente, a variável categorizada
  - 0-14 anos
  - ≥15 anos
- Sexo
  - M
  - F
- Região de Naturalidade
  - Portugal Continental
  - Ilhas
  - Estrangeiro
- Região de Residência
  - Lisboa
  - Santarém/Setúbal/Alentejo/Algarve
  - Outros Distritos do Continente
  - Ilhas
  - Estrangeiro
- Situação do Caso
  - Vivo
  - Morto
- Tipo de Diagnóstico
  - Histol. Primário
  - Histol. Metástases
  - Autópsia c/ Histologia
  - Citologia/Hematologia
  - Exames Macroscópicos
- Intervalo entre a 1ª Observação e a 1ª Consulta (dias) ou, alternativamente, a variável categorizada
  - < 0 dias
  - [0 , 180] dias
  - > 180 dias
- Intervalo entre o Diagnóstico e a 1ª Consulta (dias) ou, alternativamente, a variável categorizada
  - < -365 dias
  - [-365 , 90] dias
  - > 90 dias
- Intervalo entre os 1os Sintomas e a 1ª Consulta (dias) ou, alternativamente, a variável categorizada
  - < -365 dias
  - [-365 , 365] dias
  - > 365 dias
- Intervalo entre os 1os Sintomas e a 1ª Observação (dias) ou, alternativamente, a variável categorizada
  - ≤ 365 dias
  - > 365 dias
- Intervalo entre os 1os Sintomas e o Diagnóstico (dias) ou, alternativamente, a variável categorizada
  - ≤ 365 dias
  - > 365 dias
- Intervalo entre a 1ª Observação e o Diagnóstico (dias) ou, alternativamente, a variável categorizada
  - ≤ 60 dias
  - > 60 dias

- Topografia
  - Lábio,Cavidade Oral e Faringe (C00-C14)
  - Órgãos Digestivos (C15-C26)
  - Sist. Respirat. Órg. Intratoráx. (C30-39)
  - Ossos,Articul. Cartil. Articul. (C40-C41)
  - Sist. Hematop. Retículo Endotelial (C42)
  - Pele (C44)
  - Nervos Perif. e Sist. Nerv. Autónomo (C47)
  - Retroperitoneu e Peritoneu (C48)
  - Tec.Conjunt., Subcut. e Out.Tec.Moles (C49)
  - Mama (C50)
  - Órgãos Genitais Femininos (C51-C58)
  - Órgãos Genitais Masculinos (C60-C63)
  - Tracto Urinário (C64-C68)
  - Olho, Céreb. e Out.P.Sist.Nerv.Central (C69-C72)
  - Tiróide e Out. Glândulas Endócrinas (C73-C75)
  - Outras Localiz. Mal Definidas (C76)
  - Gânglios Linfáticos (C77)
  - Localiz. Primária Desconhecida (C80)
- Morfologia
  - Neoplas. Pouco Específicas (9990, 8000-8041)
  - Neoplasias Específicas (8042-9989)
- Comportamento do Tumor
  - Benigno, Incerto se Malig/Benig., Carc. In Situ
  - Maligno
- Lateralidade do Tumor
  - Direita, Esquerda, Bilateral
  - Não Aplicável, Desconhecida
- Estádio na Apresentação
  - Tumor Localizado
  - Com Metástases
  - Não Aplicável
  - Desconhecido
- T
  - Conhecido
  - Omisso
- N
  - Sem invasão ganglionar (N=0)
  - Com invasão ganglionar (N=1..3)
  - Omisso
- M
  - Sem Metástases à Distância (M=0)
  - Com Metástases à Distância (M=1)
  - Omisso
- Grau de Diferenciação
  - Conhecido (Bem/Moderado/Pouco/Indif.)
  - Desconhecido

A escolha dos níveis das variáveis categorizadas fez-se com base em diversos critérios: ou pela representatividade de cada um dos níveis na amostra (e.g. região de naturalidade e residência, tipo de diagnóstico, topografia), ou pelo conhecimento prévio de categorias mais problemáticas (e.g. grupo etário, morfologia, grau de diferenciação, comportamento do tumor) ou ainda pela criação de grupos que eventualmente descrevam situações menos vulgares, que possam corresponder a erros nos dados (e.g. intervalos entre datas). Intervalos negativos entre a data da 1ª observação pela doença e a data da 1ª consulta na instituição são certamente indicadores de erro, uma vez que a 1ª observação da doença não pode ocorrer após a 1ª consulta numa instituição por essa doença. Com a data dos sintomas não existem

fronteiras tão claras, dado que a doença pode ser observada e diagnosticada sem que o doente apresente sintomas; no entanto, sintomas que se manisfestem há mais de um ano, sem que ocorra uma consulta motivada pelas queixas, é uma situação menos vulgar.

O modelo de regressão logística será ajustado à 1ª amostra, definida para cada instituição, num total de 407 casos.

2.6. CONSTRUÇÃO DO MODELO MÚLTIPLO

Em função do tipo de estudo, existem diferentes objectivos para o ajustamento do modelo de regressão logística:

- no *modelo para predições,* o objectivo é construir um modelo que, de entre uma colecção de variáveis, seleccione apenas aquelas que melhor predigam ou se ajustem à resposta;
- no *modelo para factores de risco,* o objectivo é estimar a associação entre a exposição a determinados factores de risco e a resposta.

No primeiro caso, pretende-se que o modelo seja o mais parcimonioso possível e com o melhor ajustamento aos dados, enquanto que, no segundo caso, poderão obter-se modelos com mais variáveis uma vez que se avaliam efeitos de confundimento e interacções com o factor de risco em estudo. O critério de selecção das variáveis de modelos para predições prende-se mais com critérios estatísticos do que com questões de plausibilidade biológica das variáveis (Hosmer *et al.*, 1989 e Kleinbaum *et al.*, 1982).

Assim, optou-se pelo uso do método *stepwise* com eliminação regressiva, que se baseia no ajustamento do modelo saturado, com as variáveis definidas como candidatas ao modelo, e na fixação de uma regra de decisão com um valor-p de saída, p_s=0.050, isto é, com base no Teste de Razão de Verosimilhanças de Wilks saem do modelo saturado todas as variáveis cujo valor-p, dado por este teste, seja superior a p_s. Com base num critério meramente estatístico, proposto por Hosmer e Lemeshow (Hosmer *et al.*, 1989), considerou-se uma variável candidata ao modelo múltiplo se o valor-p do teste de razão de verosimilhanças for inferior a 0.250.

O valor preditivo de cada variável foi caracterizado através da razão das chances (*odds ratio* - OR) e respectivo intervalo de 95% de confiança.

2.7. DIAGNÓSTICO DO MODELO FINAL

Como em qualquer modelo ajustado, antes de se prosseguir com a inferência sobre este, dever-se-á verificar o seu nível de ajustamento. Existe um conjunto de medidas de avaliação da qualidade do ajustamento que descrevem o comportamento do modelo estimado relativamente aos dados observados, de uma forma global ou particular.

A avaliação da qualidade de ajustamento iniciou-se pela aplicação das metodologias de avaliação global - o Teste do Qui-quadrado de Pearson e o Teste de Hosmer e Lemeshow (Hosmer *et al.*, 1989) - e prosseguiu com uma avaliação cuidadosa das medidas de diagnóstico do modelo que permitem identificar observações com comportamento problemático, sendo estas medidas de três tipos (Hosmer *et al.*, 1989; Hosmer *et al.*, 1980; Lemeshow *et al.*, 1982; Hosmer *et al.*, 1988 e Hosmer *et al.*, 1997):

- medidas de resíduos
- medidas de "leverages"
- medidas de influência

As medidas de resíduos servem para identificar observações em que o valor esperado obtido pelo modelo não está próximo do valor observado da variável resposta.

Quando um indivíduo apresenta um padrão de valores para as covariáveis que é muito pouco usual, quando comparado com o dos restantes elementos da amostra, as medidas de "leverages" são as indicadas para identificar situações deste tipo.

Quando se pretende avaliar o impacte no modelo da presença ou ausência de determinado elemento na amostra devemos observar as alterações produzidas nos coeficientes estimados do modelo. A contribuição de uma observação para a alteração dos coeficientes depende tanto do valor do resíduo como do valor do "leverage". As medidas de diagnóstico destas situações são chamadas medidas de influência.

Para avaliar simultaneamente a relação da sensibilidade (proporção de casos com erros de nível 1 correctamente identificados pelo modelo) e 1-especificidade (sendo a especificidade a proporção de casos com erros de nível 0 correctamente identificados pelo modelo) dos resultados obtidos pelo modelo, faz-se variar o ponto de corte de 0 a 1 e representa-se graficamente a curva ROC (*Receiver Operating Characteristic* – Metz, 1978). Este método tem tido uma utilização cada vez mais frequente por fornecer uma medida de discriminação do modelo ajustado (Hanley *et al.*, 1982; Hanley *et al.*, 1983; Beck *et al.*, 1986; Schultz *et al.*, 1995; Brickley *et al.*, 1995; Hagen *et al.*, 1995; Fortescue *et al.*, 2000; Dreiseitl *et al.*, 2000 e Beggl *et al.*, 2000).

Hosmer (1995) considera a seguinte regra de avaliação do poder de discriminação do modelo de regressão logística baseada na área sob a curva ROC:

Tabela 1: Classificação do nível de discriminação do modelo de regressão logística

Área sob a Curva ROC	Interpretação
0,50	Não discrimina
≥ 0,70	Discriminação aceitável
≥ 0,80	Discriminação excelente
≥ 0,90	Discriminação marcante (pouco usual)

2.8. AVALIAÇÃO DA REPRODUTIBILIDADE DO MODELO

Uma vez encontrado o modelo final, com uma boa qualidade de ajustamento aos dados, pretendemos avaliar qual o verdadeiro valor discriminatório do modelo encontrado, isto é, até que ponto as predições obtidas pelo modelo correspondem à realidade numa amostra independente daquela que deu origem ao modelo.

Assim, através da segunda amostra retirada em cada instituição (204 casos) pretende-se avaliar a reprodutibilidade dos resultados através do modelo obtido na primeira amostra, avaliando a estimativa da sensibilidade, especificidade e valor preditivo positivo (proporção de casos com erros de nível 1 entre os casos assinalados pelo modelo como errados).

3. RESULTADOS

3.1. REPRESENTATIVIDADE DA AMOSTRA

Dos 611 casos seleccionados para as duas amostras a recolher nas instituições, apenas foram revistos os processos clínicos de 252 casos da *Instituição A* (94,7%) e 338 casos da *Instituição B* (98,0%). Em 11 casos os processos clínicos não foram localizados pelo arquivo clínico, em 9 o processo clínico estava muito incompleto e 1 caso esteve com o processo clínico requisitado pelo serviço responsável pelo doente, durante todo o período de revisão dos dados deste estudo.

A amostra ficou assim reduzida a 590 casos das duas instituições. Dado que o cálculo da dimensão das amostras previa 611 casos, 407 para a amostra que ajustaria o modelo de predição dos erros e 204 para a da avaliação da reprodutibilidade do modelo, decidiu-se que os 21 casos em que não foi possível rever a informação registada seriam retirados da 2ª amostra (dada a menor importância desta amostra nos objectivos do trabalho). Assim, a 2ª amostra ficou reduzida a 183 casos.

Não se rejeitou a hipótese de homogeneidade das amostras com o total de casos registados por cada uma das instituições, no que se refere à distribuição dos sexos e grupos etários. De igual modo não se encontraram diferenças estatisticamente significativas nestas variáveis relativamente aos casos em que foi possível rever os processos clínicos e os 21 casos retirados da amostra por não haver possibilidade de revisão dos dados.

Na globalidade dos dados apenas algumas variáveis interferem directamente na determinação de alguns indicadores epidemiológicos; para o cálculo das taxas de incidência dos tumores malignos estão envolvidas as variáveis sexo, idade, distrito de residência, data do diagnóstico, comportamento e localização primária do tumor (órgão). Basta que ocorra um desacordo máximo numa destas variáveis para que a observação tenha um erro com impacte na incidência .

A distribuição destes erros (nível 0 e 1), na amostra, fez-se da seguinte forma:

Tabela 2: Distribuição dos erros de registo por instituição

Erros de Registo	Total	Instituição	
		A	*B*
Nível 0	547 casos 92,7%	240 casos 95,2%	307 casos 90,8%
Nível 1	43 casos 7,3%	12 casos 4,8%	31 casos 9,2%

Como podemos verificar o nível de exactidão global foi de 92,7%, ocorrendo uma maior frequência de erros de registo com impacte na incidência na *Instituição B* do que na *A*, sendo esta diferença estatisticamente significativa (p=0,042).

Tabela 3: Distribuição dos erros de registo por amostra - a de ajustamento do modelo (1ª) e a de avaliação da reprodutibilidade (2ª)

Erros de Registo	Total	Amostra	
		1ª	*2ª*
Nível 0	547 casos 92,7%	376 casos 92,4%	171 casos 93,4%
Nível 1	43 casos 7,3%	31 casos 7,6%	12 casos 6,6%

Verifica-se que não existe uma diferença estatisticamente significativa (p=0,647) na distribuição do nível dos erros segundo a amostra.

3.2. MODELO DE PREDIÇÃO DOS ERROS DE REGISTO

3.2.1. MODELOS UNIVARIADOS

Encontraram-se algumas variáveis que individualmente estão significativamente associadas à ocorrência de erros de registo de nível 1. A variável Instituição é um dos factores mais associados onde se verifica que os casos registados pela *Instituição B* têm uma chance de ocorrência de um erro de nível 1 3,5 (valor da OR) vezes superior à dos casos da *Instituição A*. Os casos registados com idades inferiores a 15 anos são também aqueles que apresentam maior probabilidade de erro de registo (OR=6,61) relativamente aos casos com idades superiores ou iguais a 15 anos. Encontrou-se

também uma associação entre o sexo e a ocorrência de erros de nível 1 - os casos do sexo masculino têm uma chance de erro cerca de 50% menor que os casos do sexo feminino; no entanto, esta associação não revela uma significância estatística muito forte (p=0,113).

De assinalar, também, uma associação entre a ocorrência dos erros de registo e a região de naturalidade dos casos. São os casos oriundos das ilhas ou do estrangeiro aqueles que maiores probabilidades revelam de dar origem a um erro de registo, quando comparados com os casos naturais de Portugal Continental.

Para o estudo da região de residência, agruparam-se os distritos em: Lisboa, restantes distritos do ROR/Sul (Santarém, Setúbal, Évora, Beja, Portalegre e Faro), restantes distritos de Portugal Continental não pertencentes ao ROR/Sul, Ilhas e Estrangeiro. A razão pela qual se separou Lisboa dos restantes distritos do ROR/Sul deveu-se ao facto de ser conhecido que muitos casos são registados erradamente como residentes no distrito de Lisboa (onde se concentra a maioria dos hospitais com oferta de tratamento oncológico) e por isso darem origem a maiores erros de registo com impacte no cálculo das taxas de incidência por distrito. Quando avaliamos essa associação com a região de residência agrupada deste modo, verifica-se que as diferenças existentes não são estatisticamente significativas. Se englobarmos as 3 primeiras categorias numa só, denominada Portugal Continental, observa-se que a associação da região de residência com a ocorrência de erros ainda não tem significância estatística.

Quanto ao tipo de diagnóstico, são as categorias de autópsia com histologia e histologia de metástases as que apresentam maior probabilidade de ocorrência de erro de registo, quando comparadas com os casos com histologia do tumor primário. Esta variável associa-se significativamente com a ocorrência de erros de nível 1.

Existem localizações onde a incidência de erros de registo é mais elevada (e.g. C00-C14, C73-C75, C76, C77 e C80), no entanto, pela fraca representatividade destas categorias na amostra, não existe uma associação estatisticamente significativa. O mesmo acontece com a morfologia do tumor, agrupada em neoplasias pouco especificadas (e.g. 9990+8000-8041) e neoplasias especificadas (restantes códigos).

Quanto ao comportamento do tumor, verifica-se que é o grupo de tumores de comportamento benigno, incerto se benigno ou maligno e os carcinomas *in situ* que apresenta um risco 4 vezes superior de ocorrência de erros de registo, quando comparados com os tumores de comportamento maligno. Efectivamente esta variável está directamente implicada na selecção dos tumores malignos para o cálculo das incidências e é preenchida de acordo com a opção seleccionada para a histologia do tumor, aquando da introdução dos dados em computador. Muitas vezes, por exemplo, para um tumor basocelular podem surgir diversos códigos, um para maligno e outro para incerto quanto à malignidade.

Outra variável que mostrou uma associação estatisticamente significativa foi o grau de diferenciação do tumor. Os casos cujo grau de diferenciação do tumor é desconhecido têm uma chance de ocorrência de um erro de registo 3 vezes maior do que os casos cujo grau de diferenciação é conhecido.

Nenhum dos intervalos de tempo entre datas de 1os sintomas, de 1ª observação, de 1ª consulta ou de diagnóstico se mostrou directamente associado com a ocorrência de erros de registo, quer como variáveis contínuas (representadas em número de dias) quer categorizadas em intervalos de tempo.

3.2.2. CONSTRUÇÃO DO MODELO MÚLTIPLO

Identificaram-se as seguintes variáveis candidatas ao modelo saturado:

- INSTITUIÇÃO
- GRUPO ETÁRIO
- SEXO
- REGIÃO DE NATURALIDADE
- TIPO DE DIAGNÓSTICO
- COMPORTAMENTO
- GRAU DE DIFERENCIAÇÃO

Aplicando o método *stepwise* de selecção de variáveis atinge-se o seguinte modelo de efeitos principais:

Tabela 4: Modelo de efeitos principais - coeficientes do modelo (β), *odds ratios* (OR), respectivos intervalos de confiança dos OR e valor-p do Teste de Wald

Variável	Coef. β	SE(β)	OR	I.C. 95%	T. Wald valor-p
CONSTANTE	-4,83	0,64	---	---	---
INSTITUIÇÃO					
Ref.: A					
B	1,99	0,59	7,35	2,31 – 23,37	0,001
GRUPO ETÁRIO					
Ref.: + 15 anos.					
0-14 anos	2,33	0,89	10,24	1,81 – 58,13	0,009
TIPO DIAGNÓST.					
Ref.: Hist. Primário					
Hist. Metástases	2,91	0,83	18,43	3,64 – 93,21	< 0,001
Autópsia c/ Hist.	2,26	1,42	9,62	0,60 – 153,99	0,110
Citologia	-0,63	0,91	0,53	0,09 – 3,14	0,486
GRAU DIFERENC.					
Ref.: Conhecido					
Desconhecido	1,09	0,45	2,97	1,24 – 7,12	0,015

O interesse em estudar o efeito das interacções no modelo é o de avaliar o "paralelismo" entre os efeitos principais, ou seja, se a associação entre uma dada covariável e a ocorrência dos erros de registo é constante segundo os níveis de outra covariável. Uma interacção é considerada relevante, do ponto de vista quantitativo, quando o seu efeito de entrada no modelo é estatisticamente significativo (α=0,050). O critério é idêntico ao adoptado para os efeitos principais.

A inclusão de uma interacção não significativa no modelo apenas complicará as inferências a realizar e conduzirá a uma menor precisão das estimativas obtidas dos coeficientes do modelo. O critério de selecção das interacções faz-se através de *selecção propositada progressiva.*

O estudo das interacções não se deve iniciar pela inclusão de todas as interacções de interesse no modelo, dado que muitas vezes conduzem a problemas numéricos no processo inferencial.

Neste caso, o modelo de regressão logística que incluía as interacções abaixo assinaladas a negrito pôde ser estimado por haver categorias com 0 elementos, quando cruzadas as variáveis.

INST. × GR.ETARIO	GR.ETARIO × T.DIAGN.
INST. × T.DIAGN.	**GR. ETARIO × G.DIFER.**
INST. × G.DIFER.	T.DIAGN. × G.DIFER.

Ajustou-se o modelo de efeitos principais e avaliou-se o efeito de inclusão de cada interacção no modelo. Nenhuma interacção se mostrou de importância estatisticamente significativa no modelo de efeitos principais.

Chegamos, assim, ao modelo final de regressão logística para a probabilidade $\Pi(x)$ de erro de nível 1:

$$\Pi(\mathbf{x}) = \frac{e^{-4.83+1.99\times INSTIT\_2+2.33\times GR.ETARIO\_2+2.91\times T.DIAGN\_2+2.26\times T.DIAGN\_3-0.63\times T.DIAGN\_4+1.09\times G.DIFER\_2}}{1+e^{-4.83+1.99\times INSTIT\_2+2.33\times GR.ETARIO\_2+2.91\times T.DIAGN\_2+2.26\times T.DIAGN\_3-0.63\times T.DIAGN\_4+1.09\times G.DIFER\_2}}$$

onde as covariáveis x são assim definidas:

INSTIT_2 =0	→	se INSTITUIÇÃO é A
INSTIT_2 =1	→	se INSTITUIÇÃO é B
GR.ETARIO_2=0	→	se GRUPO ETÁRIO é +15 anos
GR.ETARIO_2=1	→	se GRUPO ETÁRIO é 0-14 anos
T.DIAGN_2=0	→	se TIPO DE DIAGNÓSTICO não é Histologia de Metástases
T.DIAGN_2=1	→	se TIPO DE DIAGNÓSTICO é Histologia de Metástases
T.DIAGN_3=0	→	se TIPO DE DIAGNÓSTICO não é Autópsia com Histologia
T.DIAGN_3=1	→	se TIPO DE DIAGNÓSTICO é Autópsia com Histologia
T.DIAGN_4=0	→	se TIPO DE DIAGNÓSTICO não é Citologia
T.DIAGN_4=1	→	se TIPO DE DIAGNÓSTICO é Citologia
G.DIFER_2=0	→	se GRAU DE DIFERENCIAÇÃO é Conhecido
G.DIFER_2=1	→	se GRAU DE DIFERENCIAÇÃO é Desconhecido

Suponhamos um exemplo correspondente a um caso que acorreu à *Instituição B*, com 36 anos de idade, do sexo feminino, natural de Cabo Verde e residente em Lisboa, com um carcinoma espino-celular do ovário, moderadamente diferenciado, diagnosticado por biópsia com um estádio TNM de 100. Este indivíduo apresenta, segundo o modelo ajustado, uma probabilidade estimada de erro de registo com impacte nos cálculos da incidência, dada por

$$\Pi(\mathbf{x}) = \frac{e^{-4,83+1,99\times1+2,33\times0+2,91\times0+2,26\times0-0,63\times0+1,09\times0}}{1+e^{-4,83+1,99\times1+2,33\times0+2,91\times0+2,26\times0-0,63\times0+1,09\times0}} = 0,055$$

3.2.3. AVALIAÇÃO DA QUALIDADE DE AJUSTAMENTO DO MODELO FINAL

Para a avaliação da qualidade de ajustamento do modelo obtido utilizou-se o Teste do Qui-quadrado de Pearson e o Teste de Hosmer e Lemeshow.

Tabela 5: Avaliação da qualidade do ajustamento do modelo final

Teste	Estatística	Graus de Liberdade	Valor-p
QUI-QUADRADO DE PEARSON	4,69	9	0,860
HOSMER E LEMESHOW	0,54	5	0,990

Concluíu-se que, globalmente, não existe evidência de que o modelo de regressão logística esteja mal ajustado aos dados (para um nível de significância de 5%).

Individualmente, a avaliação do impacte no modelo, pela presença ou ausência de determinado elemento na amostra, é dada no gráfico que se segue:

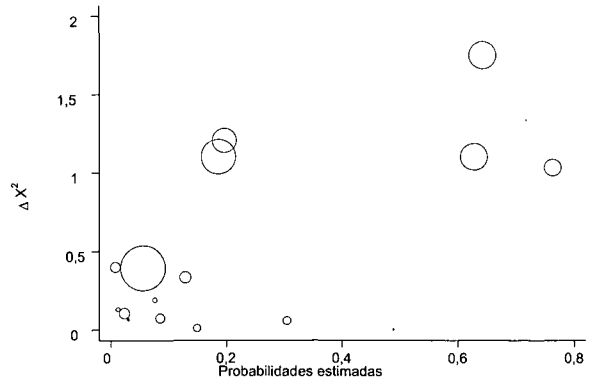


Figura 1: Gráfico de avaliação das medidas de influência de cada observação

As observações assinaladas com um círculo de maior diâmetro correspondem a observações influentes no modelo, mas pelo facto de não serem consideradas mal ajustadas (o descréscimo no valor da estatística de teste do Qui-quadrado de Pearson - ΔX^2 - devida à eliminação dos indivíduos com as convariáveis x_j é, para um nível de significância de 5%, inferior ao valor do quantil de probabilidade de um Qui-quadrado com um grau de liberdade - 3,84), decidiu-se mantê-las no modelo por forma a garantir o maior número de observações na amostra.

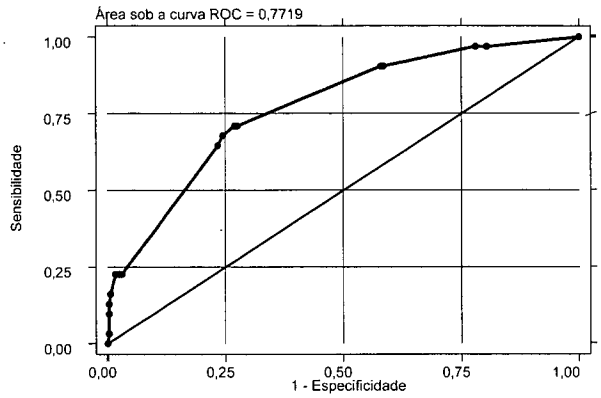


Figura 2: Curva ROC obtida pela aplicação do modelo final de regressão logística

O valor da área sob a curva ROC é de 0,772, indicando, segundo a classificação da tabela 1, que o poder de discriminação do modelo é considerado aceitável.

De facto, pelo gráfico da sensibilidade e especificidade (figura 3) verifica-se que os valores da sensibilidade do modelo (probabilidade do modelo detectar erros de registo, quando eles existam) diminuem drasticamente entre as probabilidades estimadas de 0 e 0,25.

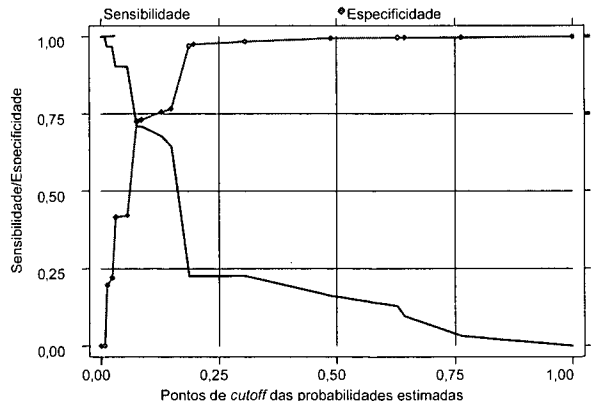


Figura 3: Sensibilidade e especificidade obtida pela aplicação do modelo final de regressão logística, para diversos valores de pontos *cutoff* das probabilidades estimadas

3.3. AVALIAÇÃO DA REPRODUTIBILIDADE DO MODELO FINAL

Aplicando o modelo final à segunda amostra, a discriminação do modelo, segundo os diferentes pontos de corte, foi a seguinte:

		Erro Observado			
	Cutoff				
Erro Estimado	**0,5 e 0,4**	0	1	Total	Sensibilidade = P(EE=1/EO=1) = **0,083**
	0	168	11	179	Especificidade = P(EE=0/EO=0) = 0,982
	1	3	1	4	Valor Preditivo Positivo = P(EO=1/EE=1) = **0,250**
	Total	171	12	183	Valor Preditivo Negativo = P(EO=0/EE=0) = 0,939

		Erro Observado			
	Cutoff				
Erro Estimado	**0,3 e 0,2**	0	1	Total	Sensibilidade = P(EE=1/EO=1) = **0,167**
	0	166	10	176	Especificidade = P(EE=0/EO=0) = 0,971
	1	5	2	7	Valor Preditivo Positivo = P(EO=1/EE=1) = **0,286**
	Total	171	12	183	Valor Preditivo Negativo = P(EO=0/EE=0) = 0,943

		Erro Observado			
	Cutoff				
Erro Estimado	**0,1**	0	1	Total	Sensibilidade = P(EE=1/EO=1) = **0,500**
	0	111	6	117	Especificidade = P(EE=0/EO=0) = 0,649
	1	60	6	66	Valor Preditivo Positivo = P(EO=1/EE=1) = **0,091**
	Total	171	12	183	Valor Preditivo Negativo = P(EO=0/EE=0) = 0,949

Mesmo assumindo um *cutoff point* de 0,1 a sensibilidade assume valores de 50%, significando que o modelo apenas assinala 50% das observações com erros de registo. O verdadeiro valor preditivo do modelo é de 9,1%, indicando que apenas esta proporção das observações estimadas como tendo erro é que realmente o têm.

4. DISCUSSÃO E CONCLUSÕES

O uso de modelos de regressão logística para a caracterização da exactidão dos dados foi apenas encontrado num estudo americano (Polissar *et al.*, 1984), embora num âmbito diferente daquele aqui estudado. Os autores fazem uso desta ferramenta para determinar diferenças estatisticamente significativas nos níveis de desacordo, ajustando simultaneamente para diferentes variáveis categorizadas.

O nosso objectivo vai um pouco mais além; pretendemos com a regressão logística encontrar um modelo que preveja a probabilidade de um determinado registo conter erros que possam alterar as estimativas das incidências por tumores malignos.

Dada a vasta dimensão de cobertura do ROR-Sul e os escassos recursos humanos de que dispõe para a implementação de procedimentos de monitorização da qualidade dos dados, pareceu-nos que este poderia ser um caminho de optimização do esforço a realizar.

Chegando ao final da revisão das duas amostras seleccionadas, foi com apreensão que nos defrontámos com apenas 43 casos (7,3%) que apresentavam desacordos de nível 1.

Na verdade, o estudo iniciou-se pressupondo uma percentagem de erros da ordem dos 20%, com um erro de previsão de 5%. Este facto levantou-nos uma primeira hipótese de que as amostras seleccionadas poderiam não ser suficientemente representativas para ajustar um modelo de predição dos erros de registo. A amostra, seleccionada aleatoriamente antes da revisão, que iria dar origem ao ajustamento do modelo de erros de registo, continha uma percentagem ligeiramente superior de erros de nível 1 (7,6%) do que a 2ª amostra de avaliação da reprodutibilidade desse modelo (6,6%).

O modelo múltiplo encontrado identificou 4 variáveis como relevantes na predição dos erros de registo: a instituição responsável pelo registo, o escalão etário, o tipo de diagnóstico e o grau de diferenciação do tumor. Não ocorreram interacções estatisticamente significativas entre estas variáveis.

A presença destas variáveis no modelo não é de todo de difícil compreensão:

- ✓ A Instituição faz sentido na medida em que dependendo do nível de formação e envolvimento das pessoas na tarefa do registo se conseguem ou não melhores resultados. Muitas vezes, se apenas uma pessoa fizer o registo dos casos, mais padronizados ficam os procedimentos e menor variabilidade encontramos. Por outro lado, a forma como os processos clínicos estão organizados na instituição pode influenciar claramente a interpretação dos dados e a sua recolha;
- ✓ O escalão etário definiu que os tumores infantis têm maior probabilidade de conter erros de nível 1. Este facto deveu-se em grande parte às crianças vindas dos PALOP cuja residência tinha sido codificada com a morada da embaixada dos seus países em Portugal. Outra explicação poderá ser também o comportamento incerto, se benigno ou maligno, dos tumores de algumas crianças que podem dar origem a erros de interpretação e codificação;
- ✓ O tipo de diagnóstico revelou que os diagnósticos feitos com base em histologias de metástases e autópsias com histologia são aqueles que maior probabilidade têm de conter erros. Estes ocorrem porque muitas vezes se regista como tumores primários o aparecimento de metástases, registando-se a localização das metástases e data do seu aparecimento como topografia e data do diagnóstico, respectivamente;
- ✓ O grau de diferenciação é o factor cuja interpretação é menos objectiva. Verificou-se que são os tumores cujo grau de diferenciação é omisso que

maiores probabilidades têm de conter erro. Poderíamos pensar que estes casos corresponderiam a diagnósticos histológicos menos claros ou mesmo a casos sem diagnóstico histológico; no entanto, o grau de diferenciação não revelou significância estatística na interacção com o tipo de diagnóstico.

O modelo revelou uma boa qualidade de ajustamento, com algumas observações influentes mas sem capacidade para questionar a hipótese de bom ajustamento.

Apesar da discriminação do modelo ser considerada aceitável, verifica-se que a sensibilidade deste diminui drasticamente quando se aumenta a probabilidade de *cutoff* de 0 a 1. Ao avaliarmos a reprodutibilidade do modelo verificamos que, considerando um ponto de corte de 0,1, o modelo apenas é capaz de predizer 50% dos erros que existem (sensibilidade) e apenas 9% dos registos estimados como contendo um erro, o têm na realidade (valor preditivo positivo). Um modelo com este desempenho não pode ser considerado uma boa ferramenta de optimização de esforços de monitorização da qualidade.

Outra fragilidade que nos pareceu clara é que dependendo da amostra seleccionada o modelo múltiplo pode não encontrar variáveis que predigam os erros de registo, dada a sua falta de representatividade.

Verificou-se que a variabilidade dos erros é bastante grande, fazendo com que a amostra deva ter uma dimensão muito maior para que seja possível caracterizar todos os erros cometidos de forma sistemática, com a prevalência necessária para serem modelados.

Outro caminho sugerido pela experiência indica que a opção dos modelos de regressão logística pode não ser a mais adequada. Outros modelos menos utilizados mas igualmente apropriados para variáveis respostas dicotómicas poderão ser alternativas a testar.

BIBLIOGRAFIA

BECK, J. R., SHULTZ, E. K., (1986), "The Use of Relative Operating Characteristic (ROC) Curves in Test Performance Evaluation", *Archives of Pathology and Laboratory Medicine*, Vol. 110, 13-20.

BEGGL, C. B., CRAMER, L. D., VENKATRAMAN, E. S., ROSAI, J., (2000), "Comparing Tumour Staging and Grading Systems: a Case Study and a Review of the Issues, Using Thymoma as a Model", *Statistics in Medicine*, Vol. 19, Nº 15, pp. 1997-2014.

BRICKLEY, M. R., PRYTHERCH, I. M., KAY, E. J., SHEPERD, J. P., (1995), "A New Method of Assessment of Clinical Teaching: ROC Analysis", *Medical Education*, Vol. 29, pp. 150-153.

DREISEITL, S., OHNO-MACHADO, L., BINDER, M., (2000), "Comparison Three-Class Diagnostic Tests by Three-Way ROC Analysis", *Medical Decision Making*, Vol. 20, Nº 3, pp. 323-331.

FORTESCUE, E. B., KAHN, K., BATES, D. W., (2000), "Prediction Rules for Complications in Coronary Bypass Surgery: a Comparison and Methodological Critique", *Medical Care*, Vol. 38, Nº 8, pp. 820-835.

GOLDBERG, J., GELFAND, H. M., LEVY, P. S., (1980), "Registry Evaluation Methods: a Review and a Case Study", *Epidemiologic Reviews*, Vol. 2, pp. 210-220.

HAGEN, M. D., (1995), "Test Characteristics: How Good Is That Test?", *Medical Decision Making*, Vol. 22, Nº 2, pp. 213-233.

HANLEY, J. A., MCNEIL, B. J., (1982), "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve", *Radiology*, Vol.143, pp. 29-36.

HANLEY, J. A., MCNEIL, B. J., (1983), "A Method of Comparing the Areas under a Receiver Operating Characteristic Curves Derived from the Same Cases", *Radiology*, Vol. 148, pp. 839-843.

HILSENBECK, S. G., (1994) "Quality Control", *Central Cancer Registries: Design, Management, and Use,* Harwood Academic Publishers, Illinois, pp. 131-177.

HILSENBECK, S. G., GLAEFKE, G. S., FEIGL, P., LANE, W. W., GOLENZER, H., AMES, C., DICKSON, C., (1985), *Quality Control for Cancer Registries*, U. S. Department of Health and Human Services, Seattle-WA.

HOSMER, D. W., LEMESHOW, S., (1989), *Applied Logistic Regression*, John Wiley & Sons, New York.

HOSMER, D., HOSMER, T., LE CESSIE, S., LEMESHOW, S., (1997), "A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model", *Statistics in Medicine*, Vol. 16, Nº 9, pp. 965-980.

HOSMER, D., LEMESHOW, S., (1980), "A Goodness-of-Fit Test for the Multiple Logistic Regression Model", *Communications in Statistics*, Series A10, pp.1043-1069.

HOSMER, D., LEMESHOW, S., KLAR, J., (1988), "Goodness-of-Fit Testing for Multiple Logistic Regression Analysis When the Estimated Probabilities are Small", *Biometrical Journal*, Vol. 30, Nº 7, pp. 1-14.

KLEINBAUM, D. G., KUPPER, L. L., MORGENSTERN, H. (1982), *Epidemiologic Research: Principles and Quantitative Methods*, Van Nostrand Reinhold, New York;

LEMESHOW, S., HOSMER, D., (1982), "A Review of Goodness-of-Fit Statistics for Use in the Development of Logistic Regression models", *American Journal of Epidemiology*, Vol. 115, pp. 92-106.

METZ, C., (1978), "Basic Principles of ROC Analysis", *Seminars in Nuclear Medicine*, Vol. 8, pp. 283-297.

MIRANDA, A. C., (1998), "Registo Oncológico Regional Sul", *Automated Data Collection in Cancer Registration*, IARC Technical Reports Nº 32, Lyon, pp. 45-47.

MUIR, C. S., DÉMARET, E., and BOYLE, P., (1985) "The Cancer Registry in Cancer Control: an Overview", *The Role of the Registry in Cancer Control*, IARC Scientific Publication Nº 66, Lyon, pp. 13-26.

PARKIN, D. M., MUIR, C. S., (1992), "Comparability and Quality of Data", *Cancer Incidence in Five Continents, Volume VI*, IARC Scientific Publications Nº 120, Lyon, pp. 45-56.

POLISSAR, L., FEIGL, P., LANE, W. W., GLAEFKE, G., DAHLBERG, S., (1984), "Accuracy of Basic Cancer Patient Data: Results from an Extensive Recoding Survey", *Journal of National Cancer Institute*, Vol. 72, Nº 5, pp. 1007-1013.

ROFFERS, S. D., (1996), "Case Completeness and Data Quality Assessments in Central Cancer registries and Their Relevance to Cancer Control", *http://www.naaccr.org/data/papers/sect94.pdf*.

SCHULTZ, E. K., (1995), "Multivariate Receiver-Operating Characteristic Curve Analysis: Prostate Cancer Screening as an Example", *Clinical Chemistry*, Vol. 41, Series 8B, pp. 1248-1255.

SKEET, R. G., (1991) "Quality and Quality Control", *Cancer Registration: Principles and Methods*, IARC Scientific Publication Nº 95, Lyon, pp. 101-107.

ZIPPIN, C., AKERS, C., (1985), "Quality Control in Cancer Registration Programs", *Current Problems in Cancer*, Vol. 9, Nº 3, pp. 49-60.

REVISTA DE ESTATÍSTICA
33ª PÁGINA
3º QUADRIMESTRE DE 2001

Mean Number of the Customers Served During a Busy Period in the M|G|∞ Queueing System

Autor:
Manuel Alberto M. Ferreira

REVISTA DE ESTATÍSTICA
35ª PÁGINA
3º QUADRIMESTRE DE 2001

MEAN NUMBER OF THE CUSTOMERS SERVED DURING A BUSY PERIOD IN THE M | G | ∞ QUEUEING SYSTEM

NÚMERO MÉDIO DE CLIENTES SERVIDOS DURANTE UM PERÍODO DE OCUPAÇÃO NO SISTEMA DE FILA DE ESPERA M | G | ∞

Autor: Manuel Alberto M. Ferreira
- Professor Associado - Departamento de Métodos Quantitativos - I.S.C.T.E.

ABSTRACT:

- We deduce an exact expression that gives the mean number of the customers served during a busy period in the $M|M|^{\infty}$ queueing system and conjecture an upper bound of this quantity for other $M|G|^{\infty}$ systems, with service times not exponential, based on the results of a Markov Renewal Process.

KEY-WORDS:

- *$M|G|\infty$, Mean, Busy Period, Markov Renewal Process.*

RESUMO:

- Deduzimos uma expressão exacta para o número médio de clientes servidos durante um período de ocupação no sistema de fila de espera $M|M|^{\infty}$ e conjecturamos um extremo superior dessa quantidade para outros sistemas $M|G|^{\infty}$, com tempos de serviço não exponenciais, com base nos resultados de um Processo Markoviano de Renovamento

PALAVRAS-CHAVE:

- *$M|G|\infty$, Média, Período de Ocupação, Processo Markoviano de Renovamento.*

In a M|G|∞ queueing system λ is the Poisson process arrival rate, α is the mean service time, $G(\cdot)$ is the service time distribution function and there are infinite servers.

In general we do not want necessarily the physical presence of infinite servers. But we only guarantee that when a customer arrives at the system it always finds immediately a server available.

Looking at the labour of the system as a sequence of idle and busy periods, it is important to know the number of the customers served during a busy period in order to manage the servers that are indicated to work in the system. We will give here some results about the mean number of the customers served during a busy period in the M|G|∞ queue system. We will call it N_B.

Be v_n, $n = 0,1,\ldots$ the mean number of entries in state n between two entries in state 0. We say that the system is in state n, $n = 0,1,\ldots$ when there are n customers being served in the queue.

The number of customers served in a busy period is equal to number of arrivals at the system in it. But, counting all entries in the various states during a busy period we are counting the arrivals and the exits. The number of arrivals equals the number of exits. And between two entries in state 0 there is only one busy period. So

$$N_B = \frac{\sum_{n=0}^{\infty} v_n}{2} \qquad \textbf{(1)}$$

For the M|M|∞ system (exponential service time), see (Ramalhoto, 1986),

$$v_n = (n+\rho)\frac{\rho^{n-1}}{n!},\ n = 0,1,\ldots \qquad \textbf{(2)}$$

where $\rho = \lambda\alpha$ is the traffic intensity.

So,

$$N_B = \frac{1}{2}\sum_{n=0}^{\infty}(n+\rho)\frac{\rho^{n-1}}{n!} = \frac{1}{2}\left(1+\sum_{n=1}^{\infty}\frac{\rho^{n-1}}{(n-1)!}+\sum_{n=1}^{\infty}\frac{\rho^{n}}{n!}\right) = \frac{1}{2}\left(1+e^{\rho}+e^{\rho}-1\right) = e^{\rho}$$

and

$$N_B^M(\rho) = e^{\rho} \tag{3}$$

Note that $N_B^M(0) = 1$, naturally, because when $\lambda \to 0$ or $\alpha \to 0$ it is difficult to have more than one customer in the system.

The mean busy period length, for any M|G|∞ system, is $\frac{e^{\rho}-1}{\lambda}$, (see, for instance, (Takács, 1962)). Let us put $\frac{e^{\rho}-1}{\lambda e^{\rho}} = A(\rho,\lambda)$. So

- $\lim_{\lambda\to 0}\frac{e^{\rho}-1}{\lambda e^{\rho}} = \lim_{\lambda\to 0}\frac{e^{\rho}\alpha}{e^{\rho}+e^{\rho}\rho} = \lim_{\lambda\to 0}\frac{\alpha}{1+\rho} = \alpha$, using l'Hospital's rule,

- $\lim_{\lambda\to\infty}\frac{e^{\rho}-1}{\lambda e^{\rho}} = \lim_{\lambda\to\infty}\frac{\alpha}{1+\rho} = 0$,

- $A_{\lambda}'(\rho,\lambda) = \frac{e^{\rho}\alpha\lambda e^{\rho} - \left(e^{\rho}+e^{\rho}\alpha\lambda\right)\left(e^{\rho}-1\right)}{\lambda^2 e^{2\rho}} = \frac{e^{2\rho}\rho - e^{2\rho} - e^{2\rho}\rho + e^{\rho} + e^{\rho}\rho}{\lambda^2 e^{2\rho}} =$
  $= \frac{-e^{\rho}+1+\rho}{\lambda^2 e^{\rho}} < 0, \quad \rho > 0,$

- $\lim_{\alpha\to 0}\frac{e^{\rho}-1}{\lambda e^{\rho}} = 0$,

- $\lim_{\alpha\to\infty}\frac{e^{\rho}-1}{\lambda e^{\rho}} = \frac{1}{\lambda}\lim_{\alpha\to\infty}\frac{e^{\rho}-1}{e^{\rho}} = \frac{1}{\lambda}$,

- $A_{\alpha}'(\rho,\lambda) = \frac{e^{\rho}\lambda\lambda e^{\rho} - \left(e^{\rho}-1\right)e^{\rho}\lambda^2}{\lambda^2 e^{2\rho}} = \frac{e^{2\rho}\lambda^2 - e^{2\rho}\lambda^2 + e^{\rho}\lambda^2}{\lambda^2 e^{2\rho}} = \frac{1}{e^{\rho}} > 0,$
  $\alpha > 0$

Then

- with α constant $A(\rho,\lambda)$ decreases with λ varying between α and 0,
- with λ constant $A(\rho,\lambda)$ increases with α varying between 0 and $\frac{1}{\lambda}$.

For other M|G|∞ systems, with service time distributions not exponentials, we do not know v_n. But in (Ramalhoto, 1986) it is suggested to consider a Markov Renewall Process as an approximation to those systems. For that Markov Renewall Process we have

$$v_n = \lambda^{n-1} \frac{B_1 \ldots B_{n-1}}{(1-\lambda B_1)\ldots(1-\lambda B_n)},\ n = 1,2,\ldots \qquad \textbf{(4)}$$

Where

$$B_n = \int_0^\infty e^{-\lambda t} \left[\frac{\int_t^\infty [1-G(x)]dx}{\alpha} \right]^n dt,\ n = 0,1,\ldots \qquad \textbf{(5)}$$

And of course,

$$v_0 = 1 \qquad \textbf{(6)}$$

But

- $B_n \leq \frac{1}{\lambda},\ n = 1,2,\ldots$ because $\alpha^{-1} \int_t^\infty [1-G(x)]dx \leq 1,$

- $B_n \le \alpha \dfrac{\gamma_s^2+1}{n+1}$, $n = 1,2,\ldots$ because $B_n \le \int_0^\infty \left[\dfrac{\int_t^\infty [1-G(x)]dx}{\alpha} \right]^n dt \le$

$\le \dfrac{1}{\alpha^n} \dfrac{n\alpha^2}{2}\left(\gamma_s^2+1\right)\dfrac{\alpha^{n-1}2}{n(n+1)} = \alpha\dfrac{\gamma_s^2+1}{n+1}$, where γ_s is the service time variation coefficient, owing to $\int_0^\infty \left[\int_t^\infty [1-G(x)]dx \right]^n dt = \dfrac{n\alpha^2}{2}\left(\gamma_s^2+1\right)\dfrac{\alpha^{n-1}b_{n-1}}{n(n+1)}$, $n \in |\mathrm{N}$, with $b_n \le 2$, $n = 0,1,\ldots$ (see (Sathe, 1985)),

- $\alpha\dfrac{\gamma_s^2+1}{n+1} \le \dfrac{1}{\lambda} \Leftrightarrow n \ge \rho\left(\gamma_s^2+1\right)-1$.

So, if $\rho \le \dfrac{1}{\gamma_s^2+1}$,

$$\nu_n \le (n+1)\frac{\rho^{n-1}\left(\gamma_s^2+1\right)^{n-1}}{n!},\ n = 1,2,\ldots \qquad (7)$$

for the Markov Renewal Process.

Then $\sum_{n=0}^{\infty}\nu_n = 1+\sum_{n=1}^{\infty}\dfrac{\left(\rho\left(\gamma_s^2+1\right)\right)^{n-1}}{(n-1)!}+\sum_{n=1}^{\infty}\dfrac{\left(\rho\left(\gamma_s^2+1\right)\right)^{n-1}}{n!} = 1+e^{\rho\left(\gamma_s^2+1\right)}+\dfrac{e^{\rho\left(\gamma_s^2+1\right)}-1}{\rho\left(\gamma_s^2+1\right)} =$

$= \dfrac{e^{\rho\left(\gamma_s^2+1\right)}\left(\rho\left(\gamma_s^2+1\right)+1\right)+\rho\left(\gamma_s^2+1\right)-1}{\rho\left(\gamma_s^2+1\right)}.\ \lim_{\rho\to 0}\dfrac{e^{\rho\left(\gamma_s^2+1\right)}\left(\rho\left(\gamma_s^2+1\right)+1\right)+\rho\left(\gamma_s^2+1\right)-1}{2\rho\left(\gamma_s^2+1\right)} =$

$\lim_{\rho\to 0}\dfrac{e^{\rho\left(\gamma_s^2+1\right)}\left(\gamma_s^2+1\right)\left(\rho\left(\gamma_s^2+1\right)+1\right)+e^{\rho\left(\gamma_s^2+1\right)}\left(\gamma_s^2+2\right)+\gamma_s^2}{2\left(\gamma_s^2+1\right)} = \dfrac{\gamma_s^2+1+\gamma_s^2+2+\gamma_s^2}{2\left(\gamma_s^2+1\right)} =$

$= \dfrac{3\gamma_s^2+3}{2\left(\gamma_s^2+1\right)} = \dfrac{3}{2} > 1\left(e^0\right)$, using l'Hospital's rule.

So, may be,

$$M_B = \frac{e^{\rho(\gamma_s^2+1)}\left(\rho(\gamma_s^2+1)+1\right)+\rho(\gamma_s^2+1)-1}{2\rho(\gamma_s^2+1)} \qquad (8)$$

is an upper bound for N_B, for M|G|∞ systems whose service times are not exponential. Note that for exponential service times **(4)** is the same that **(2)**.

REFERENCES

BOROVKOV, A. A., (1976) "Stochastic Processes in Queueing Theory".Springer-Verlag.

KENDALL, D. G., (1964) "Some recent work and further problems in the theory of queues". Theory of Prob. and its Appl. (Russian Journal). 9, 1-13.

KINGMANN, J. F. C., (1970) "Inequalities in the Theory of Queues", J. R. Statist. Soc. R. 32, 102-110.

KINGMANN, J. F. C., (1967) "Approximations for queues in heavy traffic", Queueing theory: Recent Developments and Applications. Edited by R. Gruon (Lisboa Sept. 27-October 1, 1965), 73-78, The English University Press.

MARSHALL, K. T. (1968) "Some inequalities in Queueing". Ops. Research 16, 651-665.

NEWELL, G. F., (1971) "Applications of Queueing Theory". Chapman Hall.

RAMALHOTO, M. F., (1986) Breves Considerações Sobre "Aproximações" e "Limites" Em Sistemas do Tipo G|G|∞. Technical Report. I.S.T.. C.E.A.U.L.. Lisboa.

STOYAN, D., (1976) "Approximations for M|G|s Queues", Math. Operationsforsch. Statist. 7, 587-594.

SATHE, Y. S., (1985) "Improved Bounds for the Variance of the Busy Period of the M|G|∞ Queue", A.A.P. 17, 913-914.

STOYAN, D., (1977) "Bounds and Approximations in Queueing Through Monotonicity and Continuity". Ops. Res. Vol. 25, nº 5, Sept. -Oct., 851-863.

TAKÁCS, L.,(1962) "An Introduction To Queueing Theory". Oxford University Press. New York.

REVISTA DE ESTATÍSTICA
43ª PÁGINA
3º QUADRIMESTRE DE 2001

A New Local Influence Measure in Generalized Linear Models

Autores:
Yenis Marisel González Mora
e
M. Mercedes Suárez Rancel

A NEW LOCAL INFLUENCE MEASURE IN GENERALIZED LINEAR MODELS

UMA NOVA MEDIDA DE INFLUÊNCIA LOCAL EM MODELOS LINEARES GENERALIZADOS

Autores: Yenis Marisel González Mora
- Department of Statistics, Research Operation and Computation. Faculty of Mathematics. University of La Laguna. Tenerife

M. Mercedes Suárez Rancel
- Department of Statistics, Research Operation and Computation. Faculty of Mathematics. University of La Laguna. Tenerife

ABSTRACT:

- This paper investigates the local influence assessment in Poisson generalized linear models. The assessment of local influence in generalized linear models is studied by the approach of Cook (1986). So he only deals with the local influence on the regression coefficients which are not resistant to masking and swamping effects. Suárez and González (2000) propose a new measure to detect locally influential data under perturbations to variance. Based on this measure, in this paper we propose a locally influential measure to mitigate these difficulties. We demonstrate the need of this measure. An example shows the effectiveness of the proposed method.

KEY-WORDS:

- *Generalized linear model, Masking and Swamping, Local influence.*

RESUMO:

- Este artigo procede à avaliação da influência local em modelos lineares generalizados de Poisson. A avaliação da influência local em modelos lineares generalizados é estudada pela abordagem de Cook (1986). Este autor lida apenas com o caso da influência local em coeficientes de regressão que não são robustos a efeitos masking e swamping. Suárez and González (2000) propõem uma nova medida para detectar dados localmente influentes sob perturbações da variância. Neste artigo, propomos uma medida de influência local baseada nessa medida, que minora estas dificuldades e demonstramos a necessidade de tal medida. Um exemplo ilustra a eficácia do método proposto.

PALAVRAS-CHAVE:

- *Modelos Lineares Generalizados, Masking e Swamping, influência local.*

1. INTRODUCTION

We assume the observations consist of a vector y of n independent responses from the exponential family. We study a Poisson generalized linear model, which is a particularly case of this family.

The idea of influence assessment is to monitor the sensitivity of statistical analysis subject to minor changes in the model. The works of some authors (Lawrence, 1988, Peña and Yohai, 1995) indicate that one of the attractions of the local influence concept is that it assesses the effect of joint perturbations on the data cases more easily than global influence measures Suárez and González (2000). Thus in a local sense, frequently, the results are free from masking effects that present difficulties for individual case-deletion methods.

The purpose of this study is to gain additional insight on global influence regarding the local influence analysis and its implications.

In the next section we give the general idea of generalized linear models. Section 3 we describe a general idea of local influence. Section 4 shows that local–influence analysis of perturbations of the variance is similar to the usual regression diagnostic based on Hadi's measure for detecting an influential subset. In Section 5 we extend the 'Suárez and Glez' measure to Poisson generalized linear model. In Section 6 we give a Lawrence's transformation measure. And Section 6 provides an illustrative example.

2. A POISSON GENERALIZED LINEAR MODEL

2.1. RESPONSE DISTRIBUTION

We assume the observations consist of a vector y of n independent responses from the exponential family

$$f(y;\theta;\phi) = \exp\{[\, y\theta - b(\theta)\ ]/a(\phi) + c(y;\phi)\}\, y = 0,1,2,...$$

with $\theta_i = g(\eta_i)$, $\eta_i = x_i'\beta$, where x is an n×p matrix of covariates, is a p-dimensional column vector of unknown parameters, and a(.),b(.),c(.) are known functions. The dispersion parameter ϕ is usually regarded as nuisance parameter.

Then the mean and variance of y can be written by:

$$E[y] = b'(\theta) = \mu$$

$$Var[y] = b''(\theta)a(\phi) = b''(\theta)\,\phi / w,$$

where the primes denote derivates with respect to θ.

Applying this in Poisson distribution we have:

$$f(y;\theta;\phi) = \exp\{(y \ln \lambda - \lambda) - \ln y!\} y = 0,1,2,\ldots \quad (1)$$

$\phi = 1$, $b(\theta) = \lambda$, $c(y, \phi) = -\ln y!$

$E[Y] = \lambda$

$Var[Y] = \lambda$

2.2. LINK FUNCTION

The mean μ_i of the response in the i-th observation is related to a linear predictor through a monotonic differentiable link function g

$$\eta_i = g(\mu_i) = x_i'\beta$$

Here, x_i is a fixed known vector of explanatory variables, and β is a vector of unknown parameters.

In classical linear models they are identical, and the identity link is sensible in the sense that both η and μ can take any value on the real line. However, when are dealing with counts and the distribution is Poisson, we must have $\mu>0$, so that the identity link is less attractive in part because may then be negative. Models based on independence of probabilities associated with the different classifications of cross-classified data lead naturally to considering multiplicative effects, and this is expressed by the log link, $\eta = \log \mu$ with its inverse $\mu = e^{\eta}$ .Therefore the link function for Poisson (λ) models is the log link $\eta = \log \lambda$.

2.3. MAXIMUM LIKELIHOOD FITTING

An iterative methods are required to solve the normals equations:

$$X^t s = X^t (y - \hat{y}) = 0,$$

And these lead to the iterative scheme:

$$\beta^{t+1} = \beta^t - H^{-1} s,$$

where H is the hessian matrix and s is the gradient vector of the log-likelihood function. That is:

$$s=[s_j]=[\partial L/\partial \beta_j] \text{ and } H=[h_{ij}]= [\partial^2 L/\partial \beta_i \partial \beta_j].$$

The estimated covariance matrix of the parameter estimator is given by

$$\Sigma = -H^{-1}$$

In our case, $\Sigma = (X^t V X)^{-1}$, where V is a diagonal matrix with the variance λ_i of the response Y(~ Poisson(λ)) in the i-th observation.

2.4. GOODNESS OF FIT

Two statistics that are helpful in assesing the goodness of fit of a given generalized linear model are the scaled deviance and Pearson's chi-square statistic.

The scaled deviance is defined by

$$D(y,\mu) = 2\{l(X\hat{\beta}; y) - l(\hat{\theta}; y)\},$$

where $l(\hat{\theta}; y)$ refers to the maximum of the log-likelihood function based on fitting each exactly. Therefore, the deviance in Poisson models can be written as:

$$2\{ \Sigma \, y \ln(y/\lambda) + \Sigma \, (y-\lambda)\}$$

And finally, the Pearson'chi-square statistic is defined as:

$$\chi^2 = \Sigma(y_i-\mu_i)^2 /V(\mu_i)$$

where $\hat{\mu}_i = b^{(1)}\left(g\left(x_i\hat{\beta}\right)\right)$ and $Var\left(\hat{\mu}_i\right) = b^{(2)}\left(g\left(x_i\hat{\beta}\right)\right)$.is the estimated variance function for the distribution concerned..

3. LOCAL AND DELETION INFLUENCE

We shall give a brief review of Cook's local influence approach in this section to provide some help for understanding the new formulation which will be defined in Section4.

Consider the Poisson generalized linear model defined in Section 1

$$f(y;\theta;\phi) = \exp\{(y \ln \lambda - \lambda) - \ln y!\} y = 0,1,2,...$$

with $\theta_i = ln\,\lambda$, $\eta_i = X_i\beta$. Let $\hat{\beta}$ be the estimated parameter vector of β (MLE).

Many measures have been suggested to assess the influence of observations. Cook(1986) considers a general version of Cook's distance

$$D_i = \frac{\| \hat{Y} - \hat{Y}_{(i)} \|^2}{k\sigma^2},$$

where $\hat{Y}, \hat{Y}_{(i)}$ are the n×1 vectors of the fitted values based on the full data and the data without i-th case, respectively, and k is the dimension of β.

He investigated

$$D_i(w) = \frac{\| \hat{Y} - \hat{Y}_{(w)} \|^2}{k\sigma^2},$$

where $\hat{Y}_{(w)}$ is the vector of fitted values when the i-th case has weight w and the remaining cases have weight 1.

These ideas have been extended to general cases . In generalized linear models the generalized Cook's distance(McCullagh and Nelder(1989)) is defined by

$$LD_i = \left(\hat{\beta}_{(i)} - \hat{\beta}\right)'(x'Wx)\left(\hat{\beta}_{(i)} - \hat{\beta}\right)/\hat{\phi}$$

as a measure of the effect of the i-th datum point on the parameter estimates, where $W = diag\{W_i\}$, $W_i = b^{(2)}(\hat{\theta}_i)\left[g^{(1)}(\hat{\eta}_i)\right]^2$, $\hat{\beta}_i$ denotes the estimates when that point is omitted and the superscript (2) denotes the 2nd derivative of the function.

An approximate measure of leverage is given by the diagonal element h_i of the projection matrix

$$H = W^{1/2}x(x'Wx)^{-1}x'W^{1/2}$$

Cook(1986) developed a general technique for the assement of local influence. Sensivity of the analysis was assesed though the normal curvature of the likelihood displacement surface when minor perturbations were introduced in the postulated model. Further extentions to generalized linear model were given by Thomas and Cook(1989). The likelihood displacement takes the form

$$LD(w) = 2\left[L(\hat{\beta}) - L(\hat{\beta}_w)\right],$$

where $L(\hat{\beta})$ is the likelihood for β and write $\hat{\beta}_w$ for the MLE from the perturbed model. The vector of the values w and LD(w) from the surface of interest as w varies. The direction h_{max} of the maximum curvature of the likelihood displacement surface in the postulated model(where $w = w_0$) indicates the greatest local sensitivity against perturbations. The direction of maximum curvature is used as the main diagnostic tool in the local influence method. But this measure has some practical and theoretical difficulties. For example, computability of the maximum curvature is restricted to the linear regression model; there is also a lack of invariance of the curvature under reparametrisation of the perturbation scheme; and lack of definition of the parameters.

4. DIAGNOSTIC IN POISSON GENERALIZED LINEAR MODEL

We show that local-influence analysis of perturbations of the variance is similar to Hadi's measure for detecting an influential subset. To avoid the difficulties defined in Section 3, Suárez and González(2000) suggest an alternative likelihood

displacement. Based on this likelihood displacement, we propose a quasilikelihood displacement to mitigate the swamping and masking effect:

$$LD_{(i)}(w_i) = -2\left[L(\hat{\beta}) - L\ (\hat{\beta}_w \mid w) \right] + \left[\text{var}\left(\hat{\mu}_i \right) - \text{var}\left(\hat{\mu}_{w_i} \right) \right]$$

where $\hat{\mu}_i = b^{(1)}\left(g\left(x_i\hat{\beta}\right)\right)$ and $Var(\hat{\mu}_i) = b^{(2)}\left(g\left(x_i\hat{\beta}\right)\right)$.

Therefore , the aim of the present paper is give a measure to mitigate these difficulties in Poisson generalized linear model.

Applying this displacement to Poisson model we obtain the slope:

$$LD^{**}{}_i(w_i) = \hat{y}_i - y_i - \frac{h_i^2}{m_i\ \hat{\lambda}_i},$$

Where h_i are the diagonal elements of the projection matrix H defined is section 2, $\hat{y}_i$ are the predicted values of response variable, m_i is the number of observations in the i-th covariate and λ_i are the diagonal element of the variance matrix V.

However, in Poisson generalized linear model we have no constant variance, therefore we use Box and Cox (1964) transformation in model (1) to mitigate this problem.

Then, we have:

$$LD^{**}{}_i(w_i) = \hat{y}_i - \sqrt{y_i} - \frac{h_i^2}{m_i\ \hat{\lambda}_i},$$

5. TRASFORMATION DIAGNOSTIC IN POISSON MODEL.

The aim of diagnostic in regression analysis is to appropriateness of the assumptions made in fitting a regression model to the data. Lawrence(1988) obtains diagnostics for the estimated regression parameter of the Box and Cox transformation of the response variable in the linear model:

Let $Y^{(\lambda)}$ (Poisson model defined in section1) be :

$$f(y;\theta;\phi) = \exp\{(y \ln \lambda - \lambda) - \ln y!\}\ y = 0,1,2,...$$

$$z^{(\lambda)} = y^{(\lambda)} / J(\lambda)^{1/n}, \text{ where}$$

$$J(\lambda) = \prod_{i=1}^{n} dy_i^{(\lambda)} / dy_i,$$

and we define z' as the derivate of $z^{(\lambda)}$ with respect to λ.

For the Box-Cox power family transformation we have:

$$y^{(\lambda)} = \frac{y^{\lambda} - 1}{\lambda} \text{ and } z^{(\lambda)} = \frac{y^{\lambda} - 1}{\dot{y}^{\lambda-1}\lambda}, \text{ where } \dot{y} \text{ is the geometric mean of } (y_1, y_2, \ldots, y_n).$$

The variance in the transformed model is slightly perturbed, in particular, it provides to determining those cases of the data that have strongest general influence on the estimated transformations parameters.

Now the variance of the model become:

$$\text{Var}(e) = \sigma^2 W^{-1},$$

where W is a diagonal matrix of perturbations.

Lawrence proposes a measure to assess the effect of joint perturbations on the data cases:

$$LD^* = l_{max\,i} = r_i + r'_i / \sum [\{r_j r'_j\}^2]^{1/2}$$

i= 1,...,n; where r_i and r_i' are the ith residuals from the regression of z and z' on the column of X, respectively.

The diagnostic then arises from local changes to the transformation parameter estimate caused by small perturbations; the case direction in which small perturbations have the greatest effect is the main diagnostic quantity.

6. COMPARISIONS OF LOCAL INFLUENCE MEASURES WITH A POISSON DATA.

We now compare the various local influence in the contex of a Poisson data set.

The data arise from a 5*4 factorial design (Maxwell 1961) with four replications at each factor level. The responses are the number of boys with disturbed dreams, and the factors are Rating (four levels) and Age groups (five levels).

The following model is fit to the data.

$$\text{Ln Dream} = \beta_0 + \beta_1 \text{ Rating} + \beta_2 \text{ Age} + \epsilon,$$

we have n=20 and k=3 in these data. On one hand, we apply LD,LD^*,LD^{**} and the five largest values are in Table 1.

Table 1. **Maxwell(1961) data: Five largest values based on LD(Cook's distance), and LD^* (Lawrance's measure), LD^{**} (New measure)**

Case	LD	Case	LD^*	Case	LD^{**}
20	4.031	15	0.379	20	19.927
15	1.920	20	0.339	19	17.378
2	1.649	2	0.3153	18	15.152
16	1.486	19	0.304	15	13.982
19	1.078	10	0.301	17	13.392

As can be seen from Table1, observations 20 is more local influential according to LD and LD^{**}.

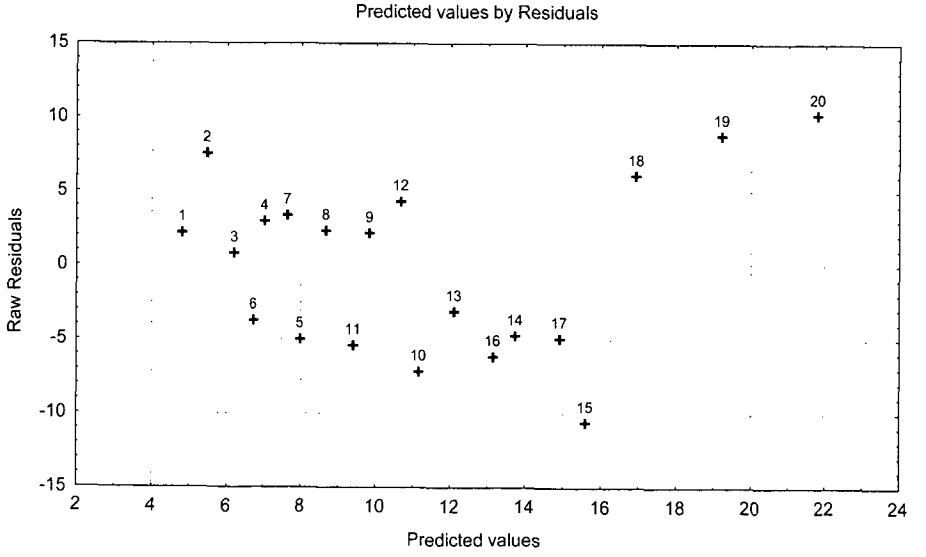


FIG. 1: A plot of Residuals versus Predicted values.

Cases 18,19,20 seem out of line with the rest of the plot.As we can see that LD, LD^* provide a useful method for investigating influence but they have some therorical and practical difficulties like swamping and masking effects.As we can observe in table 1 they are no resistant to these effects.Also,we can observe that the stranges

observations (cases:20,18,19) are detected by LD^{**} and this measure hasn't these problems .

REFERENCES

BILLOR and LOYNES, R. M. (1993), Local influence:A new approach, Communication.Stat.Theory Methods,22,1595-1611.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations,J_R.Stat Soc B,26,211-246.

COOK, R. D. (1977). Detection of influential observations in linear regression.Technometrics,19,15-18.

COOK, R. D. (1986).Assement of local influence,Journal of the Royal Statistical Society SeriesB-Methodological,48,133-169.

HADI, A. S. (1992).A new measure of overall potential influence in linear regression.computational Statistics&Data Analysis 14,1-27.

LAWRENCE, A. J. (1988),Regression transformation diagnostic using local influence, Journal of the American Statistical Association,86,1067-1072.

MAXWELL, J. A. (1961).Analysing qualitative data.London:Methuen.

MACCULLAGH, P., and NELDER, J. A. (1989),Generalized linar models.London:Chapman and Hall.J.A

NELDER, J. A. and WEDDERBURN, W. M. (1972).Generalized linear model, Journal of the Royal Statistical Society,A135,370-384.

PEÑA, D. and YOHAI, V. J. (1995).The detection of influential subsets in linear regression by using an influence matriz, Journal of the Royal Statistical Society SeriesB-Methodological,57,145-156.

SUÁREZ, M. .M. and GONZÁLEZ, M. A. (1996).Medidas basadas en influencia local.Cuadernos de Bioestadística y sus aplicaciones Informáticas, vol 14,5-17.

SUÁREZ, M. .M. and GONZÁLEZ, M. A. (2000).A connection between local and deletion influence, Sankhya:The Indian Journal of Statistics 2000, 62,series A,pt1,144-149.

SUÁREZ, M. .M. and GONZÁLEZ, M. A. (2000).Local and deletion diagnostic, Test, vol 9, No 2,345-352.

THOMAS, W and COOK, R. D. (1989). Assesing influence on regression-coefficients in generalized linear models, Biometrika,76,741-749.

REVISTA DE ESTATÍSTICA
57ª PÁGINA
3º QUADRIMESTRE DE 2001

Estimation of the False Alarm Probability based on the Successive Observations of the System

Autores:
Caballero-Águila, Raquel,
Hermoso-Carazo, Aurora
e
Linares-Pérez, Josefa

REVISTA DE ESTATÍSTICA
59ª PÁGINA
3º QUADRIMESTRE DE 2001

ESTIMATION OF THE FALSE ALARM PROBABILITY BASED ON THE SUCCESSIVE OBSERVATIONS OF THE SYSTEM

ESTIMAÇÃO DA PROBABILIDADE DE FALSO ALARME A PARTIR DE OBSERVAÇÕES SUCESSIVAS DO SISTEMA

Autores: Caballero-Águila, Raquel
- Universidad de Jaén. Dpto. de Estadística e I.O. Paraje Las Lagunillas s/n. 23071 Jaén, Spain

Hermoso-Carazo, Aurora
- Universidad de Granada. Dpto. de Estadística e I.O. Campus Fuentenueva s/n. 18071 Granada, Spain

Linares-Pérez, Josefa
- Universidad de Granada. Dpto. de Estadística e I.O. Campus Fuentenueva s/n. 18071 Granada, Spain

ABSTRACT:

- In this paper, we propose a recursive algorithm to obtain estimations of the false alarm probability in discrete linear systems with uncertain observations. We use a Bayesian approach, by specifying a Beta prior distribution for the unknown parameter and considering a quadratic loss function. By means of successive approximations of mixture distributions, we obtain a recursive algorithm which provides approximations for the Bayes estimators of the false alarm probability.

KEY-WORDS:

- *Uncertain Observations. False Alarm Probability.*

RESUMO:

- Neste artigo, propomos um algoritmo que estima de forma recursiva a probabilidade de falso alarme em sistemas lineares discretos com observações incertas. Utilizamos uma abordagem Bayesiana, em que se especifica uma distribuição *a priori* Beta para o parâmetro desconhecido e se considera uma função perda quadrática. A estimação é feita através de aproximações sucessivas de distribuições mistura, obtendo-se aproximações para os estimadores de Bayes para a probabilidade de falso alarme.

PALAVRAS-CHAVE:

- *Observações Incertas, Probabilidade de falso alarme.*

1. INTRODUCTION

In many practical situations, such as communication systems, there may be a nonzero probability (*false alarm probability*) that any observation consists of noise alone; this may be caused by an intermittent failure in the observation mechanisms. These situations can be described by an observation equation including not only an additive noise, but also a multiplicative noise component, modelled by a sequence of Bernoulli random variables; such systems are called *Systems with Uncertain Observations.*

The linear state estimation problem in linear systems with uncertain observations, under the hypotheses of mutual independence of the noises and the initial state and independence of the Bernoulli random variables, was treated by Nahi (1969). Later on, García-Ligero et al. (1997) and Caballero et al. (2000) obtained the quadratic and polynomial filters, respectively. In all these works it is assumed that, at any time, the false alarm probability or, equivalently, the probability that the signal exists in the observations, is known.

In this paper, we consider a linear discrete-time system with uncertain observations in which the uncertainties are governed by a sequence of independent Bernoulli random variables, and we assume that the probability that the observations contain the state process is unknown, but fixed throughout the time. We assume that the initial state and the additive noises of the system are gaussian and, also, that the initial state and all the noises are mutually independent.

Our aim is to obtain a recursive algorithm to estimate the false alarm probability, based on the successive observations of the system. For our purpose, we use a Bayesian approach; due to an ever-increasing computational complexity as a result of the uncertainty in the observations, the Bayes estimators of the probability are unfeasible in practice. For this reason, it becomes necessary to find approximations which are viable from a computational viewpoint.

Firstly, by using successive approximations of gaussian mixtures, we propose a method to compute approximations for the posterior densities of the unknown probability, given the observations. These approximations have also a mixture form and, hence, their computation involves an additional complexity which depends on the selected prior density. Then, we consider a Beta as the prior density and, by means of the new approximations of mixture distributions, we propose a recursive algorithm which allows us to obtain estimations of the unknown probability.

The proposed estimators of the unknown false alarm probability can be used for adapting the linear, quadratic and polynomial filtering algorithms established in Nahi (1969), Garcia-Ligero et al. (1997) and Caballero et al. (2000), respectively.

2. PROBLEM STATEMENT

Consider a discrete-time linear system with uncertain observations

$$x_{k+1} = A_k x_k + w_k, \quad k \geq 0$$

$$z_k = u_k C_k x_k + v_k, \quad k \geq 0$$

where x_k is the $n \times 1$ state vector, z_k is the $m \times 1$ observation vector at time k, and A_k, C_k are known matrices of appropriate dimensions.

The initial state, x_0, and the additive and multiplicative noises, $\{w_k;\ k \geq 0\}$ $\{v_k;\ k \geq 0\}$ and $\{u_k;\ k \geq 0\}$, satisfy:

- x_0 is a gaussian vector with zero mean and covariance matrix Σ_0.
- $\{w_k; k \geq 0\}$ is a gaussian white sequence with zero means and covariance matrices Q_k.
- $\{v_k; k \geq 0\}$ is a gaussian white sequence with zero means and covariance matrices R_k.
- $\{u_k; k \geq 0\}$ is a sequence of independent Bernoulli random variables with $P[u_k = 1] = p$, for all $k \geq 0$, being p an unknown parameter.
- The initial state x_0 and the noises $\{w_k;\ k \geq 0\}$, $\{v_k;\ k \geq 0\}$ and $\{u_k;\ k \geq 0\}$ are mutually independent.

Our aim is to obtain estimators for the probability p, based on the successive observations, $z_0, \ldots, z_k$, of the system, that can be obtained recursively. For this purpose, we use a Bayesian approach and, so, the problem is to obtain $\overline{p}_k = E\{p / Z^k\}$, the Bayes estimator of the probability p given the observations $Z^k = \{z_0, \ldots, z_k\}$, for a specific prior density and assuming a quadratic loss function.

To solve this problem we need to obtain the posterior density given the observations, for the selected prior density. Denoting by $f(p/Z^{-1})$ the prior density for p, the posterior density, $f(p/Z^k)$, can be obtained from the Bayes theorem, and it becomes

$$f(p/Z^k)=\frac{f(z_k/p,Z^{k-1})f(p/Z^{k-1})}{\int f(z_k/p,Z^{k-1})f(p/Z^{k-1})dp},\quad k\geq 0$$

where

$$f(z_k/p,Z^{k-1})=pf_1(z_k/Z^{k-1})+(1-p)f_0(z_k/Z^{k-1})$$

with $f_i(z_k/Z^{k-1})=f(z_k/u_k=i,Z^{k-1})$ and $f_i(z_0/Z^{-1})=f(z_0/u_0=i),\ i=0,1.$

Accordingly, the computation of the posterior densities $f(p/Z^k)$ requires that the densities $f_i(z_k/Z^{k-1})$, for $i=0,1$, should be calculated. From the independence hypotheses on the system, the density $f_0(z_k/Z^{k-1})$ agrees with that of the observation noise vector v_k, that is, it is the density of the gaussian distribution $N(0,R_k)$. However, as a result of the uncertainty in the observations, the determination of $f_1(z_k/Z^{k-1})$ is not simple and its computation grows in complexity as k increases. To avoid this difficulty, it seems natural to consider approximations for these densities. So, approximations for the posterior densities and, consequently, for the Bayes estimators of the parameter p are obtained.

We propose to approach this problem by approximating mixtures of gaussian distributions by gaussian distributions with their corresponding parameters. This task will be carried out in the next section.

3. ESTIMATION OF THE FALSE ALARM PROBABILITY

As we have commented in the above section, our first aim will be to obtain approximations, $\tilde{f}_1(z_k/Z^{k-1})$, which avoid the ever-increasing computational complexity of the density $f_1(z_k/Z^{k-1})$. These approximations will provide, in their turn, approximations for the posterior density, $\tilde{f}(p/Z^k)$, and for the Bayes estimators of p, $\tilde{p}_k=\int p\tilde{f}(p/Z^k)dp$.

At the first step, since $f_1(z_0/Z^{-1})$ corresponds to the gaussian distribution $N(0,C_0\Sigma_0C_0^T+R_0)$, the density $f(p/Z^0)$ can be easily computed, providing the estimator $\overline{p}_0$.

Next, we observe that $f_1(z_1 / Z^0)$ is determined by $f(x_0, w_0, z_0 / Z^{-1})$ and

$$f(x_0, w_0, z_0 / Z^{-1}) = \overline{p}_{-1} f_1(x_0, w_0, z_0 / Z^{-1}) + (1 - \overline{p}_{-1}) f_0(x_0, w_0, z_0 / Z^{-1})$$

where $f_i(x_0, w_0, z_0 / Z^{-1}) = f(x_0, w_0, z_0 / u_0 = i)$, for $i = 0, 1$, is the density of the gaussian distribution

$$N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_0 & 0 & i\Sigma_0 C_0^T \\ 0 & Q_0 & 0 \\ iC_0\Sigma_0 & 0 & iC_0\Sigma_0 C_0^T + R_0 \end{pmatrix}\right).$$

Then, we approximate $f(x_0, w_0, z_0 / Z^{-1})$ by the density of the gaussian distribution

$$N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_0 & 0 & \overline{p}_{-1}\Sigma_0 C_0^T \\ 0 & Q_0 & 0 \\ \overline{p}_{-1}C_0\Sigma_0 & 0 & \overline{p}_{-1}C_0\Sigma_0 C_0^T + R_0 \end{pmatrix}\right).$$

From this distribution, using the system equations, we obtain that $\tilde{f}_1(z_1 / Z^0)$ is the density of the distribution $N(C_1\hat{x}_{1/0},\ C_1\Sigma_{1/0}C_1^T + R_1)$, where

$$\begin{aligned} \hat{x}_{1/0} &= A_0 K_0 z_0 \\ K_0 &= \overline{p}_{-1}\Sigma_0 C_0^T \Pi_0^- \\ \Pi_0 &= \overline{p}_{-1} C_0 \Sigma_0 C_0^T + R_0 \end{aligned}$$

and

$$\Sigma_{1/0} = A_0 \Sigma_{0/0} A_0^T + Q_0$$

$$\Sigma_{0/0} = \Sigma_0 - K_0 \Pi_0 K_0^T.$$

Finally, the replacement of $\tilde{f}_1(z_1 / Z^0)$ in the expression of $f(p / Z^1)$, together with $f(p / Z^0)$, provides an approximation, $\tilde{f}(p / Z^1)$, and, from it, we obtain $\tilde{p}_1$.

The proposed procedure provides a recursive method for obtaining the densities $\tilde{f}_1(z_{k+1}/Z^k)$. In fact, let us assume that, for an arbitrary $k \geq 1$, the following approximation holds

$$f(x_k, w_k, z_k / Z^{k-1}) = \tilde{p}_{k-1} f_1(x_k, w_k, z_k / Z^{k-1}) + (1 - \tilde{p}_{k-1}) f_0(x_k, w_k, z_k / Z^{k-1})$$

where $f_i(x_k, w_k, z_k / Z^{k-1}) = f(x_k, w_k, z_k / u_k = i, Z^{k-1})$, for $i = 0, 1$, is the density of the gaussian distribution

$$N\left(\begin{pmatrix} \hat{x}_{k/k-1} \\ 0 \\ iC_k\hat{x}_{k/k-1} \end{pmatrix}, \begin{pmatrix} \Sigma_{k/k-1} & 0 & i\Sigma_{k/k-1}C_k^T \\ 0 & Q_k & 0 \\ iC_k\Sigma_{k/k-1} & 0 & iC_k\Sigma_{k/k-1}C_k^T + R_k \end{pmatrix}\right).$$

As in the first step, we approximate this mixture by the density of the gaussian distribution

$$N\left(\begin{pmatrix} \hat{x}_{k/k-1} \\ 0 \\ \tilde{p}_{k-1}C_k\hat{x}_{k/k-1} \end{pmatrix}, \begin{pmatrix} \Sigma_{k/k-1} & 0 & \tilde{p}_{k-1}\Sigma_{k/k-1}C_k^T \\ 0 & Q_k & 0 \\ \tilde{p}_{k-1}C_k\Sigma_{k/k-1} & 0 & \Pi_k \end{pmatrix}\right)$$

where $\Pi_k = \tilde{p}_{k-1}C_k\Sigma_{k/k-1}C_k^T + R_k + \tilde{p}_{k-1}(1 - \tilde{p}_{k-1})C_k\hat{x}_{k/k-1}\hat{x}_{k/k-1}^T C_k^T$.

As a consequence, $\tilde{f}_1(z_{k+1}/Z^k)$ will be the density of the gaussian distribution $N(C_{k+1}\hat{x}_{k+1/k},\ C_{k+1}\Sigma_{k+1/k}C_{k+1}^T + R_{k+1})$, where

$$\hat{x}_{k+1/k} = A_k\hat{x}_{k/k-1} + A_kK_k[z_k - \tilde{p}_{k-1}C_k\hat{x}_{k/k-1}]$$
$$K_k = \tilde{p}_{k-1}\Sigma_{k/k-1}C_k^T\Pi_k^-$$

and

$$\Sigma_{k+1/k} = A_k\Sigma_{k/k}A_k^T + Q_k$$
$$\Sigma_{k/k} = \Sigma_{k/k-1} - K_k\Pi_kK_k^T.$$

The approximation $\tilde{f}_1(z_{k+1}/Z^k)$ provides $\tilde{f}(p/Z^{k+1})$ and, from this, we obtain the estimators $\tilde{p}_{k+1}$.

This procedure provides a method for the computation of $\tilde{f}(p/Z^k)$ and the estimators $\tilde{p}_k$. However, its application presents an additional difficulty, due to the fact that this posterior density also has a mixture form. Obviously, the difficulty at the computation depends on the selected prior distribution. In the following section, we consider a *Beta* as the prior distribution, and we propose a new approximation for the Bayes estimators of p.

4. ESTIMATORS APPROXIMATION

Since p is the parameter of the Bernoulli random variables $\{u_k;\ k \geq 0\}$, and the Beta family is conjugated for the sampling of that distribution, let us specify $f(p/Z^{-1}) \equiv \beta(\alpha_0, \beta_0)$. Then $\bar{p}_{-1} = \alpha_0(\alpha_0 + \beta_0)^{-1}$, and

$$f(p/Z^0) = \delta_0 \beta(\alpha_0 + 1, \beta_0) + (1 - \delta_0)\beta(\alpha_0, \beta_0 + 1)$$

where $\delta_0 = \dfrac{\bar{p}_{-1} f_1(z_0/Z^{-1})}{\bar{p}_{-1} f_1(z_0/Z^{-1}) + (1-\bar{p}_{-1}) f_0(z_0/Z^{-1})}$.

So, as a result of the mixture form of $f(p/Z^0)$, and since each posterior density is mixture of two distributions, $\tilde{f}(p/Z^k)$ will be a mixture of 2^{k+1} distributions.

In order to avoid the computational complexity, we propose to approximate $f(p/Z^0)$ by $\hat{f}(p/Z^0)$, the density of the distribution $\beta(\alpha_0 + \delta_0, \beta_0 + 1 - \delta_0)$, and for the subsequent steps we proceed in an analogous way. So, if

$$\hat{f}(p/Z^{k-1}) = \beta\left(\alpha_0 + \sum_{i=0}^{k-1} \hat{\delta}_i, \beta_0 + \sum_{i=0}^{k-1}\left(1 - \hat{\delta}_i\right)\right)$$

where $\hat{\delta}_i = \dfrac{\hat{p}_{i-1}\hat{f}_1(z_i/Z^{i-1})}{\hat{p}_{i-1}\hat{f}_1(z_i/Z^{i-1}) + (1-\hat{p}_{i-1}) f_0(z_i/Z^{i-1})}$, the posterior distribution of p given Z^k will be mixture of two Beta distributions, with mixture parameter $\hat{\delta}_k$, and will be approximated by the distribution

$$\beta\left(\alpha_0 + \sum_{i=0}^{k} \hat{\delta}_i, \beta_0 + \sum_{i=0}^{k}\left(1 - \hat{\delta}_i\right)\right).$$

The main advantage of these approximations is that the estimators can be obtained by the following recursive relation

$$\hat{p}_k = \hat{p}_{k-1} - \frac{1}{\alpha_0 + \beta_0 + k + 1}\left[\hat{p}_{k-1} - \hat{\delta}_k\right], \; k \geq 0$$

$$\hat{p}_{-1} = \frac{\alpha_0}{\alpha_0 + \beta_0}.$$

5. NUMERICAL EXAMPLE

To test the effectiveness of the proposed estimators, we have considered the following scalar system

$$x_{k+1} = 0.5x_k + w_k, \; k \geq 0$$
$$z_k = u_k x_k + v_k, \; k \geq 0$$

where the initial state, x_0, is a random variable with distribution $N(0,1)$, and the noises $\{w_k;\ k \geq 0\}$ and $\{v_k;\ k \geq 0\}$ are gaussian white sequences with zero means and variances $E\{w_k^2\} = E\{v_k^2\} = 19/3$. The multiplicative noise $\{u_k;\ k \geq 0\}$ is a sequence of independent Bernoulli variables with unknown parameter p.

We have obtained numerical simulations for the observations of this system, considering different values of the parameter p ($p = 0.25$, $p = 0.5$ and $p = 0.75$) and, in each case, we have performed one hundred iterations of the proposed algorithm by assuming as prior distribution a Beta, $\beta\left(\sqrt{1+19/3}, \sqrt{19/3}\right)$; the parameters of this prior distribution specify the standard deviation of the first observation when the false alarm probability is zero and one, respectively.

The successive estimations of p, obtained by using the observations simulated with each value of the false alarm probability, are displayed in the below table and figures. A slow but clear decreasing and increasing tendency of the estimations can be noticed in the extreme cases, $p = 0.25$ and $p = 0.75$, respectively. In the case $p = 0.5$, since the prior estimation is very close to this value, we observe that the estimations are stabilised about it.

$p = 0.25$		$p = 0.5$		$p = 0.75$	
0.51831723324638	0.42251511734729	0.51831723324638	0.50839443315680	0.51831723324638	0.61270200596641
0.51692290437721	0.42123250358735	0.51624961308656	0.50881219639288	0.51780346913122	0.61255257876996
0.51008203131525	0.42067813657507	0.50790711705153	0.50746135975527	0.57376233220683	0.61854195262212
0.49792392924330	0.42059293230680	0.49635647191756	0.50576474230569	0.55830090888667	0.62076022231370
0.48977821354975	0.42003494831626	0.48609452342568	0.50421881359340	0.55280550095145	0.61872533935391
0.48294188029693	0.41851909475381	0.47725348546051	0.50328468346594	0.56874950435060	0.61721285895589
0.47418586368559	0.42137809765677	0.47040435212003	0.50309212789447	0.56020695149390	0.61621926555911
0.47084120586004	0.42241612164985	0.47476570826775	0.50196950260317	0.55204606851969	0.61668642159040
0.46703800606882	0.42163454979915	0.46727768368744	0.50453183456854	0.54701650238569	0.61767842465579
0.46099993258587	0.42441856302212	0.46089249551017	0.50646294790495	0.54049551891840	0.61597573738515
0.46107784505358	0.42319650016718	0.46641594336362	0.50543740162250	0.54663850631360	0.61480463031140
0.45472842375341	0.42171651899885	0.46605779494621	0.50449306147154	0.56270817235801	0.61488946941091
0.46955780885823	0.42032478224086	0.48127877160113	0.50327494332868	0.55853522670033	0.61456731903296
0.46317466717575	0.41897627158681	0.47522552414406	0.50186756253584	0.56564225256456	0.61470061295716
0.45796823523771	0.41786976437692	0.47011565780925	0.50160779961060	0.57599774731777	0.61348414584386
0.46199153354518	0.41863358999901	0.46886835871281	0.50071859294673	0.58469305260024	0.61213360387475
0.45736902312049	0.41765313244278	0.46405540469460	0.50112565888034	0.58438413276868	0.61138454485833
0.45352496399865	0.41643917322505	0.45960355419658	0.50208466880606	0.58851501593358	0.61017427956407
0.45226902640002	0.41576622143065	0.45557175786672	0.50056858441507	0.58584963283698	0.61257072600714
0.44815155826277	0.42115710574440	0.45199695325787	0.50238433665269	0.59307360811065	0.61095565602835
0.44654690067308	0.42020790947848	0.46194253289223	0.50090055646227	0.59762970227892	0.60986274975476
0.44582606943595	0.41917971201091	0.46082917707688	0.49969650500197	0.60460406963169	0.60983527075147
0.44250351452456	0.41845648688386	0.48035411576262	0.49887014604613	0.60603871931980	0.60887998250087
0.43905692380759	0.41724143885485	0.47589204492577	0.49984751318801	0.60825022581519	0.60944778589369
0.43942617880233	0.41625511874150	0.47941416453953	0.49973114043850	0.60500353419828	0.61027094601079
0.43797274378768	0.42031315704781	0.47588312075814	0.49846933766466	0.61421042422349	0.61050294881053
0.43483985819290	0.41916582236500	0.48041776533840	0.49772900506546	0.62385703099807	0.61110230710716
0.43183861756302	0.41854231044352	0.47720970078000	0.49752264538021	0.62444651964067	0.61347090918771
0.43078886751338	0.41895171813775	0.48551738946555	0.49820647399445	0.62398040402280	0.61196680772215
0.42960746395327	0.42132040401052	0.48233856475159	0.50334908657268	0.62438833657372	0.61105790539918
0.42676231716834	0.42030220538901	0.47945579743599	0.50251814524632	0.62260359429456	0.60990614249121
0.42415405937514	0.41987540671512	0.49102973294518	0.50127955271572	0.62387691282616	0.61016437140367
0.42164440242958	0.41927982374542	0.50361296866707	0.50020913912989	0.62129214720905	0.61008743230091
0.42096808050841	0.41914710004728	0.51164214977775	0.49915619699071	0.61880303725354	0.61443360282785
0.42118136551663	0.41817617569893	0.50841302267218	0.49916769902649	0.61637775459192	0.61667971271557
0.42178006540035	0.41838206124766	0.51327852682774	0.49844655227867	0.61474547676763	0.61942384653282
0.42183229484749	0.41737067649262	0.51048666936469	0.50067435077089	0.61305369487088	0.61882321966852
0.41983587045855	0.41643181688604	0.50952907178538	0.49982395153334	0.61155840092122	0.61791817230560
0.41829142946926	0.41605100798565	0.50885414141747	0.49933598718989	0.61752688293941	0.61705231344732
0.41694097593400	0.41691107656105	0.50646092083373	0.49825175922758	0.61813380060184	0.61612449467778
0.41906590950883	0.41947601859812	0.50431026074839	0.49756009255887	0.61634372027683	0.61780790385322
0.41793395897366	0.41828570969286	0.50628176563788	0.49778010837073	0.61466947929366	0.61669583416555
0.41593875592986	0.41771687302669	0.50472521056793	0.49719737098230	0.61281501490532	0.61643038684791
0.41486127448008	0.41769621688851	0.50970265338092	0.49705731813328	0.61092169420173	0.61754234657963
0.41800730267086	0.42030717424191	0.50773531588242	0.49744963824493	0.61597628271750	0.62054625633693
0.41980836763697	0.41993716379386	0.50825266187492	0.50067156353097	0.61392214854909	0.61934401777018
0.42728667798995	0.41891917157318	0.50670755978178	0.49943897568785	0.61284796987762	0.61846077869030
0.42538764641818	0.41846015479370	0.50701107913350	0.50076677831059	0.61187769334657	0.61763840968242
0.42347656534294	0.41829165612604	0.50708962191031	0.49973329741485	0.61211890302355	0.61878927763336
0.42444278421478	0.41734900510805	0.51218307264665	0.49894997686286	0.61206970263562	0.62039930264733
0.42333939490793	0.41726390448502	0.50991528635783	0.49800126976457	0.61065995693767	0.62250188891052

Table: Estimations of p with prior distribution $\beta\left(\sqrt{1+19/3}, \sqrt{19/3}\right)$

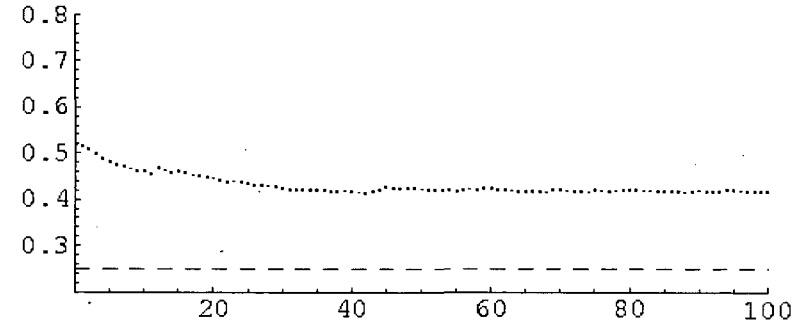


Figure 1. Estimations of $p = 0.25$

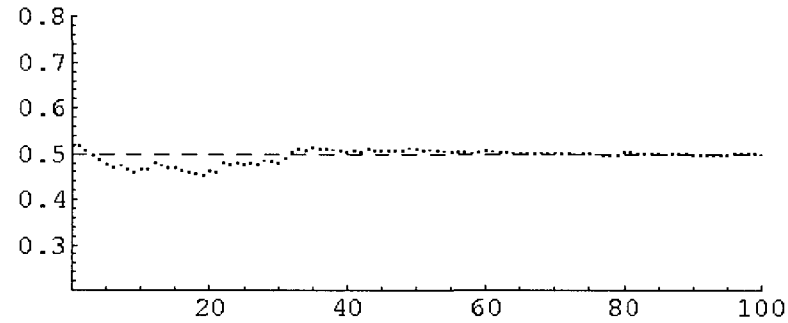


Figure 2. Estimations of $p = 0.5$

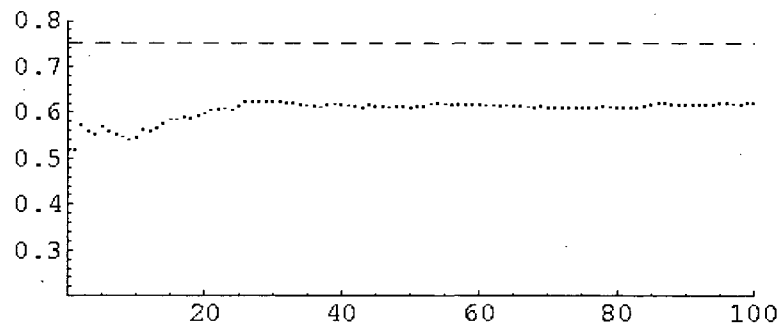


Figure 3. Estimations of $p = 0.75$

6. CONCLUSIONS

In this paper, we consider a system with uncertain observations in which the initial state and noises are mutually independent and the probability that the observations contain the state process is unknown, but fixed throughout the time. We propose a recursive estimation algorithm for that probability, based on the successive observations of the system.

The estimators obtained with this algorithm are approximations to the Bayes estimators of the false alarm probability when a Beta prior distribution is specified and a quadratic loss function is considered. These approximations are obtained by means of successive approximations of mixture distributions.

ACKNOWLEDGEMENTS

This work has been partially supported by the "Ministerio de Ciencia y Tecnología" under contract BFM2000-0602.

REFERENCES

CABALLERO, R., HERMOSO, A. and LINARES, J., *"Least Mean-Squared Error Polynomial Estimation in Systems with Uncertain Observations"*. Proceedings of the 2000 IEEE International Symposium on Information Theory, 110. Sorrento, Italy. 2000.

GARCÍA-LIGERO, M. J., HERMOSO, A. and LINARES, J., *"Second Order Polynomial Filtering for Discrete Systems with Uncertain Observation"*. Proceedings of the VIII International Symposium on Applied Stochastic Models and Data Analysis, 157-162. Anacapri, Italy. 1997.

NAHI, N. E., *"Optimal Recursive Estimation with Uncertain Observation"*. IEEE Transactions on Information Theory. IT-15, 457-462. 1969.

REVISTA DE ESTATÍSTICA
71ª PÁGINA
3º QUADRIMESTRE DE 2001

Aplicação de Métodos Estatísticos no Desporto: Análise do Campeonato de Futebol, em Portugal

Autor:
Paulo Almeida Pereira

REVISTA DE ESTATÍSTICA
73ª PÁGINA
3º QUADRIMESTRE DE 2001

APLICAÇÃO DE MÉTODOS ESTATÍSTICOS NO DESPORTO: ANÁLISE DO CAMPEONATO DE FUTEBOL, EM PORTUGAL

STATISTICAL METHODS APPLIED TO SPORTS: ANALISYS OF THE PORTUGUESE SOCCER CHAMPIONSHIP

Autor: Paulo Almeida Pereira
- Prof. Auxiliar da Universidade Católica Portuguesa, Pólo de Viseu Instituto Universitário de Desenvolvimento e Promoção Social

RESUMO:

- Métodos estatísticos, como os modelos de regressão e a análise factorial de componentes principais são aplicados ao estudo dos resultados do campeonato nacional de futebol. São desenvolvidos dois modelos, utilizando, como variável dependente, os resultados dos jogos e como variáveis independentes, dados estatísticos disponíveis para o comportamento das equipas intervenientes. Após a eliminação de *outliers* e a validação dos pressupostos dos modelos de regressão, calculam-se as estimativas dos resultados dos jogos e respectivos intervalos de confiança, a 95%, que originam classificações previstas pelos modelos, que são comparadas com os valores observados.

PALAVRAS-CHAVE:

- *Modelos de regressão, análise factorial, desporto, campeonato nacional de futebol, estimativa das classificações.*

ABSTRACT:

- Statistical methods, such as regression models and factorial analysis of principal components are applied to the study of portuguese soccer championship results. Two models are developed, using, as dependent variable, the results of soccer games and as independent variables, available statistical data for the behavior of the intervening teams. After outliers elimination and validation of the regression models assumptions, the estimation of results for the soccer games and 95% confidence intervals are calculated, which originate classification predictions by the models that are compared with the observed values.

KEY-WORDS:

- *Regression models, factorial analysis, sports, portuguese soccer championship, classification estimation.*

1. INTRODUÇÃO

O desporto gira em torno de estatísticas, tendo sempre por base um sistema numérico de resultados, por exemplo, os golos num jogo de futebol. Os acontecimentos desportivos podem, assim, constituir um tópico de investigação estatística, o que vem a acontecer há, pelo menos, três décadas (Machol *et al.*, 1976; Ladany e Machol, 1977). Devido à sua relevância, a *American Statistical Association* criou, em 1992, uma secção denominada *Statistics in Sports*.

Existem publicações, nesta área, dispersas por vários jornais de estatística, tendo surgido, recentemente, uma importante compilação de aplicações, na área da estatística, a acontecimentos desportivos (Bennett, 1998), em que são apresentados trabalhos sobre várias modalidades desportivas: Futebol Americano, Basebol, Basquetebol, Futebol, Golfe, Hóquei no Gelo, Ténis e Atletismo.

As aplicações estatísticas a eventos desportivos são também uma poderosa ferramenta para cativar a atenção dos estudantes: pela familiaridade com os desportos, o contexto da aplicação é mais facilmente compreendido, podendo ser dado o devido ênfase à análise estatística e à sua percepção crítica.

Com o desenvolvimento das tecnologias e sistemas de informação, temos presentemente uma grande disponibilidade de dados, nomeadamente estatísticos. A *Infordesporto* (Infordesporto, *site* na *Internet*) é uma dessas bases de dados, que disponibiliza uma informação completa sobre várias estatísticas relativas aos jogos de futebol do campeonato nacional, em Portugal, desde a época de 1998/99.

A análise de regressão, sendo um método estatístico que utiliza a relação entre duas ou mais variáveis quantitativas, permite estimar uma variável a partir de outras. O Modelo de Regressão Linear, discutido por alguns autores – indicam-se dois, a título de exemplo (Neter *et al.*, 1996; Draper e Smith, 1981) –, é um meio formal de exprimir essa relação estatística entre variáveis independentes, de previsão ou explicativas e uma variável dependente, de resposta ou explicada. Encontram-se algumas aplicações destes modelos a competições desportivas em publicações recentes (Smith, 1999; Szymanski e Smith, 1997; Gray, 1997; Kahane, 1997; Hamilton, 1997) que, usualmente, se cingem à análise da influência de algumas, poucas, variáveis nos resultados desportivos.

Neste trabalho, pretende aplicar-se o modelo de regressão aos campeonatos de futebol, em Portugal, de 1998/1999 e 1999/2000, de modo a estudar a relação entre os dados estatísticos para os jogos e os resultados obtidos pelas equipas intervenientes. Esta análise passa pela descrição do desenvolvimento prático de um modelo de regressão, permitindo exemplificar e detalhar os passos conducentes à validação e aplicação do modelo. O modelo permite desenvolver previsões sobre qual deveria ter sido o resultado dos jogos, de acordo com as estatísticas disponíveis, e compará-lo com o resultado efectivamente observado, originando classificações previstas pelo modelo.

A manipulação de grandes quantidades de dados estatísticos só é possível com adequados meios informáticos. O *software* utilizado neste estudo é o *SPSS - Statistical Package for Social Sciences*, para o qual se refere um texto de apoio (Pestana e Gageiro, 2000), entre vários existentes.

2. DADOS ESTATÍSTICOS

A *Infordesporto* disponibiliza uma informação estatística bastante completa sobre os jogos do campeonato nacional de futebol da I divisão, desde a época de 1998/99. Para aplicação e desenvolvimento de um modelo de regressão é necessário, em primeiro lugar, sistematizar os dados disponíveis, tendo por objectivo relacionar o resultado de um jogo de futebol (variável dependente) com o comportamento das duas equipas antagonistas, descrito pelas denominadas variáveis independentes: a informação estatística disponível para os intervenientes no jogo. No Quadro n.º 1, introduz-se a lista das estatísticas disponíveis para cada jogo, agrupadas em categorias e representadas por abreviaturas, siglas essas que serão utilizadas ao longo do texto.

Quadro n.º 1. Estatísticas disponíveis para cada jogo

	Gerais:		Guarda-redes:
RELV	estado do relvado [1]	DC	defesas completas
ESPEC	n.º de espectadores	DI	defesas incompletas
TJOGO	tempo jogado (% do tempo total)	SC	saídas completas
	Tempos de posse de bola (min.):	SI	saídas incompletas
DEF	na defesa		*Jogadores:*
MCD	no meio campo defensivo	FC	faltas cometidas
MCO	no meio campo ofensivo	FS	faltas sofridas
ATA	no ataque	P	perdas de bola
TOT	total	R	recuperações de bola
POSSE	posse de bola (% do total)	AT	Ataques
	Disciplinares:	CR	Cruzamentos
CA	cartões amarelos	RE	Remates
CV	cartões vermelhos	AS	Assistências

[1] Utilizou-se a seguinte escala ordinal: 1-mau, 2-razoável, 3-bom, 4-muito bom, 5-excelente.

Destas várias estatísticas analisadas num jogo de futebol, as estatísticas gerais e a percentagem de posse de bola apresentam um valor para cada jogo; todas as restantes apresentam dois valores, o primeiro associado a uma das equipas e o segundo associado à outra: sendo precedidas pela letra "C", quando dizem respeito à equipa que joga em casa e pelo prefixo "F", quando associadas à equipa que actua fora do seu

terreno. Deste modo, o resultado de cada jogo está relacionado com estas 42 estatísticas, denominadas variáveis independentes, para efeito do modelo de regressão.

A análise reporta-se a duas épocas (1998/1999 e 1999/2000) do campeonato nacional de futebol, constituindo cada jogo uma observação que irá integrar o modelo de regressão, pelo que existem, no total, 612 observações destas 42 variáveis. O Modelo de Regressão relaciona, estatisticamente, uma variável dependente com variáveis independentes, exprimindo a tendência da primeira para variar com as segundas, de um modo sistemático.

3. MODELO DE REGRESSÃO

A hipótese formulada neste estudo consiste em estabelecer uma relação estatística, utilizando a análise de regressão, entre o resultado de um jogo de futebol e as estatísticas disponíveis para as variáveis que estão relacionadas com o comportamento das equipas intervenientes, de modo a servir dois objectivos:

- descrição, através do desenvolvimento de um modelo válido e utilizável para caracterizar a relação entre as variáveis;
- controlo, de modo a verificar se os resultados efectivamente obtidos são, ou não, diferentes dos previstos pelo modelo.

A fórmula geral da curva de regressão para o modelo é:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad i = 1, 2, ..., n \quad (1)$$

- n é o número de observações: corresponde aos 612 jogos, das épocas de 1998/1999 e 1999/2000, do campeonato nacional de futebol;
- Y_i é a variável dependente (i representa a observação i em n observações): serão testadas duas medidas do resultado de cada jogo de futebol: a primeira, denominada VED, consiste numa escala ordinal, em que 2 = vitória, 1 = empate e 0 = derrota, definida desta forma, para garantir a simetria da variável; a segunda, representada pela sigla GMGS, é dada pela diferença, entre Golos Marcados e Golos Sofridos, para a equipa que actua em casa;
- X_{p-1} são as variáveis independentes ou explicativas: correspondem às observações das 42 estatísticas apresentadas no Quadro n.º 1, para cada jogo;
- β_k são os parâmetros do modelo, medindo a variação média do valor esperado da variável dependente Y, com o aumento de uma unidade na variável associada, quando todas as outras variáveis explicativas no modelo permanecem constantes;
- ε_i é o termo de erro aleatório, representando as variáveis com poder explicativo sobre a variável dependente omitidas pelo modelo.

No modelo, uma componente da variação na variável dependente é explicada pelas variáveis independentes que o constituem, sendo dada pelo coeficiente de determinação (r^2), mas existe uma outra componente que reflecte a nossa ignorância referente a factores explicativos não contemplados.

4. DESENVOLVIMENTO DOS MODELOS DE REGRESSÃO

4.1. DADOS ESTATÍSTICOS: VARIÁVEIS DO MODELO

Os dados em análise constituem um estudo observacional explicativo. Uma regra empírica para situações deste tipo é que devem existir 6 a 10 observações por cada variável independente, o que sucede neste estudo.

4.1.1. SELECÇÃO DAS VARIÁVEIS INDEPENDENTES A INCLUIR NOS MODELOS

É também importante reduzir o número de variáveis a incluir no modelo final (Miller, 1990), devido à dificuldade em compreender um modelo com muitas variáveis e ao aumento na variabilidade associada aos coeficientes do modelo, provocado pela correlação entre as variáveis – o que sucede, normalmente, num estudo observacional explicativo –, que diminui as capacidades descritivas e de controlo do modelo (os valores estimados para a variável dependente apresentam grande variância).

Utilizam-se procedimentos para seleccionar e testar sub-grupos de variáveis independentes importantes, com o objectivo de desenvolver um modelo com um menor número de variáveis, escolhendo as que explicam "melhor" a variável dependente.

4.1.2. ANÁLISE FACTORIAL DE COMPONENTES PRINCIPAIS

No decorrer do estudo que permite a selecção das variáveis independentes relevantes para o modelo, verifica-se a existência de uma forte correlação entre algumas variáveis, o que impede a sua utilização simultânea nos modelos de regressão. Este facto, que ocorre para qualquer uma das variáveis dependentes (VED ou GMGS), obriga à selecção de apenas uma variável, entre as várias que estão correlacionadas entre si, para integrar o modelo de regressão, situação que originaria uma perda da informação, presente nas restantes variáveis eliminadas do modelo, que se podem considerar, à partida, importantes. Para ultrapassar esta contrariedade, recorreu-se à aplicação das técnicas de análise factorial de componentes principais, de modo a reduzir o número de variáveis, minimizando a perda de informação.

As variáveis com forte correlação, entre si, são as que representam as estatísticas de ataque de qualquer das equipa (tempos de posse de bola e recuperações,

ataques, remates e cruzamentos) e as disciplinares para as equipas que actuam fora do seu terreno (cartões amarelos e vermelhos).

De acordo com as correlações existentes entre variáveis, existem cinco grupos de variáveis, apresentados no Quadro n.º 2, aos quais será aplicada a análise factorial de componentes principais, de modo a transformar as variáveis iniciais (X_1, X_2, ..., X_p), correlacionadas entre si, num menor número de variáveis não correlacionadas (F_1, F_2, ..., F_p), designadas por factores principais ou, simplesmente, por factores. Cada variável é expressa como combinação linear dos factores que lhe são comuns:

$$\begin{aligned} X_1 &= b_{11}F_1 + b_{12}F_2 + \ldots + b_{1k}F_k + U_1 \\ X_2 &= b_{21}F_1 + b_{22}F_2 + \ldots + b_{2k}F_k + U_2 \\ &\ldots \\ X_p &= b_{p1}F_1 + b_{p2}F_2 + \ldots + b_{pk}F_k + U_p \end{aligned} \qquad (2)$$

Onde b_{ij} são os coeficientes, factores ou *loadings*, que correlacionam as variáveis originais com os factores comuns e U_i representam os factores únicos, ou seja, a parte de uma variável que não é explicada pelos factores comuns.

Os primeiro e segundo grupos de variáveis são constituídos pelas estatísticas de ataque (ataques, remates, cruzamentos e recuperações), para as equipas que jogam em casa e fora, respectivamente CATA e FATA; o terceiro e quarto grupos integram os tempos de posse de bola (no meio campo defensivo, no ataque e total), também para as equipas que actuam em casa – CTAT – e fora – FTAT –, respectivamente; o quinto grupo apresenta as estatísticas disciplinares (cartões amarelos e vermelhos), da equipa que actua fora de casa – FCAV.

Da análise do quadro, verifica-se que todas as variáveis de cada grupo apresentam correlações positivas entre si e, como dado complementar, deve ser referido que todas as correlações apresentadas são significativas, para um nível de significância de 1%.

Quadro n.º 2. Grupos de variáveis sujeitas a análise factorial: correlações entre as variáveis integrantes de cada grupo

Grupo 1 – **CATA**

	CAT	CRE	CCR	CR
CAT	1,00	0,59	0,74	0,35
CRE	0,59	1,00	0,53	0,18
CCR	0,74	0,53	1,00	0,28
CR	0,35	0,18	0,28	1,00

Grupo 2 – **FATA**

	FAT	FRE	FCR	FR
FAT	1,00	0,59	0,73	0,39
FRE	0,59	1,00	0,50	0,21
FCR	0,73	0,50	1,00	0,25
FR	0,39	0,21	0,25	1,00

Grupo 3 – **CTAT**

	CMCO	CATA	CTOT
CMCO	1,00	0,49	0,77
CATA	0,49	1,00	0,70
CTOT	0,77	0,70	1,00

Grupo 4 – **FTAT**

	FMCO	FATA	FTOT
FMCO	1,00	0,51	0,77
FATA	0,51	1,00	0,64
FTOT	0,77	0,64	1,00

Grupo 5 – **FCAV**

	FCA	FCV
FCA	1,00	0,52
FCV	0,52	1,00

De modo a prosseguir com a análise factorial, deve testar-se e rejeitar a hipótese da matriz de correlações ser a matriz identidade: utilizando o teste de esfericidade de Bartlett que, para todos os grupos de variáveis, apresenta um nível de significância de 0,000, pode rejeitar-se a hipótese testada, mostrando a existência de correlação entre as variáveis.

Uma medida da adequação dos dados (MAD) para a análise factorial de componentes principais consiste na estatística Kaiser-Meyer-Olkin (KMO), que compara as correlações simples com as parciais entre as variáveis, devendo os coeficientes de correlação parciais ser pequenos. Kaiser adjectiva os valores de KMO, de acordo com a adequação dos dados para a realização da análise factorial, como:

KMO	<0,5	0,5-0,6	0,6-0,7	0,7-0,8	0,8-0,9	1,0
MAD	Inaceitável	Má	Razoável	Média	Boa	Muito boa

No Quadro n.º 3 são apresentados os valores da KMO para os cinco grupos de variáveis, que sugerem dados adequados, de forma razoável ou média, para a aplicação da análise factorial, com excepção do grupo FCAV, em que essa medida é adjectivada como má, embora sendo ainda possível realizar a referida análise.

Quadro n.º 3. Adequação dos dados para a análise factorial

Grupos de variáveis	KMO	MAD
CATA	0,715	Média
FATA	0,698	Razoável
CTAT	0,614	Razoável
FTAT	0,660	Razoável
FCAV	0,500	Má

Em cada grupo de variáveis pode, ainda, ser analisada a adequação de cada variável para utilização da análise factorial, através dos valores das correlações da diagonal principal da matriz anti-imagem, que consistem numa medida da adequação dos dados (MAD), para cada variável, introduzida no Quadro n.º 4. Valores da MAD elevados, tal como os obtidos neste estudo, permitem concluir que é possível a realização da análise factorial.

Quadro n.º 4. Adequação das variáveis para a análise factorial

CATA		FATA		CTAT		FTAT		FCAV	
Variável	MAD	Variável	MAD	Variável	MAD	Variável	MAD	Variável	MAD
CAT	0,66	FAT	0,64	CMCD	0,63	FMCD	0,66	FCA	0,50
CRE	0,81	FRE	0,80	CATA	0,67	FATA	0,77	FCV	0,50
CCR	0,70	FCR	0,69	CTOT	0,57	FTOT	0,61		
CR	0,82	FR	0,74						

Depois de verificar a possibilidade de executar adequadamente a análise factorial, prossegue-se com a extracção dos factores a partir das variáveis para os cinco grupos definidos. O primeiro passo consiste, precisamente, na definição do número de componentes (ou factores) retidos, necessários para descrever os dados, através da aplicação de três regras:

- a proporção da variância explicada pelas componentes deve ser superior a 60%;
- a variância de cada componente (valores próprios ou *eigenvalues*) retida deve ser superior à unidade, ou seja, à variância média;
- no *screeplot* (gráfico da variância explicada em função das componentes), os pontos de maior declive correspondem às componentes retidas.

De acordo com os procedimentos referidos, como pode ser observado no Quadro n.º 5, foram retidos dois factores nos dois primeiros grupos de variáveis e apenas um factor nos restantes, assinalados a *negrito*.

Quadro n.º 5. Factores (componentes) retidos a partir das variáveis originais

CATA	Valores próprios		
Factor	Total	% de σ^2	% cum.
1	**2,40**	**60,0**	**60,0**
2	**0,86**	**21,6**	**81,6**
3	0,49	12,1	
4	0,25	6,3	

FATA	Valores próprios		
Factor	Total	% de σ^2	% cum.
1	**2,40**	**59,9**	**59,9**
2	**0,85**	**21,1**	**81,0**
3	0,51	12,8	
4	0,25	6,2	

CTAT	Valores próprios		
Factor	Total	% de σ^2	% cum.
1	**2,31**	**77,1**	**77,1**
2	0,52	17,2	
3	0,17	5,7	

FTAT	Valores próprios		
Factor	Total	% de σ^2	% cum.
1	**2,28**	**76,0**	**76,0**
2	0,51	16,9	
3	0,21	7,1	

FCAV	Valores próprios		
Factor	Total	% de σ^2	% cum.
1	**1,52**	**76,0**	**76,0**
2	0,48	24,0	

No quadro apresentam-se os valores próprios, a percentagem de variância associada a cada factor e a percentagem cumulativa de variância explicada pelas componentes retidas. A variância explicada pelos factores retidos, nos vários grupos de variáveis, varia entre 76% e 82%, valores que são bastante razoáveis para a aplicação de análise factorial.

No Quadro n.º 6 apresentam-se os valores das comunalidades, que correspondem à proporção da variância de cada variável explicada pelas componentes principais retidas. Todas as variáveis apresentam uma forte relação com os factores retidos, uma vez que os valores observados para as comunalidades são elevados, o menor valor é de 65%, sendo a maioria dos valores superiores a 75%.

Quadro n.º 6. Comunalidades (Com.) associadas a cada variável

CATA		FATA		CTAT		FTAT		FCAV	
Variável	Com.	Variável	Com.	Variável	Com.	Variável	Com.	Variável	Com.
CAT	0,82	FAT	0,83	CMCD	0,74	FMCD	0,77	FCA	0,76
CRE	0,70	FRE	0,68	CATA	0,68	FATA	0,65	FCV	0,76
CCR	0,77	FCR	0,75	CTOT	0,89	FTOT	0,86		
CR	0,98	FR	0,99						

Após a aplicação e validação de todos os passos da análise factorial de componentes principais, finalmente, procede-se à determinação da matriz dos componentes, que apresenta os coeficientes, factores ou *loadings*, que relacionam as variáveis com os factores retidos.

Quadro n.º 7. Matriz de componentes

	Factor CATA			Factor FATA			Factor			Factor			Factor
Var.	1	2	Var.	1	2	Var.	CTAT	Var.	FTAT	Var.	FCAV
CAT	0,903	-0,067	FAT	0,909	-0,048	CMCD	0,860	FMCD	0,876	FCA	0,872
CRE	0,770	-0,328	FRE	0,767	-0,297	CATA	0,825	FATA	0,809	FCV	0,872
CCR	0,866	-0,127	FCR	0,843	-0,200	CTOT	0,945	FTOT	0,926		
CR	0,493	0,858	FR	0,520	0,846						

Por exemplo, para o primeiro grupo de variáveis:

$$CAT = 0{,}903CATA_1 - 0{,}067CATA_2 \qquad CRE = 0{,}770CATA_1 - 0{,}328CATA_2 \quad (3)$$
$$CCR = 0{,}866CATA_1 - 0{,}127CATA_2 \qquad CR = 0{,}493CATA_1 + 0{,}858CATA_2$$

Os valores próprios das componentes retidas correspondem à soma dos quadrados dos coeficientes das variáveis, para cada factor (soma em coluna), enquanto que as comunalidades podem ser obtidas pela soma dos quadrados dos coeficientes dos factores, para cada variável (soma em linha).

Para os dois primeiros grupos de variáveis, em que foram retidos dois factores, procede-se ainda à rotação da matriz dos componentes, de modo a extremar os valores dos coeficientes, para associar, indubitavelmente, cada variável a apenas um factor. Nos grupos de variáveis CATA e FATA, embora já exista uma separação clara entre as variáveis associadas a cada factor, realizou-se a rotação da matriz de componentes, utilizando o método *Varimax*, com a normalização de Kaiser, obtendo-se, no Quadro n.º 8, a matriz de componentes, após rotação.

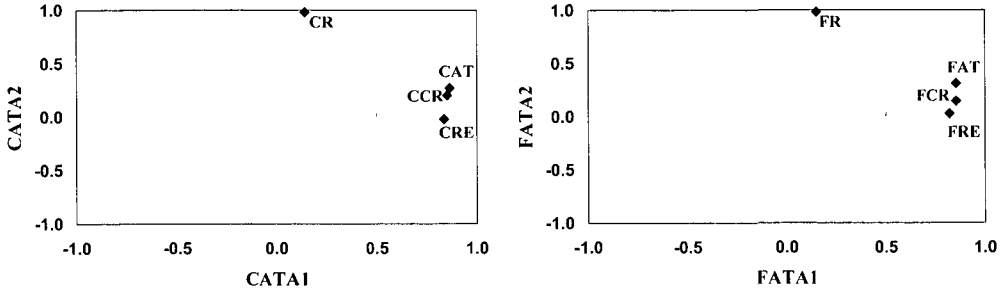
O Gráfico n.º 1 ilustra a forma como as variáveis estão associadas a cada um dos factores: para ambos os grupos de variáveis, representativos das estatísticas de ataque para as equipas que actuam em casa e fora, os ataques, remates e cruzamentos são representados pelo factor 1 e as recuperações estão ligadas ao factor 2.

No desenvolvimento dos modelos de regressão, as estatísticas de ataque são substituídas pelos factores retidos que lhes estão associados, de modo a eliminar a correlação existente entre as variáveis, não perdendo a informação por elas representada.

Quadro n.º 8. Matriz de componentes após rotação

	Factor			Factor	
Var.	CATA1	CATA2	Var.	FATA1	FATA2
CAT	0,864	0,271	FAT	0,855	0,311
CRE	0,836	-0,021	FRE	0,822	0,026
CCR	0,852	0,202	FCR	0,855	0,145
CR	0,141	0,979	FR	0,149	0,981

Gráfico n.º 1. Associação entre componentes retidos e variáveis



4.2. MODELOS DE REGRESSÃO

4.2.1. VARIÁVEIS UTILIZADAS

O objectivo deste estudo passa pelo desenvolvimento de modelos que relacionem a variável dependente, resultado de um jogo de futebol, com as variáveis independentes em análise. Foram construídos modelos, utilizando duas variáveis dependentes que exprimem, de modo diferente, o resultado de um jogo. Tal como referido anteriormente, o Modelo 1 tem como variável dependente "VED", que apresenta três resultados possíveis, para cada jogo: 2 = vitória, 1 = empate e 0 = derrota, reportados à equipa que actua no seu terreno, sendo o resultado da equipa que joga fora complementar a este; o Modelo 2 utiliza a variável dependente "GMGS", que resulta da diferença entre Golos Marcados e Golos Sofridos pela equipa que joga perante o seu público.

As 42 estatísticas disponíveis sobre cada jogo de futebol foram reduzidas; através da aplicação da análise factorial de componentes principais, previamente descrita, a algumas variáveis correlacionadas entre si. No Quadro n.º 9 estão listadas as variáveis resultantes, que serão utilizadas na construção dos modelos de regressão.

Na análise subsequente serão utilizadas estas 33 variáveis independentes, cuja nomenclatura resulta da anteriormente apresentada, no Quadro n.º 1, com a introdução do prefixo "C", quando a estatística diz respeito à equipa que actua em casa e do prefixo "F", quando se reporta à equipa que joga fora. As estatística gerais e a percentagem de posse de bola da equipa que actua no seu terreno ("POSSE") dizem respeito a cada jogo. Assinalam-se a negrito as variáveis que resultam de factores determinados na análise factorial de componentes principais.

Quadro n.º 9. Variáveis independentes para cada jogo

Gerais:		Tempos de posse:		Guarda-redes:		Jogadores:	
RELV		Casa	Fora	Casa	Fora	Casa	Fora
ESPEC		CDEF	FDEF	CDC	FDC	CFC	FFC
TJOGO		CMCD	FMCD	CDI	FDI	CFS	FFS
Disciplinares:		**CTAT**	**FTAT**	CSC	FSC	CP	FP
Casa	Fora	POSSE		CSI	FSI	***CATA1***	***FATA1***
CCA	**FCAV**					***CATA2***	***FATA2***
CCV						CAS	FAS

Para um dos jogos não estão disponíveis dados estatísticos, pelo que existem 611 observações, relativas aos jogos analisados, das variáveis em análise, com a excepção de quatro encontros, em que não estão disponíveis dados para algumas variáveis (*missing values*). Nesta situação, optou-se pela utilização dos valores médios da equipa em análise, em casa ou fora, conforme o caso, de modo a não se perder a informação estatística para outras variáveis da observação em causa.

4.2.2. CONSTRUÇÃO DOS MODELOS DE REGRESSÃO

Este constitui o passo decisivo do estudo da relação estatística entre as variáveis, embora seja descrito de forma bastante resumida.

Ambos os modelos de regressão, inicialmente, integram todas as variáveis independentes, bem como as formas funcionais não lineares da relação destas com a variável dependente, que introduzem melhorias na descrição dos dados, o que ocorre esporadicamente, para algumas variáveis, com a função quadrática, que provoca aumentos significativos no coeficiente de determinação, quando comparado com o valor do mesmo para a relação linear.

A utilização desta forma pressupõe a realização de uma transformação das variáveis, de modo a reduzir a correlação entre os termos da função quadrática, que se consegue do seguinte modo: sendo $X_{i,p}$ a observação i de uma variável deste tipo, a sua relação com a variável dependente será:

$$Y_i = \beta_0 + ... + \beta_{p1} x_{i,p} + \beta_{p2} x_{i,p}^{\ 2} + \cdots + \varepsilon_i \quad \text{com} \quad x_{i,p} = X_{i,p} - \overline{X} \quad (4)$$

Através de um processo de análise sistemática da importância de cada variável nos modelos desenvolvidos, vão sendo eliminadas, passo a passo, variáveis que não apresentam relevância, de acordo com os critérios de análise da significância das variáveis independentes, de maximização do coeficiente de determinação ajustado e utilizando o procedimento *Forward Stepwise* que, essencialmente, desenvolve uma sequência de modelos de regressão, adicionando ou retirando em cada passo uma

variável independente. A aplicação deste procedimento é descrita, pormenorizadamente, por vários autores (Freedman, 1983; Pope e Webster, 1972).

No Quadro n.º 10 são apresentados os resultados mais significativos para os modelos de regressão inicialmente construídos, bem como as variáveis independentes seleccionadas para integrar estes modelos e respectivos níveis de significância.

Quadro n.º 10. Modelos de Regressão

Modelo 1: Variável dependente - VED

Coeficiente de Determinação: r^2 = 0,334		g.l.	*SS*	*MS*
Estimativa do desvio padrão: $\sqrt{MSE}$ = 0,663	Regressão	20	130,00	6,500
F = 14,8 ⇒ Significância F =0,00	Resíduos	590	259,14	0,439

	Var. i	b_i	$s(b_i)$	*Sig. t*
Constante		4,020	0,465	0,000
Gerais	RELV	0,061	0,028	0,027
	ESPEC	$3{,}12e^{-6}$	$2{,}98e^{-6}$	0,295
Disci-	CCA	-0,025	0,019	0,181
plinares	CCV	-0,250	0,078	0,001
	FCAV	0,067	0,030	0,025
Tempos	CDEF	0,069	0,033	0,036
de	FDEF	-0,060	0,030	0,048
posse	POSSE	-4,483	0,775	0,000
Guarda-	CSC	0,022	0,010	0,027
-redes	FSC	-0,038	0,010	0,000

	Var. i	b_i	$s(b_i)$	*Sig. t*
Jogador	CFC	0,013	0,006	0,021
	CFC^2	-0,001	0,001	0,047
	CP	-0,006	0,004	0,121
	CATA1	0,108	0,042	0,009
	CATA2	0,199	0,040	0,000
	CAS	0,118	0,019	0,000
	FFC	-0,011	0,005	0,034
	FATA1	-0,096	0,040	0,018
	FATA2	-0,195	0,039	0,000
	FAS	-0,142	0,023	0,000

Modelo 2: Variável dependente - GMGS

Coeficiente de Determinação: r^2 = 0,422		g.l.	*SS*	*MS*
Estimativa do desvio padrão: $\sqrt{MSE}$ = 1,290	Regressão	18	718,93	39,941
F = 24,0 ⇒ Significância F =0,00	Resíduos	592	984,49	1,663

	Var. i	b_i	$s(b_i)$	*Sig. t*
Constante		-0,732	0,968	0,450
Gerais	RELV	0,096	0,054	0,077
	ESPEC	$7,80e^{-6}$	$5,77e^{-6}$	0,177
	TJOGO	3,658	1,492	0,014
Disci-	CCA	-0,056	0,036	0,126
plinares	CCV	-0,523	0,151	0,001
	FCAV	0,159	0,059	0,007
Guarda-	CSC	0,060	0,019	0,002
-redes	FSC	-0,099	0,018	0,000

	Var. i	b_i	$s(b_i)$	*Sig. t*
Jogador	CFC	0,035	0,010	0,001
	CP	-0,012	0,008	0,152
	CATA1	-0,014	0,069	0,835
	CATA2	0,377	0,078	0,000
	CAS	0,305	0,043	0,000
	CAS^2	0,031	0,011	0,004
	FFC	-0,037	0,010	0,000
	FATA1	-0,014	0,069	0,840
	FATA2	-0,384	0,077	0,000
	FAS	-0,340	0,044	0,000

g.l. – graus de liberdade *SS* e *MS* – somatório e média do somatório dos quadrados.

b_i e $s(b_i)$ – estimativas do coeficiente e do seu desvio padrão para a variável i.

Sig. *t* – nível de significância do teste *t* de *Student*.

Verifica-se a exclusão, do modelo 1 – VED –, das variáveis percentagem de tempo de jogo, tempos de posse de bola no meio campo defensivo e no ataque, defesas e saídas incompletas dos guarda-redes, faltas sofridas e perdas de bola da equipa que joga fora. A variável faltas cometidas pela equipa que actua em casa surge na forma quadrática.

O coeficiente de determinação indica que apenas 33,4% da variação que ocorre na variável dependente VED é explicada pelas variáveis incluídas no modelo, o teste *F*, à significância global do modelo, é validado por apresentar significância nula.

A estatística, ou variável, cuja estimativa do coeficiente apresenta valor positivo contribui positivamente para a vitória da equipa que actua em casa, tendo as estimativas negativas o efeito contrário: uma variação de uma unidade na variável independente provoca uma variação média esperada na variável dependente igual ao valor da estimativa do coeficiente. A significância do teste *t* de *Student* para cada variável indica-nos a probabilidade dessa variável tomar um valor nulo no modelo, não sendo significante. Existem três variáveis (ESPEC, CCA e CP), cujos valores da significância são superiores ao estabelecido como desejável, que é de 5%.

O modelo 2 – GMGS –, comparativamente com o anterior, passa a integrar a variável percentagem de tempo de jogo, sendo excluídos os tempos de posse de bola. A variável assistências da equipa que joga em casa aparece na forma quadrática.

O coeficiente de determinação é superior: as variáveis do modelo explicam 42,2% da variação da variável dependente GMGS e o teste *F* valida o modelo.

As variáveis apresentam estimativas para os seus coeficientes com sinal idêntico às do modelo 1, com excepção do factor CATA1, para as estatísticas de ataque da equipa que actua em casa. Existe um número apreciável de variáveis, cuja significância do teste *t* de *Student* apresenta valores da significância superiores ao estabelecido como desejável, que é de 5%, que serão analisadas nos pontos seguintes.

4.2.3. ANÁLISE DE *OUTLIERS*

Os *outliers* são casos extremos influentes em determinada análise estatística. No desenvolvimento dos modelos de regressão importa determinar o conjunto de observações que podem ser consideradas como *outliers*, de modo a equacionar a sua eliminação na construção de modelos subsequentes, com o objectivo de obter refinamentos. A análise de *outliers* (Barnett e Lewis, 1994; Rousseeuw e Leroy, 1987), com o objectivo de identificação das observações que os constituem será efectuada com a ajuda das seguintes estatísticas: resíduos estandardizados, resíduos estudantizados, *leverage*, distância de *Cook*, *dfBetas* estandardizados e *dfFit* estandardizado.

De modo a proceder a esta análise, torna-se premente introduzir a matriz *H*, cuja compreensão é facilitada pela formulação matricial do modelo.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1,p-1} \\ 1 & X_{21} & \dots & X_{2,p-1} \\ \dots & \dots & \dots & \dots \\ 1 & X_{n1} & \dots & X_{n,p-1} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix} \quad \text{ou} \quad \underset{(n\times 1)}{Y} = \underset{(n\times p)}{X} \cdot \underset{(p\times 1)}{\beta} + \underset{(n\times 1)}{\varepsilon} \qquad (5)$$

A matriz H resulta de:

$$\underset{(n\times n)}{H} = X \cdot \left(X'X\right)^{-1} \cdot X', \text{ sendo } X' \text{ a matriz transposta de } X. \qquad (6)$$

Os resíduos, para cada observação, calculam-se a partir da diferença entre os valores observados e os valores estimados pelo modelo para a variável dependente:

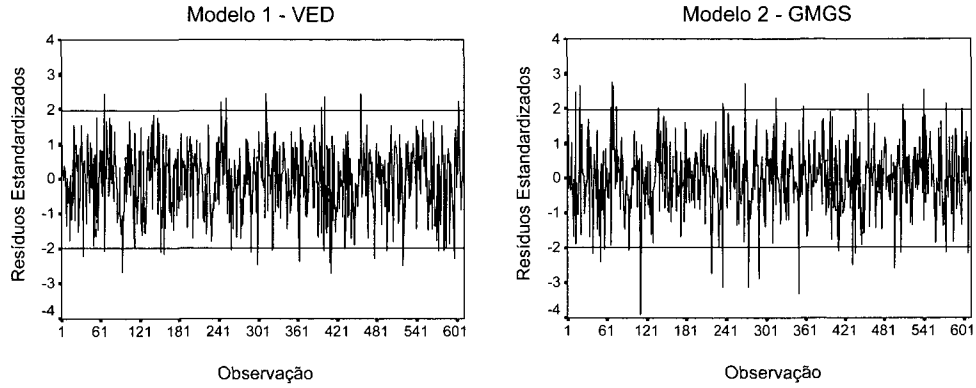
$$e_i = Y_i - \hat{Y}_i \qquad (7)$$

A partir destes valores podem calcular-se os resíduos estandardizados:

$$e_i^* = \frac{e_i}{\sqrt{MSE}} \qquad (8)$$

Considera-se como *outlier* uma observação em que o resíduo estandardizado tenha valor absoluto superior a 1,96, para um nível de significância de 5%. Nos dois modelos desenvolvidos, identificam-se algumas observações como *outliers*, representadas pelos pontos que ultrapassam os limites, no Gráfico n.º 2.

Gráfico n.º 2. Resíduos estandardizados



Um primeiro refinamento, que torna os resíduos mais eficazes no reconhecimento de *outliers*, consiste no reconhecimento do facto de que as observações podem apresentar diferentes desvios padrão entre elas. O desvio padrão de uma observação é estimado através da expressão (9), em que h_{ii} representa o elemento da diagonal principal da matriz H referente à observação i.

$$s(e_i) = \sqrt{MSE(1 - h_{ii})} \tag{9}$$

A razão entre cada resíduo e o seu desvio padrão estimado é denominada resíduo estudantizado:

$$r_i = \frac{e_i}{s(e_i)} \tag{10}$$

Uma segunda melhoria resulta do cálculo dos resíduos para a observação i, quando o modelo de regressão se baseia em todos os dados, com a excepção da observação i, obtendo-se os resíduos *deleted*:

$$d_i = Y_i - \hat{Y}_{i(i)} \tag{11}$$

Combinando as duas formas introduzidas, podem calcular-se os resíduos estudantizados *deleted*:

$$t_i = \frac{d_i}{s(d_i)} \tag{11}$$

Estes, possibilitam um melhor diagnóstico de *outliers*, que resultam das observações, cujos resíduos estudantizados *deleted* são elevados, em valor absoluto. Pode demonstrar-se que este tipo de resíduos seguem uma distribuição t de Student, pelo que é possível definir um valor crítico, a partir do qual se considera uma observação como *outlier*: para um nível de significância de 5%, esse valor crítico é também de 1,96. Assim, no Gráfico n.º 3, ilustram-se os *outliers* obtidos.

A partir da análise dos resíduos, ilustrada pelos Gráficos n.º 2 e n.º 3, a detecção de *outliers* é semelhante para os dois tipos de resíduos, sendo o seu número superior no Modelo 2 – GMGS –, bem como o valor absoluto dos resíduos.

O *leverage*, dado pelo elemento da diagonal principal (h_{ii}) da matriz H, previamente definida, representa a influência da observação i na qualidade do ajustamento feito. Quando é superior a duas vezes o seu valor médio, ou seja, $2(p+1)/n$ (utilizando a nomenclatura introduzida na formulação do modelo), a observação é considerada influente. O gráfico n.º 4 ilustra os *outliers* identificados por esta regra.

Gráfico n.º 3. Resíduos estudantizados *deleted*

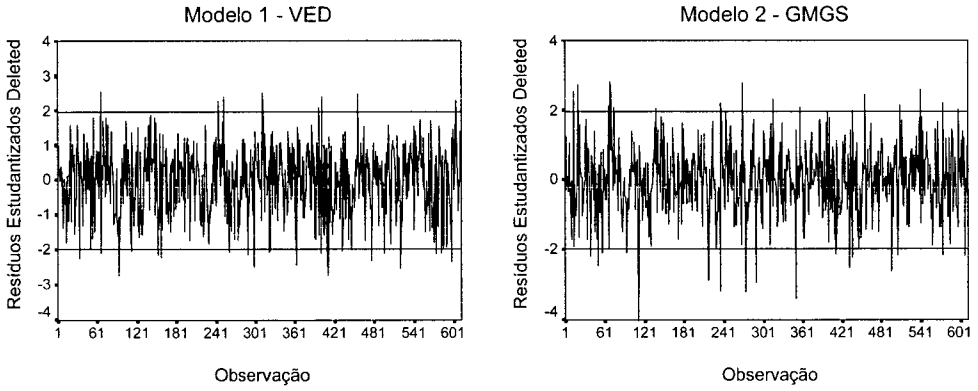
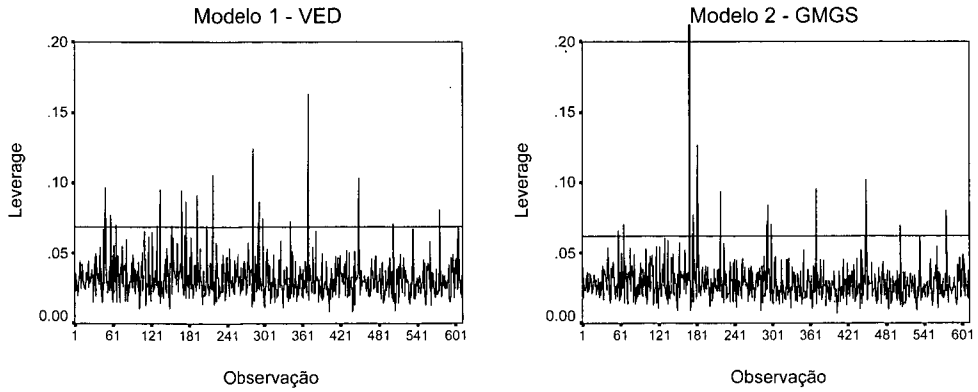


Gráfico n.º 4. *Leverage*



Após a identificação, como *outliers*, de observações, no que diz respeito aos valores das variáveis dependente e independentes, importa verificar a sua influência

no comportamento do modelo, que pode ser quantificada pela distância de *Cook*, *dfBetas* estandardizados e *dfFit* estandardizado: uma observação considera-se influente, se a sua exclusão causar alterações substanciais na função de regressão estimada.

A distância de *Cook* considera a variação provocada nos resíduos de todas as observações, quando a observação *i* é excluída do cálculo dos coeficientes de regressão, podendo ser calculada sem recorrer à estimação de uma nova função de regressão, cada vez que uma observação é excluída, através de uma expressão equivalente:

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{(p+1)\cdot MSE} \Leftrightarrow D_i = \frac{{e_i}^2}{(p+1)\cdot MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \qquad (12)$$

Uma observação é considerada influente quando a distância de *Cook* é superior a $4/(n\text{-}p\text{-}1)$. Os respectivos *outliers* são ilustradas pelo Gráfico n.º 5.

O *dfFit* estandardizado representa a diferença entre o valor estimado pelo modelo, para a observação *i*, quando todas as observações são utilizadas e o valor estimado, para a mesma observação, quando o caso *i* é excluído do cálculo da função de regressão que, tal como na equação anterior, pode ser calculado através de uma expressão equivalente, que não obriga ao cálculo da função de regressão, cada vez que uma observação é excluída do modelo.

$$dfFits_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}} \Leftrightarrow dfFits_i = t_i\sqrt{\frac{h_{ii}}{1-h_{ii}}} \qquad (13)$$

Uma observação é considerada *outlier*, quando o valor absoluto do *dfFit* estandardizado é superior a $2\sqrt{(p+1)/n}$. Os resultados, para esta medida da influência das observações, apresentam-se no Gráfico n.º 6.

Gráfico n.º 5. Distância de *Cook*

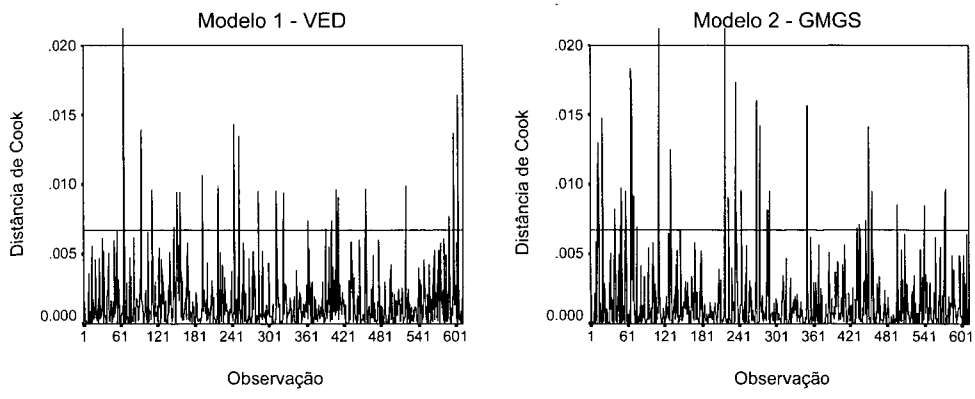
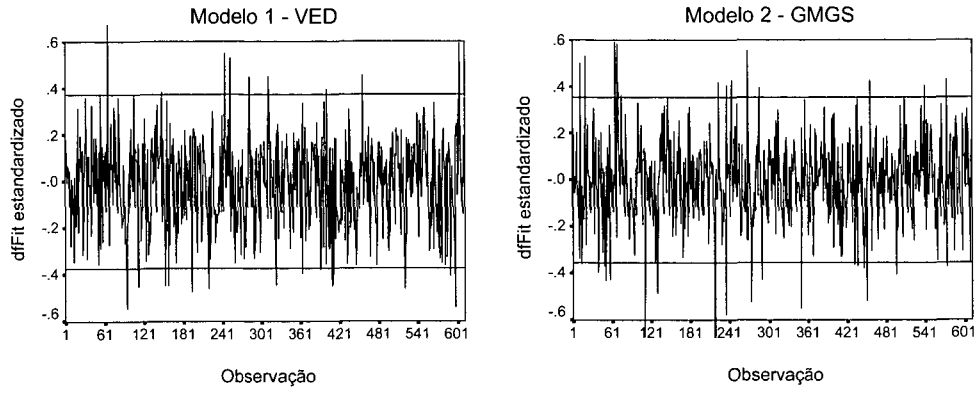


Gráfico n.º 6. dfFit estandardizado



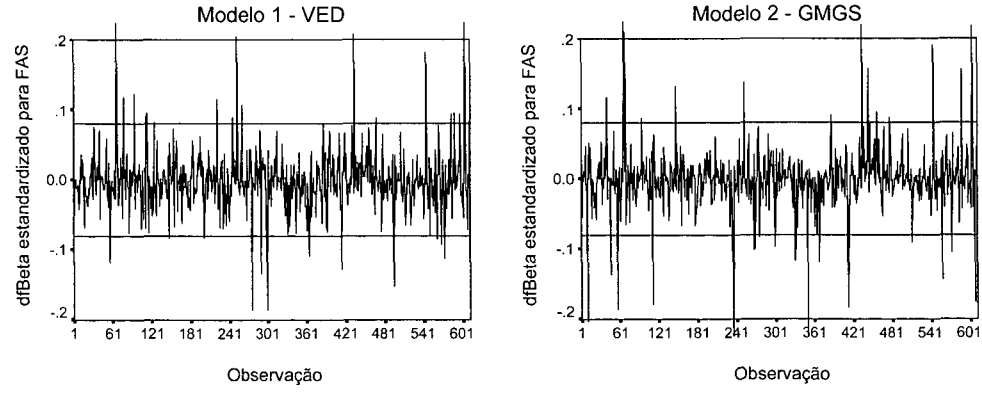
A medida da influência de uma observação *i*, em cada coeficiente da regressão β_k, resulta da diferença entre o valor estimado para o coeficiente de regressão baseado em todas as observações e o mesmo valor omitindo o caso i. O *DfBeta* estandardizado obtém-se, pelo quociente entre essa diferença e a estimativa do desvio padrão do coeficiente de regressão em análise:

$$dfBeta_i = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}} \qquad k = 0, 1, \ldots, p\text{-}1 \qquad (14)$$

Em que c_{kk} é o *k* elemento da diagonal principal da matriz $(X'X)^{-1}$.

O valor de *DfBeta* é calculado, para todas as observações, para todos os parâmetros e para a constante do modelo. As observações são consideradas *outliers* quando o valor absoluto de *DfBeta* é superior a $2/\sqrt{n}$. O Gráfico n.º 7 permite observar os pontos assim identificados, segundo o critério do *DfBeta*, para os coeficientes associados à variável CAS, a título de exemplo.

Gráfico n.º 7. *dfBeta* estandardizado para a variável CAS



A análise de *outliers* apresentada permite identificar os casos extremos considerados influentes para os modelos, que serão excluídos na construção de novas

funções de regressão. Foram considerados casos extremos influentes as observações que desrespeitam as condições impostas aos resíduos ou, então, que não estejam dentro dos limites impostos para, pelo menos, três dos restantes critérios considerados. A análise de *outliers* processou-se, deste modo, em dois passos sucessivos, de maneira a construir um modelo intermédio, cuja análise, nos mesmos termos do modelo inicial, permitiu a detecção de mais casos extremos influentes, que levaram à construção dos modelos de regressão definitivos. Os critérios estabelecidos permitem a detecção de 120 *outliers* no modelo 1 – VED – e, no modelo 2 – GMGS –, foram considerados 109 *outliers*.

É relevante observar que as observações detectadas como *outliers*, em ambos os modelos, se distribuem de modo uniforme por todas as equipas, quer actuando em casa, quer em jogos fora. Quanto às variáveis que representam os resultados dos jogos, tanto no caso da variável VED, como para a variável GMGS, há poucos empates considerados como outliers, quando comparados com a sua frequência relativa no total de jogos.

4.2.4. REFINAMENTO DOS MODELOS DE REGRESSÃO, ELIMINANDO OS *OUTLIERS*

Foram desenvolvidos novos modelos, excluindo os *outliers* detectados anteriormente, que serão utilizados para explicar as relações estatísticas entre as variáveis em análise. No Quadro n.º 11 apresentam-se os resultados significativos para os modelos de regressão definitivos, o primeiro com 491 observações e o segundo com 502, após a eliminação dos casos extremos influentes, bem como para as variáveis independentes.

As alterações mais importantes no modelo 1 – VED –, relativamente ao inicialmente construído, residem no aumento do coeficiente de determinação: a variação que ocorre na variável dependente VED, explicada pelas variáveis do modelo, aumentou praticamente para o dobro – 62,1% –; na diminuição do desvio padrão e nos níveis de significância associados às variáveis, sendo apenas dois moderadamente superiores a 5%.

As variáveis que contribuem positivamente para a vitória da equipa que actua em casa, como explicado anteriormente, pelo sinal da estimativa do respectivo coeficiente, são o estado do relvado, o número de espectadores, o tempo de posse na defesa, as saídas completas do guarda-redes, as faltas cometidas, as estatísticas de ataque e as assistências, para a equipa que actua em casa e as acções disciplinares para a equipa que joga fora. As variáveis cuja contribuição para a variável dependente é negativa são o tempo de posse na defesa, as saídas completas do guarda-redes, as faltas cometidas, as estatísticas de ataque e as assistências, para a equipa que joga fora e as acções disciplinares, a percentagem de posse de bola e as perdas de bola para a equipa que actua em casa. Analisando criticamente os resultados, não seria talvez de esperar que as faltas cometidas por uma equipa tivessem um contributo positivo para o resultado e que a percentagem de posse de bola influenciasse negativamente o resultado.

Quadro n.º 11 Modelos de Regressão

Modelo 1: Variável dependente - VED

Coeficiente de Determinação: r^2 = 0,621		g.l.	*SS*	*MS*
Estimativa do desvio padrão: $\sqrt{MSE}$ = 0,457	Regressão	20	160,91	8,045
F = 38,5 ⇒ Significância F =0,00	Resíduos	470	98,14	0,209

	Var. i	b_i	$s(b_i)$	*Sig. t*
Constante		5,757	0,366	0,000
Gerais	RELV	0,038	0,022	0,086
	ESPEC	$4{,}52e^{-6}$	$2{,}30e^{-6}$	0,050
Disci-	CCA	-0,041	0,014	0,004
plinares	CCV	-0,407	0,066	0,000
	FCAV	0,087	0,023	0,000
Tempos	CDEF	0,044	0,027	0,099
de	FDEF	-0,059	0,024	0,015
posse	POSSE	-6,387	0,611	0,000
Guarda-	CSC	0,015	0,008	0,053
-redes	FSC	-0,043	0,008	0,000

	Var. i	b_i	$s(b_i)$	*Sig. t*
Jogador	CFC	0,011	0,004	0,010
	CFC^2	-0,002	0,000	0,000
	CP	-0,006	0,003	0,053
	CATA1	0,067	0,033	0,043
	CATA2	0,202	0,032	0,000
	CAS	0,091	0,015	0,000
	FFC	-0,022	0,004	0,000
	FATA1	-0,211	0,032	0,000
	FATA2	-0,284	0,031	0,000
	FAS	-0,197	0,018	0,000

Modelo 2: Variável dependente - GMGS

Coeficiente de Determinação: r^2 = 0,620		g.l.	*SS*	*MS*
Estimativa do desvio padrão: $\sqrt{MSE}$ = 0,834	Regressão	18	546,54	30,363
F = 43,7 ⇒ Significância F =0,00	Resíduos	483	335,61	0,695

	Var. i	b_i	$s(b_i)$	*Sig. t*
Constante		-0,864	0,711	0,224
Gerais	RELV	0,083	0,039	0,034
	ESPEC	$7{,}51e^{-6}$	$4{,}30e^{-6}$	0,081
	TJOGO	4,368	1,107	0,000
Disci-	CCA	-0,083	0,026	0,001
Plinares	CCV	-0,496	0,110	0,000
	FCAV	0,120	0,042	0,004
Guarda-	CSC	0,059	0,014	0,000
-redes	FSC	-0,100	0,013	0,000

	Var. i	b_i	$s(b_i)$	*Sig. t*
Jogador	CFC	0,034	0,008	0,000
	CP	-0,012	0,006	0,039
	CATA1	-0,126	0,050	0,011
	CATA2	0,342	0,056	0,000
	CAS	0,270	0,030	0,000
	CAS^2	0,030	0,007	0,000
	FFC	-0,042	0,007	0,000
	FATA1	-0,117	0,049	0,017
	FATA2	-0,501	0,056	0,000
	FAS	-0,350	0,034	0,000

g.l. – graus de liberdade *SS* e *MS* – somatório e média do somatório dos quadrados.

b_i e $s(b_i)$ – estimativas do coeficiente e do seu desvio padrão para a variável i.

Sig. *t* – nível de significância do teste *t* de *Student*.

O novo modelo 2 – GMGS – introduz também um aumento do coeficiente de determinação, relativamente ao modelo inicial, para 62%, diminuição do desvio padrão e níveis de significância inferiores a 5% para as variáveis, com excepção apenas de uma delas.

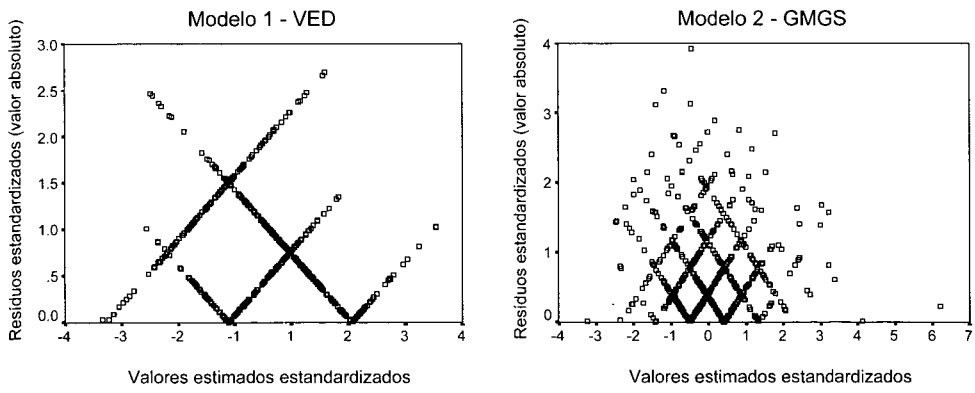
As variáveis partilhadas com o modelo 1 contribuem todas da mesma forma para o resultado do jogo, com excepção do factor CATA1, para as estatísticas de ataque (ataques, cruzamentos e remates) da equipa que actua em casa, que influencia negativamente o resultado da equipa que joga em casa, o que não seria de esperar à partida. A variável tempo jogado (% do tempo total) favorece o resultado para a equipa que actua perante o seu público.

4.2.5. VALIDAÇÃO DOS MODELOS DE REGRESSÃO

Os modelos de regressão devem cumprir determinados pressupostos, cuja verificação valida os modelos desenvolvidos. Deste modo, torna-se necessária a concretização de testes estatísticos, que incluem análise gráfica de resíduos, estudo da multicolinearidade (correlação entre variáveis independentes), análise da homocedasticidade (variância constante dos termos de erro) e medida da auto-correlação, com o objectivo de validar os modelos.

Em primeiro lugar será verificada a homocedasticidade que, etimologicamente significa variância constante. Resultando um resíduo da diferença entre os valores previstos pelo modelo e os valores observados, um dos processos alternativos para analisar a homocedasticidade consiste em observar a relação entre os resíduos estandardizados e os valores estimados estandardizados da variável dependente. No gráfico n.º 8 ilustra-se esta relação, para o valor absoluto dos resíduos estandardizados, que torna mais fácil a análise gráfica.

Gráfico n.º 8. Relação entre resíduos e valores estimados estandardizados



No modelo 1 verifica-se uma dispersão ligeiramente superior para valores estimados inferiores, em valor absoluto, derivada também do tipo de variável dependente. No modelo 2 a amplitude mantém-se aproximadamente constante em relação ao eixo horizontal, pelo que não parece existir variação da dispersão. Os

modelos parecem passíveis de ser considerados, quanto a este primeiro pressuposto, com alguns cuidados para o primeiro modelo.

Um segundo pressuposto a analisar é a inexistência de auto-correlação entre as variáveis independentes, através do teste de Durbin-Watson, que permite verificar se os termos de erro são independentes, ou seja, se o parâmetro de auto-correlação é nulo. A estatística de teste apresenta o valor de 1,85 e de 1,86 para os modelos 1 e 2, respectivamente. Para um nível de significância de 5%, o valor crítico considerado para este teste é de 1,78, pelo que não podemos rejeitar a hipótese, para nenhum dos modelos construídos, de que a auto-correlação seja nula.

Um terceiro pressuposto define que os resíduos devem seguir uma distribuição normal, podendo ser verificado pelo teste Kolmogorov-Smirnov (*K-S*), com a correcção de Lilliefors, apresentado no Quadro n.º 12.

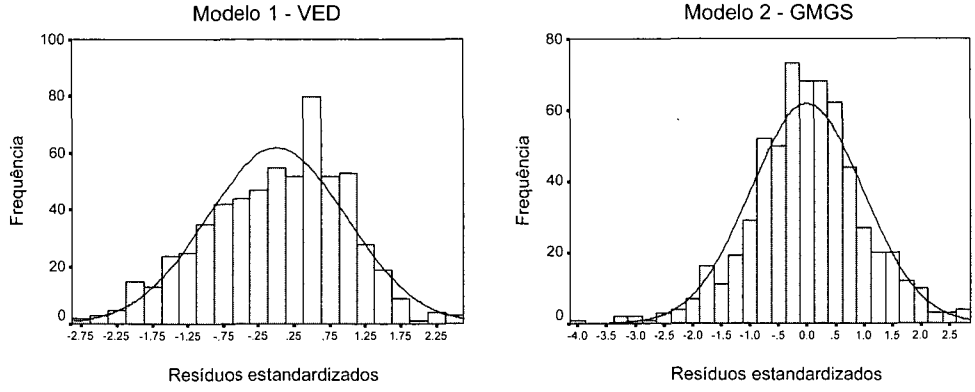
Quadro n.º 12. Teste à *normalidade* dos resíduos estandardizados

	Estatística *K-S* (Lilliefors)	Graus de liberdade	Significância
Modelo 1	0,059	611	0,000
Modelo 2	0,036	611	0,052

Exige-se, normalmente, um nível de significância de 5% para não rejeitar a hipótese dos resíduos seguirem uma distribuição normal, o que sucede apenas para o Modelo 2. Também para este pressuposto, o Modelo 1 exige uma atenção especial.

De modo a complementar o estudo da *normalidade* dos resíduos, apresenta-se o Gráfico n.º 9, que regista a diferença entre o histograma da distribuição das variáveis aleatórias residuais e a distribuição normal, observando-se alguma correspondência entre a distribuição das frequências relativas das várias classes definidas para os resíduos e a curva de distribuição normal, principalmente para o Modelo 2.

Gráfico n.º 9. Histograma dos resíduos estandardizados e distribuição normal

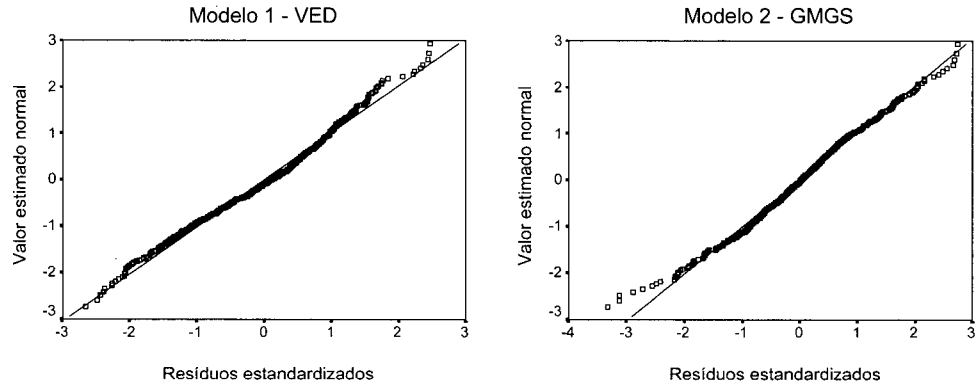


Os desvios à normalidade podem ser ainda observados no Gráfico n.º 10, em que se apresentam os gráficos *Q-Q*, de modo a ilustrar, pelos desvios à linha oblíqua, as diferenças em relação à distribuição normal. Verifica-se que estes desvios não

apresentam magnitude mais elevada para o modelo 1, pelo que podemos prosseguir com a validação dos modelos.

Finalmente, importa verificar o pressuposto da ausência de multicolinearidade, cuja intensidade pode ser estudada, sumariamente, através da análise da correlação entre as variáveis independentes. Os maiores valores absolutos observados para a correlação são entre as variáveis CATA2 e FATA2, cujas maiores componentes são dadas pelas recuperações de bola, com valores de (-0,65) e (-0,62) para os modelos 1 e 2, respectivamente, o que não indicia multicolinearidade.

Gráfico n.º 10. Gráficos *Q-Q* dos resíduos estandardizados



O factor de inflação da variância (FIV) é também uma medida da multicolineariedade, que contabiliza a inflação sofrida pela variância dos coeficientes de regressão estimados, provocada pela correlação entre variáveis. Pode ser demonstrado que este factor, para uma variável k, é:

$$(FIV)_k = \left(1 - r_k^{\,2}\right)^{-1} \qquad k = 1, 2, \ldots, p\text{-}1 \qquad (15)$$

onde r_k^2 corresponde ao coeficiente de determinação, quando a variável X_k é relacionada, através de um modelo de regressão linear, com as restantes (p-2) variáveis independentes.

Valores elevado do FIV são indicadores de multicolinearidade, considerando-se valores superiores a 10 influenciadores das estimativas dos coeficientes de regressão. No Quadro n.º 13 apresentam-se os FIV para as variáveis utilizadas nos dois modelos, cujos valores mais elevados não indiciam, de algum modo, a existência de multicolinearidade.

Quadro n.º 13. Factor de inflacção da variância

Modelo 1 - VED				Modelo 2 - GMGS			
Variável	*FIV*	*Variável*	*FIV*	*Variável*	*FIV*	*Variável*	*FIV*
Const.		CFC	1,408	Const.		CFC	1,262
RELV	1,277	CFC^2	1,206	RELV	1,279	CP	1,211
ESPEC	1,272	CP	1,244	ESPEC	1,258	*CATA1*	1,738
CCA	1,421	*CATA1*	2,505	TJOGO	1,369	*CATA2*	2,336
CCV	1,259	*CATA2*	2,303	CCA	1,399	CAS	1,617
FCAV	1,246	CAS	1,280	CCV	1,225	CAS^2	1,459
CDEF	1,848	FFC	1,309	FCAV	1,307	FFC	1,237
FDEF	1,777	*FATA1*	2,312	CSC	1,313	*FATA1*	1,695
POSSE	2,554	*FATA2*	2,195	FSC	1,255	*FATA2*	2,287
CSC	1,333	FAS	1,203			FAS	1,146
FSC	1,290						

b_i e $s(b_i)$ - estimativas do coeficiente e do seu desvio padrão para a variável *i*.

A análise de ambos os modelos construídos permite concluir que podem ser aplicados para os dados estudados, uma vez que cumprem, de um modo geral, os pressupostos analisados, pelo que a sua utilização para a previsão dos resultados dos jogos de futebol pode ser realizada.

5. PREVISÃO DOS RESULTADOS DOS JOGOS

Um dos objectivos dos modelos de regressão é, precisamente, a previsão da variável dependente a partir dos valores das variáveis independentes. Com base nos modelos desenvolvidos é possível cumprir este objectivo, calculando a estimativa do resultado de cada jogo ($\hat{Y}_i$), a partir do comportamento das equipas intervenientes: dados estatísticos, relevantes para o modelo, observados para cada jogo (X_{i1}, X_{i2}, ..., $X_{i,p-1}$) e das estimativas dos coeficientes dos modelos de regressão (b_0, b_1, ..., b_{p-1}), apresentadas nos quadros n.º 12 e 13.

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_{p-1} X_{i,p-1} \qquad i = 1, 2, ..., n \quad (16)$$

Estas estimativas, dos resultados dos jogos, utilizando os modelos desenvolvidos, podem ser, então, comparadas com os resultados efectivamente observados.

É também premente efectuar uma análise de sensibilidade aos resultados dos jogos previstos pelos modelos, para o que se calculam intervalos de confiança para as

estimativas do previsões, sendo utilizado um nível de confiança de 95%. Sendo X_h o vector coluna constituído pelos valores das variáveis independentes, para uma observação em particular, os limites inferior (LI) e superior (LS) do intervalo de confiança, a 95%, podem ser obtidos através da seguinte expressão.

$$\hat{Y}_i \pm t(0{,}025, n-p) \cdot \sqrt{MSE \cdot \left[X'_h \cdot (X'X)^{-1} \cdot X_h\right]} \qquad (17)$$

Sendo t o valor da distribuição t de *Student* para uma probabilidade de 0,025 e para n-p graus de liberdade e X'_h a matriz transposta de X_h.

Após o cálculo da previsão para o valor da variável dependente, bem como dos seus limites de confiança, que apresentam valores contínuos, importa transformá-los em valores que representam o resultado de um jogo de futebol, em termos de vitória, empate ou derrota de uma das equipas, objectivo atingido através de análise que se apresenta seguidamente.

Nas duas épocas estudadas, foram disputados 612 jogos, tendo ocorrido 304 vitórias da equipa que joga em casa (VC), 174 empates (E) e 134 vitórias das equipas que actuam fora (VF). Esta distribuição dos três resultados possíveis, para o período em análise: 49,7% VC, 28,4% E e 21,9% VF apresenta valores semelhantes aos verificados em outros períodos de tempo. Se calcularmos, por exemplo, a distribuição dos resultados para as seis épocas de futebol, em que o campeonato decorreu em condições semelhantes às actuais (com 18 equipas e três pontos atribuídos à vitória), desde 94/95 até 99/2000, verificam-se os seguintes valores: 50,3% VC, 26,4% E e 23,3% VF. Esta constância da distribuição de resultados consubstancia o passo seguinte: a elaboração de regras de decisão para definir os resultados dos jogos em função dos valores previstos pelos modelos para as variáveis dependentes utilizadas.

As regras de decisão, apresentadas no Quadro n.º 14, consistem em que:

- Para o Modelo 1 – VED, considera-se vitória da equipa da casa quando o valor previsto para a variável dependente é superior a 1,37, vitória da equipa que joga fora quando o valor previsto é inferior a 0,90 e empate nos restantes casos.
- No Modelo 2 – GMGS, a valores previstos, para a variável dependente, superiores a 0,50 corresponde uma vitória caseira, inferiores a –0,25 indicam vitória forasteira e aos restantes está associado o empate.

Deste modo, a distribuição dos três resultados possíveis, previstos pelos modelos desenvolvidos, é idêntica à distribuição observada para os 612 jogos em análise: as diferenças percentuais observadas, nas frequências relativas de cada resultado, são mínimas.

Quadro n.º 14. Regras de decisão para definir os resultados dos jogos ($\hat{Y}_i$)

	Frequência	Modelo 1 - VED		Modelo 2 - GMGS	
Resultado	Observada	Regra de decisão	Frequência	Regra de decisão	Frequência
VC	49,7%	$\hat{Y}_i > 1{,}37$	49,8%	$\hat{Y}_i > 0{,}50$	50,1%
E	28,4%	$0{,}90 < \hat{Y}_i < 1{,}37$	28,2%	$-0{,}25 < \hat{Y}_i < 0{,}50$	28,2%
VF	21,9%	$\hat{Y}_i < 0{,}90$	22,1%	$\hat{Y}_i < -0{,}25$	21,8%

As equipas intervenientes em cada jogo são representadas, em todos os quadros e gráficos seguintes, por uma abreviatura com três dígitos, de acordo com o Quadro n.º 15.

Quadro n.º 15 – Abreviaturas dos nomes das equipas

ACA	Académica	CAM	Campomaiorense	MAR	Marítimo
ALV	Alverca	CHA	Desp. Chaves	POR	Porto
BEI	Beira-mar	EST	Estrela Amadora	RIO	Rio Ave
BEL	Belenenses	FAR	Farense	SAL	Salgueiros
BEN	Benfica	GIL	Gil Vicente	SCL	Santa Clara
BOA	Boavista	GUI	Guimarães	SET	Vit. Setúbal
BRA	Sp. Braga	LEI	U. Leiria	SPO	Sporting

Estabelecidas as regras de decisão, torna-se possível efectuar a previsão do resultado de cada um dos jogos, utilizando os modelos de regressão. No Quadro n.º 16 são apresentadas as previsões calculadas com ambos os modelos, a título de exemplo, para os jogos entre os quatro primeiros classificados de cada uma das épocas estudadas e comparadas com os resultados efectivamente observados. Os jogos são apresentados por ordem cronológica.

Importa explicar que os resultados dos jogos (Res.) são representados pelos pontos atribuídos à equipa que joga em casa, cujos significados são, de acordo com a simbologia já utilizada, "3" – VC, "1" – E e "0" – VF.

Na primeira coluna é identificado o jogo em análise, bem como o resultado verificado, em termos de golos marcados, golos sofridos e pontos obtidos, relativos à equipa que actua em casa. Para os dois modelos desenvolvidos são calculadas as estimativas do valor da variável dependente e dos seus limites de confiança, a 95%, bem como os resultados correspondentes, de acordo com as regras de decisão estabelecidas e a análise da sua concordância (☑) ou não (☒) com o resultado efectivamente observado.

Analise-se, a título de exemplo, o último jogo apresentado da temporada de 1999/2000, que poderia decidir o título: Sporting-Benfica (0-1). Ambos os modelos preconizam uma vitória do Sporting, através da estimativa da variável dependente, resultado do jogo, a partir dos dados estatísticos. Apenas o limite de confiança inferior, para um grau de confiança de 95%, permite chegar à conclusão de que o

resultado poderia ser um empate, mas nenhum dos modelos estima o resultado que realmente se verificou, uma vitória do Benfica.

Quadro n.º 16 Resultados dos jogos: valores observados e previstos

Campeonato de 1998/1999

Valor Observado		Modelo 1 – VED						Modelo 2 - GMGS					
Jogo (GM-GS)	Res.	$\hat{Y}_i$	Res.	LI	Res.	LS	Res.	$\hat{Y}_i$	Res.	LI	Res.	LS	Res.
POR-BOA (0-2)	0	-0,44	0 ☑	-0,67	0 ☑	-0,21	0 ☑	-1,69	0 ☑	-2,06	0 ☑	-1,32	0 ☑
BOA-BEN (2-1)	3	0,03	0 ☒	-0,25	0 ☒	0,31	0 ☒	-1,86	0 ☒	-2,38	0 ☒	-1,34	0 ☒
BOA-SPO (2-2)	1	-0,08	0 ☒	-0,31	0 ☒	0,14	0 ☒	-2,03	0 ☒	-2,47	0 ☒	-1,59	0 ☒
POR-BEN (3-1)	3	2,25	3 ☑	2,03	3 ☑	2,48	3 ☑	2,61	3 ☑	2,21	3 ☑	3,02	3 ☑
POR-SPO (3-2)	3	1,38	3 ☑	1,18	1 ☒	1,58	3 ☑	0,73	3 ☑	0,37	1 ☒	1,09	3 ☑
SPO-BEN (1-2)	0	0,44	0 ☑	0,17	0 ☑	0,71	0 ☑	-0,19	1 ☒	-0,62	0 ☑	0,25	1 ☒
BOA-POR (0-0)	1	1,06	1 ☑	0,89	0 ☒	1,23	1 ☑	-0,36	0 ☒	-0,68	0 ☒	-0,09	1 ☑
BEN-BOA (0-3)	0	1,05	1 ☒	0,71	0 ☑	1,40	3 ☒	0,74	3 ☒	0,15	1 ☒	1,32	3 ☒
BEN-POR (1-1)	1	1,60	3 ☒	1,41	3 ☒	1,78	3 ☒	0,09	1 ☑	-0,24	1 ☑	0,42	1 ☑
SPO-BOA (1-1)	1	1,64	3 ☒	1,44	3 ☒	1,83	3 ☒	0,70	3 ☒	0,42	1 ☑	0,97	3 ☒
SPO-POR (1-1)	1	1,53	3 ☒	1,36	1 ☑	1,69	3 ☒	1,51	3 ☒	1,27	3 ☒	1,75	3 ☒
BEN-SPO (3-3)	1	1,17	1 ☑	0,89	0 ☒	1,44	3 ☒	0,19	1 ☑	-0,31	0 ☒	0,68	3 ☒

Campeonato de 1999/2000

Valor Observado		Modelo 1 - VED						Modelo 2 - GMGS					
Jogo (GM-GS)	Res.	$\hat{Y}_i$	Res.	LI	Res.	LS	Res.	$\hat{Y}_i$	Res.	LI	Res.	LS	Res.
BOA-POR (1-1)	1	1,20	1 ☑	1,07	1 ☑	1,34	1 ☑	0,22	1 ☑	-0,03	1 ☑	0,46	1 ☑
SPO-BOA (2-0)	3	1,80	3 ☑	1,62	3 ☑	1,99	3 ☑	1,30	3 ☑	0,97	3 ☑	1,64	3 ☑
BEN-BOA (1-1)	1	0,97	1 ☑	0,73	0 ☒	1,20	1 ☑	-0,15	1 ☑	-0,54	0 ☒	0,25	1 ☑
POR-SPO (3-0)	3	1,91	3 ☑	1,65	3 ☑	2,17	3 ☑	2,35	3 ☑	1,95	3 ☑	2,75	3 ☑
POR-BEN (2-0)	3	2,06	3 ☑	1,84	3 ☑	2,28	3 ☑	1,04	3 ☑	0,68	3 ☑	1,40	3 ☑
BEN-SPO (0-0)	1	1,14	1 ☑	0,81	0 ☒	1,47	3 ☒	0,68	3 ☒	0,08	1 ☑	1,28	3 ☒
POR-BOA (1-0)	3	1,55	3 ☑	1,41	3 ☑	1,70	3 ☑	1,13	3 ☑	0,89	3 ☑	1,37	3 ☑
BOA-SPO (0-1)	0	0,71	0 ☑	0,53	0 ☑	0,89	0 ☑	-0,29	0 ☑	-0,58	0 ☑	0,00	1 ☒
BOA-BEN (1-1)	1	0,58	0 ☒	0,40	0 ☒	0,76	0 ☒	-0,81	0 ☒	-1,12	0 ☒	-0,50	0 ☒
SPO-POR (2-0)	3	1,96	3 ☑	1,70	3 ☑	2,22	3 ☑	0,40	1 ☒	-0,08	1 ☒	0,87	3 ☑
BEN-POR (1-0)	3	1,15	1 ☒	0,98	1 ☒	1,31	1 ☒	0,21	1 ☒	-0,09	1 ☒	0,51	3 ☑
SPO-BEN (0-1)	0	1,59	3 ☒	1,36	1 ☒	1,82	3 ☒	0,58	3 ☒	0,18	1 ☒	0,99	3 ☒

Os modelos permitem efectuar as previsões da variável dependente, resultado do jogo, para todas as observações, à excepção de um deles, ACA-SAL (0-1), para o qual não estão disponíveis dados estatísticos, tal como já foi referido anteriormente.

Aplicando os modelos a todos os jogos, podem calcular-se as classificações finais das equipas estimadas pelos modelos, através do somatório de pontos conseguidos por cada equipa e compará-las com as classificações efectivamente verificadas, como pode observar-se no Quadro n.º 17.

Para uma análise mais detalhada, dividem-se os pontos totais em pontos obtidos em casa e fora e apresentam-se, para cada modelo, os limites inferiores e superiores da classificação final, tendo por base os limites de confiança, a 95%, para o pior e melhor resultado, respectivamente, de cada equipa, em cada jogo.

A título de informação adicional, verifica-se que o modelo 1 – VED estima resultados diferentes dos observados para 246 jogos e relativamente ao modelo 2 – GMGS, existem 256 jogos nestas condições. Os dois modelos originam previsões diferentes, entre ambos, em 140 jogos.

Quadro n.º 17. Classificações finais observadas e estimadas pelos modelos

Campeonato de 1999/2000

	Pontos em casa			Pontos fora			Total de pontos						
								Mod. 1-VED			Mod. 2-GMGS		
	Obs.	VED	GMGS	Obs.	VED	GMGS	Obs.	LI	Mod.	LS	LI	Mod.	LS
SPO	42	47	43	35	26	24	77	60	73	75	47	67	80
POR	49	45	45	24	17	20	73	55	62	71	55	65	72
BEN	44	38	42	25	20	24	69	43	58	67	59	66	74
BOA	33	22	23	22	20	19	55	33	42	54	33	42	52
GIL	35	34	33	18	15	12	53	39	49	58	38	45	51
MAR	31	39	37	19	25	20	50	47	64	71	46	57	68
GUI	38	33	35	10	12	15	48	38	45	61	38	50	59
EST	25	35	36	20	18	20	45	44	53	67	41	56	63
BRA	26	29	27	17	13	11	43	33	42	52	32	38	55
LEI	27	23	22	15	7	9	42	25	30	48	25	31	46
ALV	31	20	23	10	8	16	41	22	28	45	26	39	54
BEL	26	24	16	14	15	14	40	28	39	55	18	30	44
CAM	27	26	26	9	8	13	36	24	34	45	29	39	53
FAR	24	24	23	11	10	12	35	25	34	40	25	35	41
SAL	20	21	16	14	17	15	34	28	38	49	26	31	46
RIO	27	37	35	6	8	11	33	37	45	51	38	46	55
SET	21	31	25	12	17	16	33	37	48	59	32	41	49
SCL	22	36	30	9	8	14	31	33	44	52	37	44	58

Campeonato de 1998/1999

	Pontos em casa			Pontos fora			Total de pontos						
								Mod. 1-VED			Mod. 2-GMGS		
	Obs.	VED	GMGS	Obs.	VED	GMGS	Obs.	LI	Mod.	LS	LI	Mod.	LS
POR	48	46	48	31	25	25	79	62	71	74	61	73	77
BOA	42	35	32	29	30	29	71	49	65	72	36	61	74
BEN	38	36	39	27	26	25	65	56	62	75	49	64	72
SPO	40	37	42	23	25	18	63	48	62	71	53	60	75
SET	36	40	37	17	23	19	53	50	63	64	49	56	72
LEI	31	26	23	21	20	15	52	26	46	58	25	38	55
GUI	35	37	41	15	15	15	50	42	52	60	42	56	65
EST	32	35	40	13	15	11	45	38	50	54	34	51	61
BRA	28	35	47	14	16	12	42	39	51	64	44	59	71
MAR	26	23	26	15	10	12	41	25	33	48	22	38	50
FAR	26	24	27	13	15	15	39	32	39	55	29	42	55
SAL	28	18	24	10	9	10	38	22	27	36	22	34	45
CAM	26	24	23	11	11	11	37	27	35	50	21	34	45
ALV	25	18	21	10	14	14	35	24	32	45	26	35	45
RIO	20	24	18	15	17	13	35	29	41	51	23	31	50
BEI	24	27	24	9	22	17	33	34	49	67	33	41	56
CHA	19	20	22	6	8	7	25	20	28	33	20	29	44
ACA	14	15	19	7	15	21	21	28	30	49	24	40	48

De modo a sistematizar a análise de como a classificação estimada pelos modelos difere da observada, no Quadro n.º 18 indicam-se as alterações classificativas resultantes das estimativas obtidas pelos modelos, com a indicação de manutenção (=), subida ($\nearrow n$) ou descida ($\searrow n$) de n posições na tabela.

Quadro n.º 18. Comparação das classificações estimadas com as observadas

Classificações da época de 1999/2000								Classificações da época de 1998/1999							
Observada		Mod. VED			Mod. GMGS			Observada		Mod. VED			Mod. GMGS		
SPO	77	SPO	73	=	SPO	67	=	POR	79	POR	71	=	POR	73	=
POR	73	MAR	64	↗4	BEN	66	↗1	BOA	71	BOA	65	=	BEN	64	↗1
BEN	69	POR	62	↘1	POR	65	↘1	BEN	65	SET	63	↗2	BOA	61	↘1
BOA	55	BEN	58	↘1	MAR	57	↗2	SPO	63	BEN	62	↘1	SPO	60	=
GIL	53	EST	53	↗3	EST	56	↗3	SET	53	SPO	62	↘1	BRA	59	↗4
MAR	50	GIL	49	↘1	GUI	50	↗1	LEI	52	GUI	52	↗1	SET	56	↘1
GUI	48	SET	48	↗10	RIO	46	↗9	GUI	50	BRA	51	↗2	GUI	56	=
EST	45	GUI	45	↘1	GIL	45	↘3	EST	45	EST	50	=	EST	51	=
BRA	43	RIO	45	↗7	SCL	44	↗9	BRA	42	BEI	49	↗7	FAR	42	↗2
LEI	42	SCL	44	↗8	BOA	42	↘6	MAR	41	LEI	46	↘4	BEI	41	↗6
ALV	41	BOA	42	↘7	SET	41	↗6	FAR	39	RIO	41	↗4	ACA	40	↗7
BEL	40	BRA	42	↘3	ALV	39	↘1	SAL	38	FAR	39	↘1	LEI	38	↘6
CAM	36	BEL	39	↘1	CAM	39	=	CAM	37	CAM	35	=	MAR	38	↘3
FAR	35	SAL	38	↗1	BRA	38	↘5	ALV	35	MAR	33	↘4	ALV	35	=
SAL	34	CAM	34	↘2	FAR	35	↘1	RIO	35	ALV	32	↘1	SAL	34	↘3
RIO	33	FAR	34	↘2	LEI	31	↘6	BEI	33	ACA	30	↗2	CAM	34	↘3
SET	33	LEI	30	↘7	SAL	31	↘2	CHA	25	CHA	28	=	RIO	31	↘2
SCL	31	ALV	28	↘7	BEL	30	↘6	ACA	21	SAL	27	↘6	CHA	29	↘1

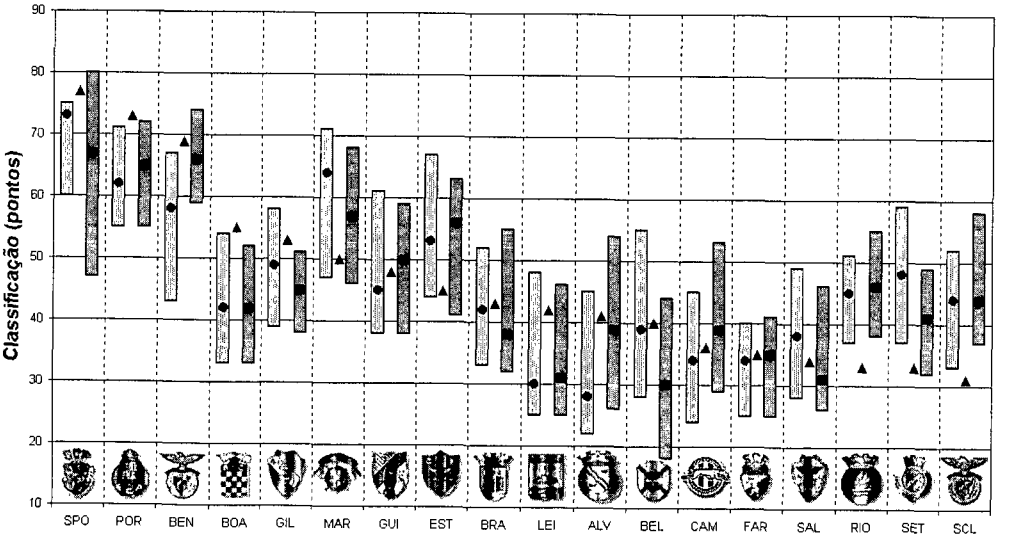
Ambos os modelos confirmam, através dos seus resultados, os vencedores de ambos os campeonatos e indicam subidas significativas, em 1999/2000, para o Marítimo, Est. Amadora, Setúbal, Rio Ave e Santa Clara, que obtiveram uma classificação inferior à estimada pelos modelos e descidas significativas para o Boavista, Braga, Leiria, Alverca e Belenenses, que obtiveram resultados acima dos estimados. Na temporada anterior, em 1998/99, as subidas mais significativas estimadas pelos modelos são do Braga, Beira-mar e Académica, sendo as descidas relevantes para o Leiria, Marítimo e Salgueiros.

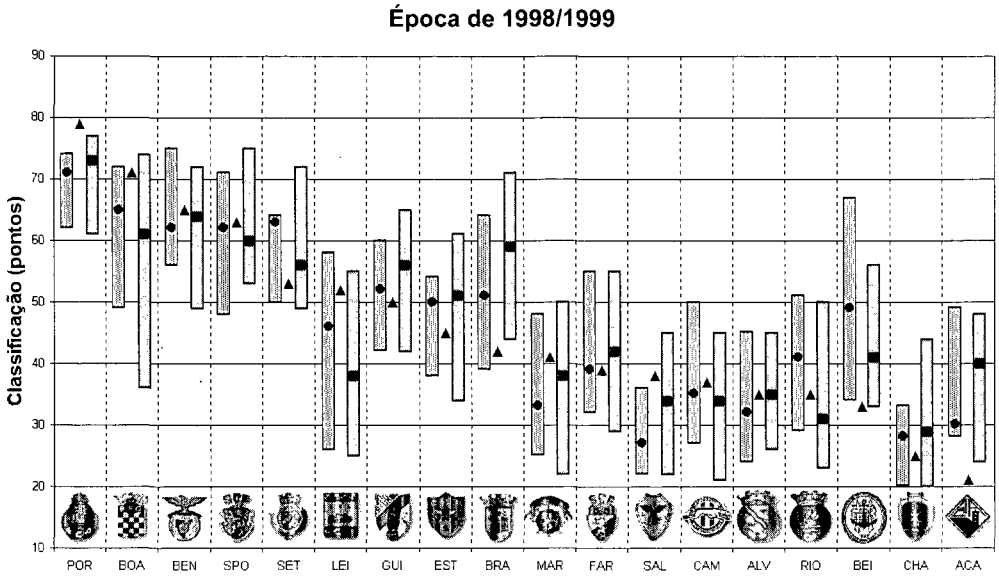
No entanto, pode colocar-se uma questão: se a regra de decisão fosse outra ou se o modelo tivesse uma estrutura diferente, qual a influência nos resultados previstos pelo modelo? Uma abordagem à resposta a esta questão pode ser obtida através de uma análise de sensibilidade aos resultados, utilizando os limites dos intervalos de confiança, a 95%, para as previsões pelos modelos, dos valores estimados das classificações finais, para cada temporada, já apresentados no Quadro n.º 17 e ilustrados pelo Gráfico n.º 11. Estes limites dão uma ideia da variação que pode ocorrer nas classificações finais estimadas pelos modelo, por variações pontuais do resultado estimado de cada jogo.

A análise deste gráfico permite visualizar, facilmente, a comparação entre os valores observados para a pontuação obtida na classificação final, por cada equipa e os mesmos valores previstos pelos modelos de regressão linear, já interpretada anteriormente, para as equipas com maiores diferenças. Os intervalos de confiança, previstos por ambos os modelos, apresentam alguma concordância entre si e incluem, na maioria dos casos, os valores observados das classificações.

Gráfico n.º 11. Análise de sensibilidade à classificação final estimada

Época de 1999/2000





Classificações: ▲ – observada.

● – prevista pelo Modelo 1 – VED – e ▯ - respectivo intervalo de confiança, a 95%.

■ – prevista pelo Modelo 2 – GMGS – e ▯ - respectivo intervalo de confiança, a 95%.

6. CONSIDERAÇÕES FINAIS

Os procedimentos que dão origem aos modelos de regressão permitem exemplificar o desenvolvimento e aplicação de um modelo de regressão para estudos observacionais explicativos, ilustrando os vários passos que vão sendo tomados.

Os modelos desenvolvidos, após a eliminação de *outliers*, respeitam os pressupostos exigidos para modelos de regressão, apresentando a sua aplicação, aos resultados dos jogos de futebol do campeonato nacional, alguma sustentabilidade. No entanto, partindo de outros prismas, no que diz respeito à análise dos resultados dos jogos de futebol, em função das estatísticas disponíveis, poderão ser elaborados outros modelos, com variações mais ou menos relevantes, mas com resultados que também podem ser significativos.

Será importante verificar a aplicação dos modelos a um conjunto superior de observações e as alterações por elas provocadas nas estimativas dos coeficientes dos modelos e até nas variáveis que os integram, através do estudo de futuras épocas do campeonato nacional de futebol.

É de referir que os limites dos intervalos de confiança, a 95%, provocam alterações significativas nos resultados dos modelos, devido a estes apresentarem uma variabilidade considerável, quanto à previsão dos resultados, facto que se deve, certamente à aleatoriedade associada a um desafio de futebol, que não passa de um

jogo: o resultados de muitos partidos pode variar por pequenas alterações nos factores analisados e, certamente, devido a inúmeros factores não estudados.

Existe um grande número de variáveis que influenciam a variável dependente em estudo e que não fazem parte dos modelos desenvolvidos, como sejam o comportamento do árbitro, a estabilidade psicológica, a motivação, o orçamento de cada equipa, os cantos, as grandes penalidades convertidas e falhadas, os momentos em que são marcados os golos, entre muitas outras.

As variáveis incluídas neste modelo são de fácil análise e percepção, pelo que os resultados podem ser facilmente entendidos por um leitor menos familiarizado com estes métodos estatísticos. A extrapolação desta aplicação para o estudo de outros casos é assim de fácil desenvolvimento.

BIBLIOGRAFIA

BARNETT, V.; LEWIS, T., "*Outliers in Statistical Data*", 3rd Edition, John Wiley & Sons, New York, 1994.

BENNETT, Jay (Editor), "*Statistics in Sport*", *Arnold Applications of Statistics Series*, Oxford University Press, London, 1998.

DRAPER, N.R.; SMITH, H., "*Applied Regression Analysis*", 2.nd Edition, John Wiley & Sons, New York, 1981.

FREEDMAN, D.A., "*A Note on Screening Regression Equations*", *The American Statistician*, 37, pp. 152-155, Alexandria, 1982.

GRAY, Philip K., "*Testing market efficiency: Evidence from the NFL sports betting market*", *The Journal of Finance*, 52 (4), pp. 1725-1737, Cambridge, 1997.

LADANY, S. P. e MACHOL, R. E. (Editores), "*Optimal Strategies in Sports*", North-Holland, New York, 1977.

INFORDESPORTO, *http://www.infordesporto.pt.*

KAHANE, Leo, "*Team roster turnover and attendance in major league baseball*", *Applied Economics*, 29 (4), p. 425, London, 1997.

MACHOL, R. E., LADANY, S. P. e MORRISON, D. G. (Editores), "*Management Science in Sports*", North-Holland, New York, 1976.

MILLER, A.J., "*Subset Selection in Regression*", Ed. Chapman and Hall, London, 1990.

NETER, J.; KUTNER, M.H.; NACHTSHEIM, C.J.; WASSERMAN, W., "*Applied Linear Statistical Models*", *Fourth Edition*, Irwin, Chicago, 1996.

PESTANA, M. Helena; GAGEIRO, João N., "*Análise de dados para Ciências Sociais - A complementaridade do SPSS*", Edição Revista, Edições Sílabo, Lisboa, 2000.

POPE, P.T.; WEBSTER, J.T., "*The Use of an F-Statistic in Stepwise Regression*", *Technometrics*, 14, pp. 327-340, 1972.

ROUSSEEUW, P.J.; LEROY, A.M., "*Robust Regression and Outlier Detection*", Ed. John Wiley & Sons, New York, 1987.

SMITH, Tyler, "*Can the NCAA Basketball tournament seeding be used to predict margin of victory?*", *The American Statistician*, 53 (2), pp. 94-98, Alexandria, 1999.

SZYMANSKI, S., SMITH, R., "*The English football industry: profit, performance and industrial structure*", *International Reviews of Applied Economics*, 11, pp. 135-154, 1997.

REVISTA DE ESTATÍSTICA
107ª PÁGINA
3º QUADRIMESTRE DE 2001

Classificação dos Fundos de Investimento Imobiliário: Aplicação da Análise de Clusters

Autor:
Raul M. Silva Laureano

REVISTA DE ESTATÍSTICA
109ª PÁGINA
3º QUADRIMESTRE DE 2001

CLASSIFICAÇÃO DOS FUNDOS DE INVESTIMENTO IMOBILIÁRIO: APLICAÇÃO DA ANÁLISE DE CLUSTERS

PROPERTY FUNDS CLASSIFICATION: AN APPLICATION OF CLUSTER ANALYSIS

Autor: Raul M. Silva Laureano
- Assistente, Investigador UNIDE - ISCTE

RESUMO:

- Sendo os fundos de investimento imobiliário, mais um produto financeiro à disposição dos investidores/aforradores, concorrendo com tantos outros produtos, então quanto mais e melhor se conhecer as suas características, nomeadamente a composição das carteiras, com mais confiança se pode investir capitais no produto, atendendo sempre ao perfil do investidor. Assim, torna-se indispensável a classificação deste tipo de fundos de investimento segundo a composição da sua carteira.

PALAVRAS-CHAVE:

- *Produtos financeiros, fundos de investimento, imobiliário, análise de carteiras, análise de clusters, análise de variância.*

ABSTRACT:

- Being the Property Funds another financial product available for investors, among so many financial products, the better one knows their characteristics, namely the portfolio composition, the higher the confidence of investors according to their profile. For this reason, it is important to classify these products according their portofolio composition.

KEY-WORDS:

- *Finacial Products, Investment Funds, Real Estate, Portfolio Analysis, Cluster Analysis, Analysis of Variance.*

1. INTRODUÇÃO

Dependendo o futuro de muitas famílias da boa aplicação das suas poupanças ao longo da vida, e sendo o investimento em imobiliário considerado um investimento sem risco a médio e longo prazo, então o investimento indirecto em imobiliário via fundos de investimento (imobiliário), pode ser um investimento interessante, atendendo sempre aos objectivos do investidor.

Pretende-se com o presente artigo apresentar para um produto financeiro, fundos de investimento imobiliário (FII), com características muito próprias, nomeadamente quanto ao risco assumido, uma metodologia que de certa forma, permita não só aos investidores e potenciais investidores neste tipo de produto, mas também às pessoas ligadas ao sector imobiliário, via fundos de investimento ou não, classificar os FII em diferentes categorias consoante a composição das respectivas carteiras.

Numa primeira parte procura-se de forma muito resumida apresentar os FII numa óptica financeira, ou seja, como produto financeiro, e não como "agente económico" exercendo a sua actividade no sector imobiliário, promovendo empreendimentos e dinamizando o mercado de arrendamento urbano. Posteriormente, explica-se como a análise de clusters permite detectar grupos de fundos homogéneos, ou seja, com carteiras semelhantes, aplicando-se esta técnica aos fundos abertos existentes em 31 de Dezembro de 1993. Por fim, aplicam-se outras técnicas estatísticas para validar os resultados obtidos, ou seja, para se verificar se as carteiras dos fundos de cada um dos grupo encontrados são, de facto, diferentes.

2. FUNDOS DE INVESTIMENTO IMOBILIÁRIO (FII)

"Os fundos de investimento imobiliário são instituições de investimento colectivo que têm por fim o investimento de capitais recebidos do público em carteiras diversificadas de valores fundamentalmente imobiliários, segundo um princípio de divisão de riscos"[1]. Os fundos são divididos em partes iguais, as unidades de participação, e podem ser abertos, se as unidades de participação são em número variável, ou podem ser fechados, caso contrário.

Os participantes, aforradores particulares ou entidades colectivas que investiram no fundo, têm direito à sua quota-parte nos rendimentos e nas mais-valias realizadas, que poderão ser objecto de distribuição periódica (fundos de distribuição) ou capitalizados no valor da unidade de participação (fundos de capitalização) em conjunto com a valorização em função dos preços de mercado ou de outros que se convencionem adequados, do conjunto de valores integrantes do património do fundo.

[1] Decreto-lei nº294/95 de 17 de Novembro

Ao rácio, calculado em cada momento, entre o património líquido do fundo e o número de unidades de participação em circulação, designa-se por valor patrimonial líquido (VPL). O VPL, como o seu nome indica, é líquido das despesas correntes suportadas pelo fundo.

A composição do património dos FII e suas limitações é apresentado no quadro 1.

Quadro nº 1. Composição do património e suas limitações

Limitações	Fundos abertos	Fundos fechados
Um mínimo de 6% do valor líquido global do fundo será constituído por numerário, depósitos bancários, títulos da dívida pública e aplicações nos mercados interbancários	Sim	
Os valores imobiliários[2] não podem representar menos de 75% do valor líquido global do fundo	Sim[3]	Sim[4]
Até 10% do seu valor líquido global, o património do fundo poderá ser constituído por terrenos destinados à execução de programas de construção	Sim[4]	Sim[4]
As participações superiores a 50% no capital de sociedades que se dediquem exclusivamente à aquisição, venda, arrendamento, gestão e exploração de imóveis não podem representar mais de 25% do valor líquido global do fundo	Sim[4]	
Não podem ser aplicados num único empreendimento mais de 20% do valor líquido global do fundo	Sim[4]	Sim[4]

3. OBTENÇÃO DE GRUPOS DE FUNDOS: ANÁLISE DE CLUSTERS

Para se detectar grupos homogéneos de fundos de investimento imobiliário quanto à composição da carteira, recorre-se a uma análise de clusters.

Esta análise é composta por um conjunto de procedimentos de estatística multivariada que visam organizar um conjunto de indivíduos (fundos) para os quais é conhecida informação detalhada (composição das carteiras), em grupos homogéneos (clusters), de tal modo que os fundos pertencentes a um mesmo grupo sejam tão

---

2 Consideram-se valores imobiliários os direitos de propriedade sobre bens imóveis que possam ser adquiridos para fundos e as participações superiores a 50% no capital social de sociedades cujo objecto seja exclusivamente aquisição, venda, arrendamento, gestão e exploração de imóveis.

3 As percentagens devem ser respeitadas a partir do início do terceiro exercício do fundo, devendo ser regularizadas no prazo máximo de um ano as situações de desconformidade resultantes da alteração dos valores venais dos bens ou do exercício do direito de reembolso pelos participantes dos fundos abertos.

semelhantes quanto possível e sempre mais semelhantes aos fundos do mesmo grupo do que aos fundos dos restantes grupos.

A análise de clusters implica a tomada de seis importantes decisões:

1. a selecção de fundos a serem agrupados;
2. a definição de um conjunto de variáveis a partir das quais será obtida a informação necessária ao agrupamento dos fundos;
3. a definição do método de análise;
4. a definição de uma medida de semelhança ou distância entre cada dois fundos;
5. a escolha de um critério de agregação ou desagregação dos fundos, isto é, a definição de um algoritmo de classificação;
6. a validação dos resultados encontrados.

1ª decisão

Uma vez que o objectivo deste estudo é encontrar grupos de fundos com carteiras semelhantes em 31 de Dezembro de 1993 de modo a facilitar ao aforrador a selecção do fundo em que vai investir, é aconselhável agrupar apenas os fundos abertos, pois são aqueles que se encontram mais acessíveis aos aforradores.

2ª decisão

Para caracterizar as carteiras dos fundos utilizou-se as seguintes treze variáveis, referentes a 31 de Dezembro de 1993:

- valor total das aplicações em carteira (milhões de contos);
- % da carteira aplicada em imobiliário (%);
- % da componente mobiliária da carteira aplicada em numerário e depósitos bancários (%);
- % da componente mobiliária da carteira aplicada em títulos de dívida pública de curto prazo (%);
- % da componente mobiliária da carteira aplicada em títulos de dívida pública de médio e longo prazo (%);
- % da componente mobiliária da carteira aplicada em outros produtos (%);
- % da componente imobiliária da carteira aplicada em terrenos (%);
- % da componente imobiliária da carteira aplicada em construções acabadas (%);
- % da componente imobiliária da carteira aplicada em construções em curso (%);
- idade do fundo (meses);
- número de distritos em que o fundo tem imóveis (unidades);

- % da componente imobiliária da carteira aplicada nos três maiores distritos em que o fundo tem investimentos (%);
- % da componente imobiliária da carteira aplicada no maior distrito em que o fundo tem investimentos (%).

Como nem todas as variáveis se encontram definidas na mesma unidade, a análise estava sujeita a que as variáveis com maiores valores tivessem um maior peso. Para evitar esta situação estandardizaram-se as variáveis anteriormente referidas.

3ª decisão

Existem vários métodos que podem ser utilizados na formação dos grupos, entre os quais se destacam os métodos de optimização-partição e os métodos hierárquicos que se subdividem em aglomerativos e divisivos. O método escolhido é o método hierárquico aglomerativo, que considera um número n de indivíduos, relativamente aos quais dispomos de informações sobre variáveis que os caracterizam, informações essas que são processadas parcialmente ou não com base em determinadas técnicas de aglomeração, de modo a serem obtidos conjuntos tão homogéneos quanto possível.

No início da formação de grupos, admite-se que cada indivíduo constitui um grupo. Posteriormente, são agregados os indivíduos em função de determinados procedimentos seleccionados, terminando o processo com a formação de um grupo que compreende a totalidade dos indivíduos. Este método implica a utilização de uma medida de distância entre os indivíduos e um critério de associação dos mesmos em grupos.

4ª decisão

A medida de distância escolhida foi o quadrado da distância Euclidiana, que traduz a distância entre dois indivíduos i e j e que é igual ao somatório dos quadrados das diferenças entre valores de i e j para todas as variáveis.

5ª decisão

Escolhida a medida de distância, surge o problema do critério de agregação dos indivíduos a seleccionar. Não existindo um critério ideal, a solução é utilizar vários critérios e comparar os resultados obtidos. Se estes forem semelhantes, conclui-se que se obtiveram grupos homogéneos com elevado grau de confiança. O critério adoptado para apresentar os resultados foi Complete Linkage ou critério do vizinho mais afastado que consiste em agrupar dois grupos num só, de acordo com a distância entre os seus elementos mais afastados ou menos semelhantes.

6ª decisão

Por fim surge o problema da validação dos resultados obtidos, ou seja, é necessário escolher o número óptimo de grupos a partir do dendrograma e da árvore de agrupamento resultantes do método escolhido.

Figura nº 1. Dendrograma

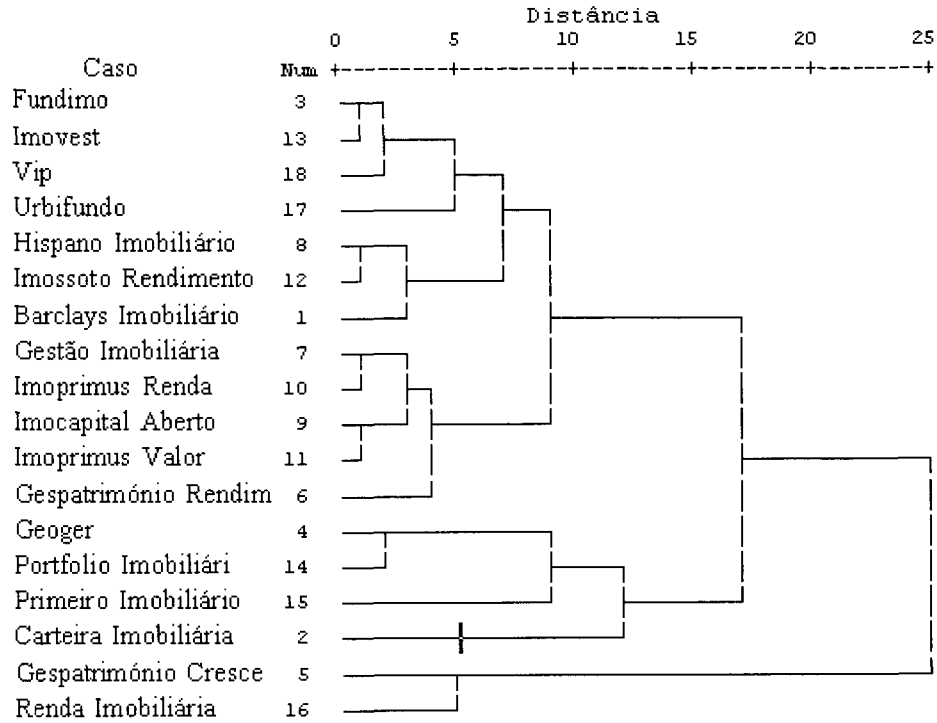
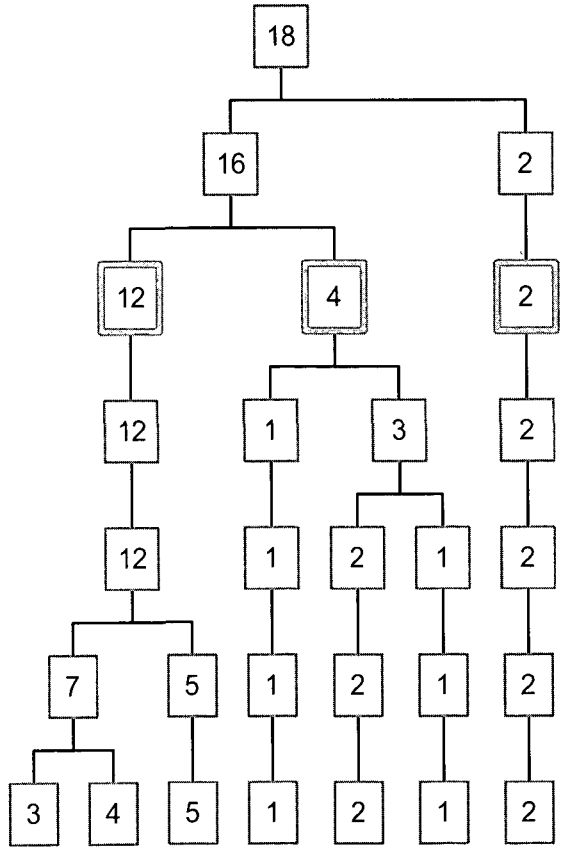


Figura nº 2. Árvore de agrupamento



Pelo conhecimento dos fundos em análise e pelo que se pode retirar do dendrograma e da árvore de agrupamento, o número de grupos homogéneos a reter é de três.

Quadro nº 2. Composição dos grupos homogéneos

Grupo 1 12 fundos	Grupo 2 4 fundos	Grupo 3 2 fundos
Fundimo	Geoger	Gespatrimónio Crescente
Imovest	Portfólio Imobiliário	Renda Imobiliária
Vip	Primeiro Imobiliário	
Urbifundo	Carteira Imobiliária	
Hispano Imobiliário		
Imosotto Rendimento		
Barclays Imobiliário		
Gestão Imobiliária		
Imoprimus Renda		
Imocapital Aberto		
Imoprimus Valor		
Gespatrimónio Rendimento		

4. VALIDAÇÃO DOS GRUPOS

Para se ter a certeza de que os três grupos de fundos são de facto homogéneos, ou seja, são compostos por fundos com composições de carteiras semelhantes, é necessário proceder a uma análise estatística: análise de variância (Teste F)[4] e, no caso de não ser aplicável por falta de verificação dos seus pressupostos, o Teste Kruskal-Wallis. Esta análise visa concluir se as médias das variáveis, consideradas na análise de clusters, para cada grupo considerado, assumem valores significativamente diferentes. As hipóteses a considerar neste Teste F (igualdade de médias) são:

$$Ho: \mu_1 = \mu_2 = \mu_3$$

$$H1: \mu_i \neq \mu_j \qquad \text{para algum par } (i, j) \text{ com } i \neq j$$

em que cada um dos μ é a média de cada grupo no que se refere à variável em estudo.

No entanto, esta análise implica a verificação de dois pressupostos:

- os grupos tenham sido tirados de populações com as mesmas variâncias (Teste Levene);

[4] Como os grupos formados a partir da análise de clusters não constituem amostras de indivíduos escolhidos aleatoriamente e de forma independente, os resultados obtidos para as probabilidades são aproximados

- as populações tenham distribuição normal ou aproximadamente normal (Teste Kolmogorov-Smirnov).

Para todos os testes, considera-se que:

não rejeito Ho $\Rightarrow$ Probabilidade do teste acontecer $> \alpha$ com $\alpha = 0.05$

rejeito Ho $\Rightarrow$ Probabilidade do teste acontecer $< \alpha$

em que α representa o nível de significância.

Por outro lado, serão interpretados os valores do teste Scheffe, cuja robustez é considerada superior na avaliação das diferenças estatisticamente significativas das médias entre dois grupos. Este teste indica quais os pares de grupos que têm média significativamente diferente para uma determinada variável.

Os resultados obtidos para as treze variáveis (não estandardizadas) caracterizadoras da composição das carteiras dos fundos, são os constantes do quadro 3.

Quadro nº 3. Validação dos grupos

Variável	**Teste F** **Prob: 0.000 < α = 0.05**	**Diferenças entre grupos**
valor total das aplicações em carteira (milhões de contos)	Médias: G1 23.9545 G2 3.4189 G3 236.1558	G1 G2 G3 G1 G2 G3 * *
Variável	**Teste F** **Prob:0.0024< α = 0.05**	**Diferenças entre grupos**
% da carteira aplicada em imobiliário (%)	Médias: G1 63.4140 G2 87.1010 G3 0.0000	G1 G2 G3 G1 * G2 * G3
Variável	**Teste F** **Prob:0.0003 < α = 0.05**	**Diferenças entre grupos**
% da componente mobiliária da carteira aplicada em numerário e depósitos bancários (%)	Médias: G1 87.3859 G2 20.6522 G3 90.3453	G1 G2 G3 G1 * G2 G3 *

Variável	Teste F Prob:0.0001 < α = 0.05	Diferenças entre grupos
% da componente mobiliária da carteira aplicada em títulos de dívida pública C/prazo (%)	Médias: G1 5.1821 G2 70.6014 G3 0.0934	G1 G2 G3 G1 G2 * * G3

Variável	Teste F Prob:0.9598 > α = 0.05	Diferenças entre grupos
% da componente mobiliária da carteira aplicada em títulos de dívida pública M/L prazo (%)	Médias: G1 7.1421 G2 8.7464 G3 6.8787	G1 G2 G3 G1 G2 G3

Variável	Teste F Prob:0.0670 >α = 0.05	Diferenças entre grupos
% da componente mobiliária da carteira aplicada em outros produtos (%)	Médias: G1 0.2899 G2 0.0000 G3 2.6826	G1 G2 G3 G1 G2 G3

Variável	Teste F Prob:0.0839 >α = 0.05	Diferenças entre grupos
% da componente imobiliária da carteira aplicada em terrenos (%)	Médias: G1 2.8621 G2 31.8865 G3 0.0000	G1 G2 G3 G1 G2 G3

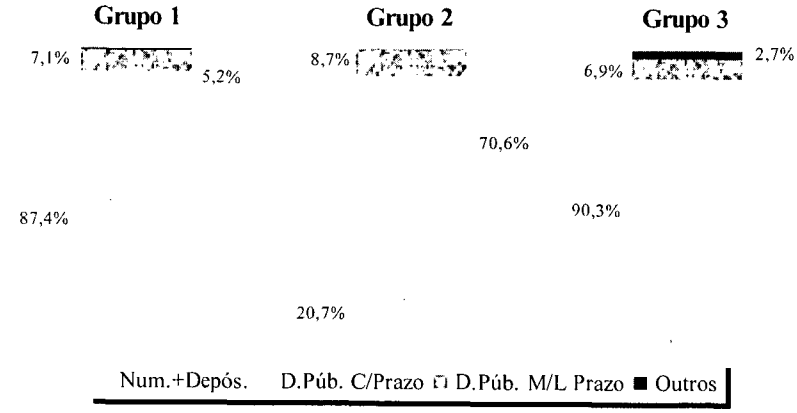
Variável	Teste F Prob:0.0001 < α = 0.05	Diferenças entre grupos
% da componente imobiliária da carteira aplicada em construções acabadas (%)	Médias: G1 92.2009 G2 39.4672 G3 0.0000	G1 G2 G3 G1 * * G2 G3

Variável	**Teste F** **Prob:0.0440 < α = 0.05**	**Diferenças entre grupos**
% da componente imobiliária da carteira aplicada em construções em curso (%)	Médias: G1 4.9371 G2 28.6463 G3 0.0000	G1 G2 G3 G1 G2 G3
Variável	**Teste F** **Prob:0.3252 > α = 0.05**	**Diferenças entre grupos**
Idade do fundo (meses)	Médias: G1 38.8333 G2 37.7500 G3 9.0000	G1 G2 G3 G1 G2 G3
Variável	**Teste F** **Prob:0.0649 > α = 0.05**	**Diferenças entre grupos**
Número de distritos em que o fundo tem imóveis (unidades)	Médias: G1 8.9167 G2 2.0000 G3 0.0000	G1 G2 G3 G1 * G2 * G3
Variável	**Teste F** **Prob: 0.000 < α = 0.05**	**Diferenças entre grupos**
% da componente imobiliária da carteira aplicada nos três maiores distritos (%)	Médias: G1 88.9633 G2 100.0000 G3 0.0000	G1 G2 G3 G1 * G2 * G3
Variável	**Teste F** **Prob: 0.0007< α = 0.05**	**Diferenças entre grupos**
% da componente imobiliária da carteira aplicada no maior distrito (%)	Médias: G1 67.2308 G2 83.3250 G3 0.000	G1 G2 G3 G1 * G2 * G3

Assim, verifica-se que, em termos médios:

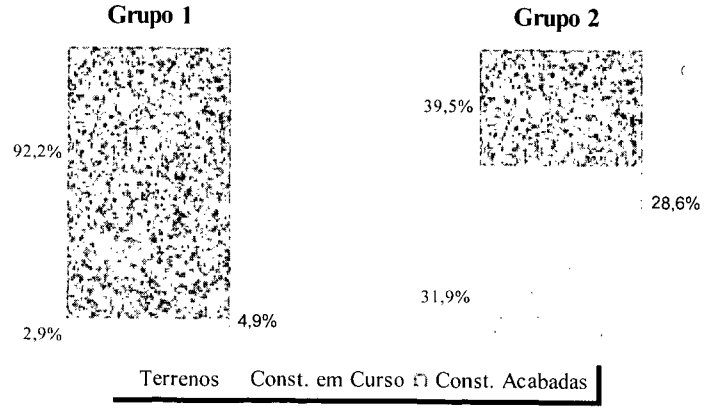
- o grupo 3 é composto pelos dois maiores fundos, apresentando uma média para o valor total das aplicações em carteira de cerca de 236 milhões de contos. O grupo 1 tem em média cerca de 24 milhões de contos e o grupo 2 apenas cerca de 3.4 milhões de contos;
- quanto à componente imobiliária, esta não existe no grupo 3. O grupo 2 é o único que satisfaz o rácio de estrutura que diz que os valores imobiliários não podem representar menos de 75% do valor líquido global do fundo, apresentando o rácio o valor de 87.1%. O grupo 1 tem cerca de 63.4% do total das aplicações investidos em valores imobiliários;
- no que respeita à componente mobiliária da carteira, o gráfico nº 1 demonstra a situação de cada um dos grupos. Nos grupos 1 e 3 são os depósitos bancários que predominam, enquanto no grupo 2 os títulos de dívida pública de curto prazo são os preferidos.

Gráfico nº 1. Carteira mobiliária por grupos homogéneos



- A carteira imobiliária do grupo 1 é constituída, quase na sua totalidade por construções acabadas, enquanto a carteira do grupo 2 encontra-se repartida pelos três estádios de desenvolvimento do produto imobiliário.

Gráfico nº 2. Carteira imobiliária por grupos homogéneos



- quanto à diversificação geográfica das aplicações imobiliárias, o grupo 1 está representado em 9 distritos, representando os 3 maiores distritos cerca de 89% do total das aplicações imobiliárias e o maior distrito cerca de 67%. O grupo 2 tem aplicações imobiliárias em apenas dois distritos, representando o maior cerca de 83%.
- por fim os fundos que constituem os grupos 1 e 2 estão em actividade há cerca de 38 meses, enquanto os dois fundos do grupo 3 apenas estão em actividade há 9 meses.

Resumindo as carteiras de cada um dos grupos temos:

- o grupo 1 é um grupo de dimensão média e em que a componente imobiliária se aproxima dos 65%. Em termos de aplicações mobiliárias é dada nítida preferência aos depósitos bancários (87%) em detrimento dos títulos de dívida pública. Em termos de imobiliário 92% estão aplicados em construções acabadas, tendo uma diversificação geográfica bastante elevada.
- o grupo 2 é um grupo de pequena dimensão com uma componente imobiliária de cerca de 87%. Este grupo dá preferência aos títulos de dívida pública para as suas aplicações mobiliárias (cerca de 79%). É um grupo com forte potencialidade futura visto que cerca de 60% da componente imobiliária estão aplicados em terrenos e em construções em curso, bens estes que não geram receitas a curto prazo. Em termos geográficos é um grupo pouco diversificado.
- o grupo 3 é o grupo dos fundos «mobiliários», ou seja, não tem bens imóveis em carteira, o que poderá ser justificado pelos poucos meses em que se encontram em actividade os dois fundos que constituem este grupo. Dos cerca de 236 milhões de contos que cada fundo gere em média 90% estão aplicados em depósitos bancários.

5. CONCLUSÕES

Procurou-se com o presente artigo apresentar uma metologia, análise de clusters, tendo em vista a classificação dos fundos de investimento imobiliário abertos segundo a composição das suas carteiras.

Utilizou-se treze variáveis para caracterizar as carteiras, em 31 de Dezembro de 1993, desde o valor total das aplicações, percentagens aplicadas em cada componente mobiliária e imobiliária, concentração geográfica da componente imobiliária até à idade do fundo.

Detectaram-se três grupos de fundos em que as respectivas carteiras possuem características diferentes, provavelmente proseguindo objectivos de gestão, também, diferentes. E quanto às rendibilidades, serão elas também diferentes?

Sendo objectivo de qualquer investidor gerir o binómio risco/rendibilidade, procurando maximizar a rendibilidade para um determinado risco que pretende assumir, então não interessa saber apenas que existem fundos com carteiras diferentes, interessa sim saber o que se pode esperar de cada tipo de carteira em termos de rendibilidade e risco.

É o que se passa com os fundos de investimento mobiliário, que já se encontram classificados segundo as suas carteiras em, por exemplo, fundos de acções, de obrigações, de tesouraria e por aí fora. Classificações estas que indicam à partida ao investidor o que pode esperar ao aplicar as suas poupanças neste ou naquele fundo.

Para outro artigo fica o estudo das rendibilidade dos fundos e da relação que poderá existir entre a composição da carteira e a respectiva rendibilidade. Ficará, também, para outra oportunidade estudar a estabilidade temporal dos grupos, sabendo desde já que o grupo composto pelos fundos sem componente imobiliária não poderá se manter, pois a legislação não permite fundos sem componente imobiliária a não ser no início da actividade.

BIBLIOGRAFIA

GONÇALVES, MANUEL VALADAS, (1998), "Fundos de Investimento Imobiliário", Editora Rei dos Livros.

LAUREANO, RAUL M. S., (1995), "Fundos de Investimento Imobiliário em Portugal em 1993", Dissertação de Mestrado, ISCTE, Lisboa.

PEIXOTO, ÁLVARO. C. e ANTUNES, P. E., (1994), "Guia do Investidor em Fundos de Investimento", APDMC - Associação Portuguesa para o Desenvolvimento do Mercado de Capitais, Junho.

PESTANA, MARIA HELENA e GAGEIRO, JOÃO NUNES, (2000), "Análise de Dados para Ciências Sociais: A Complementaridade do SPSS", Edições Sílabo, 2ªEdição Revista e aumentada.

REIS, ELIZABETH, (2000), "A Análise de Clusters e as Aplicações às Ciências Empresariais: Uma visão Crítica da Teoria dos Grupos Estratégicos", in Reis, Elizabeth e Ferreira, Manuel (Ed.), Temas em Métodos Quantitativos, nº1, Edições Sílabo, pp.205-238.

REIS, ELIZABETH, (2001), "Estatística Multivariada Aplicada", Edições Silabo, 2ª Edição revista e corrigida.

RELATÓRIOS DE ACTIVIDADE e CONTAS ANUAIS das Sociedades Gestoras de Fundos de Investimento Imobiliário e dos respectivos fundos.

REGULAMENTOS de gestão dos Fundos de Investimento Imobiliário.

DECRETO-LEI nº 294/95 de 17 de Novembro.

DECRETO-LEI nº 323/97 de 26 de Novembro.

ANEXO

Variáveis utilizadas na análise de clusters

Fundo	Total aplicações (milhões de contos)	% carteira aplicada em imobiliário	Componente mobiliária			
			% numerário e depósitos bancários	% títulos dívida pública curto prazo	% títulos dívida pública médio e longo prazo	% outros produtos
Barclays Imobiliário	55,79601	1,58%	97,15%	0,00%	2,85%	0,00%
Carteira Imobiliária	2,392576	90,82%	1,26%	72,57%	26,17%	0,00%
Fundimo	33,97689	73,60%	100,00%	0,00%	0,00%	0,00%
Geoger	2,018574	94,69%	5,42%	94,58%	0,00%	0,00%
Gespatrimónio Crescente	195,1385	0,00%	81,25%	0,00%	13,39%	5,37%
Gespatrimónio Rendimento	23,54675	88,87%	69,43%	0,00%	30,57%	0,00%
Gestão Imobiliária	17,5492	46,35%	82,97%	0,00%	17,03%	0,00%
Hispano Imobiliário 2	1,240036	73,99%	100,00%	0,00%	0,00%	0,00%
Imocapital Aberto	34,25316	70,19%	97,48%	0,00%	2,52%	0,00%
Imoprimus Renda	19,50788	22,81%	85,56%	0,00%	14,44%	0,00%
Imoprimus Valor	42,95262	77,25%	87,59%	0,00%	12,41%	0,00%
Imossoto Rendimento	4,127536	64,22%	89,43%	10,57%	0,00%	0,00%
Imovest	19,46365	71,70%	98,18%	0,00%	1,82%	0,00%
Portfolio Imobiliário	6,053029	90,34%	0,14%	91,04%	8,82%	0,00%
Primeiro Imobiliário	3,211355	72,55%	75,78%	24,22%	0,00%	0,00%
Renda Imobiliária	277,1731	0,00%	99,44%	0,19%	0,37%	0,00%
Urbifundo	10,74903	93,25%	96,52%	0,00%	0,00%	3,48%
Vip	24,29099	77,16%	44,33%	51,62%	4,05%	0,00%

Variáveis utilizadas na análise de clusters (continuação)

Fundo	Componente imobiliária			idade (meses)	Componente imobiliária		
	% terrenos	% construções acabadas	% contruções em curso		Nº distritos com imóveis	% nos 3 maiores distritos	% no maior distrito
Barclays Imobiliário	0,00%	94,86%	5,14%	20	4	92,21%	60,04%
Carteira Imobiliária	100,00%	0,00%	0,00%	11	1	100,00%	100,00%
Fundimo	0,00%	99,47%	0,53%	80	8	95,74%	87,39%
Geoger	0,00%	86,34%	13,66%	75	3	100,00%	65,63%
Gespatrimónio Crescente	0,00%	0,00%	0,00%	7	0	0,00%	0,00%
Gespatrimónio Rendimento	0,00%	100,00%	0,00%	19	19	74,58%	63,01%
Gestão Imobiliária	12,26%	84,68%	3,07%	24	16	80,39%	52,35%
Hispano Imobiliário 2	0,00%	100,00%	0,00%	15	2	100,00%	62,04%
Imocapital Aberto	10,61%	77,47%	11,92%	37	18	80,62%	61,37%
Imoprimus Renda	0,00%	100,00%	0,00%	6	12	80,62%	52,71%
Imoprimus Valor	0,70%	99,30%	0,00%	27	16	69,90%	26,51%
Imossoto Rendimento	0,00%	100,00%	0,00%	35	1	100,00%	100,00%
Imovest	0,00%	88,04%	11,96%	78	7	93,50%	75,88%
Portfolio Imobiliário	0,00%	71,53%	28,47%	26	3	100,00%	67,67%
Primeiro Imobiliário	27,55%	0,00%	72,45%	39	1	100,00%	100,00%
Renda Imobiliária	0,00%	0,00%	0,00%	11	0	0,00%	0,00%
Urbifundo	6,70%	67,25%	26,05%	51	3	100,00%	65,47%
Vip	4,08%	95,35%	0,57%	74	1	100,00%	100,00%

REVISTA DE ESTATÍSTICA
125ª PÁGINA
3º QUADRIMESTRE DE 2001

Pós-estratificação e Without Replacement Bootstrap: Aplicações ao Inquérito às Empresas/Harmonizado

Autora:
Ana Cristina M. Costa

REVISTA DE ESTATÍSTICA
127ª PÁGINA
3º QUADRIMESTRE DE 2001

PÓS-ESTRATIFICAÇÃO E WITHOUT REPLACEMENT BOOTSTRAP: APLICAÇÕES AO INQUÉRITO ÀS EMPRESAS/HARMONIZADO

POST-STRATIFICATION AND WITHOUT REPLACEMENT BOOTSTRAP APPROACHES IN A BUSINESS SURVEY

Autora: Ana Cristina M. Costa
- Assistente no Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa

RESUMO:

- Neste artigo analisam-se alguns métodos de estimação que visam o tratamento dos problemas da base de sondagem e da ocorrência de não respostas nos inquéritos por amostragem. Estes erros não amostrais têm repercussões nas estimativas obtidas, uma vez que as propriedades dos estimadores se deterioram. Na inferência clássica das sondagens, são de destacar os estimadores de pós-estratificação. As propriedades teóricas destes estimadores requerem ainda alguma investigação para planos de sondagem complexos, embora estes métodos sejam frequentemente utilizados. É então abordada a metodologia *Bootstrap* para a estimação da variância dos estimadores. Apresentam-se algumas aplicações, dos métodos de pós-estratificação e do algoritmo *Without Replacement Bootstrap* (*BWO*), proposto por Sitter (1992), aos dados do Inquérito às Empresas/Harmonizado (IEH) de 1997, conduzido pelo Instituto Nacional de Estatística. Os estimadores considerados são discutidos em termos de precisão, sendo efectuadas recomendações quanto às suas aplicações no âmbito do IEH.

PALAVRAS-CHAVE:

- *Pós-estratificação; problemas da base de sondagem; não resposta; reponderação; métodos de ajustamento; Bootstrap.*

ABSTRACT:

- This paper approaches issues related with frame problems and nonresponse in surveys. These nonsampling errors affect the accuracy of the estimates whereas the estimators become biased and less precise. We analyse some estimation methods that deal with those problems, in the design-based perspective, and give an especial focus to the poststratification procedures. For complex sampling designs the theoretical properties of the estimators need further research. We then address the Bootstrap methodology for variance estimation. Some applications of the poststratification estimators and the *Without Replacement Bootstrap* (*BWO*) algorithm, proposed by Sitter (1992), are also presented, using data from the 1997 Annual Business Survey, conducted by Portugal's National Statistics Institute. The

precision of the analysed estimators is discussed and some recommendations are made regarding their applications under this survey.

KEY-WORDS:

- *Poststratification; frame problems; nonresponse; reweighting; adjustment methods; Bootstrap.*

1. INTRODUÇÃO

O Inquérito às Empresas / Harmonizado (IEH), conduzido pelo Instituto Nacional de Estatística de Portugal (INE), é realizado anualmente e tem cobertura nacional. O desenho do IEH corresponde a um plano de amostragem aleatória estratificada sem reposição. As estimativas dos totais das diversas variáveis de interesse têm sido obtidas através do estimador de Horvitz-Thompson. A base de amostragem é constituída a partir do Ficheiro Geral de Unidades Estatísticas (FGUE) do INE.

A aplicação dos métodos de pós-estratificação ao IEH é essencialmente motivada pelo problema das *mudanças de estrato*. As respostas obtidas no inquérito sugerem que determinadas empresas não se mantêm nos estratos iniciais. Este problema resulta da informação auxiliar que constava do FGUE, e que serviu de base à estratificação, se encontrar desactualizada ou incorrecta. Por outro lado, o IEH apresenta também não resposta total. Os problemas da base de sondagem e também as não respostas têm repercussões nas estimativas obtidas, uma vez que as propriedades dos estimadores se deterioram.

O presente artigo surge na sequência da investigação efectuada no âmbito de um projecto que teve por principal objectivo identificar métodos de estimação que permitissem lidar com o problema das *mudanças de estrato*.

Os estimadores de pós-estratificação ajustam o coeficiente de extrapolação de cada elemento da amostra, por forma a que esta reflicta a estrutura actual da população e tenha também em conta a ocorrência de não respostas. Estes estimadores inserem-se numa classe de métodos de tratamento de não respostas usualmente designados por métodos de recomposição ou métodos de ajustamento. Espera-se, portanto, que com estes métodos seja possível melhorar as estimativas das diversas variáveis de interesse.

Na inferência clássica das sondagens (*design-based*), as propriedades teóricas dos estimadores de pós-estratificação requerem ainda alguma investigação para planos de sondagem complexos; sendo de salientar a abordagem condicional efectuada por Rao (1985). É então abordada a metodologia *Bootstrap* para a estimação da variância dos estimadores.

Apresentam-se e discutem-se os resultados obtidos através da aplicação, aos dados do IEH de 1997, de diversos métodos de ajustamento e do algoritmo *Without Replacement Bootstrap* (*BWO*), proposto por Sitter (1992). Os estimadores considerados são comparados em termos de precisão, sendo efectuadas recomendações quanto às suas aplicações no âmbito do IEH.

2. ENQUADRAMENTO TEÓRICO

2.1. ESTIMADORES DE PÓS-ESTRATIFICAÇÃO

Como o próprio nome indica, a pós-estratificação consiste em estratificar a amostra depois de esta ter sido recolhida, utilizando informação auxiliar que se encontre disponível na fase de estimação. Naturalmente, tal como nos planos de sondagem aleatória estratificada, os pós-estratos devem ser o mais homogéneos possível e, portanto, a variável que define os pós-estratos deverá estar fortemente correlacionada com as variáveis de interesse. Nos métodos de pós-estratificação, assume-se que as dimensões dos pós-estratos na população são conhecidas. Estes métodos consistem, então, em ajustar os pesos iniciais por forma a que a distribuição da amostra reponderada, para certas características da população, esteja de acordo com a distribuição conhecida do número de elementos da população com essas características.

Quando se recorre a duas ou mais variáveis auxiliares para pós-estratificar a amostra, podem ocorrer duas situações. Se a dimensão de todos os pós-estratos resultantes (do cruzamento dessas variáveis) for conhecida na população, o problema reduz-se ao caso em que se utiliza apenas uma variável de pós-estratificação e, portanto, os métodos de pós-estratificação são directamente aplicáveis.

No entanto, tal informação nem sempre se encontra disponível. Por vezes, dispõem-se apenas das dimensões marginais na população. Ou seja, a única informação auxiliar que existe diz respeito à dimensão da população nas categorias definidas por cada uma das variáveis, tomadas isoladamente. Para lidar com este problema, Deming e Stephan (1940) introduziram um método designado *raking ratio*. Posteriormente, Deville e Särndal (1992) desenvolveram uma família de *estimadores de calibração*, onde os pesos iniciais são ajustados através de um conjunto de *equações de calibração*. Por forma a que os pesos ajustados se aproximem o mais possível dos pesos de inclusão, é escolhida uma *função distância*. Uma extensão destes métodos, designada *generalized raking*, deve-se a Deville, Särndal e Sautory (1993). O método *raking ratio* proposto por Deming e Stephan constitui um caso particular destes métodos.

Os métodos de pós-estratificação têm sido analisados sob diversos pontos de vista por vários autores, veja-se por exemplo: Williams (1962); Holt e Smith (1979); Rao (1985); Valliant (1993); Leonard *et al.* (1994) e Rao (1994).

Em seguida, apresenta-se de forma abreviada a abordagem considerada por Williams (1962) e Rao (1985), sobre a forma dos estimadores de pós-estratificação para um plano de sondagem genérico e para o plano de sondagem aleatória estratificada.

Suponhamos que, na fase de estimação, se dispõe de informação auxiliar que permita dividir a amostra em L pós-estratos. Sejam $n_1, \ldots, n_i, \ldots, n_L$ as dimensões amostrais dos pós-estratos e s_i o conjunto dos elementos da amostra que pertencem ao pós-estrato i (i=1, ..., L). Uma vez que a estratificação da amostra é efectuada depois

de esta ter sido recolhida, as dimensões amostrais dos pós-estratos são variáveis aleatórias, contrariamente à amostragem estratificada convencional. Nos métodos de pós-estratificação assume-se, também, que as dimensões $N_1, \ldots, N_i \ldots, N_L$ dos pós-estratos na população são conhecidas.

Para um plano genérico de amostragem, um **estimador de pós-estratificação do total da população**, τ, é dado por:

$$\hat{\tau}_{PS} = \sum_{i=1}^{L} \sum_{k \in s_i} \frac{N_i}{\hat{N}_i} w_k y_k \qquad (1)$$

Sendo $w_k = 1/\pi_k$ o peso de inclusão (ou coeficiente de extrapolação) do indivíduo k, subjacente ao desenho da amostra (*design-weight*) e $\hat{N}_i$ o usual estimador centrado de domínios da dimensão do *i-ésimo* pós-estrato:

$$\hat{N}_i = \sum_{k \in s_i} \frac{1}{\pi_k} \qquad (2)$$

Esta forma de apresentar o estimador de pós-estratificação permite-nos evidenciar o ajustamento pelo quociente dos pesos iniciais w_k.

Sendo N a dimensão da população, um **estimador de pós-estratificação da média da população**, μ, é dado, obviamente, por:

$$\hat{\mu}_{ps} = \frac{1}{N} \hat{\tau}_{PS} \qquad (3)$$

Em seguida, apresentam-se em mais detalhe os estimadores de pós-estratificação para o plano de sondagem aleatória estratificada.

2.1.1. SONDAGEM ALEATÓRIA ESTRATIFICADA

Seja U a população em estudo de dimensão N conhecida. Suponhamos que foi retirada de U uma amostra aleatória s, de dimensão n, através de um plano de sondagem aleatória estratificada, $s = (s_1, \ldots, s_h, \ldots, s_H)$, no qual foi utilizada a sondagem aleatória simples sem reposição em cada estrato.

Suponhamos que, na fase de estimação, se dispõe de informação auxiliar que permita dividir a amostra s em L pós-estratos, definidos por forma a que sejam o mais homogéneos possível. Como se referiu anteriormente, supõe-se que as dimensões dos pós-estratos na população são conhecidas.

Neste caso, os estratos iniciais podem cruzar os pós-estratos e, portanto, as dimensões amostrais resultantes da intersecção dos estratos iniciais com os pós-estratos são aleatórias.

Antes de apresentarmos o estimador de pós-estratificação genérico (1), para o plano de sondagem em análise, vamos considerar alguma notação adicional. Seja $N_{\bullet h}$ a dimensão do estrato inicial *h* na população (h=1, ..., H); $N_{i\bullet}$ a dimensão do pós-estrato *i* na população (i=1, ..., L); $n_{\bullet h}$ o número de elementos da amostra pertencentes ao estrato inicial *h* (h=1, ..., H); s_{ih} o conjunto de elementos da amostra que pertencem simultaneamente ao estrato inicial *h* (h=1, ..., H) e ao pós-estrato *i* (i=1, ..., L); e n_{ih} a dimensão (aleatória) de s_{ih}.

Um **estimador de pós-estratificação de** τ, para o plano de sondagem aleatória estratificada, é dado por:

$$\hat{\tau}_{ps,str} = \sum_{i=1}^{L}\sum_{h=1}^{H}\sum_{k\in s_{ih}} \frac{N_{i\bullet}}{\hat{N}_{i\bullet}}\frac{N_{\bullet h}}{n_{\bullet h}} y_k \qquad (4)$$

com $\hat{N}_{i\bullet}$ dado por

$$\hat{N}_{i\bullet} = \sum_{h=1}^{H}\frac{N_{\bullet h}}{n_{\bullet h}} n_{ih} \qquad (5)$$

Rao (1985) apresenta um caso particular do estimador $\hat{\tau}_{ps,str}$ em que se consideram apenas H=2 estratos iniciais e L=2 pós-estratos, com o objectivo de ilustrar como é difícil investigar as propriedades condicionais do estimador (1) numa sondagem complexa. Mesmo para esta situação simples, Rao (1985) demonstra que o valor esperado do estimador (1), i.e. o valor esperado do estimador (4), condicionado sobre as dimensões amostrais observadas nos pós-estratos ($n_{1\bullet}$, $n_{2\bullet}$) não é tratável na abordagem condicional.

Williams (1962) sugeriu um estimador da variância do estimador de pós-estratificação genérico (1) que não revela boas propriedades na abordagem condicional, mesmo no caso em que o desenho da amostra corresponde a um plano de sondagem aleatória simples sem reposição, tal como demonstra Rao (1985). Este autor propõe um estimador alternativo que pode ser preferível tanto na abordagem condicional, como não condicional.

Denote-se por $\hat{V}(\hat{\tau}_{\pi}) = v(y_k)$ a função que define o estimador da variância do estimador usual de τ (*estimador de Horvitz-Thompson* ou *estimador-*π). Ou seja, no caso da sondagem aleatória estratificada sem reposição, tem-se

$$v(y_k) = \sum_{h=1}^{H} N_{\bullet h}^2 \left(\frac{1}{n_{\bullet h}} - \frac{1}{N_{\bullet h}}\right) s_h^2 = \hat{V}(\hat{\tau}_{\pi_{str}}) \qquad (6)$$

onde,

$$s_h^2 = \frac{1}{n_{\bullet h} - 1} \sum_{k \in s_h} (y_k - \bar{y}_h)^2 \qquad (7)$$

O estimador da variância de $\hat{\tau}_{PS}$ proposto por Rao (1985), para um plano de sondagem genérico, e que denotar-se-á por $\hat{V}_{rao}(\hat{\tau}_{ps})$, é dado por:

$$\hat{V}_{rao}(\hat{\tau}_{ps}) = v(z_k) \qquad (8)$$

onde, $v(z_k)$ se obtém a partir de $\hat{V}(\hat{\tau}_\pi)$ substituindo-se y_k por:

$$z_k = \sum_i \frac{N_i}{\hat{N}_i} (y_{ik} - \frac{\hat{\tau}_i}{\hat{N}_i} \mathbb{1}_{k \in s_i}) = \sum_i \frac{N_i}{\hat{N}_i} (y_{ik} - \hat{\mu}_i \mathbb{1}_{k \in s_i}) \qquad (9)$$

sendo, $\mathbb{1}_{k \in s_i}$ a variável indicatriz que toma o valor 1 se o elemento *k* pertence ao pós-estrato *i* e toma o valor zero no caso contrário;

$$y_{ik} = y_k \mathbb{1}_{k \in s_i} = \begin{cases} y_k, & \text{se k pertence ao pós-estrato i} \\ 0, & \text{caso contrário} \end{cases} \qquad (10)$$

e $\hat{\mu}_i = \hat{\tau}_i / \hat{N}_i$, sendo $\hat{N}_i$ dado por (2) e

$$\hat{\tau}_i = \sum_{k \in s_i} \frac{y_k}{\pi_k} \qquad (11)$$

As propriedades deste estimador são, também, difíceis de investigar. No caso mais simples do plano de sondagem aleatória simples sem reposição o estimador (8) conduz a um estimador da variância condicionalmente válido, dadas as dimensões amostrais dos pós-estratos (Rao 1985, 1994). Särndal, Swensson e Wretman (1989, citados por Rao 1994) justificam também o estimador (8) numa abordagem *model-assisted* adequada a planos de sondagem com uma etapa. Assim, é de esperar que para planos de sondagem complexos este estimador tenha também boas propriedades condicionais.

No caso da sondagem aleatória estratificada, para amostras grandes tais que os pós-estratos têm uma dimensão razoável em cada estrato inicial, espera-se que o estimador (8) seja um bom estimador de $V(\hat{\tau}_{ps,str})$, na abordagem condicional.

Quando ocorrem não respostas, as propriedades do estimador são ainda mais difíceis de analisar. No entanto, poderá também ter boas propriedades numa abordagem condicional se os elementos tiverem valores semelhantes para as variáveis de interesse e as probabilidades de resposta forem iguais, em cada pós-estrato.

Alternativamente ao estimador proposto por Rao (1985), podem-se considerar métodos de re-amostragem (*resampling*), como o Bootstrap, para se estimar a variância de $\hat{\tau}_{ps,str}$.

2.2. ESTIMAÇÃO NA PRESENÇA DE NÃO RESPOSTAS

Um problema da maioria das sondagens consiste na falta de obtenção, total ou parcial, de resposta aos questionários. A não resposta total (ou *unit nonresponse*) ocorre quando há ausência total de resposta ao questionário. A não resposta parcial (ou *item nonresponse*) ocorre quando há ausência de resposta apenas para uma parte do questionário.

Na presença de não respostas os estimadores usuais são enviesados. Uma discussão detalhada sobre os efeitos estatísticos da não resposta pode ser obtida em Lessler e Kalsbeek (1992).

Existem diversos métodos que permitem lidar com o problema da não resposta, tanto na fase de planeamento e recolha dos dados, como na fase de estimação. Os métodos de pós-estratificação permitem, não só lidar com os problemas das bases de sondagem, mas também lidar com o problema das não respostas. Referências bibliográficas relevantes sobre outros métodos podem ser obtidas em Lessler e Kalsbeek (1992) e Azevedo (1999).

Os estimadores de pós-estratificação inserem-se numa classe de métodos de tratamento de não respostas usualmente designados por **métodos de recomposição** ou **métodos de ajustamento**. Estes procedimentos consistem em reponderar a amostra, i.e. ajustar os pesos de inclusão, por forma a que os pesos ajustados tenham em consideração as não respostas.

De um modo geral, estes métodos são utilizados no tratamento das não respostas totais. Podem também ser utilizados no caso das não respostas parciais apesar de exigirem mais trabalho, uma vez que é necessário calcular diferentes ponderadores para as diferentes variáveis de interesse.

2.2.1. INTRODUÇÃO AOS MÉTODOS DE AJUSTAMENTO DAS NÃO RESPOSTAS

Suponhamos que a população U, de dimensão N, pode ser dividida em duas sub-populações: seja U_1 a sub-população, de dimensão N_1, correspondente aos elementos para os quais se obteria resposta se fossem seleccionados para a amostra; e seja U_0 a sub-população, de dimensão N_0, correspondente aos elementos de U para os quais não se obteria resposta se fossem seleccionados para a amostra. No que se segue,

utiliza-se o índice *1* para designar os elementos respondentes (na população ou na amostra) e o índice *0* (zero) para designar os elementos não respondentes (na população ou na amostra).

Um conceito fundamental, para os métodos de ajustamento, é o de **probabilidade de resposta**, que se denota por p_k, e é dado por

$$p_k = P(\mathbb{I}_{k\in U_1} = 1)\,,\ k = 1, 2, \ldots, N \qquad (12)$$

onde, $\mathbb{I}_{k\in U_1}$ é a variável indicatriz que toma o valor 1 se o elemento *k* pertence à sub-população U_1 e toma o valor zero no caso contrário (i.e., se pertence a U_0).

Seja $w_k = 1/\pi_k$ o peso de inclusão, ou peso inicial, do elemento *k*. Na presença de não resposta, a amostra é constituída por $n_1<n$ elementos respondentes. Nestas condições, o estimador de Horvitz-Thompson é enviesado. É possível obter-se um estimador centrado se o ponderador utilizado tiver em consideração a probabilidade de inclusão (π_k) e a probabilidade condicional de que o *k-ésimo* elemento torna-se respondente, se for seleccionado para a amostra, ou seja, quando o ponderador tem também em consideração a probabilidade de resposta ($p_k>0$, $\forall k=1,\ldots,N$) [Lessler e Kalsbeek, 1992, 182]. Obtém-se, desta forma, o estimador centrado

$$\hat{\tau}^{*}_{HT} = \sum_{k=1}^{n_1} w^{*}_k y_k \qquad (13)$$

onde, $w^*_k = 1/(\pi_k p_k)$.

Os métodos de ajustamento das não respostas, na inferência clássica das sondagens (*design-based*), consistem então estabelecer estimadores que ajustam os pesos iniciais w_k, através de diferentes estimadores das probabilidades de resposta p_k (geralmente, desconhecidas). Para mais detalhes sobre os métodos de ajustamento, veja-se Little (1986) e Lessler e Kalsbeek (1992). Nas secções que se seguem, apresentam-se os métodos de ponderação em classes e de pós-estratificação.

2.2.2. MÉTODO DE AJUSTAMENTO POR PONDERAÇÃO EM CLASSES

O método de ajustamento por ponderação em classes consiste em estimar as probabilidades de resposta através da divisão da amostra obtida (incluindo respondentes e não respondentes) em *H* subconjuntos mutuamente exclusivos e exaustivos, designados **classes** ou **células de ajustamento**. Assume-se que, em cada célula *h* (h=1,...,H), os elementos têm valores semelhantes para a variável de interesse *Y* e que todas as probabilidades de resposta são iguais.

Lessler e Kalsbeek (1992) referem que se o plano de sondagem for multi-etápico, é usual escolherem-se as unidades amostrais das primeiras etapas para definir as classes; no caso de uma sondagem aleatória estratificada, é usual utilizarem-

se as variáveis de estratificação. Estes autores referem ainda que, idealmente, as variáveis que definem as células devem estar fortemente associadas à variável de interesse, mas não devem estar mutuamente associadas.

Seja $s_h = s_{1h} \cup s_{0h}$ o conjunto de elementos da amostra pertencentes à *h-ésima* célula de ajustamento (de dimensão amostral n_h); onde, s_{1h} é o subconjunto de s_h correspondente aos elementos respondentes (de dimensão n_{1h}) e s_{0h} é o subconjunto constituído pelos elementos não respondentes (de dimensão n_{0h}). Sejam ainda w_{hk} e p_{hk}, respectivamente, o peso de inclusão e a probabilidade de resposta do *k-ésimo* elemento da *h-ésima* célula de ajustamento.

O estimador genérico de p_{hk}, utilizado neste método, é dado por:

$$\hat{p}_{hk} = \frac{\sum_{k=1}^{n_{1h}} w_{hk}}{\sum_{k=1}^{n_h} w_{hk}}, \quad k \in s_h,\ h = 1, \ldots, H \tag{14}$$

O ponderador ajustado, a utilizar nos estimadores por ponderação em classes, é então

$$w_{hk}^{(pc)} = \frac{\sum_{k=1}^{n_h} w_{hk}}{\sum_{k=1}^{n_{1h}} w_{hk}} w_{hk}, \quad k \in s_h,\ h = 1, \ldots, H \tag{15}$$

Este ponderador ajustado pode ser escrito como

$$w_{hk}^{(pc)} = \frac{\hat{N}_h}{\hat{N}_{1h}} w_{hk}, \quad k \in s_h,\ h = 1, \ldots, H \tag{16}$$

onde,

$$\hat{N}_h = \sum_{k=1}^{n_h} w_{hk} \tag{17}$$

$$\hat{N}_{1h} = \sum_{k=1}^{n_{1h}} w_{hk} \tag{18}$$

Um **estimador do total da população por ponderação em classes** é

$$\hat{\tau}_{pc} = \sum_{h=1}^{H}\sum_{k=1}^{n_{1h}} w_{hk}^{(pc)} y_{hk} = \sum_{h=1}^{H}\sum_{k=1}^{n_{1h}} \frac{\hat{N}_h}{\hat{N}_{1h}} w_{hk} y_{hk} \qquad (19)$$

onde, $\hat{N}_h$ e $\hat{N}_{1h}$ são dados, respectivamente, por (17) e (18) e y_{hk} é o valor da variável de interesse para o elemento k da *h-ésima* célula de ajustamento.

2.2.3. MÉTODOS DE AJUSTAMENTO POR PÓS-ESTRATIFICAÇÃO

Utilizando-se a notação apresentada anteriormente, suponhamos que a amostra pode ser pós-estratificada em L pós-estratos e se conhecem as dimensões $N_1,\ldots,N_i,\ldots,N_L$ dos pós-estratos na população.

Na secção 2.1 apresentou-se o estimador de pós-estratificação genérico com o intuito de lidar com os problemas da base de sondagem, na ausência de não respostas. Quando se pretende lidar simultaneamente com os erros da base de sondagem e com o enviesamento provocado pela presença de não respostas, podem-se combinar os métodos de pós-estratificação e de ponderação em classes.

Uma forma de combinar esses dois métodos consiste em assumir-se que os pós-estratos correspondem exactamente às células de ajustamento (do método de ponderação em classes). Obtém-se, assim, um ponderador ajustado dado por

$$w_{ik}^{(ps)} = \frac{N_i}{\hat{N}_i} w_{ik}^{(pc)} = \frac{N_i}{\hat{N}_i}\frac{\hat{N}_i}{\hat{N}_{1i}} w_{ik} = \frac{N_i}{\hat{N}_{1i}} w_{ik},\ k \in s_i,\ i = 1, \ldots, L \qquad (20)$$

com, $\hat{N}_{1i}$ dado por (18) ou seja,

$$\hat{N}_{1i} = \sum_{k=1}^{n_{1i}} w_{ik} \qquad (21)$$

Utilizando-se estes pesos ajustados, obtém-se um **estimador de pós-estratificação do total da população**, na presença de não respostas, dado por

$$\hat{\tau}_{ps} = \sum_{i=1}^{L}\sum_{k=1}^{n_{1i}} w_{ik}^{(ps)} y_{ik} = \sum_{i=1}^{L}\sum_{k=1}^{n_{1i}} \frac{N_i}{\hat{N}_{1i}} w_{ik} y_{ik}. \qquad (22)$$

onde, $\hat{N}_{1i}$ é dado por (21) e y_{ik} é o valor da variável de interesse para o elemento *k* do *i-ésimo* pós-estrato (célula de ajustamento).

Até ao momento, assumiu-se que as células de ajustamento da não resposta correspondem exactamente aos pós-estratos. No entanto, uma das abordagens mais utilizadas, para lidar com a não resposta total, consiste em obter os ponderadores ajustados pelo método de ajustamento em classes e, em seguida, ajustar esses ponderadores através da pós-estratificação. Ou seja, usualmente, as células de ajustamento são definidas separadamente para cada um dos métodos de ajustamento.

Assim, o primeiro passo consiste em ajustar os pesos iniciais nas células de ajustamento *h* do método de ponderação em classes, através de (15):

$$w_{hk}^{(pc)} = \frac{\hat{N}_h}{\hat{N}_{1h}} w_{hk}, \; k \in s_h, \; h = 1, \ldots, H \tag{23}$$

com, $\hat{N}_h$ e $\hat{N}_{1h}$ definidos, respectivamente, por (17) e (18); e w_{hk} o peso inicial do indivíduo *k* pertencente à *h-ésima* célula de ajustamento da não resposta.

Em seguida, estes ponderadores são ajustados novamente através da pós-estratificação da amostra:

$$w_{ik}^{(pc,ps)} = \frac{N_i}{\hat{N}_{1i}^{*}} w_{ik}^{(pc)}, \; k \in s_i, \; i = 1, \ldots, L \tag{24}$$

onde, $w_{ik}^{(pc)}$ é o peso ajustado por ponderação em classes, (23), do elemento *k* pertencente ao pós-estrato *i*; e $\hat{N}_{1i}^{*}$ é agora dado por

$$\hat{N}_{1i}^{*} = \sum_{k=1}^{n_{1i}} w_{ik}^{(pc)} \tag{25}$$

Um **estimador de pós-estratificação do total da população**, com **ajustamento da não resposta por ponderação em classes**, é dado por

$$\hat{\tau}_{pc,ps} = \sum_{i=1}^{L} \sum_{k=1}^{n_{1i}} w_{ik}^{(pc,ps)} y_{ik} \tag{26}$$

onde, $w_{ik}^{(pc,ps)}$ é o ponderador ajustado definido por (24).

Para planos de sondagem complexos, os estimadores dos métodos de ajustamento por ponderação em classes e por pós-estratificação são difíceis de analisar. O enviesamento e a variância dos estimadores da média da população, por ponderação em classes e por pós-estratificação, para um plano de sondagem aleatória simples sem reposição foram considerados por Thomsen (1973, 1978, citado por Little 1986), Kalton (1983, citado por Lessler e Kalsbeek 1992) e por Oh e Scheuren (1983, citados por Little 1986). É de salientar que estes autores verificaram que o estimador de pós-estratificação deverá ter erro quadrático médio inferior ao do estimador por ponderação em classes para esse plano de sondagem (mesmo considerando diferentes abordagens).

2.3. ESTIMAÇÃO DA VARIÂNCIA PELO MÉTODO *BOOTSTRAP WITHOUT REPLACEMENT (BWO)*

Os métodos de re-amostragem (*resampling*) baseiam-se na ideia de que a amostra obtida é representativa da população alvo, podendo extrair-se novas e repetidas amostras a partir da amostra original, com o objectivo de estimar variâncias ou intervalos de confiança. Alguns exemplos destes métodos são o *Jackknife*, proposto por Quenouille (1949), e o *Bootstrap*, introduzido por Efron (1979). Referências bibliográficas relevantes sobre extensões destes métodos a dados de sondagens podem ser encontradas em Shao e Tu (1995).

A aplicação da metodologia Bootstrap a dados de sondagens requer algumas alterações ao algoritmo "puro" (também designado na literatura por bootstrap *naïf* ou *naive*) proposto por Efron (1979). Os primeiros trabalhos nesta área devem-se a Gross (1980) e a Chao e Lo (1985). O método proposto por estes autores para a sondagem aleatória simples sem reposição designa-se, geralmente, na literatura por *Bootstrap Without Replacement* ou *Without Replacement Bootstrap* (*BWO*). Sitter (1992) propôs uma extensão deste método à sondagem aleatória estratificada.

O **algoritmo BWO proposto por Sitter (1992)** é o seguinte:

1º – Construir de forma independente para cada estrato h (h = 1, ..., H) uma pseudo-população, replicando-se os elementos da amostra original desse estrato k_h vezes, sendo

$$k_h = \frac{N_h}{n_h}\left(1-\frac{1-f_h}{n_h}\right);\ f_h = n_h/N_h,\ h = 1, \ldots, H \qquad (27)$$

2º – Seleccionar n'_h unidades de cada estrato h através de tiragens aleatórias sem reposição, sendo $n'_h = n_h - (1 - f_h)$, h = 1, ..., H; por forma a obter-se uma amostra bootstrap. A partir da amostra bootstrap, calcular a réplica bootstrap do estimador, $\hat{\theta}^*$.

3º – Repetir o 2º passo um grande número de vezes, B, por forma a obterem-se $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*b}, \ldots, \hat{\theta}^{*B}$ réplicas bootstrap do estimador.

4º – Estimar $V(\hat{\theta})$ por

$$\hat{V}^{*}_{BWO} = E^{*}[\hat{\theta}^{*} - E^{*}(\hat{\theta}^{*})]^2 \tag{28}$$

onde, E* denota o valor esperado relativamente às amostras bootstrap; ou, pela aproximação de Monte Carlo

$$\hat{V}^{*}_{BWO} \approx \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{*b} - \hat{\theta}^{*}(\cdot)\right)^2 \tag{29}$$

onde,

$$\hat{\theta}^{*}(\cdot) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b} \tag{30}$$

Note-se que este método pressupõe que k_h e n'_h sejam inteiros. No entanto, tal não ocorre a menos que f_h seja igual a "0" (zero) ou igual a 1. Sitter (1992) sugere um procedimento que designa por *randomization between bracketing integers* para contornar este problema e refere as propriedades do estimador bootstrap, subjacente a este algoritmo, para o caso linear. Neste caso, o método conduz a estimadores bootstrap consistentes. No caso não linear, o método parece ser promissor, como verificou Sitter (1992) através de estudos por simulação.

3. APLICAÇÕES AO INQUÉRITO ÀS EMPRESAS / HARMONIZADO DE 1997

Foram realizadas aplicações práticas[5], com os dados do IEH de 1997 (IEH97), dos métodos de ajustamento das não respostas (secção 2.2) e do método de Bootstrap BWO proposto por Sitter (1992), apresentado na secção 2.3.

O desenho do IEH corresponde a um plano de amostragem aleatória estratificada sem reposição. A base de amostragem é constituída a partir do Ficheiro Geral de Unidades Estatísticas (FGUE) do INE.

Para efeitos de apuramento, o universo é estratificado pelos escalões definidos pelas variáveis: *Escalões de NUTS II* – Nomenclatura das Unidades Territoriais para Fins Estatísticos; *Escalões de Classificação Portuguesa das Actividades Económicas*

[5] Resultados detalhados deste estudo e aplicações aos dados de 1996 podem ser encontrados em Machado e Costa (1998, 2001).

(CAE - REV. 2); *Escalões de número de pessoas ao serviço* e *Escalões de forma jurídica.*

O inquérito é realizado por amostragem para as unidades estatísticas (empresas) com menos de 100 pessoas ao serviço e de forma exaustiva para as unidades estatísticas com 100 e mais pessoas ao serviço[6].

Designem-se por *Pequenas e médias empresas* as empresas consideradas para inquirição com recurso à teoria de amostragem e por *Grandes empresas* as consideradas para inquirição exaustiva. No presente estudo consideram-se apenas as empresas do Continente inquiridas por amostragem, ou seja, as *Pequenas e médias empresas* do Continente.

A título ilustrativo, foram seleccionadas três variáveis: *Número médio de pessoas ao serviço* (total – remunerado e não remunerado), *Vendas* e *Prestações de serviços*. Para estas variáveis não se verifica a ocorrência de não respostas parciais (quando se eliminam da amostra as empresas identificadas com não resposta total).

3.1. ASPECTOS METODOLÓGICOS

Os estimadores considerados foram: o estimador de ponderação em classes, o estimador de pós-estratificação e o estimador de pós-estratificação com ajustamento da não resposta por ponderação em classes. Obtiveram-se também estimativas através do estimador de Horvitz-Thompson que, apesar de neste caso ser enviesado, não deixa de ser uma referência.

No método de ajustamento por ponderação em classes consideram-se os estratos iniciais como sendo as células de ajustamento das não respostas. Supõe-se então que, em cada estrato, os elementos têm valores semelhantes para as variáveis de interesse e que as probabilidades de resposta são iguais.

Nas aplicações do método de ajustamento por pós-estratificação, obtiveram-se as estimativas correspondentes à pós-estratificação da amostra através de três técnicas: segundo a variável *Escalões de número de pessoas ao serviço* (Quadro 1); segundo a variável *Escalões de volume de vendas* (Quadro 2); segundo o cruzamento dos escalões das variáveis *Escalões de número de pessoas ao serviço* e *Escalões de volume de vendas* (Quadro 3). Em qualquer dos casos, pressupõe-se que a(s) variável(s) de pós-estratificação está estreitamente relacionada com as variáveis de interesse. Neste método considera-se que os pós-estratos correspondem às células de ajustamento da não resposta e, portanto, assume-se que as probabilidades de resposta são iguais, em cada pós-estrato.

No método de pós-estratificação com ajustamento da não resposta por ponderação em classes consideram-se os estratos iniciais como sendo as classes de ajustamento da não resposta. Supõe-se, portanto, que os elementos têm valores

[6] Uma descrição mais detalhada da metodologia desta sondagem pode ser obtida em Instituto Nacional de Estatística (1997).

semelhantes para as variáveis de interesse e que as probabilidades de resposta são iguais, em cada estrato inicial.

Neste caso, a amostra foi pós-estratificada através de duas técnicas: segundo a variável *Escalões de número de pessoas ao serviço* (Quadro 1) e segundo o cruzamento dos escalões das variáveis *Escalões de número de pessoas ao serviço* e *Escalões de volume de vendas* (Quadro 3).

Quadro 1. Dimensões dos pós-estratos na população segundo a variável Escalões de número de pessoas ao serviço

Escalão	*Pessoas ao serviço*	Dimensão do pós-estrato na população
0	0	116581
1	1 a 9	610786
2	10 a 19	17965
3	20 a 49	9846
4	50 a 99	2141

Quadro 2. Dimensões dos pós-estratos na população segundo a variável *Escalões de volume de vendas*

Escalão	*Volume de vendas* (milhares de escudos)	Dimensão do pós-estrato na população
1	≤ 30 000	16678
2	> 30 000	740641

Quadro 3. Dimensões dos pós-estratos na população segundo as variáveis *Escalões de número de pessoas ao serviço* e *Escalões de volume de vendas*

Pós-estrato	*Escalões de número de pessoas ao serviço*	*Escalões de volume de vendas*	Dimensão do pós-estrato na população
1	0	1	3455
2	0	2	113126
3	1	1	13213
4	1	2	597573
5	2	1	10
6	2	2	17955
7	3	1	-
8	3	2	9846
9	4	1	-
10	4	2	2141

Para estimar a variância do estimador de pós-estratificação utilizou-se o estimador proposto por Rao (1985), dado por (8) para o caso da ausência de não respostas num plano de sondagem aleatória estratificada. Relativamente aos restantes estimadores propostos, não foi possível encontrar na literatura estimadores da variância para o plano de sondagem subjacente ao IEH (sondagem aleatória estratificada sem reposição). Para contornar este problema, foi utilizado o método de Bootstrap BWO, proposto por Sitter (1992).

As estimativas da variância foram obtidas através das aproximações de Monte Carlo. Para tal, foram retiradas 1000 amostras bootstrap, da população bootstrap construída segundo o referido algoritmo (*c.f.* secção 2.3).

3.2. RESULTADOS E CONCLUSÕES

As estimativas da média obtidas através do estimador de pós-estratificação por *Escalões de volume de vendas* (*PS por EVVN*) parecem indicar que este é extremamente enviesado. Este resultado não é de estranhar uma vez que, neste caso, se consideram apenas dois pós-estratos e, portanto, não se deve verificar o pressuposto de que os pós-estratos são homogéneos. Por esta razão, optou-se por não se aplicar o método Bootstrap BWO a este estimador.

As estimativas bootstrap da média obtidas para os estimadores considerados parecem indicar que o estimador de ponderação em classes é também bastante enviesado. Assim, as estimativas dos desvio-padrão deste estimador subestimam o verdadeiro valor do erro, uma vez que não contêm a contribuição do enviesamento. Aliás, como foi referido na secção 2.2, há evidências teóricas de que o estimador de pós-estratificação tem um erro quadrático médio inferior ao desse estimador e portanto deverá ser preferível.

Quanto aos restantes estimadores por pós-estratificação, as diferenças observadas parecem prender-se com o tipo de variáveis utilizadas para pós-estratificar a amostra. Dado que não é possível estimar o respectivo enviesamento, e consequentemente o seu erro quadrático médio, é mais difícil precisar qual das formas de pós-estratificação é mais adequada (por *Escalões de número de pessoas ao serviço* ou pelo cruzamento destes com os *Escalões de volume de vendas*). No entanto, pela análise dos desvios padrão e dos coeficientes de variação estimados por bootstrap verifica-se que, para as variáveis em estudo, há evidências de que o estimador de pós-estratificação por *Escalões de número de pessoas ao serviço* deverá ser mais adequado.

É também de salientar que as estimativas da variância do estimador de pós-estratificação, obtidas através do estimador proposto por Rao (1985), não diferem muito das obtidas por Bootstrap. Este era o resultado esperado dado que, no caso do IEH97, a dimensão dos pós-estratos em cada estrato inicial é bastante razoável. No entanto, tal poderá não ocorrer se forem consideradas outras variáveis (de análise ou de pós-estratificação) ou se os pressupostos assumidos não se verificarem, dado que há evidências de que as estimativas obtidas através do estimador proposto por Rao (1985) subestimam o valor da verdadeira variância.

Como foi referido anteriormente, os pós-estratos devem ser o mais homogéneos possível e, portanto, a variável que define os pós-estratos deve estar fortemente correlacionada com as variáveis de interesse. Sob este pressuposto e caso as dimensões dos pós-estratos sejam conhecidas, poder-se-ão considerar outras variáveis de pós-estratificação. Neste caso, seria também interessante analisar a utilização do estimador de pós-estratificação com ajustamento da não resposta por ponderação em classes.

A utilização dos estimadores de pós-estratificação tem por principal objectivo lidar com os problemas na base de sondagem. Dado que a taxa de não resposta total no IEH é muito elevada (cerca de 27%), fica também como sugestão para futuras investigações a utilização destas técnicas em simultâneo com outros métodos de tratamento das não respostas, nomeadamente os métodos de imputação.

BIBLIOGRAFIA

AZEVEDO, Áurea Sofia Pimenta (1999). *Estimação na Presença de Não Respostas – Aplicação ao Inquérito às Empresas (Harmonizado) do Instituto Nacional de Estatística.* Dissertação de Mestrado, Instituto Superior de Estatística e Gestão de Informação - Universidade Nova de Lisboa.

CHAO, M. T. e LO, S. H. (1985). A Bootstrap method for finite populations. *Sankhyä* A 47, 399-405.

DEMING, W. E. e STEPHAN, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics* 11, 427-444.

DEVILLE, J. C. e SÄRNDAL, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87, No. 418, 376-382.

DEVILLE, J. C., SÄRNDAL, C. E. e SAUTORY, O. (1993). Generalised raking procedures in survey sampling. *Journal of the American Statistical Association* 88, No. 423, 1013-1020.

EFRON, B. (1979). Bootstrap methods: another look at the Jackknife. *Annals of Mathematical Statistics* 7, 1-26.

GROSS, S. (1980). "Median estimation in sample surveys." Proceedings of the Section *on* Survey Research Methods, American Statistical Association, 181-184.

HOLT, D. e SMITH, T. M. F. (1979). Post stratification. *Journal of the Royal Statistical Society* A 142, 33-46.

INSTITUTO NACIONAL DE ESTATÍSTICA (1997). *Inquérito ás Empresas / Harmonizado – Dossier Global do Projecto.* Departamento de Estatísticas das Empresas, INE/DEE, Junho 1997.

LEONARD, K. A., *et al.* (1994). "Approximating the variance of the survey regression estimator using

poststratification." Proceedings of the 1994 Joint Statistical Meetings, Survey Research Methods Section, Vol. I, 222-227.

LESSLER, J. T. e KALSBEEK, W. D. (1992). *Nonsampling Error in Surveys*. Wiley Series in probability and Mathematical Statistics, John Wiley & Sons, New York.

LITTLE, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review* 54, No. 2, 139-157.

MACHADO, J. F. e COSTA, A. C. (1998). *Mudanças de Estrato nos Inquéritos às Empresas / Harmonizados - Relatório Intercalar*. Contrato Programa INE/ISEGI, Instituto Superior de Estatística e Gestão de Informação, Univ. Nova de Lisboa, Outubro de 1998.

MACHADO, J. F. e COSTA, A. C. (2001). *Mudanças de Estrato nos Inquéritos às Empresas / Harmonizados - Relatório Final*. Contrato Programa INE/ISEGI, Instituto Superior de Estatística e Gestão de Informação, Univ. Nova de Lisboa, Junho de 2001.

QUENOUILLE, M. (1949). Approximate tests of correction in time series. *Journal of the Royal Statistical Society* B 11, 18-44.

RAO, J. N. K. (1985). Conditional inference in survey sampling. *Survey Methodology* 11, No. 1, 15-31.

RAO, J. N. K. (1994). "Resampling methods for complex surveys." Proceedings of the 1994 Joint Statistical Meetings, Survey Research Methods Section, Vol. I, 35-41.

SHAO, J. e TU, D. (1995). *The Jackknife and Bootstrap*. Springer-Verlag, New York.

SITTER, R. R. (1992). Comparing three Bootstrap methods for survey data. *The Canadian Journal of Statistics* 20, No. 2, 135-154.

VALLIANT, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association* 88, nº 421, 89-96.

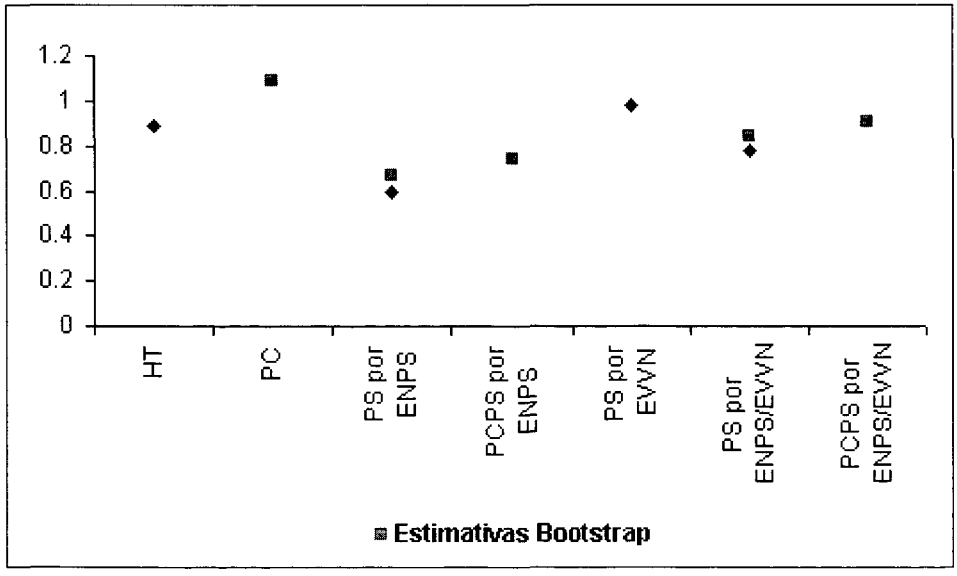
WILLIAMS, W. H. (1962). The variance of an estimator with post-stratified weighting. *Journal of the American Statistical Association* 57, 622-627.

ANEXO

Quadro A 1. Estimativas obtidas para a variável *Nº médio de pessoas ao serviço*

Estimador	Desvio padrão do estimador da média	Coeficiente de variação da média (%)	*Estimativas bootstrap* Desvio padrão do estimador da média	*Estimativas bootstrap* Coeficiente de variação da média (%)
Horvitz-Thompson (**HT**)	0.0173	0.89	-	-
Ponderação em classes (**PC**)	-	-	0.0173	1.09
Pós-estratificação por *Escalões de número de pessoas ao serviço* (**PS por ENPS**)	0.0173	0.60	0.0200	0.67
Pós-estratificação por *Escalões de número de pessoas ao serviço* com ajustamento da não resposta por ponderação em classes (**PCPS por ENPS**)	-	-	0.0224	0.74
Pós-estratificação por *Escalões de volume de vendas* (**PS por EVVN**)	0.0872	0.98	-	-
Pós-estratificação por *Escalões de número de pessoas ao serviço* e por *Escalões de volume de vendas* (**PS por ENPS/EVVN**)	0.0346	0.78	0.0400	0.84
Pós-estratificação por *Escalões de número de pessoas ao serviço* e por *Escalões de volume de vendas* com ajustamento da não resposta por ponderação em classes (**PCPS por ENPS/EVVN**)	-	-	0.0436	0.91

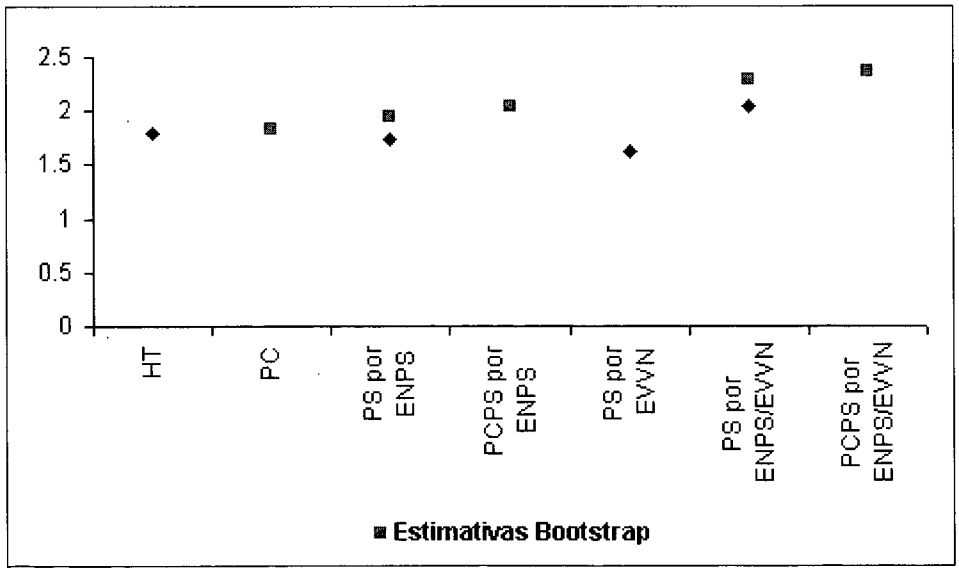
Gráfico A 1: Coeficientes de variação estimados do estimador da média (%), para a variável *Nº médio de pessoas ao serviço*



Quadro A 2. Estimativas obtidas para a variável *Vendas*

Estimador	Desvio padrão do estimador da média	Coeficiente de variação da média (%)	*Estimativas bootstrap* Desvio padrão do estimador da média	*Estimativas bootstrap* Coeficiente de variação da média (%)
Horvitz-Thompson (**HT**)	377.65	1.79	-	-
Ponderação em classes (**PC**)	-	-	283.65	1.84
Pós-estratificação por *Escalões de número de pessoas ao serviço* (**PS por ENPS**)	528.50	1.74	626.12	1.94
Pós-estratificação por *Escalões de número de pessoas ao serviço* com ajustamento da não resposta por ponderação em classes (**PCPS por ENPS**)	-	-	680.90	2.04
Pós-estratificação por *Escalões de volume de vendas* (**PS por EVVN**)	2018.08	1.62	-	-
Pós-estratificação por *Escalões de número de pessoas ao serviço* e por *Escalões de volume de vendas* (**PS por ENPS/EVVN**)	1630.53	2.04	1928.42	2.29
Pós-estratificação por *Escalões de número de pessoas ao serviço* e por *Escalões de volume de vendas* com ajustamento da não resposta por ponderação em classes (**PCPS por ENPS/EVVN**)	-	-	1918.28	2.36

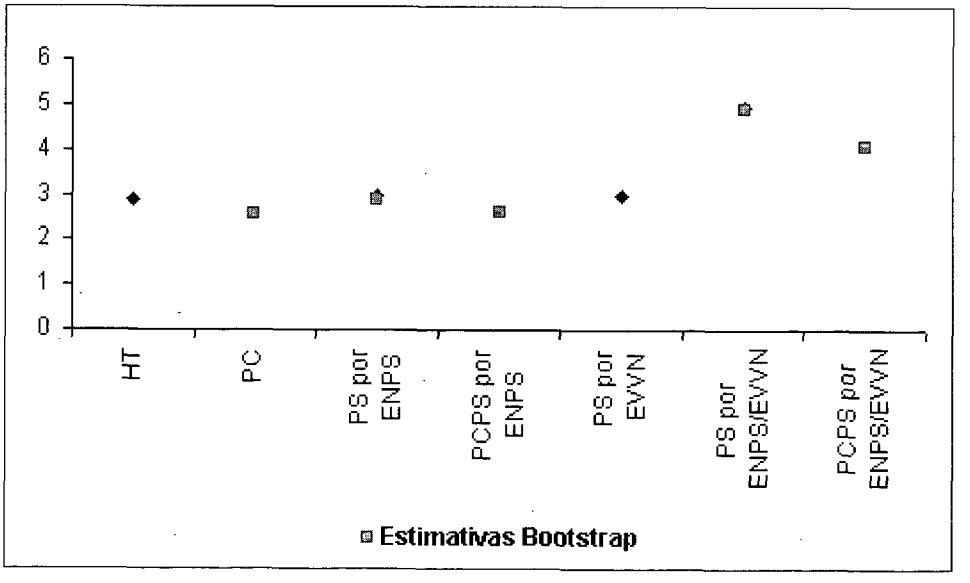
Gráfico A 2. Coeficientes de variação estimados do estimador da média (%), para a variável *Vendas*



Quadro A 3. Estimativas obtidas para a variável *Prestações de serviços*

Estimador	Desvio padrão do estimador da média	Coeficiente de variação da média (%)	*Estimativas bootstrap* Desvio padrão do estimador da média	*Estimativas bootstrap* Coeficiente de variação da média (%)
Horvitz-Thompson (**HT**)	165.17	2.88	-	-
Ponderação em classes (**PC**)	-	-	104.10	2.56
Pós-estratificação por *Escalões de número de pessoas ao serviço* (**PS por ENPS**)	238.51	2.97	243.30	2.87
Pós-estratificação por *Escalões de número de pessoas ao serviço* com ajustamento da não resposta por ponderação em classes (**PCPS por ENPS**)	-	-	221.17	2.59
Pós-estratificação por *Escalões de volume de vendas* (**PS por EVVN**)	906.20	3.00	-	-
Pós-estratificação por *Escalões de número de pessoas ao serviço* e por *Escalões de volume de vendas* (**PS por ENPS/EVVN**)	828.23	4.94	850.26	4.88
Pós-estratificação por *Escalões de número de pessoas ao serviço* e por *Escalões de volume de vendas* com ajustamento da não resposta por ponderação em classes (**PCPS por ENPS/EVVN**)	-	-	661.02	4.09

Gráfico A 3. Coeficientes de variação estimados do estimador da média (%), para a variável *Prestações de serviços*



REVISTA DE ESTATÍSTICA
149ª PÁGINA
3º QUADRIMESTRE DE 2001

Exclusão Social e Pobreza(s) em Portugal: uma primeira abordagem aos dados do Painel dos Agregados Familiares da União Europeia (1994 – 1997)

Autores:
Rui Branco
e
Cristina Gonçalves

EXCLUSÃO SOCIAL E POBREZA(S) EM PORTUGAL: UMA PRIMEIRA ABORDAGEM AOS DADOS DO PAINEL DOS AGREGADOS FAMILIARES DA UNIÃO EUROPEIA (1994 - 1997)

SOCIAL EXCLUSION AND POVERTY IN PORTUGAL: A FIRST APPROACH TO THE EUROPEAN COMMUNITY HOUSEHOLD PANEL (1994-1997)

Autores: Rui Branco
- Técnico Superior do Instituto Nacional de Estatística — Gabinete de Estudos e Conjuntura — Serviço de Estudos Demográficos e Sociais

Cristina Gonçalves
- Técnica Superior do Instituto Nacional de Estatística — Gabinete de Estudos e Conjuntura — Serviço de Estudos Demográficos e Sociais

RESUMO:

- Este estudo pretende contribuir para a definição, medição e caracterização das situações de pobreza em Portugal e recorre aos dados oficiais mais recentes do Painel dos Agregados Familiares da União Europeia (1994 a 1997).
  Baseia-se numa análise multidimensional da pobreza onde se consideraram quatro abordagens de natureza relativa de identificação: monetária; condições de vida; subjectiva e múltipla.
  Recorrendo à perspectiva longitudinal determinaram-se indicadores de avaliação da pobreza persistente e da prevalência da pobreza.
  A pobreza segundo as condições de vida foi a única abordagem em que se verificou uma redução contínua e notória das taxas de pobreza anuais.

PALAVRAS-CHAVE:

- *Painel dos Agregados Familiares da União Europeia; Índice de Pobreza Monetário; Índice de Pobreza segundo as Condições de Vida; Índice de Pobreza Subjectiva; Índice de Pobreza Múltipla; Linhas de Pobreza; Pobreza Persistente; Prevalência da Pobreza; Exclusão social; Pobreza; Portugal.*

ABSTRACT:

- This study aims to contribute for the definition, measurement and characterisation of poverty in Portugal for that it explores de European Community Household Panel (1994 a 1997).
  It is based in a multidimensional analysis of poverty where four relative identification criterions are used: monetary, living conditions, subjective and multiple.

Recurring to the longitudinal perspective indicators on the persistence and prevalence of poverty were determined.
Living conditions approach was the only one that continuously and notoriously registered decreasing poverty rates

KEY-WORDS:

- *European Community Household Panel; Monetary Poverty Index; Living Conditions Poverty Index; Subjective Poverty Index; Multiple Poverty Index; Poverty Thresholds; Persistent Poverty; Poverty Prevalence; Social Exclusion; Poverty; Portugal.*

1. INTRODUÇÃO

A pobreza e a exclusão social não são fenómenos recentes nem exclusivos dos países em vias de desenvolvimento. As recentes mudanças sociais, económicas e políticas que têm ocorrido um pouco por todo o planeta, têm coincidido com um incremento generalizado da importância destes fenómenos nos países desenvolvidos, sendo frequente estabelecer-se uma relação entre esta realidade e a capacidade de resposta dos modelos sociais vigentes.

O presente documento é ainda uma versão preliminar e parcelar de um estudo mais vasto que se encontra em elaboração no Gabinete de Estudos e Conjuntura do INE, com o objectivo de definir, medir e caracterizar as situações actuais de exclusão e pobreza em Portugal ao longo da última década.

Os dados mais recentes divulgados pelo EUROSTAT relativos ao Painel dos Agregados Familiares da União Europeia (PAUE) são a base da informação estatística explorada neste trabalho. De momento, estão disponíveis os dados relativos às primeiras 4 vagas do painel (1994 a 1997 [7]), sendo de esperar que a breve trecho sejam disponibilizados dados posteriores a 1997.

O termo exclusão social é habitualmente utilizado num sentido amplo, englobando condições de vida e de participação civil, social e política e está intimamente ligado ao conceito de cidadania. O de pobreza contempla, por seu lado, as condições sócio-económicas de vida e, como tal, pode ser medido e analisado sob diversas frentes.

Ambos os fenómenos são complexos e pluri-dimensionais. Por esta razão, existem diversas condicionantes na sua medição e identificação, quer pela insuficiência das fontes de informação, quer pelas dificuldades de conceptualização. Quanto ao fenómeno da pobreza a sua análise pode ser feita sob diferentes abordagens consoante o tipo de informação e de variáveis que se considerem, designadamente: pobreza absoluta e relativa; pobreza objectiva e subjectiva ou pobreza monetária e não monetária.

Atendendo às restrições concretas impostas pela informação disponível e à prossecução do objectivo de explorar a informação estatística mais recente, ao longo deste estudo apresentam-se medidas de pobreza relativas, monetárias e não monetárias abrangendo aspectos considerados objectivos, bem como, auto-avaliações dos indivíduos face à pobreza; medidas que no seu conjunto se complementam e permitem uma aproximação ao conceito mais abrangente de exclusão social.

No segundo capítulo apresentam-se algumas abordagens da pobreza baseadas em medidas relativas calculadas para o total da população residente em Portugal. Na perspectiva de uma análise multidimensional calcularam-se quatro medidas de pobreza relativa: uma primeira, de natureza monetária, que se designou de Índice de Pobreza Monetário (IPM); e a que se juntaram dois índices não monetários: o Índice de Pobreza segundo as Condições de Vida (IPCV), que incide sobre as condições dos alojamentos e a privação de bens correntes e duradouros e o Índice de Pobreza Subjectiva (IPS) que reflecte o modo como cada um avalia

[7] Destaca-se que a informação monetária disponibilizada em cada vaga se refere ao ano anterior ao do inquérito, toda a restante informação refere-se ao momento do inquérito.

a sua situação económica. Numa tentativa de complementar esta análise calculou-se uma quarta medida que se designou Índice de Pobreza Múltipla (IPMult) e que resulta da incidência simultânea de pobreza identificada pelos índices anteriores.

No terceiro capítulo, e aproveitando as potencialidades dos dados em Painel, apresentam-se indicadores relativos à persistência e à prevalência da pobreza no período de 1994 a 1997. Nomeadamente estuda-se a evolução das situações de pobreza duradoura para cada uma das abordagens consideradas e simultaneamente identificam-se as populações que durante pelo menos um dos anos do período estiveram sujeitas a cada um dos quatro tipos de pobreza.

Num quarto capítulo destacam-se algumas características das diversas sub-populações identificadas de acordo com as várias abordagens de pobreza estudadas. O enfoque baseia-se em características demográficas (como o sexo), geográficas (como a distribuição por NUTS II) económicas (como actividade económica, sector de actividade, profissão e rendimento) sociais (como a participação social) ou ainda a auto-avalição do estado de saúde.

A terminar sintetizam-se as principais conclusões e incluem-se anexos com as notas sobre a metodologia utilizada na construção dos índices de pobreza.

2. QUATRO ABORDAGENS DA POBREZA AJUSTADAS AO PAUE

Com base no Painel dos Agregados Familiares da União Europeia propõem-se quatro formas de identificar populações que em termos relativos são menos favorecidas. Recorre-se a informação monetária, a dados sobre as condições de vida como o acesso a bens e ainda a informação baseada em auto-avaliações feitas pelos inquiridos, nomeadamente, quanto à sua situação para fazer face às despesas e encargos.

Obtiveram-se quatro índices de pobreza cuja construção se encontra descrita nas notas metodológicas anexas a este trabalho.

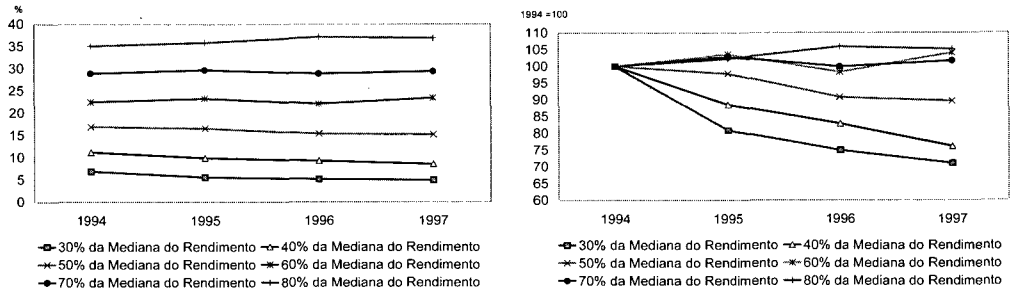
Neste estudo, privilegia-se o contributo estatístico tendo-se adaptado algumas das formas de operacionalização de conceitos correntes na literatura à informação disponível, sempre com a preocupação de evitar preconceitos quanto às possíveis abordagens de análise.

A operacionalização proposta baseia-se, muito sumariamente, numa visão de complementaridade entre diversos indicadores fundados em informação diversa de cariz objectivo e subjectivo. Acrescenta-se apenas que com a apresentação unicamente de medidas relativas não se pretende emitir nenhum juízo de valor quanto a uma eventual superioridade destas face à alternativa absoluta; nem tão pouco a utilização exclusiva de dados representativos no âmbito da totalidade do país, pretende desvalorizar o contributo de estudos específicos e localizados, fundamentais em muitos casos para detectar realidades imperceptíveis na lógica dos grandes números.

2.1. ÍNDICE DE POBREZA MONETÁRIA

A linha de pobreza monetária correntemente em vigor no EUROSTAT – os 60% da mediana do rendimento líquido total – identifica 22,5% dos portugueses como sendo pobres segundo este critério monetário[8] em 1994 e 23,4% em 1997. Estes valores traduzem uma ligeira oscilação no período não sendo notória uma tendência clara em tão curta série temporal. Sublinhe-se que representam uma das mais elevadas taxas de pobreza no espaço da união europeia[9].

Figura 2.1. Indivíduos com rendimento inferior a várias percentagens da mediana do rendimento por adulto equivalente – em percentagem da população e em índice face a 1994, (1994 a 1997)



Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997.

Como pesquisa complementar apuraram-se outras linhas baseadas numa percentagem diferente da mediana. Do Figura 2.1 surge evidente que ocorreu uma redução contínua no número de indivíduos sujeitos às situações de mais baixos rendimentos por adulto equivalente. Uma redução tanto maior quanto menor a percentagem da mediana do rendimento utilizada para determinar o limiar de pobreza.

Os indivíduos classificados como tendo um rendimento líquido por adulto equivalente inferior a 30% da mediana eram pouco mais de 70% (em 1997) dos que eram assim classificados em 1994.

Para percentagens acima dos 60% da mediana a situação aproximou-se mais da evolução registada para a linha de pobreza oficial dos 60%. Destaca-se, contudo, que a 80% da mediana se verificaram dois anos de agravamento sucessivo no número de indivíduos com rendimentos inferiores a este limiar (1995 e 1996) tendo depois estabilizado com o valor final para 1997 a indicar que existiam mais 5,1% de indivíduos nesta situação do que em 1994.

---

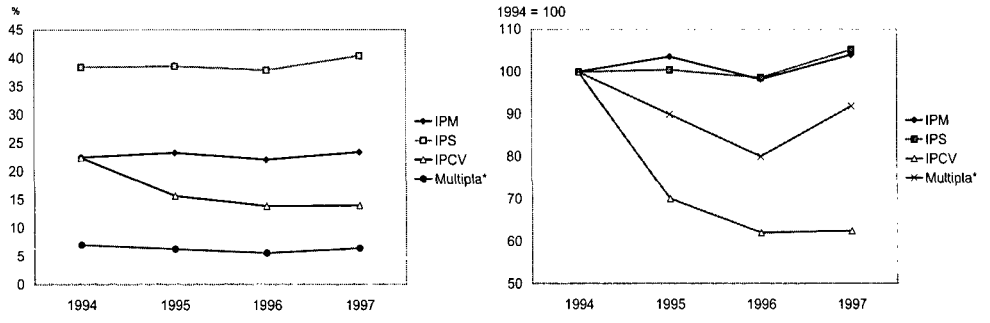
8 Também designada de Índice de Pobreza Monetário (IPM) no decurso deste trabalho.

9 De acordo com a publicação "Living conditions in Europe – Statistical pocketbook" (2000) do EUROSTAT para dados da 3ª vaga (1996), Portugal (22%) era de entre os 13 países para os quais havia informação disponível o que registava a mais elevada percentagem de indivíduos cujo rendimento por adulto equivalente era inferior a 60% da mediana do rendimento líquido anual. Era seguido de perto pela Grécia (21%) encontrando-se a Dinamarca no extremo oposto com 11% de pobres monetários.

2.2. ÍNDICE DE POBREZA SEGUNDO AS CONDIÇÕES DE VIDA

Para se estabelecer um Índice de Pobreza segundo as Condições de Vida (IPCV) utilizaram-se 29 variáveis[10] recolhidas com base no PAUE, e como se descreve detalhadamente nos anexos, foram atribuídos índices de privação aos agregados de acordo com a existência ou não de privação face aos 29 aspectos considerados.

Figura 2.2. Índices de pobreza em percentagem da população e em índice face a 1994 (1994 a 1997)



Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

O contributo de cada privação detectada para o *score* final de cada agregado é idêntico à frequência total de agregados que diziam não ter falta do respectiva variável em estudo. Apurada a situação de um agregado face à globalidade das variáveis, as frequências de privação foram adicionadas e obteve-se o *score* final de privação desse agregado que foi depois sujeito a uma hierarquização em conjunto com os *scores* de todos os agregados inquiridos.

Chegou-se, então, a uma distribuição social dos agregados ou dos indivíduos[11] face às condições de vida. Neste ponto traçou-se a linha de pobreza no valor do *score* que identificou como pobre uma percentagem de indivíduos semelhante à apurada para os 60% de mediana do rendimento no ano de 1994, ou seja, 22,5%. Assim, em 1994 a percentagem de pobres segundo o IPCV foi idêntica à do IPM por imposição exógena do investigador.

Uma vez que o valor do *score* foi fixado[12] para os anos seguintes tornou-se possível ter uma ideia da evolução em termos de tendência deste tipo de pobreza. Por

[10] Que aumentaram para 30 a partir de 1996, inclusive.

[11] Atribuindo o *score* do agregado a cada um dos indivíduos que o compõem é possível proceder à análise com base nos indivíduos, opção que se seguiu neste estudo.

[12] A opção resulta de uma analogia à dialéctica «preços constante» *versus* «preços correntes» típica das questões monetárias. Mantendo-se as mesmas variáveis e metodologia esta foi considerada a forma mais proveitosa de operacionalizar o IPCV. Um indivíduo poderia sair de uma situação de pobreza se diminuísse o número de itens em que sofre de privação. É

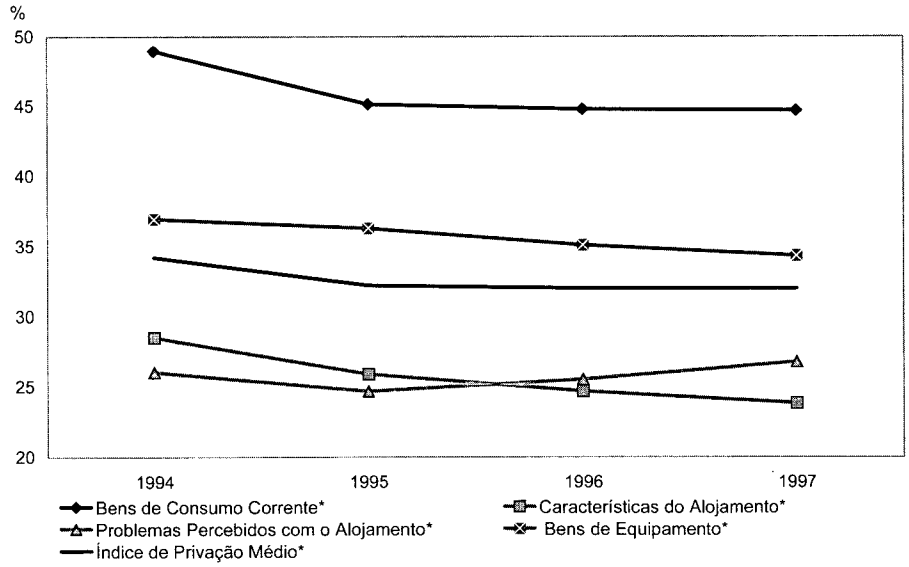
outro lado, ao se ter imposto um valor próximo do obtido pelo IPM surgiram populações de dimensão semelhante – mais facilmente comparáveis -, classificadas como pobres segundo conceitos distintos.

A análise da evolução do IPCV nos quatro anos (Figura 2.2) mostra que há diferenças claras em relação ao IPM com o primeiro a registar uma redução clara no número de indivíduos classificados como pobres ao longo de período. Tomando 1994 como ano de referência (1994 =100) verifica-se que em 1995 o IPCV caiu para 70,1, em 1996 diminuiu novamente para 62,1 tendo estabilizado depois com 62,5 em 1997. O fosso entre IPM e IPCV aumentou consecutivamente no período em análise.

Na Figura 2.3 apresentam-se indicadores de privação média que resultam da agregação dos quatro tipos de variáveis utilizadas na construção do IPCV, bem como, um indicador de privação média global[13].

Verifica-se que os níveis mais elevados de privação média se registam para os bens de consumo corrente[14]: em 1994 a privação média nas seis variáveis consideradas neste índice atingia os 48,9% indivíduos[15] diminuindo continuamente no período para um mínimo de 44,7% em 1997.

Figura 2.3. Índices de privação para os vários tipos de variáveis consideradas para o IPCV (1994 a 1997)



Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

possível ainda imaginar uma conjugação favorável que poderia levar um indivíduos a sair de uma situação de pobreza mesmo sem diminuir o número de itens de privação: seria necessário a sociedade, em geral, aumentar o seu nível de privação reduzindo assim as frequências de penalização para um mesmo conjunto de itens de privação identificados para esse indivíduo (consequência do carácter relativo da medida).

[13] Os quatro tipos de variáveis foram: Bens de consumo corrente; Características do alojamento; Problemas percebidos com o alojamento e Bens de equipamento.

[14] As variáveis que compõem estes índices encontram-se indicadas no anexo II.

[15] Convém sublinhar que se optou por ponderar todas as variáveis de igual forma.

Os níveis de privação média diminuíram ainda continuamente em dois dos restantes índices calculados: bens de equipamento e características do alojamento tendo sido para este conjunto de variáveis que ocorreu a maior diminuição dos níveis de privação no período. Curiosamente, quando o alojamento foi avaliado pela percepção subjectiva dos indivíduos através de perguntas como "O alojamento tem falta de espaço?" "Há ruído proveniente dos vizinhos ou do exterior?", entre outras, ocorreu um agravamento dos níveis de privação entre os anos extremo do período. A melhoria ao nível dos equipamentos básicos disponíveis no alojamento não foi acompanhada de uma melhoria ao nível dos problemas percebidos pelos indivíduos quanto a outras características do alojamento e do seu meio envolvente.

Como seria de esperar o índice de privação médio global registou um comportamento semelhante ao verificado através do IPCV.

2.3. ÍNDICE DE POBREZA SUBJECTIVA

A terceira abordagem de classificação de pobres baseia-se exclusivamente na resposta dos agregados à pergunta:

"Tendo em conta a totalidade do rendimento mensal do seu agregado, qual das seguintes opções melhor define a capacidade do agregado para fazer face aos encargos e despesas?

1. Com grande dificuldade
2. Com dificuldade
3. Com alguma dificuldade
4. Com alguma facilidade
5. Com facilidade
6. Com grande facilidade".

Os agregados e respectivos indivíduos que respondessem ter dificuldade ou muita dificuldade em fazer face aos encargos e despesas foram considerados pobres em termos subjectivos tendo sido quantificados pelo Índice de Pobreza Subjectivo (IPS). Atentando novamente à Figura 2.2 verifica-se que entre todos os índices calculados este é o que classifica a maior percentagem de indivíduos como pobres: em 1994 eram 38,4% do total dos indivíduos, valor que atingiu os 40,4% volvidos quatro anos.

A evolução ao longo dos quatro anos é, aliás, muito semelhante à que ocorre com o IPM; será interessante verificar se na presença de uma série temporal mais extensa se detecta uma correlação forte entre estas duas formas de medir a pobreza, ou se o número de indivíduos que acumulam estes dois tipos de pobreza é elevado.

2.4. ÍNDICE DE POBREZA MÚLTIPLA

A última "abordagem" de medição de pobreza proposta é já uma alternativa composta resultando da verificação simultânea dos três tipos de pobreza anteriores por um mesmo agregado/indivíduo – Índice de Pobreza Múltipla (IPMult).

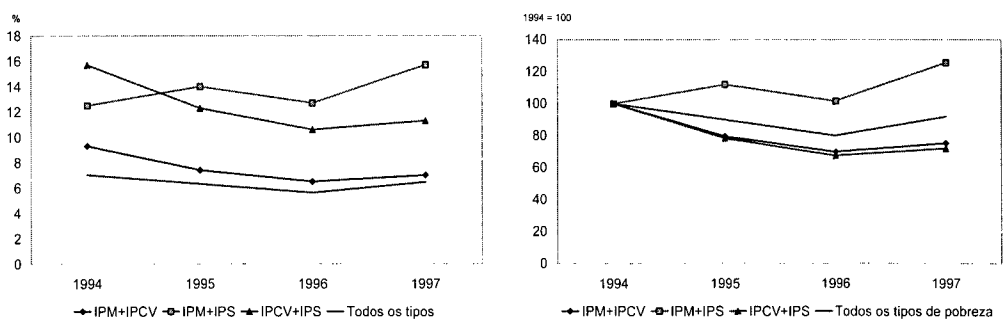
O factor de escala desta medida está fortemente influenciado pela escala dos restantes índices[16], no entanto, há um valor acrescentado nesta operacionalização que põe em pé de igualdade as naturezas mais ou menos distintas das restantes abordagens: o IPMult identifica inequivocamente situações de pobreza múltipla e portanto selecciona indivíduos em situações particularmente desfavoráveis.

Na Figura 2.2 verifica-se que o IPMult oscila ligeiramente no período sendo o valor mais elevado precisamente o de 1994 onde 7% dos indivíduos eram assim classificados. Nos dois anos seguintes registam-se quebras na ordem dos 10 pontos percentuais, se tomarmos o ano de 1994 como referência (1994 =100). No entanto, em 1997 há um incremento deste tipo de pobreza sendo o número de pobres equivalente a 91,9% dos que o eram em 1994.

A análise da pobreza múltipla é complementada pela construção de índices que combinam os três tipos de pobreza base em grupos de dois.

Na Figura 2.4, particularmente na análise em números índice (1994 =100), que é imune ao factor de escala, pode verificar-se que a combinação da pobreza monetária com a pobreza subjectiva registou um evolução irregular mas claramente no sentido do seu incremento, dado que o índice ultrapassou sempre os 100 nos anos posteriores a 1994 registando um valor máximo em 1997 onde existiam mais 25,6% de indivíduos sujeitos a este tipo de pobreza múltipla do que em 1994. Destaque-se que a evolução deste tipo de pobreza múltipla é concordante com a que se verificou no IPM e no IPS que estão na sua génese, contudo, registou um acréscimo global no período claramente superior ao registado individualmente em qualquer um dos índices. Acrescenta-se ainda que a percentagem de pobres monetários que eram simultaneamente pobres subjectivos foi sempre superior a 50% em todos os anos do período tendo atingido o valor máximo em 1997 onde 67,1% dos pobres monetários afirmavam ter dificuldades ou muitas dificuldades em fazer face Às despesas e aos encargos.

Figura 2.4. Índices de pobreza múltipla em percentagem dos indivíduos e em índice face a 1994 (1994 a 1997)



Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

[16] Aliás os restantes índices também o estão pelas hipóteses seguidas para determinação dos limiares de pobreza.

Nos remanescentes tipos de pobreza múltipla registaram-se sempre valores inferiores aos de 1994 ao longo do período. Em 1997 existiam menos 24,7% de pobres múltiplos segundo as condições de vida e segundo a abordagem monetária e menos 28,0% de pobres múltiplos segundo as condições de vida e segundo a abordagem subjectiva do que em 1994. A redução do IPCV teve um impacto claro nestes dois índices de pobreza múltipla verificando-se que a redução de incidência do IPCV teve uma amplitude muito próxima da verificada neste dois tipos de pobreza múltipla onde o IPCV está indirectamente incluído.

3. A POBREZA PERSISTENTE E A PREVALÊNCIA DA POBREZA

Quantos indivíduos classificados como pobres segundo um determinado tipo de pobreza se mantiveram nessa situação nos anos seguintes? Quantos indivíduos foram classificados como pobres pelo menos um vez no decurso de um determinado período de tempo segundo cada um dos tipos de pobreza em estudo? A resposta a estas duas perguntas quantificará os pobres persistentes e a prevalência da pobreza, respectivamente.

São inúmeras as possibilidades de exploração da informação contida no PAUE que se iniciou em 1994. A análise da persistência e da prevalência da pobreza que aqui se fará sumariamente é apenas uma pequena parcela do estudo longitudinal possível e cujo desenvolvimento decorre um pouco por todos os países participantes no PAUE.

3.1. POBREZA PERSISTENTE

Verifica-se que, como seria de esperar, a pobreza persistente diminui com o avançar no tempo face ao momento de referência (1994), conforme evidencia a Figura 3.1.

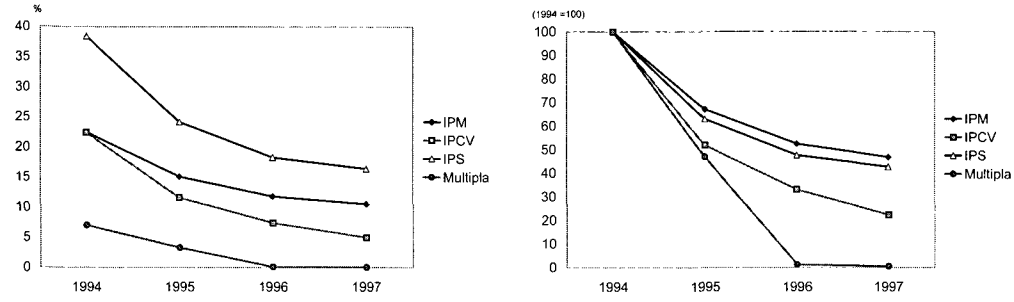
Ressalta que dos pobres múltiplos de 1994 menos de metade o eram ainda em 1995 e menos de 2% se mantinham com esse estatuto volvidos dois anos. Considerando os restantes tipos de pobreza individualmente, o cenário é contudo mais preocupante.

O IPM é entre os três índices de pobreza o que sofre menor erosão longitudinal, constatando-se que 46,8% dos indivíduos que eram pobres monetário em 1994 permaneceram nessa situação de pobreza durante pelo menos quatros anos consecutivos.

Uma evolução muito próxima é registada pela pobreza subjectiva registando-se em ambos os casos uma diminuição acentuada no ritmo de saídas desse tipo de pobreza com o afastamento do período de referência. Entre 1994 e 1995 o número de pobres persistentes diminuiu 32,9% no IPM e 37,1% no IPS, entre 1995 e 1996 as

diminuições foram de 21,8% e 24,4%, respectivamente e entre 1996 e 1997 foram de apenas 10,8% no IPM e de 10,3% no IPS.

Figura 3.1. Pobreza persistente em percentagem dos indivíduos e em índice face a 1994 (1994 a 1997)



Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

Quanto ao IPCV é notória uma redução mais rápida no número de indivíduos que permanecem continuamente neste tipo de pobreza: em um ano quase metade (48,2%) dos indivíduos que sofriam deste tipo de pobreza deixaram de ser considerados pobres e nos anos seguintes, embora o ritmo de saídas tenha diminuído (-36,2% em 1995/96 e –32,4% em 1996/97), manteve-se elevado. Em consequência, apenas 22,3% dos indivíduos que eram pobres segundo o IPCV em 1994 continuavam a sê-lo passados quatro anos.

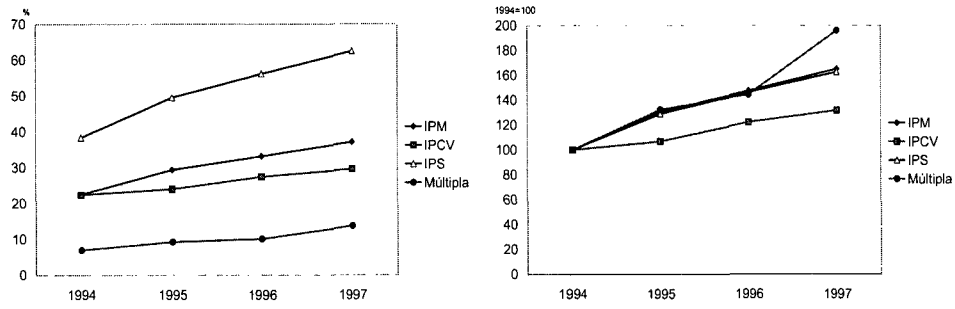
3.2. PREVALÊNCIA DA POBREZA

Quanto à prevalência da pobreza, a Figura 3.2 mostra que em quatro anos o número de agregados que afirmaram ter dificuldades ou grandes dificuldades em fazer face às despesas e aos encargos levou a que cerca de 70% dos indivíduos tivessem passado pela pobreza subjectiva durante pelo menos um ano no período. É portanto relativamente frequente existirem situações de pobreza subjectiva duradoura (conclusão que advém da análise da persistência) mas é também comum haver elevada rotatividade nos elementos que passam por este tipo de pobreza como se pode concluir pelo que se vê na Figura 3.2 e atendendo à evolução do número anual de pobres subjectivos em cada ano (Figura 2.2).

Considerando a evolução dos vários tipos de pobreza face ao ano base (1994=100) verifica-se que o IPCV é o que vai registando o menor crescimento de novos indivíduos a passar por este tipo de pobreza existindo um comportamento quase coincidente nos restantes tipos de pobreza, comportamento do qual apenas o IPMult se destaca no ano de 1997 com um maior incremento coincidente com um aumento mais acentuado no número de indivíduos sujeitos a este tipo de pobreza nesse ano.

Destaca-se que, apesar do valor anual mais elevado de pobres múltiplos ter sido de 7,0% (1994), 13,8% dos indivíduos que se mantiveram no painel entre 1994 e 1997 passaram durante pelo menos um ano por uma situação de pobreza múltipla; uma percentagem que alcançou os 37,2% para o IPM e 29,6% para o IPCV.

Figura 3.2. Prevalência da pobreza em percentagem dos indivíduos e em índice face a 1994 (1994 a 1997)



Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

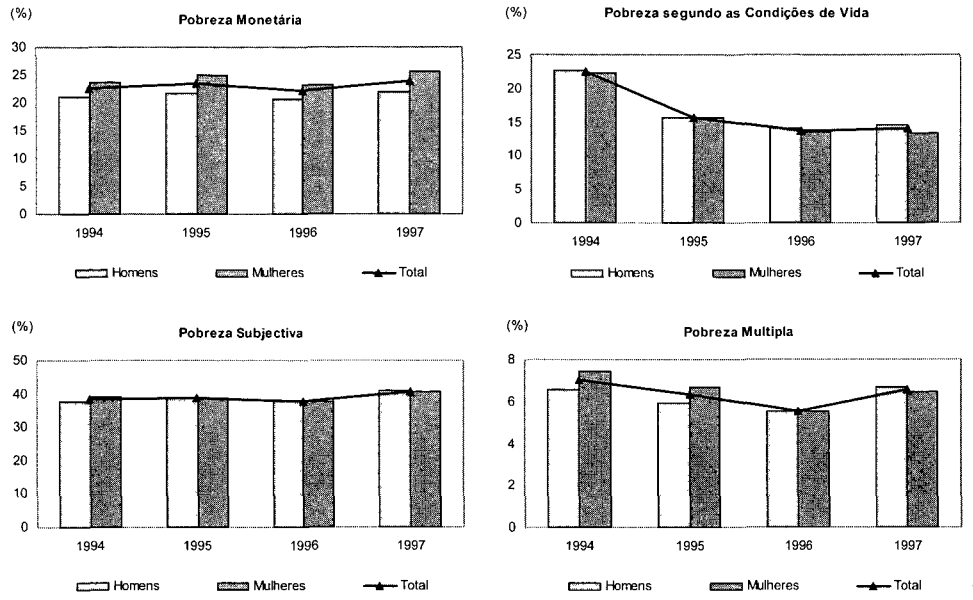
4. ALGUMAS CARACTERÍSTICAS DAS POPULAÇÕES POBRES

Após a identificação da população pobre, segundo as quatro abordagens exploradas nos capítulos anteriores, importa agora conhecer melhor as suas características individuais e as do agregado a que pertencem.

4.1. SEXO

As mulheres registam taxas de pobreza monetária mais elevadas que os homens, superioridade que se acentua no último ano analisado (1997): 25,6% contra 21,9% de homens (Figura 4.1).

Figura 4.1. Taxas de Pobreza segundo o sexo, 1994-1997



Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

Nos restantes tipos de pobreza as disparidades são mais ténues, não parecendo existir uma relação directa entre a incidência da pobreza e o sexo dos indivíduos.

4.2. GRUPOS ETÁRIOS

Numa análise por idades, destacam-se dois grupos etários particularmente vulneráveis à pobreza (Quadro 4.1): os idosos (65 e mais anos) e o grupo dos mais jovens (0-24 anos). O primeiro tradicionalmente um grupo populacional muito permeável à pobreza e a situações de exclusão social e o segundo considerado um fenómeno mais recente, sobre o qual se começam agora a debruçar alguns estudos.

O primeiro detém os índices de pobreza monetária mais elevados, com diferenças bastante significativas face à taxa de pobreza da população total (acima dos 10 pontos percentuais), encontrando-se os restantes grupos etários abaixo desse limiar, excepto em 1997. Neste último ano o grupo dos 0-24 anos atinge pela primeira vez uma taxa de pobreza monetária superior à verificada para o total da população.

Quadro 4.1. Hierarquização das taxas de pobreza, por tipo de pobreza e segundo o grupo etário, 1994-1997

IPM (%)							
1994		**1995**		**1996**		**1997**	
65 e + anos	39,6	65 e + anos	39,2	65 e + anos	37,1	65 e + anos	37,5
Total	22,5	Total	23,4	Total	22,1	0-24 anos	26,2
45-64 anos	21,6	0-24 anos	23,3	0-24 anos	22,0	Total	23,8
0-24 anos	20,9	45-64 anos	21,6	45-64 anos	20,2	45-64 anos	20,0
25-44 anos	16,4	25-44 anos	17,1	25-44 anos	16,0	25-44 anos	17,1
IPCV (%)							
1994		**1995**		**1996**		**1997**	
65 e + anos	24,7	65 e + anos	18,5	0-24 anos	15,8	0-24 anos	18,0
0-24 anos	24,0	0-24 anos	16,5	65 e + anos	14,0	65 e + anos	14,7
Total	22,5	Total	15,6	Total	13,7	Total	14,0
25-44 anos	21,7	25-44 anos	15,4	25-44 anos	13,4	25-44 anos	12,3
45-64 anos	19,7	45-64 anos	12,9	45-64 anos	10,8	45-64 anos	9,7
IPS (%)							
1994		**1995**		**1996**		**1997**	
65 e + anos	43,2	65 e + anos	41,8	65 e + anos	42,7	0-24 anos	45,4
0-24 anos	38,9	0-24 anos	41,2	0-24 anos	40,5	65 e + anos	42,7
Total	38,4	Total	38,7	Total	37,8	Total	40,6
25-44 anos	37,3	25-44 anos	37,8	25-44 anos	35,1	25-44 anos	38,2
45-64 anos	36,1	45-64 anos	33,9	45-64 anos	34,0	45-64 anos	35,3
IPMult (%)							
1994		**1995**		**1996**		**1997**	
65 e + anos	11,2	65 e + anos	8,5	0-24 anos	7,5	0-24 anos	9,6
Total	7,0	0-24 anos	7,1	65 e + anos	6,2	65 e + anos	7,3
0-24 anos	6,9	Total	6,3	Total	5,5	Total	6,6
45-64 anos	6,3	45-64 anos	5,2	25-44 anos	4,2	25-44 anos	4,9
25-44 anos	5,6	25-44 anos	5,1	45-64 anos	3,7	45-64 anos	3,8

Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

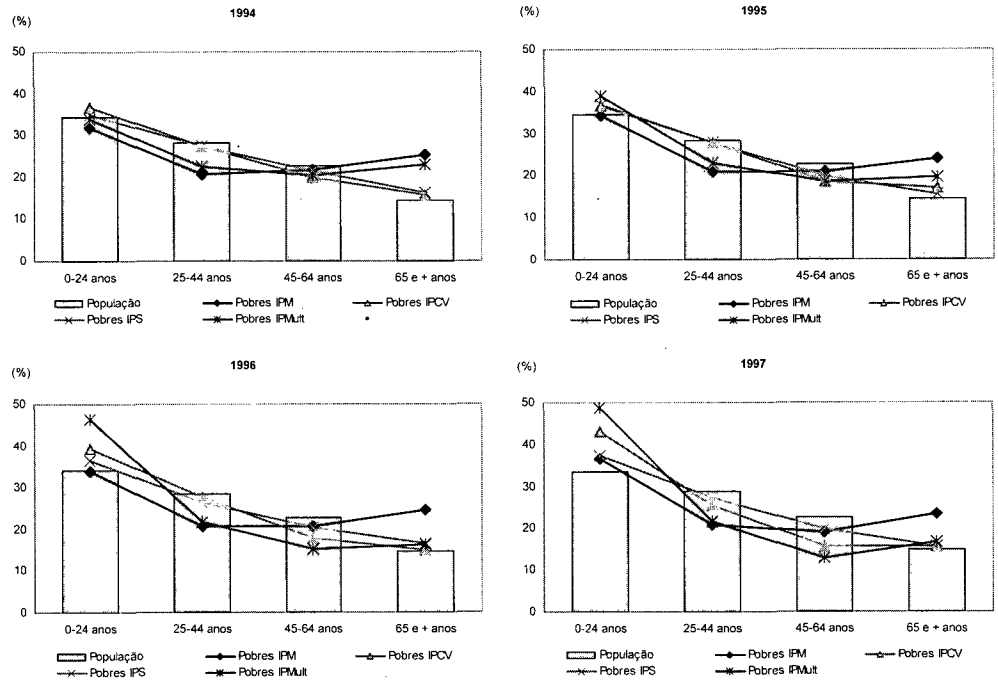
No índice de pobreza segundo as condições de vida registam-se taxas de pobreza mais baixas e existem menos diferenças entre os grupos etários. Quer neste tipo de pobreza, quer no da pobreza subjectiva, que observa valores bastante mais elevados, acontece que os mais idosos cedem a sua posição cimeira aos indivíduos mais novos, a partir de 1996, no primeiro e de 1997, no segundo.

O índice de pobreza múltipla ilustra bem a deterioração da situação dos mais jovens no que se refere a situações de carência aqui analisadas. Estes indivíduos passam a registar, a partir de 1996, as maiores percentagens neste critério.

Esta premissa é confirmada pela análise da distribuição do total de pobres por grupo etário que, como se verifica, se agrava ao longo do período no grupo dos jovens. Como se pode verificar na figura 4.2, enquanto que a proporção de jovens pobres só ultrapassava a dos jovens totais nos índices de pobreza segundo as condições de vida e

subjectiva, no último ano analisado isso acontece em todas as abordagens de pobreza e com diferenças percentuais na ordem dos 10-15 pontos percentuais.

Figura 4.2. Distribuição da população e de pobreza, segundo o tipo, por grupo etário, 1994-1997



Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

Por outro lado, não é de admirar que o peso relativo dos pobres seja superior ao da própria população no conjunto dos indivíduos com 65 e mais anos, já que tradicionalmente este é um dos grupo populacionais mais vulneráveis a situações de pobreza.

Como é igualmente visível, os restantes grupos populacionais (25-44 anos e 45-64 anos) são os menos vulneráveis à pobreza, ora um, ora outro, conforme o tipo de pobreza que se trate.

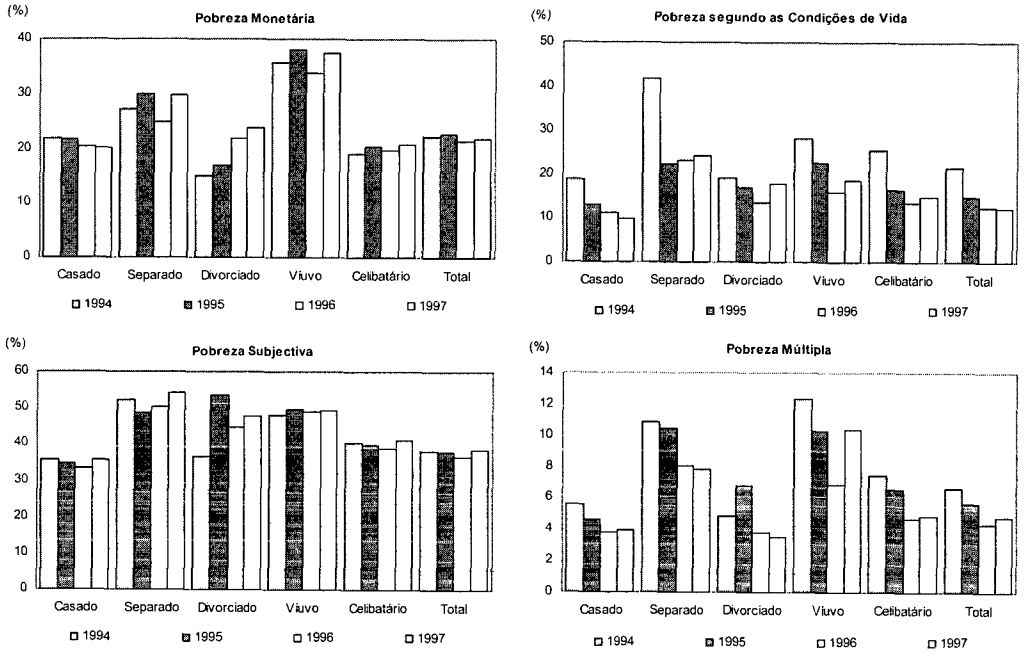
4.3. ESTADO CIVIL

Os indivíduos viúvos atingiram as maiores proporções de pobres monetários ao longo do período, logo seguidos dos separados. Os divorciados, que detinham a taxa mais baixa em 1994, sofrem um aumento importante do número de pobres, ultrapassando os casados a partir de 1996.

Os divorciados classificados como pobres na forma subjectiva registam taxas de pobreza relativamente elevadas, quando comparadas com as outras abordagens à pobreza. A figura 4.3 mostra ainda que os indivíduos viúvos e separados registam também os índices mais elevados nos outros tipos de pobreza.

A situação de casado é, regra geral, a menos atingida por qualquer tipo de pobreza, pois, é a que regista proporções inferiores quando comparados com o total de efectivos, sucedendo o oposto com os separados e viúvos.

Figura 4.3. Taxas de Pobreza segundo o estado civil, 1994-1997



Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

A estes resultados contrapõem-se outros que ilustram a incidência da pobreza nos indivíduos que declarando-se não serem casados, vivem numa união consensual. Esta situação parece só apresentar vantagem relativamente ao rendimento disponível e, que neste trabalho se traduzem por índices de pobreza monetária, já que nos outros tipos de pobreza estes registam quase sempre percentagens de pobres superiores relativamente aos que não vivem numa união consensual.

4.4. AGREGADO FAMILIAR

Segundo os dados do Painel a dimensão média das famílias foi de 3,8 indivíduos em 1994, registando um aumento contínuo até atingir os 4,1 em 1997.

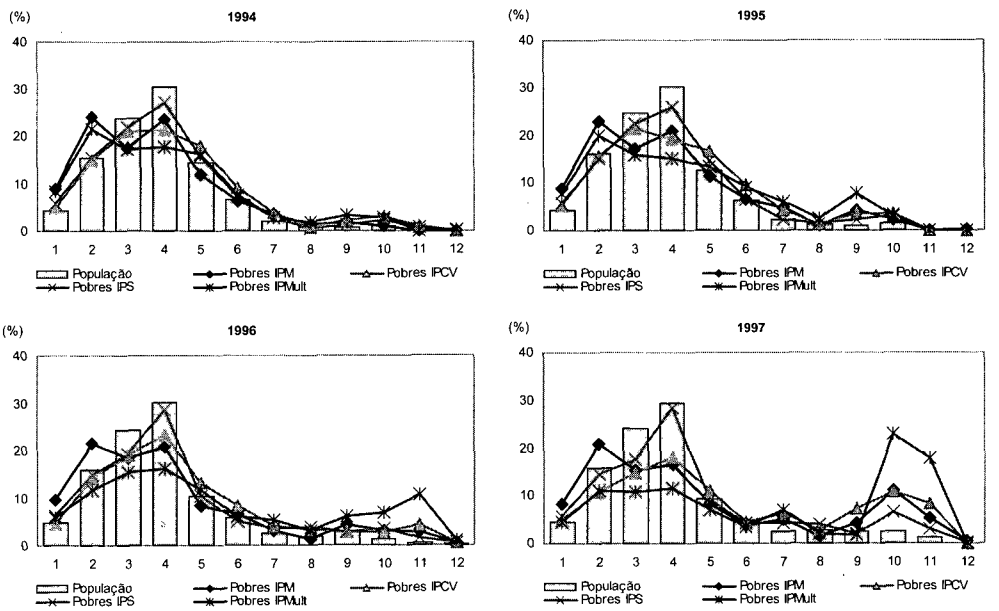
Como facilmente se deduz, as famílias numerosas têm pouca expressão, prevalecendo as de dimensão mais reduzida.

Relativamente à incidência da pobreza, é exactamente nos agregados com maior número de elementos, nomeadamente entre sete e dez, que os pobres atingem valores mais acentuados [17], como está bem patente na figura 4.4.

Nos agregados constituídos por um ou dois elementos, os índices atingem igualmente valores relativos mais elevados, sobretudo no tipo de pobreza monetário. Uma boa parte destes agregados respeitam provavelmente a famílias unipessoais e monoparentais, já consideradas noutros estudos como um dos grupos mais vulneráveis à pobreza.

A observação da distribuição dos indivíduos pela dimensão do agregado e a incidência dos pobres em cada grupo, permite confirmar aquelas afirmações: o peso relativo da pobreza, para qualquer abordagem considerada, é mais baixo nos agregados com três a cinco elementos, verificando-se o oposto nos de maior e menor dimensão, tal como já se referiu atrás.

Figura 4.4. Distribuição da população e de pobreza, segundo o tipo, por dimensão da família, 1994-1997



Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

Analisando agora a pobreza por tipo de família pode confirmar-se que as constituídas por um único elemento e as monoparentais se encontram numa situação bastante desvantajosa relativamente às constituídas por casais sem crianças ou com uma criança.

[17] Os agregados com onze ou mais elementos não mereceram aqui grande atenção devido a problemas de representatividade provocados pela reduzida expressão na amostra.

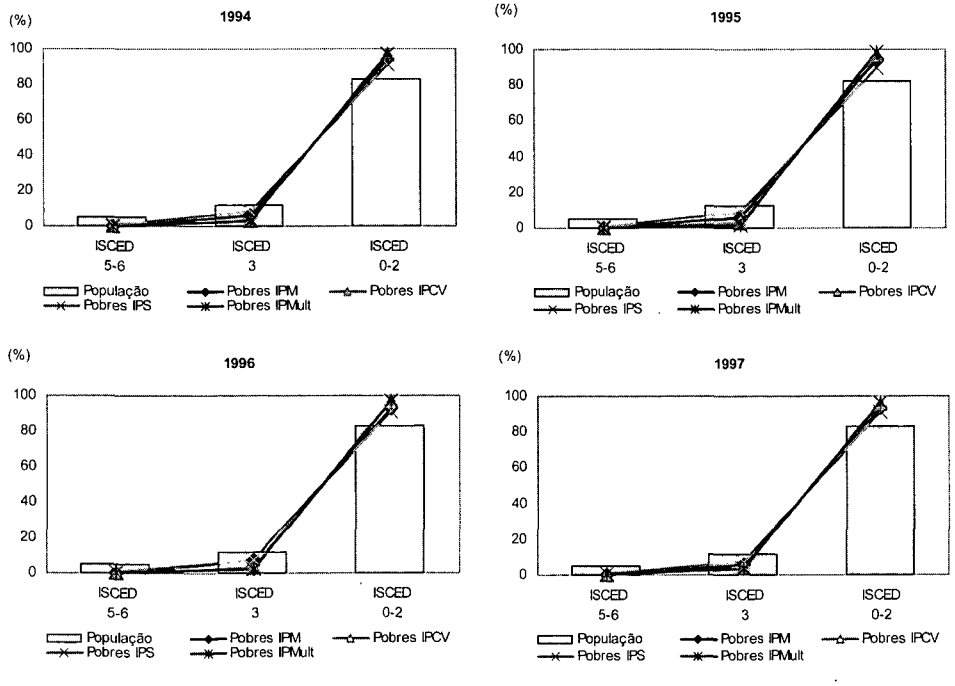
Ainda assim, é importante referir que a idade dos membros é uma condição importante à maior ou menor vulnerabilidade a situações de pobreza; pois, como se pode verificar relativamente às famílias unipessoais o risco de pobreza é maior nos indivíduos com 65 e mais anos. Da mesma forma, o conjunto das famílias constituídas por casal sem filhos, em que pelo menos um dos membros do casal é idoso (65 e mais anos), regista taxas de pobreza elevadas.

Os casais com três ou mais crianças (com menos de 16 anos) estão também bastante expostos aos diversos tipos de pobreza, assim como as famílias monoparentais com criança(s) com menos de 16 anos. A maior incidência da pobreza nestes grupos populacionais pode ajudar a perceber em parte a razão do grupo etário dos 0 aos 24 anos ser o segundo, e em certos critérios o primeiro, em termos de taxas de pobreza registadas no período.

4.5. NÍVEL EDUCACIONAL

É geralmente aceite que o nível educacional é um factor preponderante no projecto de vida dos indivíduos. Desta forma é fácil perceber que os indivíduos detentores de níveis escolares mais baixos são os mais expostos à pobreza, registando ao longo do período as taxas de pobreza mais elevadas, independentemente do tipo de pobreza subjacente.

FIGURA 4.5. DISTRIBUIÇÃO DA POPULAÇÃO E DE POBREZA, SEGUNDO O TIPO, POR NÍVEL ESCOLAR, 1994-1997



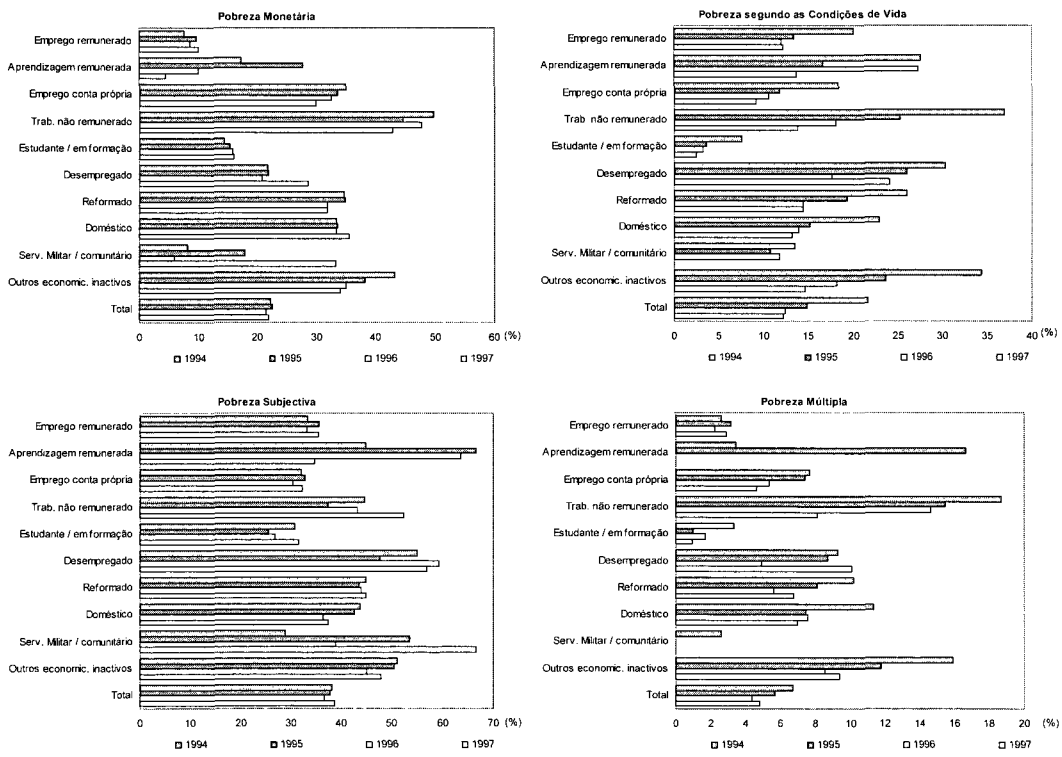
Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

Analisando a distribuição da população e da pobreza por nível escolar dos indivíduos, verifica-se que a proporção dos pobres no grupo dos indivíduos com grau escolar até ao 3º ciclo (ISCED 0-2) é bastante superior ao total dos indivíduos; esta proporção é preocupantemente muito próxima dos 100%, em alguns casos [18] (Figura 4.5). Pela mesma ordem de ideias os indivíduos detentores dos níveis 5-6 do ISCED são os menos vulneráveis aos diversos tipos de pobreza seleccionados nesta análise e durante o período em estudo.

4.6. ACTIVIDADE ECONÓMICA

Causa ou consequência, ou acumulação das duas, a condição perante a actividade económica é um dos factores com maior influência na vulnerabilidade à pobreza entre os indivíduos.

Figura 4.6. Taxas de Pobreza segundo condição perante a actividade económica, 1994-1997



Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

[18] No presente trabalho foi utilizada a classificação internacional de escolaridade das Nações Unidas: *International Standard Classification of Education* (ISCED). Nesta classificação, o **nível 0** corresponde à educação pré-escolar (a não frequência escolar também se enquadra neste nível); o **nível 1** aos 1º e 2º ciclos do ensino básico; o **nível 2** ao 3º ciclo do ensino básico; o **nível 3** ao ensino secundário; o **nível 4** não encontra correspondência no sistema educativo nacional (corresponde a um ensino pós-secundário que não é ensino superior); o **nível 5** corresponde ao ensino superior que engloba bacharelato, licenciatura, DESE, pós-licenciatura e mestrado; e o **nível 6** ao grau de doutoramento.

Grosso modo, são os trabalhadores não remunerados, entre os activos, e os reformados e domésticos, entre os inactivos, os indivíduos mais vulneráveis à pobreza, como está bem patente na figura 4.6. De referir ainda que os desempregados atingem igualmente taxas muito elevadas na pobreza subjectiva, o que não é de estranhar tendo em conta a questão inerente a este critério [19].

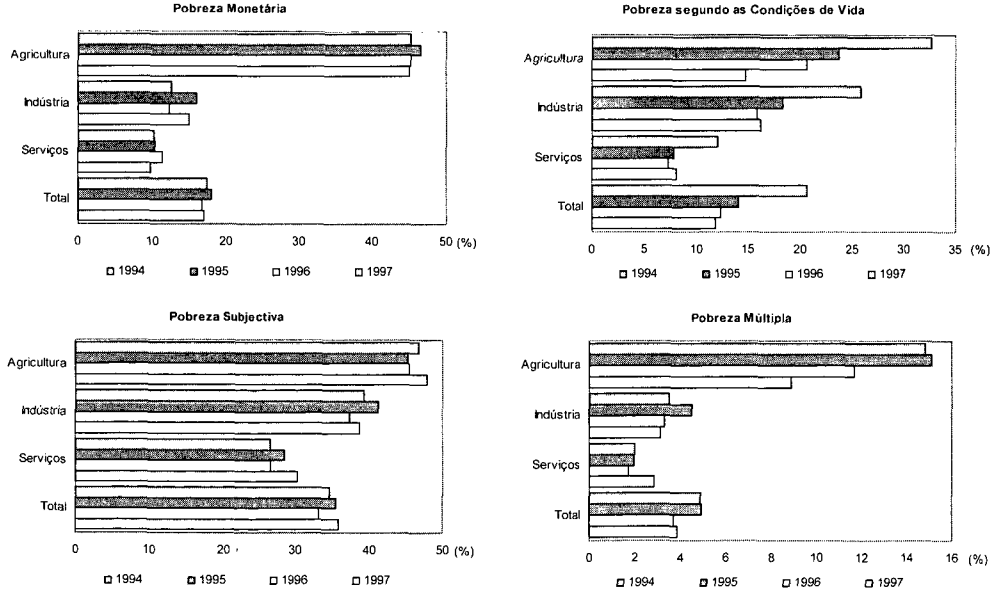
Os estudantes destacam-se, por seu turno, pois de entre os não activos são um dos grupos que regista as mais baixas taxas de pobreza. Os valores particularmente baixos nos critérios de pobreza segundo as condições de vida e múltipla podem indiciar que grande parte destes está inserido em famílias nucleares, com um ou dois filhos, que como se viu, apresentam menor propensão para situações de pobreza.

4.7. SECTOR DE ACTIVIDADE E GRUPO PROFISSIONAL

Os trabalhadores agrícolas ou, de um modo geral, os indivíduos activos que exercem a sua actividade no sector primário, sobressaem pela negativa nesta análise, em especial, na pobreza dos tipos monetária e múltipla. De referir ainda que ao nível das condições de vida e pobreza múltipla, verifica-se uma ligeira melhoria entre 1994 e 1997, enquanto que relativamente á pobreza subjectiva há um ligeiro agravamento no mesmo período (Figura 4.7).

Na posição oposta surgem os indivíduos do sector dos serviços, com a menor proporção de indivíduos pobres ao longo do período.

Figura 4.7. Taxas de Pobreza segundo sector de actividade económica, 1994-1997



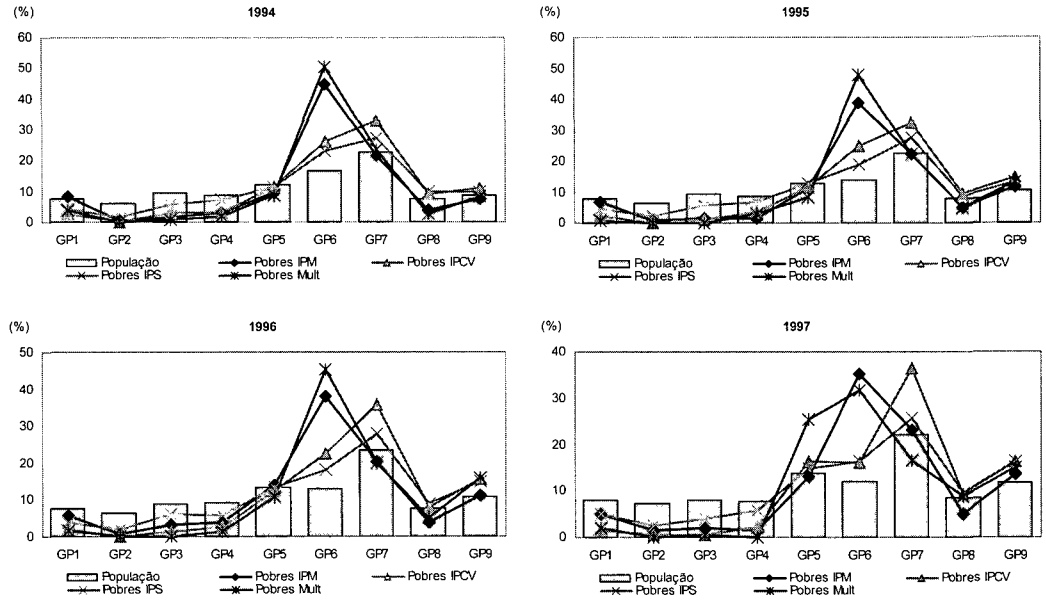
Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

[19] Os diminutos efectivos envolvidos em determinadas categorias, tais como, em regime de aprendizagem remunerada e a cumprir serviço militar ou comunitário (não activos), não permitiram uma análise mais aprofundada.

Esta análise pode ser confirmada na figura 4.8, na qual estão representados os grupos profissionais a que pertencem os indivíduos inquiridos e que são economicamente activos [20]. Como está bem patente, o Grupo 6 (Agricultores e Trabalhadores Qualificados da Agricultura e Pescas) é aquele em que a proporção de pobres monetários é maior, comparativamente ao número de efectivos abrangidos. Nos outros tipos de pobreza é igualmente um dos mais elevados.

De um modo geral, é bem visível que a pobreza está directamente relacionada com a profissão do indivíduo, pois o aumento das taxas de pobreza é inversamente proporcional à qualificação das profissões exercidas pelos mesmos.

FIGURA 4.8. DISTRIBUIÇÃO DA POPULAÇÃO E DE POBREZA, SEGUNDO O TIPO, POR GRUPO PROFISSIONAL, 1994-1997



Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

As últimas observações sobre a actividade económica respeitam ao tipo de empresa a que pertence o indivíduo e ao regime de trabalho.

A incidência da pobreza é bastante superior nos indivíduos que exercem a sua actividade nas empresas privadas, em qualquer dos critérios analisados.

---

[20] As profissões identificadas pelos inquiridos foram, numa fase posterior agrupadas em classes relativamente às quais se estabeleceu uma correspondência com a classificação nacional de profissões 94, a 1 dígito: Grupo 1: quadros superiores da administração pública dirigentes e quadros superiores de empresa; Grupo 2 - especialistas das profissões intelectuais e científicas; Grupo 3 - técnicos e profissionais de nível intermédio; Grupo 4 - pessoal administrativo e similares; Grupo 5 - pessoal dos serviços e vendedores; Grupo 6 - agricultores e trabalhadores qualificados da agricultura e pescas; Grupo 7 - operários, artífices e trabalhadores similares; Grupo 8 - operadores de instalações e máquinas e trabalhadores da montagem; Grupo 9 - trabalhadores não qualificados

Por outro lado, e como seria de esperar o regime de *part-time* é igualmente mais desvantajoso para os indivíduos, uma vez que está directamente ligado com o valor monetário da remuneração.

4.8. RENDIMENTO

A principal fonte de rendimento dos indivíduos durante as quatro vagas do inquérito foi a respeitante a salários e vencimentos; embora tenha perdido importância relativa no período representou sempre mais de 60% da principal fonte dos inquiridos (68,7% em 1994 e 63,3% em 1997).

As pensões (entre 15 e 17%) e os rendimentos de trabalho por conta própria (entre 10 e 13%), surgem nas posições seguintes. As restantes categorias (tais como subsídios de desemprego ou outros, outros benefícios sociais e rendimentos de propriedade, investimentos ou poupanças) são pouco significativos, não ultrapassando em conjunto os 7%, em cada ano do período.

A hierarquização das taxas de pobreza, constante do Quadro 4.2, permite visualizar que os indivíduos cuja principal fonte de rendimento assume a forma de transferência social, seja ela qual for, ocupam os lugares cimeiros da pobreza, independentemente do critério subjacente.

Relativamente ao Índice de Pobreza Monetário, os indivíduos que vivem de outros benefícios sociais registam taxas de pobreza duas ou três vezes superiores às registadas para o conjunto total de pobres. Da mesma forma, os que têm como principal rendimento as pensões, estão também muito acima dos valores médios de pobreza encontrados para o total. Os indivíduos que dependem de salários e vencimentos são, como seria de se esperar, os menos vulneráveis à pobreza.

Nos tipos de pobreza segundo as condições de vida e subjectiva as diferenças entre os valores que assumem as taxas são mais ténues. Ainda assim, observa-se que na pobreza segundo as condições de vida as taxas registam uma diminuição contínua (ainda que ligeira nalguns casos), enquanto que os que foram classificados como pobres subjectivos registam um aumento em 1997, chegando a categoria de outros benefícios sociais a atingir valores quase duplo da média.

Consequentemente, os indivíduos que acumulam todos os tipos de pobreza são em especial os pensionistas, os beneficiários de subsídios estatais ou outras formas de apoio social, encimando a hierarquização das taxas no período de 1994 a 1997.

Quadro 4.2. Hierarquização das taxas de pobreza, por tipo de pobreza e segundo a principal fonte de rendimento, 1994-1997

IPM (%)

1994		1995		1996		1997	
Outros benefícios sociais	66,5	Outros benefícios sociais	71,7	Outros benefícios sociais	66,1	Outros benefícios sociais	53,1
Pensões	53,8	Pensões	48,5	Pensões	46,4	Pensões	43,9
Subsídios desempr. ou outros	52,2	Subsídios desempr. ou outros	39,6	Subsídios desempr. ou outros	26,7	Subsídios desempr. ou outros	34,8
Rend. Prop., investim. ou poupanças	46,8	Rend. Prop., investim. ou poupanças	30,5	Rendim. trab. conta própria	26,7	Rendim. trab. conta própria	28,7
Rendim. trab. conta própria	26,5	Rendim. trab. conta própria	23,7	Total	22,2	Rend. Prop., investim. ou poupanças	27,2
Total	22,8	Total	23,6	Rend. Prop., investim. ou poupanças	19,0	Total	23,9
Salários ou vencimentos	12,2	Salários ou vencimentos	14,9	Salários ou vencimentos	13,4	Salários ou vencimentos	15,6

IPCV (%)

1994		1995		1996		1997	
Subsídios desempr. ou outros	38,0	Outros benefícios sociais	31,2	Outros benefícios sociais	26,9	Outros benefícios sociais	20,5
Outros benefícios sociais	33,3	Subsídios desempr. ou outros	28,1	Subsídios desempr. ou outros	16,5	Pensões	16,4
Pensões	28,0	Pensões	20,2	Pensões	14,7	Salários ou vencimentos	15,6
Total	22,5	Salários ou vencimentos	15,9	Salários ou vencimentos	14,5	Total	13,9
Salários ou vencimentos	22,0	Total	15,6	Total	13,8	Subsídios desempr. ou outros	5,3
Rendim. trab. conta própria	13,0	Rendim. trab. conta própria	5,2	Rendim. trab. conta própria	7,3	Rendim. trab. conta própria	3,5
Rend. Prop., investim. ou poupanças	12,7	Rend. Prop., investim. ou poupanças	4,8	Rend. Prop., investim. ou poupanças	0,9	Rend. Prop., investim. ou poupanças	0,5

IPS (%)

1994		1995		1996		1997	
Outros benefícios sociais	56,5	Subsídios desempr. ou outros	75,6	Subsídios desempr. ou outros	58,3	Outros benefícios sociais	74,2
Subsídios desempr. ou outros	54,4	Outros benefícios sociais	55,8	Outros benefícios sociais	51,0	Subsídios desempr. ou outros	56,1
Pensões	49,2	Pensões	46,0	Pensões	46,5	Pensões	45,9
Total	38,3	Total	38,5	Total	37,9	Total	40,6
Rend. Prop., investim. ou poupanças	37,6	Salários ou vencimentos	37,5	Salários ou vencimentos	37,9	Salários ou vencimentos	39,4
Salários ou vencimentos	36,3	Rend. Prop., investim. ou poupanças	31,7	Rendim. trab. conta própria	24,3	Rendim. trab. conta própria	30,6
Rendim. trab. conta própria	27,1	Rendim. trab. conta própria	27,7	Rend. Prop., investim. ou poupanças	16,2	Rend. Prop., investim. ou poupanças	18,3

IPMult (%)

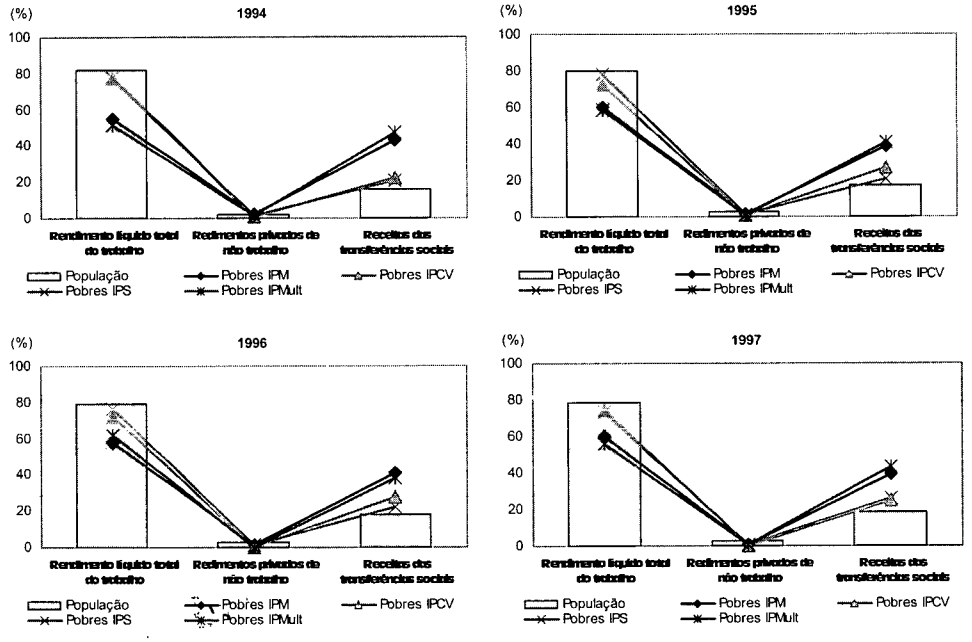
1994		1995		1996		1997	
Subsídios desempr. ou outros	29,4	Outros benefícios sociais	25,8	Outros benefícios sociais	19,6	Outros benefícios sociais	18,2
Outros benefícios sociais	23,0	Subsídios desempr. ou outros	21,9	Pensões	8,6	Pensões	10,5
Pensões	17,5	Pensões	12,4	Total	5,5	Total	6,6
Rend. Prop., investim. ou poupanças	7,6	Total	6,3	Salários ou vencimentos	4,6	Salários ou vencimentos	6,3
Total	7,1	Salários ou vencimentos	4,6	Rendim. trab. conta própria	4,1	Rendim. trab. conta própria	1,5
Rendim. trab. conta própria	5,3	Rend. Prop., investim. ou poupanças	4,3	Subsídios desempr. ou outros	3,9	Rend. Prop., investim. ou poupanças	0,5
Salários ou vencimentos	3,8	Rendim. trab. conta própria	2,3	Rend. Prop., investim. ou poupanças	0,9	Subsídios desempr. ou outros	0,0

Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

Relativamente à proporção dos pobres em cada categoria do rendimento líquido total, salienta-se a sobrerepresentação dos pobres entre o total de indivíduos que dependem essencialmente das receitas das transferências sociais (Figura 4.9).

A proporção é bastante elevada na pobreza monetária e múltipla (rondando os 40% ao longo dos quatro anos) e um pouco mais baixa na pobreza segundo as condições de vida e subjectiva (entre os 20 e os 27%). Sublinhe-se que relativamente à população total, os indivíduos cuja principal fonte de rendimento advém das transferências sociais nunca alcança os 20% ao longo do período.

Figura 4.9. Distribuição da população e do rendimento líquido total por tipo de pobreza, 1994-1997



Fonte: Cálculos do GEC/SEDS com base no PAUE 1994-1997

Parece assim ser clara a relação entre ter como principal fonte de rendimento qualquer receita de transferências sociais e a permeabilidade a situações de pobreza; por outras palavras, os pobres identificados pelas quatro abordagens propostas estão sobrerepresentadas no grupo de indivíduos seleccionados pelas medidas de política social vigentes (beneficiários de transferências sociais).

4.9. LOCALIZAÇÃO GEOGRÁFICA

De acordo com os dados do Painel, foi possível desagregar os indivíduos por localização geográfica a nível de NUTS II [21] (Quadro 4.3).

[21] Nomenclatura das Unidades Territoriais para fins Estatísticos (NUTS), nível II: Norte, Centro, Lisboa e Vale do Tejo, Alentejo, Algarve, Região Autónoma dos Açores e Região Autónoma da Madeira.

As Regiões Autónomas são as NUTS onde reside a maior percentagem de pobres monetários entre 1994 e 1997; Madeira em 1994 e 1997 e Açores, nos anos intermédios.

Pode verificar-se igualmente que o Norte é a região com maior nível de privação no que se refere às condições de vida, entre 1994 e 1996, trocando de posição com a Região Autónoma da Madeira no último ano em análise. Esta última é igualmente a mais afectada no que respeita à acumulação de todos os tipos de pobreza explorados no presente estudo, registando taxas bastante superiores às encontradas para o total do país. Os indivíduos do Alentejo sobressaem, por seu turno, no que se refere à pobreza subjectiva, notando-se um agravamento contínuo deste tipo de pobreza durante os quatro anos, um comportamento oposto ao que se verifica para a região Centro.

Quadro 4.3. Hierarquização das taxas de pobreza, por tipo de pobreza e localização geográfica, 1994-1997

IPM (%)							
1994		1995		1996		1997	
R. A. Madeira	35,7	R. A. Açores	41,2	R. A. Açores	41,8	R. A. Madeira	42,3
R. A. Açores	34,1	R. A. Madeira	39,6	R. A. Madeira	40,2	R. A. Açores	38,7
Algarve	33,9	Algarve	36,1	Algarve	34,4	Algarve	34,3
Centro	29,2	Centro	32,0	Centro	32,6	Centro	29,4
Alentejo	29,1	Alentejo	30,9	Alentejo	27,6	Alentejo	25,9
Total	22,5	Norte	26,0	Total	22,1	Norte	25,2
Norte	22,5	Total	23,4	Norte	21,9	Total	23,8
Lisboa e V.Tejo	12,9	Lisboa e V.Tejo	12,0	Lisboa e V.Tejo	12,5	Lisboa e V.Tejo	15,8
IPCV (%)							
1994		1995		1996		1997	
Norte	29,7	Norte	20,3	Norte	19,9	R. A. Madeira	21,1
Alentejo	27,0	R. A. Madeira	19,6	R. A. Madeira	17,6	Norte	19,1
R. A. Açores	24,0	Alentejo	17,9	R. A. Açores	13,9	Alentejo	16,3
R. A. Madeira	22,9	R. A. Açores	16,7	Total	13,7	Total	13,9
Total	22,5	Centro	16,5	Centro	12,6	R. A. Açores	12,2
Centro	20,8	Total	15,6	Alentejo	10,4	Lisboa e V.Tejo	10,1
Lisboa e V.Tejo	13,8	Lisboa e V.Tejo	10,6	Lisboa e V.Tejo	8,8	Centro	10,1
Algarve	13,8	Algarve	6,7	Algarve	8,3	Algarve	9,2
IPS (%)							
1994		1995		1996		1997	
Centro	45,8	R. A. Açores	46,2	Alentejo	47,6	Alentejo	50,6
Alentejo	41,0	Norte	44,7	R. A. Açores	47,2	R. A. Madeira	47,5
Norte	39,3	Alentejo	41,2	Norte	40,5	Norte	43,8
Total	38,4	Total	38,7	Centro	37,8	R. A. Açores	41,2
Algarve	34,5	Centro	38,7	Total	37,8	Total	40,6
Lisboa e V.Tejo	34,4	Algarve	37,6	Lisboa e V.Tejo	34,1	Lisboa e V.Tejo	38,4
R. A. Açores	28,2	Lisboa e V.Tejo	32,5	Algarve	32,3	Centro	36,0
R. A. Madeira	22,9	R. A. Madeira	28,4	R. A. Madeira	28,7	Algarve	32,7
IPMult (%)							
1994		1995		1996		1997	
R. A. Madeira	10,2	Centro	9,5	R. A. Madeira	10,3	R. A. Madeira	15,8
Alentejo	10,0	Alentejo	8,5	Centro	7,3	Norte	8,2
Centro	9,4	R. A. Madeira	8,5	R. A. Açores	7,1	Total	6,6
Norte	7,8	Norte	8,4	Norte	7,0	Alentejo	6,6
R. A. Açores	7,8	R. A. Açores	7,0	Algarve	5,7	R. A. Açores	5,8
Total	7,0	Total	6,3	Total	5,5	Lisboa e V.Tejo	5,3
Algarve	4,7	Algarve	4,7	Alentejo	5,0	Centro	4,9
Lisboa e V.Tejo	3,7	Lisboa e V.Tejo	2,2	Lisboa e V.Tejo	2,8	Algarve	4,7

Fonte: Cálculos do GEC/SEDS com base no PAUE 1994-1997

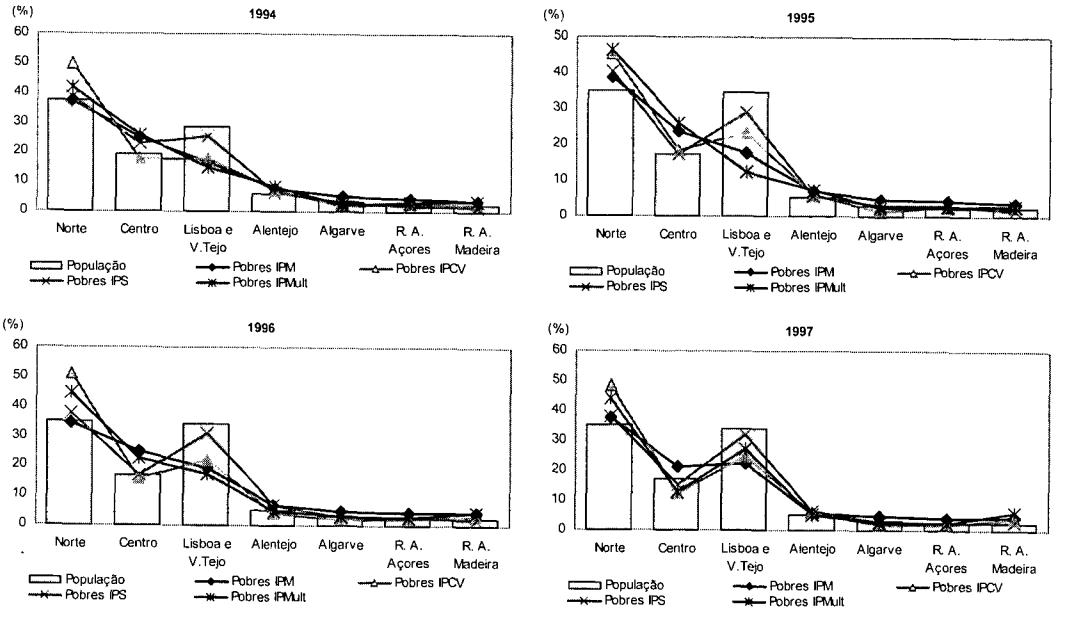
A região de Lisboa e Vale do Tejo apresenta as taxas de pobreza monetária e múltipla mais baixas ao longo do período, excepto em 1997 nos pobres múltiplos, ano em que o Centro e o Algarve apresentam taxas mais baixas.

Relativamente à pobreza segundo as condições de vida, o Algarve destaca-se como a região menos afectada em todos os anos; uma situação similar ao que sucede com a Região Autónoma da Madeira relativamente ao Índice de Pobreza Subjectivo onde, com excepção de 1997, é a região menos desfavorecida. Efectivamente, de 1996 para 1997, a Madeira regista um aumento muito forte na percentagem de indivíduos pobres subjectivos que poderá estar relacionado com o aumento verificado na pobreza monetária em 1997 (relativo a 1996).

Passando para uma análise sobre a distribuição dos pobres e da população total em cada NUTS, que permite avaliar melhor o peso do conjunto dos pobres em cada categoria, verifica-se que, a região de Lisboa e Vale do Tejo é a única cuja proporção dos indivíduos pobres se situa sempre abaixo da proporção total, em qualquer critério de pobreza ao longo dos quatro anos (Figura 4.10).

Ao contrário, todas as outras regiões, especialmente Norte, Açores, Madeira e Alentejo encontram-se na situação oposta, isto é, o quinhão de pobres é quase sempre superior ao da população.

Figuras 4.10. Distribuição da população e de pobreza, segundo o tipo, por NUTS II, 1994-1997



Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

4.10. ESTADO DE SAÚDE

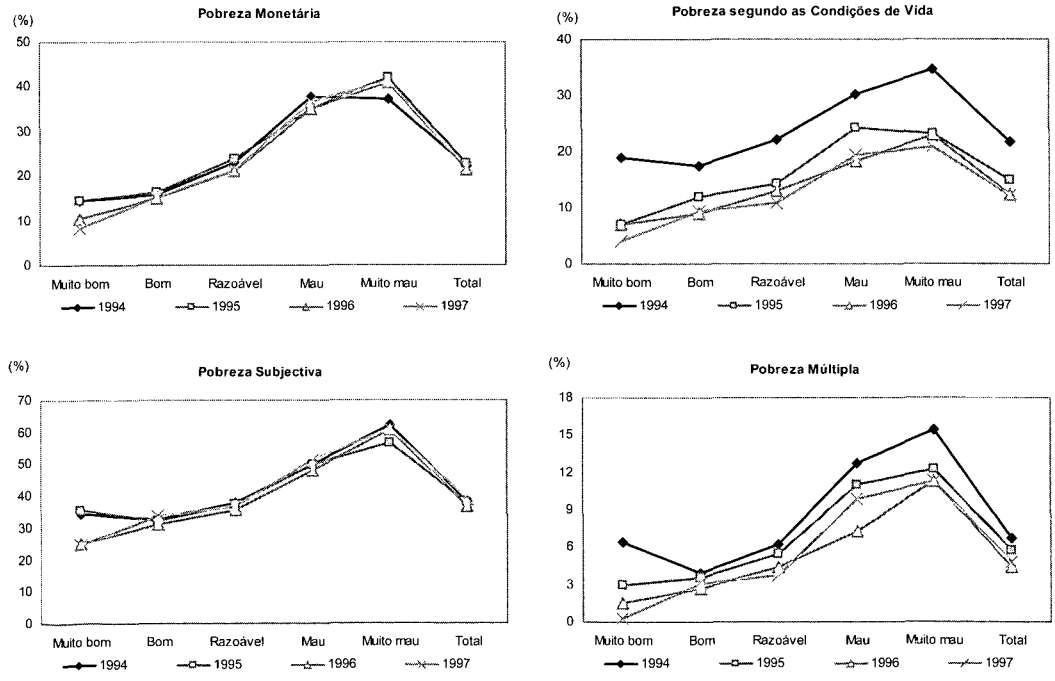
O estado de saúde dos indivíduos é outra das características sobre a qual se pode conjecturar se é causa ou efeito para situações de pobreza, ou ainda, como muitas vezes parece, uma mistura de ambas as situações.

A falta de uma alimentação equilibrada, de um bem estar físico ou psicológico, de condições adequadas de alojamento, entre outros, são factores que propiciam um bom estado de saúde. Por outro lado, problemas com a saúde podem conduzir a situações de emprego precário ou à saída precoce do mercado de trabalho, que por sua vez estão directamente relacionados com escassez de recursos financeiros, e, eventualmente à pobreza.

Os dados do Painel mostram que a maior parte da população inquirida não tem problemas com a saúde. A percentagem dos indivíduos que afirmam que a sua saúde é boa ou muito boa ronda os 50% entre 1994 e 1997; este valor ascende aos 80% se lhes juntarmos os que responderam ter uma saúde razoável.

A incidência da pobreza é tanto maior quanto pior é o estado de saúde, conforme é evidenciado na figura 4.11.

Figuras 4.11. Taxas de Pobreza segundo o estado de saúde dos indivíduos, 1994-1997



Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

4.11. PARTICIPAÇÃO SOCIAL

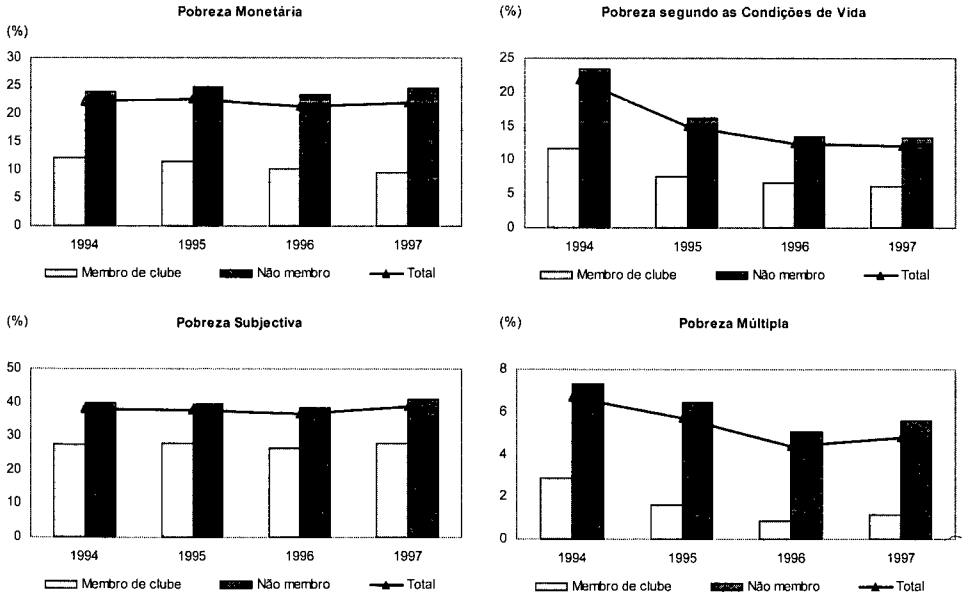
Para finalizar esta breve caracterização dos indivíduos inquiridos no PAUE e, entre eles, os que já vivenciaram situações de pobreza, seleccionou-se uma questão que reflecte o grau de participação social de cada respondente.

A questão foi seleccionada por poder constituir uma aproximação, ainda que superficial, à mobilização da sociedade civil. Os movimentos sociais e de solidariedade social funcionam muitas vezes como motores de desenvolvimento humano e social na medida em que se estabelecem redes de apoio social mais fortes e uma maior proximidade entre os cidadãos, pelo menos os que partilham gostos e interesses comuns.

A questão pretendia saber se o indivíduo era membro de alguma organização cultural ou social, como por exemplo, clube desportivo ou recreativo (incluindo os sócios), associações de bairro ou de partidos políticos.

Como se pode verificar, os indivíduos não membros de qualquer instituição do género registaram taxas de pobreza superiores, com diferenças bastante evidentes.

Figuras 4.12. Taxas de Pobreza segundo a participação social, 1994-1997



Fonte: Cálculos do GEC/SEDS com base no PAUE 1994 a 1997

5. CONCLUSÕES

- Do estudo feito com base nas quatro abordagens de medição da pobreza propostas neste trabalho não se infere uma tendência clara quanto à evolução da pobreza em Portugal entre 1994 e 1997.
- A pobreza monetária manteve-se relativamente estável terminando o período com uma taxa de pobreza ligeiramente superior à de 1994. Ao longo dos quatro anos verificou-se, contudo, uma redução muito expressiva do número de indivíduos com rendimentos monetários por adulto equivalente muito inferiores ao limiar de pobreza.
- Analisando a situação quanto à pobreza monetária em termos europeus verifica-se que para o ano em que há informação tratada pelo EUROSTAT (1996) Portugal regista a mais elevada taxa de pobreza.
- A pobreza segundo as condições de vida é a única abordagem em que se verifica uma redução contínua e notória das taxas de pobreza anuais. Este facto ficou a dever-se essencialmente às melhorias ocorridas no período quanto à posse de bens de consumo corrente, bens de equipamento e características do alojamento. Apenas a qualidade percebida de alguns aspectos relacionados com o alojamento e o seu meio envolvente avaliados pelos próprios indivíduos registaram uma deterioração nos quatro anos.
- A abordagem subjectiva que se baseia exclusivamente na autoavaliação das dificuldades para fazer face a encargos e despesa é de todas as medidas a que classifica como pobres a maior proporção de indivíduos apesar se de restringir esta classificação aos indivíduos que afirmam claramente ter dificuldades ou grandes dificuldades – esta proporção aumentou ligeiramente nos quatro anos.
- As situações cumulativas de pobreza apurada segundo as três abordagens anteriormente mencionadas são sintetizadas através de vários índices de pobreza múltipla que agrupam todos ou apenas dois tipos de pobreza.
- Nos quatro anos em estudo apenas as situações de acumulação de pobreza monetária com pobreza subjectiva registaram crescimento no período. Nas restantes hipóteses de acumulação de pobreza ocorreu uma diminuição do número de indivíduos sujeitos anualmente, com excepção do último ano do período onde, ainda assim, os valores ficaram sempre aquém dos níveis máximos registados em 1994.
- Particular atenção mereceu ainda outro tipo extremamente preocupante de posição face à pobreza: a sujeição a situações de longos períodos de pobreza. Esta análise foi feita através da evolução da pobreza persistente. Verificou-se que praticamente não há indivíduos que permaneçam em situações de pobreza múltipla por mais de dois anos e que a redução anual no número de pobres persistentemente sujeitos a privação ao nível das condições de vida é acentuada.
- O mesmo não sucede com as situações de pobreza monetária e subjectiva: o ritmo de saídas destas situações de pobreza diminui fortemente com o passar dos anos indiciando que existirão porções importantes de população sujeitas

a situações duradouras de dificuldades monetárias: volvidos 4 anos face ao período de referência quase metade dos pobres monetários e subjectivos que o eram em 1994 mantém essa condição.

- A pobreza subjectiva agrega o maior grupo de indivíduos que alguma vez experimentou uma situação de pobreza ao longo dos quatro anos em que se baseia este estudo: mais de 60% da população foi pobre em termos subjectivos durante pelo menos um dos anos em causa.
- Em termos de prevalência, a pobreza múltipla foi o tipo de pobreza que atingiu a menor percentagem de indivíduos ao longo dos quatro anos mas foi também neste tipo de pobreza que ocorreu a maior rotatividade de pobres dado que o número de indivíduos que entre 1994 e 1997 esteve sujeito a pobreza múltipla durante pelo menos um ano era em 1997 quase duplo (13,8%) da maior percentagem de pobres múltiplos registado num único ano (7,0% verificado em 1994).
- Relativamente às características principais dos indivíduos pobres pode concluir-se, em primeiro lugar, que a situação mais desvantajosa que era tradicionalmente atribuída às mulheres só é comprovada de forma clara relativamente à pobreza monetária.
- Como seria de se esperar, os idosos registam as maiores taxas de pobreza, mas tendem a ser acompanhados pelos indivíduos mais jovens, que atingem inclusive as maiores taxas de pobreza nos últimos anos do período para todas as abordagens, com excepção da monetária.
- Relativamente ao estado civil, viúvos e casados destacam-se por se encontrarem em posições opostas no que respeita às maiores e menores taxas de pobreza.
- Como seria também de se esperar, as famílias monoparentais, os idosos, a viver só ou em casal (ambos idosos) e as famílias alargadas, são as categorias mais vulneráveis a qualquer tipo de pobreza.
- Baixos níveis escolares, profissões pouco qualificadas ou exercidas no sector agrícola, parecem ser factores que estão associados a situações de pobreza, designadamente do tipo monetário. Os trabalhadores não remunerados e os desempregados, entre os indivíduos activos, os reformados e domésticos, entre os não activos, registam as maiores taxas de pobreza.
- As questões do rendimento resumem-se numa ideia chave: os indivíduos que dependem das transferências sociais registam taxas de pobreza muito elevadas quando comparadas com os indivíduos cuja principal fonte de rendimento advém do trabalho; as taxas dos pensionistas pobres e dos que auferem outros benefícios sociais são bom exemplo disso.
- A estes factores acrescem os indivíduos com um estado de saúde precário e os que não registam qualquer tipo de participação social, ambos com maiores percentagem de pobres.
- As Regiões Autónomas dos Açores e Madeira são as mais desfavorecidas geograficamente já que aí residem um maior número de pobres monetários. O Norte, o Centro e o Alentejo também se associam à regiões mais pobres nos outros tipos de pobreza.

BIBLIOGRAFIA

ALMEIDA, João Ferreira; CAPUCHA, Luís Antunes; COSTA, António Firmino da; MACHADO, Fernando Luís; NICOLAU, Isabel; REIS, Elizabeth (1992) "Exclusão Social: Factores e Tipos de Pobreza em Portugal", Celta Editora, Oeiras.

ANDREB, Hans-Jurgen et all (1997), Empirical Poverty Research in a Comparative Perspective, Ashgate, Suffolk.

ATKINSON, A.B. (1987) "On the Measurement of Poverty", Econometrica, Vol.55, Nº 4 – July 1987, 749-764.

BRANCO, Rui (2000), "Non-monetary Poverty Measures", Instituto Nacional de Estatística, Seminar on International Comparisons of Poverty, Bratislava, 5-6 June 2000, Session 1: Measuring Poverty; comparative studies in France and in Central Europe.

BRANCO, Rui; GONÇALVES, Cristina (2001) "Metodologia de Cálculo do Índice de Pobreza segundo as Condições de Vida", Gabinete de Estudos e Conjuntura/ Serviço de Estudos Demográficos e Sociais, Março/2001.

BRANCO, Rui, GONÇALVES, Cristina, "Envelhecimento Demográfico – Aspectos Demográficos, Económicos e Sociais da População Idosa em Portugal", Instituto Nacional de Estatística, I Congresso Português de Demografia, org. ISCTE/INESLA, Tróia, Grândola, 21-23 Setembro 2000, Sessão Plenária: População e Envelhecimento.

BRANCO, Rui, GONÇALVES, Cristina, "Population Ageing – demographic, economic and social aspects of older persons in Portugal", Conferência Europeia de População 2001, Helsínquia – Finlândia, 7 a 9 de Junho de 2001, Tema F: Envelhecimento Demográfico.

CHRISTOPHER, Karen; ENGLAND, Paula; McLANAHAN, Sara; PHILIPS, Katherin Ross; SMEEDING, Timothy M. (2000) "Gender Inequality in Poverty in Affluent Nations: The Role of Single Motherhood and the State", JCPR Working Paper 108, 01-08-2000. URL: http://www.jcpr.org/wp/Wpprofile.cfm?ID=126

COSTA, A. Bruto da (1993), "Pobres Idosos" in Estudos Demográficos nº 31, Instituto Nacional de Estatística - Gabinete de Estudos Demográficos, Lisboa.

COSTA, A. Bruto da (1998), "Exclusões Sociais" Cadernos Democráticos, colecção Fracturas nº 2, Lisboa.

EUROSTAT (1998) "Recommendations on Social Exclusion and Poverty Statistics", 31st Meeting of the Statistical Programme Committee", Luxembourg.

EUROSTAT (2000) "Rapport d'étape sur la mise en œuvre des recommandations de la TF sur les "statistiques de l'exclusion sociale et de la pauvreté", approuvé par la 31e réunion du CPS les 26 et 27 novembre 1998", 37eme réunion du Comité du Programme Statistique, Porto, 31 mai 2000.

EUROSTAT (2000), "Living Conditions in Europe – Statistical pocketbook", França

EUROSTAT (2000), "European Social Statistics – Income, poverty and social exclusion", Luxemburgo.

FERREIRA, Leonor Vasconcelos (2000) "Distribuição do Rendimento e Pobreza: A Região Norte no Contexto Nacional entre 1990 e 1995", Instituto Nacional de Estatística – Direcção Regional do Norte, Estatísticas & Estudos Regionais, Set-Dez 2000, N°.24

FALL, Madior; HORECKÝ, Marián; ROHÁCOVÀ, Eva (1997) "La pauvreté en Slovaquie et en France: quelques éléments de comparaison", Economie et Statistique N° 308-309-310, INSEE.

FLEURBAEY, Herpin, MATINEZ, Verger, "Can poverty be measured?", Studies n° 21, November 1998, INSEE.

HAGENAARS, A.J.M. (1986) "The Perception of Poverty", North-Holland, Amsterdam

HAGENAARS, Aldi J. M.; VAN PRAAG, Bernard M. S. (1985) "A Synthesis of Poverty Line Definitions", The Review of Income and Wealth, n° 2,

INE (1992), "Inquérito aos Orçamentos Familiares 1989/1990", Instituto Nacional de Estatística, Lisboa.

INE (1992), "Metodologia do Inquérito aos Orçamentos Familiares 1989/1990", Série Estudos n° 62, Instituto Nacional de Estatística, Lisboa.

INE (1997), "Inquérito aos Orçamentos Familiares: Metodologia 1994/1995", Série Estudos n° 74, Instituto Nacional de Estatística – Dep. Estatísticas Sócio-Económicas, Lisboa.

INE (1997), "Inquérito aos Orçamentos Familiares: Resultados 1994/1995", Instituto Nacional de Estatística - Dep. Estatísticas Sócio-Económicas, Lisboa.

INE (1999), "As Gerações Mais Idosas", Série Estudos n° 83, Instituto Nacional de Estatística - Gabinete de Estudos e Conjuntura, Lisboa.

LOLLIVIER, S., VERGER, D. (1997a) "Condition de vie - Pauvreté d'existence, monétaire ou subjective sont distinctes" (or "Living-Conditions Poverty, Resources Poverty, and Subjective Poverty are Distinct") Economie et Statistique N°308-309-310, INSEE. URL: http://www.insee.fr/va/produits/pub/studies/artis/is027.htm

LOLLIVIER, S., VERGER, D. (1997b) "Une approche de la pauvreté par les conditions de vie", F 9701, INSEE.

LOLLIVIER, Stéfan; VERGER, Daniel (1997), "Pauvreté d'existence, monétaire ou subjective sont distinctes", Economie et Statistique, n° 308-309-310, 1997 – 8/9/10 (pp. 113-142), INSEE, Seminar on International Comparisons of Poverty, Bratislava, 5-6 June 2000.

MARTIN-GUZMAN, M. Pilar (s.d.) "The Construction and Use of Level of Living Indicators", Training of European Statisticians, Course ESO-C-02-E, Instituto Nacional de Estadística, Madrid.

MEDINA, Fernando (1999) "Equivalence Scales: A Brief Review of concept and Methods", ECLAC, apresentado no 3° Encontro do Grupo de Peritos em Estatísticas sobre Pobreza (Grupo do Rio), Lisboa, 22-24 de Novembro de 1999.

MEJER, Lene (1999) "Statistics on Social Exclusion: The EU Methodological Approach", Eurostat, apresentado no 3º Encontro do Grupo de Peritos em Estatísticas sobre Pobreza (Grupo do Rio), Lisboa, 22-24 de Novembro de 1999.

MINUJIN, Alberto (1999) "Putting Children into Poverty Statistics", UNICEF, apresentado no 3º Encontro do Grupo de Peritos em Estatísticas sobre Pobreza (Grupo do Rio), Lisboa, 22-24 de Novembro de 1999.

NUNES, Celso Luís Pereira (1999), Tese de Mestrado em Estudos Económicos e Sociais: "Linhas de Pobreza para Portugal Continental: 1989/90 e 1994/95", Universidade do Minho, Escola de Economia e Gestão, Gualtar.

RAVALLION, Martin e LOKSHIN, Michael; «Identifying Welfare Effects from Subjective Questions»,World Bank.

RAVALLION, Martin e LOKSHIN, Michael; «Subjective Economic Welfare», Development Research Group, The World Bank.

RODRIGUES, Carlos Farinha (1996) "Medida e Decomposição da Desigualdade em Portugal (1980/81 – 1989/90), Revista de Estatística – 3º Quad 1996, nº 3, Instituto Nacional de Estatística, Lisboa.

RODRIGUES, Carlos Farinha (1999) "Income Distribution and Poverty in Portugal 1994/1995 - A comparison between European Community Household Panel and the Household Budget Survey", June 1999, CISEP, ISEG / Universidade Técnica de Lisboa.

RODRIGUES, Carlos Farinha (1999) "Repartição do Rendimento e Pobreza em Portugal (1994/95) (Uma Comparação entre Painel de Agregados Familiares e o Inquérito aos Orçamentos Familiares), Revista de Estatística – 1º Quad 99, nº 10, Instituto Nacional de Estatística, Lisboa.

SOARES, Regina; D'UVA, Teresa Bago (2000), "Income, Inequality and Poverty", Instituto Nacional de Estatística, Seminar on International Comparisons of Poverty, Bratislava, 5-6 June 2000, Session 2: Analysis of Poverty in Europe.

SHORT, Kathleen (1999) "Experimental Poverty Thresholds for the United States 1990 to 1998", U.S. Census Bureau/ HHESD/ Poverty Measurement Research apresentado no 3º Encontro do Grupo de Peritos em Estatísticas sobre Pobreza (Grupo do Rio), Lisboa, 22-24 de Novembro de 1999.

VERGER, Daniel (2000), "Les Mesures de la Pauvreté – Trois types d'approche: Approches monétaire, en termes de conditions de vie et subjective. Les résultats français", INSEE, Seminar on International Comparisons of Poverty, Bratislava, 5-6 June 2000.

VERGER, Daniel (2000), "Ressources disponibles et Consommation des familles modestes: la multiplicité ds approches de la pauvreté", INSEE, Seminar on International Comparisons of Poverty, Bratislava, 5-6 June 2000.

ANEXO I: NOTAS METODOLÓGICAS

I. Metodologia para cálculo do Índice de Pobreza Monetária

O *Índice de Pobreza Monetária* (IPM) tem por base um conceito de receita líquida total que incorpora rendimentos monetários e sobre o qual se identifica a linha de pobreza. A fonte utilizada foram as vagas de 1994 a 1997 do Painel dos Agregados Familiares da União Europeia (PAUE) disponibilizadas pelo EUROSTAT. Sublinhe-se que a informação monetária recolhida em cada vaga se refere a rendimentos do ano anterior, ou seja, na prática estão disponíveis dados referentes ao período entre 1993 e 1996.

Descrição sumária das várias etapas operacionais até se obter o IPM.

1. Apuramento da **receita líquida total** de cada Agregado Doméstico Privado (ADP) considerando o conjunto de ordenados e salários, rendimentos do trabalho por conta própria, rendimentos privados excluindo os do trabalho, pensões e outras transferências sociais (todos os dados tratados a preços correntes).
2. Cálculo e posterior imputação a cada indivíduo da **receita líquida total por adulto equivalente**, tendo em conta a escala de equivalência da OCDE Modificada aplicada a cada agregado:

1º adulto	= 1
Restantes adultos	= 0,5
Crianças < 14 anos	= 0,3

3. A **determinação da linha de pobreza** tem subjacente o critério estabelecido pelo EUROSTAT, traçando a linha nos 60% da mediana do valor da receita líquida total por adulto equivalente, atendendo à distribuição da receita pelos indivíduos.
4. O **valor do Índice** determina-se pela percentagem de indivíduos que tem rendimento inferior ao estabelecido pela linha de pobreza monetária.

II. Metodologia para cálculo do Índice de Pobreza segundo as Condições de Vida

O *Índice de Pobreza segundo as Condições de Vida* (IPCV) incorpora informação relativa essencialmente a quatro vertentes: capacidade do ADP em adquirir bens de consumo corrente; características do alojamentos; problemas percebidos do alojamento e posse de bens de equipamento. É tanto mais pobre um indivíduo ou agregado quanto maior for a acumulação de privação no conjunto dos itens considerados para a elaboração do índice. O PAEU (1994 a 1997) foi a fonte utilizada.

Descrição sumária das várias etapas operacionais até se obter o IPCV.

1. **Selecção das variáveis** com base nos seguintes critérios: Teste de consenso: não existindo nenhum estudo prévio/inquérito de opinião sobre que variáveis os indivíduos privilegiavam para formar um limiar mínimo de condições de vida, este passo de selecção foi ultrapassado com

base nas variáveis inquiridas – que resultaram de vários debates no seio de grupos de trabalho do EUROSTAT –, bem como, com base na avaliação da relevância atribuída por parte dos autores deste trabalho às variáveis disponíveis face às características da sociedade actual e tendo em conta a realidade que se pretendia estudar.

Teste de frequência: tal como é preconizado por alguns autores desta temática, foi considerado irrelevante dado que a metodologia adoptada tem em consideração um ponderador final de privação que agrupa informação de todas as variáveis. Assim, o peso relativo das variáveis é proporcional à sua verificação junto dos agregados ou indivíduos, ou seja, a contribuição de uma variável que esteja representada apenas num número ínfimo de observações tem um peso igualmente ínfimo no cálculo do ponderador global de privação.

⇨ **Variáveis consideradas comuns a todas as vagas do PAUE:**

- *Capacidade de adquirir bens de consumo corrente:* ter aquecimento adequado em casa; pagar uma semana de férias, por ano, for a de casa; substituir móveis usados por novos; comprar roupa nova com alguma frequência; ter uma refeição de carne ou peixe pelo menos de 2 em 2 dias; convidar familiares ao amigos para refeição ou bebida em sua casa pelo menos uma vez por mês;
- *Características do alojamento:* cozinha separada; banheira ou chuveiro; retrete no interior; água corrente; aquecimento central ou termo-acumulador; varanda (que permita colocar cadeira para apanhar sol), terraço ou jardim.
- *Problemas percebidos com o alojamento:* alojamento com falta de espaço; ruído proveniente dos vizinhos ou do exterior; muito escuro/falta de luminosidade; falta de condições para aquecimento adequado; telhado mete água; fundações, paredes ou chão húmidos; caruncho nos caixilhos das janelas ou no soalho; poluição ou outro problema ambiental provocado pelo trânsito ou indústria; crime ou vandalismo na área de residência;
- *Bens de equipamento:* automóvel; TV a cores; videogravador; micro-ondas; máquina de lavar loiça; telefone; segunda habitação.

⇨ **Variáveis consideradas apenas no PAUE de 1996 e 1997**: computador pessoal.

⇨ **Variável construída**: número de indivíduos por assoalhada (penalização para alojamentos onde residam mais de 1 indivíduo de um mesmo ADP por assoalhada.

2. **Atribuição de ponderações de penalização**: **1** (penalização) e **0** (sem penalização)
3. **Recodificação** das variáveis nas ponderações anteriores e apuramento das respectivas frequências.
4. **Ponderação** das variáveis a incluir no *score* pelas respectivas frequências de não privação. O *score* afecto a cada ADP/indivíduo inclui, desta forma, as

penalizações acumuladas por cada bem em falta, corrigidas da respectiva importância determinada pela sociedade via frequência de não privação.

5. **Construção do *Score* individual**. Verifica-se que quanto maior for o número de variáveis que o ADP/ indivíduo possui, menor é o grau de privação e vice-versa.
6. **Cálculo**: hieraquização das observações dos ADP/individuos por grau de privação e determinação dos percentis.
7. O critério escolhido para a **determinação da linha de pobreza** foi o de (tomando o ano de 1994 como referência), seleccionar uma sub-população com dimensão muito próxima da identificada como pobre pelo Índice de Pobreza Monetária. Ou seja, apurou-se qual era o nível de privação de entre as observações do *score* hierarquizadas que mais se aproximava de classificar 22,5% de indivíduos como pobres segundo as condições de vida. Apurado o valor que produzia esse efeito para 1994, foi mantido invariável nos anos seguintes permitindo assim detectar-se qual a tendência de evolução que se seguiu (fosse ela de agravamento ou de melhoria face à pobreza). Em 1996 o valor de referência foi recalculado mantendo a mesma proporcionalidade face ao valor estabelecido em 1994. Este cálculo deveu-se à incorporação da variável adicional "computador pessoal" que passou o ser recolhida pelo PAUE e que se decidiu incorporar no *score*.

III. Metodologia para cálculo do Índice de Pobreza Subjectiva

O *Índice de Pobreza Subjectiva* (IPS) teve por base a seguinte pergunta: "Um agregado pode ter diferentes fontes de rendimento e mais do que um membro do agregado pode contribuir para o orçamento do mesmo. Tendo em conta a totalidade do rendimento mensal do seu agregado, qual das seguintes opções melhor define a capacidade do agregado para fazer face aos encargos e despesas?

1. Com grande dificuldade / 2. Com dificuldade / 3. Com alguma dificuldade / 4. Com alguma facilidade / 5. Com facilidade / 6. Com grande facilidade ".

Classificaram-se como agregados pobres nesta abordagem subjectiva todos os que afirmassem ter "dificuldade" ou "grande dificuldade" em fazer face aos encargos e despesas.

IV. Metodologia para o cálculo do Índice de Pobreza Múltipla

O *Índice de Pobreza Múltipla* (IPM) resulta da acumulação dos três tipos de pobreza apresentados anteriormente, por cada ADP/indivíduo.

A utilização desta medida de pobreza é justificada quer pela análise multidimensional que se pretendeu dar a este trabalho, quer pelo auxílio que presta no estudo de tendências da pobreza em determinadas sub-populações.

ANEXO II:

Scores para construção do Índice de Pobreza segundo as Condições de Vida – PAUE 1994 a 1997
Dados ponderados para indivíduos

Variável Desdobramento(s)	Nome variável	Legenda	Pond.	1994	1995	1996	1997
				%			
Capacidade de adquirir bens de consumo corrente							
Ter aquecimento adequado em casa	hf003	no	1	71,4	64,7	66,5	65,3
		yes	0	28,6	35,3	33,5	34,7
Pagar uma semana de férias, por ano, fora de casa	hf004	no	1	61,0	58,8	62,3	63,3
		yes	0	39,0	41,2	37,7	36,7
Substituir móveis usados por novos	hf005	no	1	73,1	75,1	72,4	72,7
		yes	0	26,9	24,9	27,6	27,3
Comprar roupa nova com alguma frequência	hf006	no	1	50,4	45,7	43,2	40,6
		yes	0	49,6	54,3	56,8	59,4
Ter uma refeição de carne ou peixe pelo menos de 2 em 2 dias	hf007	no	1	8,8	6,3	6,4	7,3
		yes	0	91,2	93,7	93,6	92,7
Convidar amigos/familiares para refeição ou bebida em sua casa pelo menos uma vez por mês	hf008	no	1	28,9	20,3	17,9	19,0
		yes	0	71,1	79,7	82,1	81,0
*Índice de Privação Médio para Bens de Consumo Corrente**				48,9	45,2	44,8	44,7
Características do alojamento							
Cozinha separada	ha008	no	1	1,4	1,3	1,6	1,3
		yes	0	98,6	98,7	98,4	98,7
Banheira ou chuveiro	ha009	no	1	15,1	12,2	10,2	10,1
		yes	0	84,9	87,8	89,8	89,9
Retrete no interior	ha010	no	1	13,8	10,5	9,3	9,2
		yes	0	86,2	89,5	90,7	90,8
Água corrente	ha011	no	1	21,1	17,7	16,5	14,4
		yes	0	78,9	82,3	83,5	85,6
Aquecimento central ou termo-acumulador	ha012	no	1	90,8	89,5	89,2	87,0
		yes	0	9,2	10,5	10,8	13,0
Varanda (que permita colocar cadeira para apanhar sol), terraço ou jardim	ha013	no	1	28,9	24,2	21,3	21,0
		yes	0	71,1	75,8	78,7	79,0
*Índice de Privação Médio para Caracteristicas do alojamento**				28,5	25,9	24,7	23,8
Variáveis construídas							
Nº de indivíduos por assoalhada							
Um ou menos		<=1	0	65,5	67,9	66,7	66,5
Mais de um	100.0	>1	1	34,5	32,1	33,3	33,5
Problemas percebidos do alojamento							
Alojamento com falta de espaço	ha014	no	0	61,7	66,6	67,2	67,2
		yes	1	38,3	33,4	32,8	32,8
Ruído proveniente dos vizinhos ou do exterior	ha015	no	0	82,0	83,6	76,3	75,5
		yes	1	18,0	16,4	23,7	24,5
Muito escuro/falta de luminosidade	ha016	no	0	82,8	82,8	83,2	81,3
		yes	1	17,2	17,2	16,8	18,7
Falta de condições para aquecimento adequado	ha017	no	0	59,1	61,5	61,1	59,0
		yes	1	40,9	38,5	38,9	41,0
Telhado que mete água	ha018	no	0	80,9	84,7	83,8	83,2
		yes	1	19,1	15,3	16,2	16,8
Fundações, paredes ou chão humidos	ha019	no	0	64,3	67,4	66,3	62,4
		yes	1	35,7	32,6	33,7	37,6
Caruncho nos caixilhos das janelas ou no soalho	ha020	no	0	72,1	72,0	72,2	71,1
		yes	1	27,9	28,0	27,8	28,9
Poluição ou outro problema ambiental provocado pelo trânsito ou indústria	ha021	no	0	81,4	80,9	82,2	80,6
		yes	1	18,6	19,1	17,8	19,4
Crime ou vandalismo na área de residência	ha022	no	0	81,3	78,4	77,9	78,7
		yes	1	18,7	21,6	22,1	21,3
*Índice de Privação Médio para Problemas Percebidos com o alojamento**				26,0	24,7	25,5	26,8
Bens de equipamento							
Automóvel	hb001	no-cannot afford	1	28,4	26,4	23,0	22,8
		no - other reason	0	9,4	7,7	8,5	7,8
		yes	0	62,1	65,9	68,5	69,4
TV a cores	hb002	no-cannot afford	1	11,0	7,7	6,5	5,1
		no - other reason	0	2,0	1,0	0,9	0,5
		yes	0	87,0	91,3	92,6	94,4
Videogravador	hb003	no-cannot afford	1	33,7	31,4	30,0	28,2
		no - other reason	0	15,4	12,2	12,6	11,5
		yes	0	50,9	56,4	57,4	60,3
Micro-Ondas	hb004	no-cannot afford	1	45,9	47,6	43,9	43,5
		no - other reason	0	42,7	36,7	37,8	35,0
		yes	0	11,4	15,7	18,3	21,6
Máquina de lavar loiça	hb005	no-cannot afford	1	50,1	50,9	49,1	47,9
		no - other reason	0	31,7	29,7	31,6	30,6
		yes	0	18,2	19,4	19,3	21,6
Telefone	hb006	no-cannot afford	1	18,5	16,8	16,1	15,7
		no - other reason	0	5,0	3,9	3,1	2,3
		yes	0	76,4	79,3	80,8	82,0
Segunda habitação	hb007	no-cannot afford	1	71,2	73,3	70,1	71,6
		no - other reason	0	19,9	16,9	20,6	19,1
		yes	0	8,9	9,7	9,3	9,3
Computador pessoal	hb008	no-cannot afford	1	n.a.	n.a.	42,1	40,0
		no - other reason	0	n.a.	n.a.	40,6	39,8
		yes	0	n.a.	n.a.	17,4	20,2
*Índice de Privação Médio para Bens de Equipamento**				37,0	36,3	35,1	34,4
Índice de Privação Médio* =				34,2	32,2	32,0	32,0

*Somatório das frequências de privação acumuladas/nº de variáveis de privação consideradas

REVISTA DE ESTATÍSTICA
189ª PÁGINA
3º QUADRIMESTRE DE 2001

Survey on National Consumer Price Indexes

Autor:
Rui Evangelista

SURVEY ON NATIONAL CONSUMER PRICE INDEXES

INQUÉRITO SOBRE ÍNDICES DE PREÇOS NO CONSUMIDOR

Autor: Rui Evangelista([22])
- Técnico Superior de Estatística no Departamento de Síntese Económica de Conjuntura - Núcleo de Estatísticas de Preços no Consumidor

ABSTRACT:

- This text presents the main findings of a survey, carried out by the prices division of the Portuguese National Statistical Institute, on national Consumer Price Indexes. The survey had two goals: firstly, it attempted to gather basic information on other countries' CPIs; secondly, it attempted to identify common preferences in relation to future developments and on the use of certain techniques such as new quality adjustment methods.

  The survey results enabled us to find out some interesting trends, particularly on the use of new data sources and on the use of quality adjustment methods. From the answers given, this latter issue is, and will continue to be, highly regarded in the next years.

KEY-WORDS:

- *Consumer Price Index, data sources, quality adjustment methods.*

RESUMO:

- Este texto apresenta os principais resultados de um inquérito sobre Índices de Preços no Consumidor levado a cabo pelo Núcleo de Estatísticas de Preços no Consumidor do Instituto Nacional de Estatística. O inquérito teve dois objectivos: em primeiro lugar, reunir informação de base sobre Índices de Preços no Consumidor de outros países e, em segundo lugar, identificar 'preferências' comuns sobre futuros desenvolvimentos e sobre o uso de algumas técnicas como, por exemplo, novos métodos de ajustamento da qualidade.

  Os resultados obtidos permitem-nos identificar algumas tendências interessantes, especialmente em relação ao uso de novas fontes de informação e sobre o uso de

---

[22] This article is based on a text written for the benefit of the statistical agencies that have requested some feedback on the survey main findings. I would like to express my gratitude for the helpful suggestions given by colleagues of some of the contacted statistical offices. Thanks also to Ana Cabral, Cristina Fernandes and Daniel Santos for the comments provided.

métodos de ajustamento da qualidade. A partir das respostas obtidas, este último ponto é, e continuará a ser, altamente considerado nos próximos anos.

PALAVRAS-CHAVE:

- *Índice de Preços no Consumidor, fontes de informação, métodos de ajustamento da qualidade.*

1. INTRODUÇÃO

The prices division of the Portuguese National Statistical Institute is currently working on the revision of its Consumer Price Index. In this context, the department has decided to carry out a survey of national CPIs. This text presents and summarises the main findings of the survey.

The underlying idea of the survey is that, by gathering information on other countries' Consumer Price Indexes, additional guidance could be given on the present CPI revision process. Thus, the general objectives of the survey were twofold. Firstly, it attempted to gather basic information on other CPIs. Secondly, it attempted to identify common preferences (or trends) in relation to CPI developments and on the use of certain techniques/procedures such as new quality adjustment methods.

This text is organised as follows. Part two describes the scope and overall objectives of the survey. It provides information on the followed sampling strategy and survey response rates. Part three presents the main empirical results generated by the survey. Some data analysis is carried out in this section. Finally, a few considerations are made in part four of this text.

1. METHODOLOGY

2.1. SURVEY MEDIUM

This survey was sent out via e-mail only. However, respondents were given the option to send their answers to the questionnaire either by e-mail or post.

2.2. DESIGN OF THE QUESTIONNAIRE

Hill and Hill (1998) provided the initial theoretical and practical guidelines for the development of the survey instrument. In June/July, a small pilot was carried out with the objective of obtaining some feedback on the clarity of the questionnaire. This pilot helped introduce some changes to the original questionnaire. Appendix A shows the questionnaire used in this survey.

2.3. SAMPLING STRATEGY AND RESPONSE TO THE QUESTIONNAIRE

This survey uses non-probability sampling and therefore cannot ensure that participants are representative of all the statistical offices in the world. The sample was, however, predefined so that Europe, particularly the countries belonging to the European Union, would be well represented.

Countries were grouped into seven 'geographical areas', guidelines for the definition of which were provided in the 1997 United Nations statistical yearbook (UN, 1997). From all of the seven sample groups a total of 72 countries were chosen and sent an e-mail containing a questionnaire and a covering letter. The names of the surveyed countries are given in appendix B. Table one provides information on the chosen sample groups.

Table 1. Sample groups

Geographical area identification label	Geographical area	Number of surveyed countries
G1	European Union	15
G2	Eastern Europe	6
G3	Other European Countries	11
G4	Africa	6
G5	Asia and Oceania	17
G6	Latin America and the Caribbean	14
G7	Northern America	3

A total of 43 responses were received, representing an overall response rate of almost 60%. However, two of these responses were not usable, since several parts of the questionnaire had not been completed. Table two details response rates by geographical area.

Table 2. Response rate by geographical area

Geographical area identification label[1]	No. of usable responses	Geographical area response rate (%)	Proportion of total usable answers (%)
G1	14	93.33	34.15
G2	6	100	14.63
G3	7	63.64	17.07
G4	2	33.33	4.88
G5	8	47.06	19.51
G6	3	21.43	7.32
G7	1	33.33	2.44
Total	*41*	-	*100*

Note: See table one for definitions of geographical areas.

Two preliminary remarks can be drawn from tables one and two. Firstly, it should be noted that our sample of respondents rests heavily on the answers given by countries belonging to the G1, G2 and G3 groups. While these areas account for 44% of total sampled countries, they represent almost 66% of total usable answers. Secondly, response rates seem to differ across areas, with the G1, G2 and G3 sample groups displaying higher response rates than other groups. In presence of such an outcome, the 'survey-generated' information should be interpreted with caution. In analysing the data, one should bear in mind that the overall results might be biased due to the influence that the G1, G2 and G3 groups have on the total sample.

3. SURVEY RESULTS

This is a presentation of some of the data from our survey on national CPIs. For a better understanding of the replies analysed below, please consult the questionnaire in appendix A. Below is a selection of results that seem worth mentioning.

3.1. OVERALL CPI CHARACTERISATION

- Only two of the respondents answered that their CPI is released on a quarterly basis. All other respondents opted for the 'monthly' option for defining their price index periodicity.
- 68.3% of all respondents considered that their national CPIs could be best defined as a fixed base index. The countries defining their CPIs as 'a chain index with annual links' belong to the G1, G2 and G3 geographical areas.
- 63.4% of all respondents declared that the cost-of-living index (COLI) concept does not provide the theoretical framework for their national CPI. This proportion climbs to 78.6% for the G1 geographical area (in the G1, G2 and G3 areas this figure reaches 74.1%).
- The total number of countries claiming to produce a seasonally adjusted CPI was 34.1%. In the G1 group, this percentage is 21.4%.

3.2. PRICE DATA SOURCES

The table below summarises the answers given in relation to the combination of data sources used for price collection.

Table 3. **Used sources on prices**

Sources / Areas	Percentage of respondents that use... Agents/price collectors	Mail/telephone surveys	Internet	Mail catalogues/ Magazines	Other
G1	100%	93%	86%	79%	29%
G2	100%	50%	17%	-	17%
G3	71%	57%	43%	71%	29%
G5	88%	63%	25%	38%	38%
G4 + G6 + G7	83%	83%	50%	50%	-
All sample	90%	73%	51%	54%	24%

Note: n = 41.

The results shown in table three allow us to say that:

- By and large, the most used source of information on prices comes from the work developed by pricing agents. The percentage of respondents that use this source range from 71% (G3 group) to 100% (G1 and G2 groups). Furthermore, answers given to the Part B, question 2 of the questionnaire, show that this source is, without a doubt, the *main* source of price information for the compilation of national CPIs.
- The percentage of countries stating to use the internet for the collection of prices, is much higher in the G1 area than in any other group of countries.
- 'Other' quoted sources of information include scanner data and the purchase of private data sources/services by statistical offices. Contracting out the task of price collection seems to be a reality, at least for some areas of national CPIs.

Further analysis of the answers given in part B of the questionnaire shows that:

- 68.3% of all respondents revealed that the main source of information on prices is not collected throughout the whole month.
- 42.9% of all respondents declared to collect their prices within the first fifteen days of the month (main source of information only). The percentage of countries that collect their prices in the last fifteen days of the week is

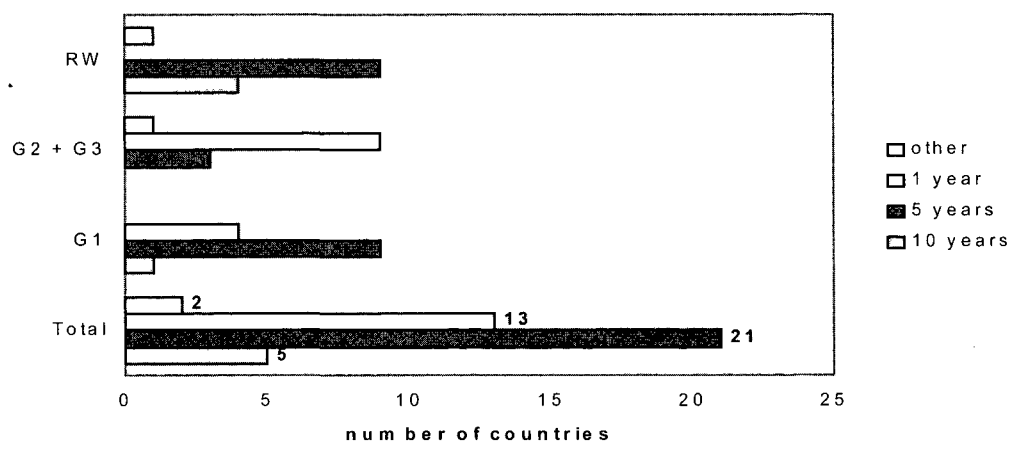
merely of 10.7%. The proportion of respondents declaring to collect prices in an overlapping period is 46.4%.

- Only 7.3% of all respondents (three countries) intend to change the frequency of their price collection. Of these three countries, two are planning to increase it and one to decrease it.
- The number of countries that is currently using scanner data for obtaining information on prices is very low. Only 4.9% of all respondents use this data source.
- However, it is worth noting that six countries declared that they are either studying the possibility of using scanner data or going to use it in the near future. Three of these countries belong to the G1 area.

3.3. WEIGHT DATA SOURCES

- Household Budget Surveys are the main source for CPI weight updating. National Accounts data account for a smaller proportion of the given answers on this question and are only regarded as the main source for weight updating in the G1 and G2 groups.
- 36.6% of all respondents stated that their main source for weight updating is available on an annual basis. In the G1 area, this figure is 57.1%.
- 51.2% of all respondents answered that, on average, weights are updated every five years. Graphic one splits the respondents into four groups according to the answers given to question three' in part C of the questionnaire.

Grap. 1: Average CPI weight updating



- Weight updating is most frequently done on a five year basis. The RW([23]) group has the highest number of countries changing their weight structure on a ten year basis. Moreover, none of the respondents belonging to this group change their weights on an annual basis.
- 70.7% of all respondents are *not* planning to reduce the time lag between the introduction of different weight structures. This figure drops, however, to 64,3% in the G1 group. In the 'G2 and G3' group this figure is 76.9%.
- One year is the period which is most quoted by the respondents that are planning to reduce the time lag mentioned above: 58.3% of the statistical offices wishing to change it have quoted this period as the desired frequency.

Table four summarises the responses given on the 'mix' of weight updating data sources.

Table 4. Additional weight data sources

Sources / Areas	Percentage of respondents that use... Governmental Institutions	Mail/ telephone surveys	Internet	Magazines	Other	Percentage of countries that use 4 or mores sources
G1	86%	57%	64%	50%	71%	57%
G2	67%	17%	17%	17%	67%	17%
G3	71%	57%	43%	43%	71%	43%
G5	50%	25%	13%	13%	13%	13%
G4 + G6 + G7	50%	33%	-	-	33%	-
All sample	68%	42%	34%	29%	54%	32%

Note: n = 41.

- Considered as a whole, the G1 area has the highest percentage of countries using four or more weight data sources (57%). The corresponding percentage is substantially lower in all the other groups, ranging from 13% to 43%.
- Countries having their CPIs chained annually tend to combine their National Accounts and Household Budget Surveys data for weight updating. While the former data source is used in the shaping of weights at higher levels, the latter is used for the more detailed level of aggregation.

[23] For the sake of simplicity, it was decided to include all non-European countries in the 'RW group'.

3.4. QUALITY ADJUSTMENT METHODS AND CPI BIAS MEASUREMENT

Part D of the questionnaire sought to assess the present use of quality adjustment methods used by statistical offices. It also aimed sought to obtain information on the introduction of new quality adjustment methods and CPI bias measurement.

The table below provides information about which quality adjustment methods are currently being used by our sample of respondents.

Table 5. Used quality adjustment methods by commodities/group of commodities

Method	Percentage of countries that reported the use of...					Percentage of countries that use more than one method
Group	Overlap pricing	Option pricing	Hedonic pricing	Other	None	
Food at home, fast food and restaurants	78%	2%	-	29%	17%	27%
Household appliances	78%	17%	5%	37%	10%	42%
Radio/TV, computers, furniture and telephone	80%	24%	5%	37%	2%	41%
Clothing and footwear	73%	5%	7%	39%	10%	34%
New and used cars	68%	34%	5%	34%	10%	46%
Health	49%	2%	-	27%	41%	21%
Entertainment	68%	10%	2%	34%	20%	32%

Note: n = 41.

The data allow us to say that:

- The overlapping pricing method accounts for the highest percentages in all commodities/group of commodities covered by the survey. This seems to be the most known used technique for the adjustment of quality change.
- The group 'Radio/TV, computers, furniture and telephone' has the lowest percentage of answers in the 'none' option. This is one of the groups that has the highest percentage of countries that use more than one method for the adjustment of quality changes. This supports the idea that sees this group of commodities as the one with the most frequent problems in terms of quality change.

- 'New and used cars' is the group that accounts for the highest percentage for the use of the option pricing method (34%). This seems to be the group in which this method is more readily accepted.
- From the responses given, it seems that the hedonic pricing method is, at present, not widely used. Only six countries use the hedonic regression method for the adjustment of quality change in the seven CPI areas covered by the survey. Table six presents the countries that are presently using hedonics in their CPI compilation.

Table 6. The use of the Hedonic Price Method in national CPI

Country	Group of commodities	Commodities on which this method is applied	Relevant literature
USA	· Household appliances · Clothing and footwear · Radio / TV, computers, furniture and telephone	Clothing, computers, televisions	Moulton (2001) Fixler et al. (1999) Moulton et al. (1998)
France	· Household appliances · Clothing and footwear · Entertainment	Dishwashers Man shirts with sleeves, Bestselers	Bascher and Lacroix (1999)
Colombia	· Radio / TV, computers, furniture and telephone · New and used cars	-	Delgado (1998)
Sweden	· Clothing and footwear	Clothing	Norberg and Ribe (2001)
Japan	· Radio / TV, computers, furniture and telephone	-	-
Finland	· New and used cars	New and used cars	Vartia (1997)

From the answers given to the questionnaire, there are some reasons to believe that the scenario on quality adjustment will change considerably in the near future. Thus:

- 60.0% of all respondents plan to introduce new quality adjustment methods in their CPIs (69.2% in the G1 area);
- 87.5% of those respondents planning to introduce new quality adjustment methods will introduce them during the next five years([24]).
- The hedonic regression method was referred to by 82.6% of the total number of respondents that answered to the open-ended question on which new quality adjustment methods were to be introduced in their national CPIs.

Somewhat surprisingly, the survey revealed that only a small proportion of statistical offices had tried to quantify the magnitude of their CPI bias. Thus:

[24] 25.0% are planning to introduce new quality adjustment methods during the next year.

- At the time of the survey, only 24.4% of the surveyed institutions had made (or had seen other institutions make) some sort of CPI bias estimation.
- CPI biases were last estimated between 1996-2000 (the most frequent year is 1999). It is interesting to note that all these estimations were made in the same year or subsequent to the release of the 'Boskin Report' (Boskin et al., 1996).

In the last question of the questionnaire, respondents were asked to rank, by order of importance, six possible CPI improvements. A total of 38 answers were received([25]). Table seven shows the results.

Table 7 CPI Improvements: obtained scores by geographical region

Area	Improvement (a)	Improvement (b)	Improvement (c)	Improvement (d)	Improvement (e)	Improvement (f)
G1	49 (3)	58 (4)	62 (5)	37 (1)	44 (2)	44 (2)
G2 + G3	28 (2)	46 (6)	39 (4)	31 (3)	42 (5)	24 (1)
RW	45 (4)	39 (3)	76 (6)	38 (2)	63 (5)	33 (1)

n = 38.

Notes: (a) Implementation of new Statistical and sampling techniques.
(b) More frequent weight updates.
(c) Estimation of the size of various CPI bias.
(d) Implementation of new technologies for obtaining readily available information on prices and consumed quantities of goods.
(e) Introduction and/or expanded use of the hedonic pricing method.
(f) Introduction/development of new sources of information for collecting prices.
The ranking of each improvement within each geographical group is shown in brackets.

From table seven we see that:

- The introduction and/or development of the hedonic pricing method – improvement (e) -, is given more importance in the G1 group than in the rest of the other considered groups.
- The importance of more frequent weight updates – improvement (b) - seems to be more highly regarded in the RW group than in European countries (as mentioned earlier, this sample group shows the highest number of countries with a weight updating of ten years and the lowest number of countries (i.e. zero) with an average weight updating of one year).

[25] Three countries did not answer this question arguing that were unable to differentiate between the six provided improvements.

3.5. CPI METHODOLOGIES

The survey provided the opportunity to gather information on other countries' CPIs. Thus, we have decided to include, at the end of the questionnaire, a request for publications on CPIs methodologies.

Table eight provides an overview of the publications that have been sent to our national statistical office.

Table 8. Methodologies

Institution	Publication
Australian Bureau of Statistics	Guide to the Consumer Price Index: 14th Series
Central Statistics Office (Ireland)	CPI: Introduction of updated Series (base: mid-November 1996 as 100)
DANE (Colombia)	(a) Metodologia Indice de Precios al Consumidor (IPC - 98); (b) Guia de Uso del IPC (97 - 98).
Ministere des Affaires Economiques (Belgium)	L'indice des Prix a la Consommation: base 96
National Statistical Service of Greece	Revised CPI (1994 = 100)
State Institute of Statistics (Turkey)	Urban Places CPI (1994 = 100)
Statistics Finland	The CPI 1995 = 100: Handbook for Users

It should be borne in mind that table eight only covers the publications which have been sent to us by mail. Some additional information could also be extracted from the internet because some countries have given their CPI website addresses. Moreover, in the absence of a written methodology in English, some other countries have chosen to send us papers and recent CPI press releases.

4. CONCLUDING REMARKS

This survey has highlighted some points that seem worth mentioning. Firstly, it is interesting to note that the countries having their CPIs chained annually tend to use and combine their National Accounts and Household Budget Surveys data for the purposes of weight updating. The availability and quality of these two data sources seems to be an important point to take into account if one wants to build a CPI with annual links.

Secondly, the COLI concept does not provide the theoretical framework for national CPIs. Although this seems to be best applied for Europe as a whole, almost 64% of the sample do not base the building of the CPI on the cost of living concept.

This, of course, raises the question of whether or not theoretical constructs – such as the COLI concept - are important in CPI design.

Thirdly, the survey has disclosed 'unusual' price data sources which use may be promising in the near future. Prices gathered and/or collected by private companies and scanner data are examples of 'uncommon price data sources' that were referred in the survey. As the complexity of an economy grows, the more difficult the collection of prices will become. In order to tackle this problem, some extra attention should be devoted to the use of alternative means of price collection.

Finally, the issue of quality adjustment is, especially in Europe, highly regarded. More than half of all respondents plan to introduce new quality adjustment methods in the next five years. In this context, the introduction and development of explicit quality adjustment methods – such as hedonics – assume particular importance.

However, it is somewhat surprising that, despite this interest on quality adjustment methods, only a few countries have attempted to quantify quality and other possible CPI bias. Only 10 out of the 41 surveyed countries have taken such an empirical approach to this important issue.

REFERENCES

BASHER J. and LACROIX, T. (1999). Dish-washers and PCs in the French CPI: Hedonic Modeling, From Design to Practice, Paper presented at the fifth meeting of the International Working Group on Price Indices, Reykjavik, Iceland, August, Institut National de la Statistique et des Études Économiques.

BOSKIN, M.J, DULBERGER, E.R., GORDON, R.J., GRILICHES, Z. and JORGENSON, D. (1996). Toward a more Accurate Measure of the Cost of Living, Final Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index.

DELGADO, E. E. (1998). Metolodologia Indice de Precios al Consumidor IPC-98, DANE - Departamento Administrativo Nacional de Estadística, Colombia.

FIXLER, D., FORTUNA, C., GREENLEES and LANE, W. (1999). The Use of Hedonic Regressions to Handle Quality Change: The Experience in the U.S. CPI, Paper presented at the fifth meeting of the International Working Group on Price Indices, Reykjavik, Iceland, August, Bureau of Labour Statistics.

HILL, M. M. and HILL, A. (1998). A Construção de um Questionário, Working Paper n.º 98/11, Dinâmica – Centro de Estudos Sobre a Mudança Sócio-económica, Instituto Superior de Ciência do Trabalho e da Empresa, Universidade Técnica de Lisboa.

MOULTON, B. R. (2001). The Expanding Role of Hedonic Methods in the Official Statistics of the United States, Bureau of Economic Analysis, http://www.bea.doc.gov/bea/about/expand3.pdf.

MOULTON, B. R., LAFLEUR, T. J. and MOSES, K. E. (1998). Research on Improved Quality Adjustment in the CPI: The Case of Televisions, Bureau of Labour Statistics, http://www4.statcan.ca/secure/english/ottawagroup/pdf/18moul98.pdf.

NORBERG, A. and RIBE, M. (2001). Hedonic Methods in Price Statistics: Swedish Applications, Paper presented at the Deutsche Bundesbank Symposium on Hedonic Methods in Price Statistics, Wiesbaden, Germany, 21-22 June, Statistics Sweden.

UNITED NATIONS (1997). Statistical Yearbook, Department of Economic and Social Affairs, Statistics Division, New York.

VARTIA, Y. (1997). Quality Differences and Prices of New Cars: Price per Quality Unity Approach, Preliminary Version of a paper for a planning meeting at Statistics Finland, February 1997, University of Helsinki.

APPENDIX A: CONSUMER PRICE INDEX QUESTIONNAIRE

Please fill in the questionnaire and send it back to the following e-mail address: rui.evangelista@ine.pt. Alternatively, your answers can be sent to:

Gabinete de Estudos e Conjuntura do Instituto Nacional de Estatística
C/O: Rui Evangelista
Av. António José de Almeida; n.º 5
1000 – 043 Lisboa
Portugal

If you prefer to send us the questionnaire by e-mail, please replace the relevant box '❒' with a cross ('x').

Part A) Overall CPI characterisation

1. Periodicity

❒ monthly

❒ quarterly

❒ other

2. Which one of the following formulas is used for the aggregation of prices in your national CPI?

❒ arithmetic mean

❒ geometric mean

❒ harmonic mean

❒ other
(please specify below)

_____

_____

3. Which formula is used in the compilation of higher level indexes?

❒ Laspeyres

❒ Paasche

❒ other
(please specify below)

_____

4. Your national CPI can be best defined as:

❒ a fixed base index

❒ a chain index with annual links

❒ other (please specify below)

_____

5. What is the *price* reference period for your national CPI?

❒ a whole year

❒ a single month

❒ other

6. What is the *weight* reference period for your national CPI?

❒ a whole year

❒ a single month

❒ other (please specify below)

_____

7. Does the concept of the cost-of-living index provide the theoretical framework for your national CPI?

❒ yes ❒ no

8. Does your institution produce a seasonally adjusted CPI?

❒ yes ❒ no

Part B) Price data sources

1. Which data sources does your institution use for collecting prices?

❒ Agents/price collectors

❒ Mail/telephone surveys

❒ Internet

❒ Mail catalogues/Magazines

❒ other (please specify below)

_____

2. Please state, which is the main source of information for collecting prices

---

3. For the main source of information only (as stated in the Part B, Q2):

a) When are prices collected?

❒ throughout the whole month

❒ on a() specific day(s)/week(s) of the relevant month

b) If you have marked 'on a() specific day(s)/week(s) of the relevant month' please identify the month period in which they are collected.

❒ within the first fifteen days of the month

❒ within the last fifteen days of the month

c) Is your Statistical Office planning to change the frequency of price collection?

❒ yes

❒ no

If '*yes*', will it be....

❒ increased

❒ decreased

4. Does your institution use scanner data for obtaining information on prices?

❒ yes

❒ no

Part C) Weight data sources

1. Which is the main source of information for your national CPI weights?

❒ Household Budget Surveys

❒ National Accounts

❒ Other
(please specify below)

---

2. Availability of the above mentioned main source

❒ quinquennial

❒ biennial

❒ annual

❒ other
(please specify below)

---

3. On average, CPI weights are updated on...

❒ a ten-year basis

❒ a five-year basis

❒ a two-year basis

❒ an annual basis

4. Are you planning to reduce the time lag between the introduction of two different weight structures?

❒ yes ❒ no

If 'yes', what will be the desired frequency?

_____ year(s).

5. Which other data sources do you use for updating the index weight structure?

❒ Information gathered by governmental regulatory institutions
❒ Mail/telephone surveys
❒ Data available on the Internet
❒ Magazines
❒ other (please specify below)

Part D) Quality adjustment methods and CPI bias measurement

1. Which methods are currently used for the adjustment of quality changes of the following commodities/group of commodities?

Food at home, fast food and restaurants

❒ Overlap pricing ❒ Option pricing ❒ hedonic pricing ❒ other ❒ none

Household appliances ([26])

❒ Overlap pricing ❒ Option pricing ❒ hedonic pricing ❒ other ❒ none

Radio / TV, computers, furniture and telephone

❒ Overlap pricing ❒ Option pricing ❒ hedonic pricing ❒ other ❒ none

Clothing and footwear

❒ Overlap pricing ❒ Option pricing ❒ hedonic pricing ❒ other ❒ none

New and used cars

❒ Overlap pricing ❒ Option pricing ❒ hedonic pricing ❒ other ❒ none

---

[26] This group includes commodities such as refrigerators, freezers, washing machines, dishwashers, stoves, microwave ovens, toasters, etc.

Health ([27])

❒ Overlap pricing ❒ Option pricing ❒ hedonic pricing ❒ other ❒ none

Entertainment ([28])

❒ Overlap pricing ❒ Option pricing ❒ hedonic pricing ❒ other ❒ none

2. Are you planning to introduce new quality adjustment methods in your CPI?

❒ yes ❒ no

If '*yes*', which ones?

1._____

2._____

3._____

4._____

When?

❒ During the next year

❒ During the next five years

❒ In more than five years

3. Has your institution ever made any CPI bias estimation?

❒ yes ❒ no

4. Has the size of your national CPI bias ever been estimated by other institution?

❒ yes ❒ no

If marked '*yes*' in at least one of the two above questions (that is, either in Q3 or Q4 or both), please mention when it was last estimated

_____

[27] This group includes commodities like pharmaceuticals, dental services, hospital treatments, etc.

[28] This group includes commodities such as newspapers, books, sporting equipment, package holidays etc.

5. Taken into account your national CPI present structure, how would you rank, by order of importance, the following improvements?

(Please give 1 to the most important item; 2 to the item which is, in your opinion, the next important, and continue in this fashion up until you give 6 as the least important item.)

Improvement	**Importance**
(a)Implementation of new statistical and sampling techniques	___
(b)More frequent weight updates	___
(c) Estimation of the size of various CPI bias	___
(d) Implementation of new technologies for obtaining readily information on prices and consumed quantities of goods	___
(e) Introduction and/or expanded use of the hedonic pricing method	___
(f) Introduction/development of new sources of information for collecting prices	___

We would be very grateful if you could send us recent methodological publications on your CPI together with this questionnaire.

Thank you very much for answering this questionnaire

APPENDIX B: SURVEYED COUNTRIES

European community

Austria	Netherlands	Germany
Italy	Finland	Sweden
Belgium	Portugal	Greece
Luxembourg	France	UK
Denmark	Spain	Ireland

Eastern Europe

Belarus	Czech Republic	Poland
Bulgaria	Hungary	Slovakia

Other European Countries

Croatia	Faeroe Islands	Latvia
Malta	Slovenia	Turkey
Estonia	Iceland	Lithuania
Norway	Switzerland	

Africa

Benin	Botswana	Malawi
South Africa	Seychelles	Tunisia

Asia and Oceania

Australia	Mongolia	Indonesia
Jordan	Hong Kong Special Administrative Region of China	Singapore
Armenia		Israel
Korea	New Zealand	Thailand
China	India	Japan
Malaysia	Philippines	
Cyrus		

Latin America and the Caribbean

Argentina	Jamaica	Colombia
El Salvador	Brazil	Saint Lucia
Aruba	Mexico	Dominican Republic
Ecuador	Chile	Uruguay
Bolivia	Peru	

Northern America

Canada	Greenland	USA

INFORMAÇÕES

ACTIVIDADES E PROJECTOS IMPORTANTES DO SISTEMA ESTATÍSTICO NACIONAL

IMPORTANTS ACTIVITIES AND PROJECTS OF THE NATIONAL STATISTICAL SYSTEM

PRINCIPAIS OBJECTIVOS E ACÇÕES NO ÂMBITO DO SISTEMA ESTATÍSTICO NACIONAL 2002

O processo de planeamento da actividade estatística abrange a definição dos objectivos estratégicos, através das Linhas Gerais da Actividade Estatística Nacional, de periodicidade quadrienal; a concretização destes objectivos e a apresentação das principais acções a desenvolver a médio prazo, no Programa Estatístico de Médio Prazo, com a mesma periodicidade; e a planificação anual das acções a executar, no Plano de Actividades do INE e das Entidades com delegação de competências do INE.

As Linhas Gerais da Actividade Estatística Nacional e respectivas prioridades para 1998-2002 foram aprovadas pelo Conselho Superior de Estatística em Maio de 1997 (Deliberação nº 125/97). O Programa Estatístico de Médio Prazo 1998-2002, que obteve parecer favorável do Conselho Superior de Estatística, em Novembro de 1997 (Deliberação nº 135/97), concretiza os principais objectivos estratégicos e acções de desenvolvimento para os vários domínios da actividade estatística.

Na sequência dos dois documentos acima referidos, foi elaborado o Plano de Actividades do INE e das Entidades com delegação de competências do INE para 2002, sobre o qual o Conselho Superior de Estatística emitiu parecer favorável em Dezembro de 2001 (Deliberação nº 223/01).

1. PRINCIPAIS OBJECTIVOS A DESENVOLVER EM 2002

No ano de 2002, o Sistema Estatístico Nacional inicia uma nova etapa do seu desenvolvimento, marcada pela introdução de novos patamares de exigência na função coordenação e pela importância que é atribuída aos objectivos relacionados com o incremento da coerência e integração dos produtos estatísticos oficiais.

Na função coordenação do sistema, pretende-se implementar novas iniciativas potenciadoras da cooperação inter-institucional no seio do SEN e com outros prestadores de serviços públicos de informação. Este posicionamento visa optimizar a capacidade de resposta do SEN às crescentes exigências no grau de cobertura, na qualidade e na oportunidade da informação estatística produzida, favorecendo, para o

efeito, as condições de auscultação das necessidades específicas dos utilizadores, iniciada em 2001 com o lançamento de um inquérito ao grau de satisfação dos utilizadores da informação produzida pelo INE.

Neste âmbito, o Instituto Nacional de Estatística, enquanto órgão central do SEN, procederá ao desenvolvimento de instrumentos que contribuam para este desiderato. A organização dos produtos estatísticos em domínios lógicos e coerentes permitirá implementar um modelo de gestão estratégica de subsistemas de informação estatística que possibilitará a introdução de significativos ganhos de eficácia no desempenho da função de coordenação.

A coerência e integração constituem objectivos centrais das iniciativas que se pretendem empreender, em particular, no robustecimento metodológico e na consistência da informação de suporte à elaboração do sistema de contas nacionais e do sistema integrado de indicadores de conjuntura.

Para o ano 2002 serão de assinalar ainda os seguintes objectivos:

- Difundir os resultados provisórios e definitivos dos Censos 2001, facto que constitui um resultado histórico na capacidade de tratamento da informação, só possível devido aos avanços tecnológicos em áreas como as da leitura óptica e das técnicas de reconhecimento de caracteres;
- Desenvolver medidas para concretizar as obrigações nacionais decorrentes do Plano de Acção relativo às necessidades estatísticas da União Económica e Monetária, por forma a atingir um nível correspondente à média dos três Estados Membros com melhor desempenho, tanto em termos de prazos como de qualidade da informação;
- Intensificar a incorporação dos desenvolvimentos propiciados pelas novas tecnologias de informação e comunicação que permitam racionalizar os procedimentos de recolha, tratamento e difusão da informação estatística, permitindo ganhos na fiabilidade e actualidade da informação, bem como diminuir a sobrecarga sobre os respondentes;
- Aprofundar a articulação inter-institucional ao nível regional no sentido de promover o desenvolvimento do sistema de informação regional compatível com as necessidades de planeamento e avaliação de programas de âmbito regional;
- Consolidar os sistemas de meta-informação estatística harmonizados ao nível do Sistema Estatístico Nacional e ao nível do Sistema Estatístico Europeu;
- Desenvolver o sistema de informação geográfica do INE, em particular na definição de uma rotina de manutenção e actualização permanente da Base Geográfica de Referenciação da Informação e no lançamento de operações piloto tendentes à criação de uma Base Geográfica de Referenciação de Edifícios;
- Implementar novos projectos com destaque para o Sistema de Informação das Operações Urbanísticas, o Sistema Integrado de Informação sobre as Empresas, o Sistema Integrado de Informação sobre as Famílias, o Sistema

Integrado de Informação sobre as Cidades e o Sistema Integrado de Informação sobre o Ambiente.

Para assegurar a consecução dos objectivos enunciados permanecem relevantes as seguintes medidas de política:

- Remoção das dificuldades que persistem no acesso, por parte do INE e dos seus órgãos delegados, a todas as fontes administrativas de informação relevante para a produção das estatísticas oficiais;
- Garantia, no âmbito da aplicação da Lei nº 67/98 – Lei da Protecção de Dados Pessoais - , do acesso e utilização para fins estatísticos de dados pessoais e dos correspondentes ficheiros e suportes informáticos;
- Materialização da contratualização das relações entre o Estado e o INE na parte correspondente à produção de estatísticas oficiais legalmente obrigatórias e de outras estatísticas incluídas no Programa Estatístico Comunitário;
- Resolução do problema das instalações do INE, por via da construção de um edifício, já projectado, no terreno de que o Instituto dispõe e é seu património próprio.

2. OBJECTIVOS E ACÇÕES A DESENVOLVER EM 2002

A. Novas necessidades estatísticas da União Europeia

Plano de Acção da União Económica Monetária

O Plano de Acção aprovado em 29 de Setembro de 2000 pelo ECOFIN, inventaria as principais estatísticas a desenvolver face às prioridades políticas relevantes para todos os Estados Membros da UEM. No âmbito do Plano de Acção deverão ser desenvolvidos cinco grupos de indicadores:

- Contas Nacionais Trimestrais (principais agregados);
- Estatísticas Trimestrais das Administrações Públicas;
- Estatísticas do Mercado de Trabalho;
- Indicadores de Curto Prazo;
- Estatísticas do Comércio Internacional.

Todas as operações estatísticas directamente relacionadas com o Plano de Acção e as respectivas fontes de informação são consideradas de prioridade absoluta, razão pela qual têm sido objecto de um acompanhamento regular rigoroso.

O objectivo a alcançar pelo INE é o cumprimento dos prazos estipulados para a produção dos indicadores incluídos no Plano de Acção e o seu envio atempado ao EUROSTAT. Pretende-se que, no final de 2002, Portugal consiga estar ao lado dos

países mais avançados nesta matéria, o que significa que em relação a cada um dos indicadores, se terá de verificar uma convergência para o padrão correspondente à média dos três Estados Membros com melhor desempenho, tanto em termos de prazos como de qualidade da informação.

Foi realizado no INE um trabalho de análise dos indicadores do Plano de Acção, no que respeita à sua cobertura, prazos de disponibilidade, responsabilidades, recursos necessários e principais problemas a resolver, de modo a tornar possível o envio atempado ao EUROSTAT das estatísticas nacionais, de acordo com os padrões definidos neste Plano de Acção.

O cumprimento destas metas implicará três tipos de acções:

- Simplificação das metodologias, nomeadamente das Contas Nacionais Trimestrais e do Comércio Internacional;
- Informatização dos processos, com o desenvolvimento de novas soluções informáticas para as estatísticas de base e índices do Comércio Internacional, Dormidas na hotelaria e Painel Trimestral das Empresas;
- Reforço de recursos humanos nas seguintes áreas: Contas Nacionais Trimestrais, Indicadores de Curto Prazo e Estatísticas do Mercado de Trabalho.
- Releva-se o objectivo de se reduzir o prazo de disponibilidade dos cinco grupos de indicadores.

Neste contexto, as iniciativas a desenvolver serão as seguintes:

- *Contas Nacionais Trimestrais*

  O INE tem disponibilizado as CNT, em 2001, num prazo de 120 dias após o fim do trimestre de referência, de acordo com o Regulamento Comunitário SEC 95. Considera-se ser uma meta a atingir a antecipação do prazo de disponibilização para 70 dias após o fim do trimestre de referência, a partir do final do ano 2002.

- *Estatísticas Trimestrais das Administrações Públicas*

  Estas estatísticas são transmitidas de acordo com as exigências do Plano de Acção.

- *Estatísticas do Mercado de Trabalho*

  As questões que se colocam têm a ver com quatro fontes essenciais:
  - Índice de Custo do Trabalho
  - Inquérito ao Emprego
  - Taxa de desemprego (mensal)
  - População e Emprego (Quadro 1 do SEC 95)

  Os dados estatísticos relativos às três primeiras fontes cumprem os requisitos definidos no Plano de Acção.

  A informação relativa ao Quadro 1 do SEC 95 não tem sido transmitida ao EUROSTAT: Está em curso a realização de estudos metodológicos para

obtenção dos indicadores solicitados, prevendo-se o início da transmissão de resultados no final do ano de 2002.

- *Indicadores de Curto Prazo*

  São disponibilizados indicadores sobre a Indústria, Construção (com excepção das licenças de construção), Comércio a Retalho e Serviços (neste último caso, com excepção do volume de vendas), em conformidade com os prazos definidos no Plano de Acção. Prevê-se a transmissão de todos os indicadores solicitados a partir do 2º semestre de 2002.

- *Estatísticas do Comércio Internacional*

  Todos os dados são transmitidos, embora com algum atraso no que se refere às estatísticas intra--comunitárias. Prevê-se uma melhoria no tocante aos prazos de disponibilização em duas fases: 70 dias após o final do período de referência, a partir do 2º semestre de 2002, e 50 dias a partir do final do ano de 2002.

Indicadores Estruturais

No Conselho Europeu de Lisboa (Março de 2000), foi definido para a União Europeia um objectivo estratégico para 2010: *"tornar-se na economia do conhecimento mais competitiva e dinâmica do mundo, capaz de garantir um crescimento económico sustentável, com mais e melhores postos de trabalho e uma maior coesão social"*.

O Conselho considerou, então, ser necessário discutir e avaliar os progressos obtidos na consecução deste objectivo, e incumbiu a Comissão de elaborar um Relatório de Síntese anual, com base num conjunto de indicadores estruturais globais e relativos ao emprego, inovação e investigação, reforma económica e coesão social e ambiente.

Os indicadores estruturais reportam-se a 3 níveis:

- Nível 1 – O nível superior é utilizado na elaboração do relatório de síntese anual da Comissão.
- Nível 2 – O nível intermédio será utilizado para avaliação da implementação das "Orientações Gerais de Política Económica", definidas com base em objectivos específicos de crescimento, emprego, inovação e coesão social.
- Nível 3 – O nível mais detalhado de indicadores será utilizado para acompanhar e avaliar o progresso das várias iniciativas políticas e planos de acção dos diferentes serviços da Comissão Europeia.

Os três níveis estão interligados – os indicadores do nível 1 constituem um subconjunto dos de nível 2, os quais por sua vez, são um subconjunto dos de nível 3.

A Comissão pretende que os indicadores sejam em número limitado, estáveis ao longo do tempo e de fácil compreensão.

Para o Relatório de Síntese de 2002 a Comissão apresentou uma revisão à anterior lista de indicadores a qual, entre outras alterações de natureza específica, passou a incluir um novo domínio sobre o Ambiente:

Indicadores Económicos Globais

- PIB per capita (em padrões de poder de compra) e taxa de crescimento real do PIB
- Produtividade do trabalho (por pessoa empregada e por hora trabalhada)
- Crescimento do emprego
- Taxa de inflação
- Crescimento do custo unitário real do trabalho
- Défice do sector público
- Dívida da Administração Pública

Emprego

- Taxa de emprego (total e por sexo)
- Taxa de emprego, 55-64 anos
- Desigualdades salariais entre homens e mulheres
- Incidência fiscal nos trabalhadores de baixos salários
- Formação ao longo da vida
- Acidentes de trabalho
- Taxa de desemprego

Inovação e Investigação

- espesas em Recursos Humanos (Despesa pública em educação)
- Despesa em I&D
- Nível de acesso à Internet (alojamentos e empresas)
- Doutoramentos em ciência e tecnologia
- Patentes
- Capital de risco
- Despesa em Tecnologias de Informação e Comunicações

Reforma Económica

- Níveis relativos de preços e convergência de preços
- Preços nas indústrias de rede
- Estrutura de mercado nas indústrias de rede
- Contratos públicos
- Apoios estatais sectoriais e ad hoc
- Aumentos de capital no mercado bolsista
- Investimento das empresas

Coesão Social

- Distribuição de rendimento
- Risco da pobreza
- Risco persistente de pobreza

- Coesão regional (variação do desemprego – NUTS II)
- Abandono escolar precoce
- Taxa de desemprego de longa duração
- Alojamentos com pessoas desempregadas

Ambiente

- Emissões de gases de efeito de estufa
- Intensidade energética da economia
- Intensidade de tráfego – veículo-quilómetro por unidade do PIB
- Transporte de passageiros e de mercadorias por modo de transporte
- Qualidade do ar
- Produção e destino final de resíduos sólidos municipais
- Contributo das energias renováveis para a produção de electricidade

A compilação dos indicadores pelo EUROSTAT e respectiva confirmação pelos Institutos Nacionais de Estatística dos Estados Membros, decorre até ao início de Janeiro de cada ano.

Os dados compilados são actualizados periodicamente, sendo utilizados nos Relatórios de Síntese da Comissão aos Conselhos da Primavera.

O INE acompanha este processo através da participação nos Grupos de Trabalho do EUROSTAT e do Comité de Política Económica, neste último conjuntamente com o Departamento de Prospectiva e Planeamento do Ministério do Planeamento.

B Outros Objectivos e Principais Acções a desenvolver

Tendo sido desenvolvidos, com um maior detalhe, os objectivos relacionados com o Plano de Acção relativo às necessidades estatísticas da União Económica e Monetária e os Indicadores Estruturais, apresentam-se seguidamente outros objectivos a atingir no âmbito do Plano de Actividades, bem como as principais acções a desenvolver em 2002.

Como objectivos gerais, salientam-se:

- Adequar, de modo crescente, a produção e disponibilização de informação estatística às necessidades dos diferentes utilizadores, nomeadamente através do CSE, e dos resultados do inquérito ao grau de satisfação dos utilizadores de informação produzida pelo INE;
- Intensificar a incorporação dos desenvolvimentos propiciados pelas novas tecnologias de informação e comunicação que permitam racionalizar os procedimentos de recolha, tratamento e difusão da informação estatística permitindo ganhos na fiabilidade e actualidade da informação, bem como diminuir a sobrecarga sobre os respondentes;
- Incrementar a coerência e a integração dos produtos estatísticos oficiais, através do aprofundamento do modelo de gestão orientado para a coordenação estratégica de Sistemas de informação. São de destacar o

Sistema de Informação das Operações Urbanísticas, o Sistema Integrado de Informação sobre as Empresas, o Sistema Integrado de Informação sobre as Famílias, o Sistema Integrado de Informação sobre as Cidades e o Sistema Integrado de Informação sobre o Ambiente;

- Promover o desenvolvimento de sistemas de integração de informação nas diferentes dimensões: curto prazo e estrutural, transversal e longitudinal, intra e inter subsistemas de informação estatística;
- Garantir gradualmente a disponibilização de séries longas compatibilizadas para os principais produtos estatísticos em particular no domínio da informação económica e social.

No sentido de alcançar estes objectivos, o Plano de Actividades para 2002 tem inscritos 350 modelos estatísticos, dos quais 248 da responsabilidade do INE (70,9%), e 102 de outras entidades intervenientes na produção estatística oficial (29,1%). Dos 350 modelos estatísticos, 48 são de prioridade absoluta (13,7%), 259 são de 1ª prioridade (74%), e 43 são de 2ª prioridade (12,3%). Do total dos modelos, 142 são obrigatórios por Legislação Nacional ou Comunitária (40,6%). Apresentam-se ainda 17 modelos estatísticos novos (4,9%), considerando-se os restantes 333 de desenvolvimento corrente (95,1%).

O Plano de Actividades para 2002 apresenta a caracterização de todos os modelos estatísticos nele inscritos, ventilados por 30 áreas estatísticas, a seguir enumeradas:

⇒ Administrações Públicas
⇒ Agricultura, Produção Animal e Silvicultura
⇒ Ambiente
⇒ Ciência e Tecnologia
⇒ Comércio Internacional
⇒ Comércio Interno e Outros Serviços
⇒ Condições de Vida das Famílias
⇒ Conjuntura Económica
⇒ Contas Nacionais e Regionais
⇒ Cultura, Desporto e Recreio
⇒ Deficiência e Reabilitação
⇒ Demografia
⇒ Educação
⇒ Emprego e Salários
⇒ Empresas
⇒ Estatísticas Gerais
⇒ Formação Profissional
⇒ Habitação, Construção e Obras Públicas
⇒ Indústria e Energia
⇒ Iniciativas de Produção e Estudos Regionais
⇒ Instituições Financeiras e Seguros
⇒ Justiça
⇒ Pesca
⇒ Preços
⇒ Protecção Social
⇒ Relações e Condições do Trabalho
⇒ Saúde
⇒ Sociedade da Informação
⇒ Transportes e Comunicações
⇒ Turismo e Restauração

1. Produção Estatística

1.1. Estatísticas e Indicadores Económicos

Principais objectivos:

- Consolidar os calendários de disponibilização das Contas Nacionais portuguesas o que envolve a continuidade do cumprimento dos prazos dos denominados exercícios dos défices orçamentais excessivos e recursos próprios PNB e IVA, e das Contas Trimestrais, e a convergência tendencial para os prazos de disponibilização das Contas Nacionais definitivas a que o INE é obrigado no quadro do Regulamento SEC 95;
- Evoluir tendencialmente para uma situação de efectivo cumprimento dos prazos impostos pelo Regulamento SEC, em matéria de Contas Regionais;
- Adaptar o processo de produção de Contas Nacionais (com relevo para as Contas Trimestrais), às exigências do "Plano de Acção para a UEM" decidido pelo ECOFIN, que impõe calendários ainda mais exigentes de disponibilização de contas nalguns domínios;
- Reduzir os prazos de disponibilidade dos indicadores de curto prazo no quadro do Plano de Acção da UEM;
- Desenvolver um Sistema Integrado de Informação sobre as Empresas;
- Desenvolver um Sistema de Informação sobre as Operações Urbanísticas;
- Promover o desenvolvimento de novos indicadores relevantes, num ambiente de integração na economia europeia;
- Desenvolver os trabalhos ligados à mudança de base do IPC;
- Desenvolver o Sistema de Informação sobre a Agricultura.

Acções a desenvolver:

- Reestruturação do processo de produção das Contas Nacionais definitivas, particularmente no segmento das contas por ramo de actividade, tendo por objectivo disponibilizar no futuro aquelas contas com o desfasamento de três anos imposto pelo Regulamento SEC 95;
- Retoma da elaboração de Contas Nacionais provisórias, como forma de cumprir alguns dos prazos do Regulamento SEC 95, que envolvem desfasamentos, mesmo em sede das contas anuais, de somente nove a doze meses;
- Implementação dos trabalhos conducentes à estruturação de um sistema de informação sobre Turismo e de um estudo preliminar sobre a Conta Satélite do Turismo;
- Desenvolvimento das Contas Económicas da Agricultura e implementação das Contas Económicas da Silvicultura e da Pesca;
- Adequação das metodologias definidas para os Indicadores de curto prazo ao Regulamento da Comissão Europeia sobre Conceitos e definições;

- Novos desenvolvimentos do sistema de produção de indicadores qualitativos (base amostral, ponderações e modelo de difusão);
- Passagem do IPC a um índice encadeado anualmente, o que permite acomodar os ajustamentos necessários entre duas bases do índice (esquema de ponderação e novos produtos); novos instrumentos de verificação da qualidade do indicador (Índices de Qualidade Implícitos); reformulação dos métodos de agregação dos índices regionais para o cálculo dos índices nacionais; novos métodos de ajustamento de qualidade e re-análise da periodicidade de recolha de preços para alguns sub-índices;
- Análise do modelo de difusão do IPC/IPCH introduzindo agregados especificamente vocacionados para a interpretação e análise da inflação;
- Aplicação dos novos procedimentos de recolha de informação do Carvão e do Aço, após o final do Tratado CECA;
- Realização de novas operações estatísticas na área do Comércio Interno: Inquérito aos Estabeleci-mentos dos Centros Comerciais e Inquérito aos Estabelecimentos Comerciais – Unidades Comerciais de Dimensão Relevante (reformulação do Inquérito aos Estabelecimentos Comerciais – Grandes Superfícies Retalhistas Alimentares);
- Alteração da metodologia de estimação do valor estatístico Intrastat, para as empresas isentas de declaração;
- Produção de índices mensais de valores unitários do Comércio internacional, no âmbito do Plano de Acção/UEM;
- Participação nos trabalhos de informatização da transmissão de dados da Exportação, da DGAIEC para o INE;
- Conclusão da implementação e teste da utilização de formulários Web do Intrastat;
- Consolidação do IDEP e de outros procedimentos automatizados de recolha, crítica e validação de informação de base do Comércio internacional;
- Estudo e implementação de novos modelos de ajuda à análise de resultados das Estatísticas do Comércio Internacional e da Produção Industrial;
- Recolha de dados no âmbito do questionário AUVIS – *Audio-visual Statistics*;
- Reformulação das operações estatísticas dos Transportes Ferroviários e Aéreos - decorrentes da eventual aprovação de Regulamentos comunitários, e das Comunicações;
- Consolidação do sistema de estatísticas de base relativo à Directiva do Turismo, designadamente através da execução do Inquérito às Moradias Turísticas e da reformulação dos Inquéritos à Hotelaria e Estabelecimentos Similares;
- No âmbito do Inquérito às Empresas Harmonizado (IEH) recurso a processos de estimação com base na informação amostral proveniente de empresas com menos de 20 pessoas ao serviço;

- Aprofundamento dos métodos de controlo de qualidade (dados macro) do IEH;
- Desenvolvimento e implementação de processos alternativos à inquirição directa do IEH, com recurso às bases de dados administrativas e de Associações Profissionais, visando a racionalização do processo de produção com a consequente diminuição de custos *lato sensu*;
- Desenvolvimento de estudos com vista à reformulação do processo de inquirição às entidades do sistema financeiro;
- Produção e difusão da informação estatística sobre os Serviços Financeiros, que privilegie a passagem dos dados contabilísticos a variáveis do domínio estatístico-económico, visando a reformulação das Estatísticas Monetárias e Financeiras;
- Participação no primeiro projecto harmonizado sobre Demografia das empresas, em colaboração com o EUROSTAT;
- Definição e aplicação de um modelo de exploração do Painel trimestral de empresas (PTE) para a obtenção de resultados antecipados face às novas exigências de calendário das Contas Nacionais Trimestrais;
- Consolidação do processo de controlo de qualidade do PTE, designadamente a implementação do sistema de reconhecimento e tratamento semi-automatizado de valores anómalos;
- Realização dos trabalhos para o EUROSTAT, no contexto do Regulamento sobre Estatísticas estruturais das empresas, para os anexos específicos de vários subsectores da área financeira;
- Desenvolvimento das acções associadas à consolidação do projecto Indicadores de Preços na Habitação;
- Implementação dos Inquéritos aos Projectos de Obras de Edificação e de Demolição de Edifícios, Utilização de Obras Concluídas, Conclusão de Obras, Alterações de Utilização dos Edifícios, Trabalhos de Remodelação de Terrenos e Operações de Loteamento Urbano;
- Reformulação metodológica do projecto "previsões agrícolas";
- Concepção do Inquérito à Estrutura das Explorações Agrícolas 2003;
- Realização do Inquérito Base às Plantações de Árvores de Fruto 2002;
- Concepção do Inquérito à Floricultura 2002.

1.2. Estatísticas e Indicadores Sociais

Principais objectivos:

- Difundir os resultados provisórios e definitivos dos Censos 2001, facto que constitui um resultado histórico na capacidade de tratamento da informação, só possível devido aos avanços tecnológicos em áreas como as de leitura óptica e das técnicas de reconhecimento de caracteres;

- Ajustar as estimativas da população aos resultados dos Censos 2001;
- Constituir um ficheiro exaustivo de endereços postais de alojamentos alicerçado nos questionários dos Censos 2001;
- Automatizar a recolha e registo de dados das estatísticas registrais correntes;
- Desenvolver o Sistema de Informação Estatística sobre as Famílias, em articulação com as necessidades específicas de cada área estatística;
- Desenvolver o Sistema Integrado de Informação sobre as Cidades;
- Desenvolver o Sistema de Informação sobre o Trabalho, Emprego e Formação Profissional;
- Reformular o Sistema de Informação das Estatísticas da Justiça.

Acções a desenvolver:

- Promover trabalhos conducentes à progressiva articulação dos ficheiros estatísticos existentes à base geográfica de referenciação de informação no contexto da implementação do Sistema de Informação Geográfica do INE (INESIG);
- Preparação de uma nova base de amostragem de unidades de alojamento (Amostra-Mãe), derivada da informação recolhida através dos Censos 2001 e desenvolvimento de um modelo de actualização permanente;
- Concepção de um sub-sistema de informação para as "Instituições Particulares sem fins lucrativos", com repercussões nas áreas do ambiente, cultura, protecção social, saúde, etc;
- Concepção de um novo "Inquérito às Condições de Vida", com execução de fase piloto, em articulação com o Regulamento Comunitário SILC (*Statistics on Income and Living Conditions*), para substituição do "Painel Europeu de Agregados Familiares";
- Estudo da incorporação no Inquérito às Condições de Vida de um "Inquérito Anual ao Consumo" e de outros módulos temáticos;
- Desenvolvimento das "Estatísticas do livro", em articulação com o Instituto Português do Livro e da Leitura, em ligação com o circuito associado ao ISBN (*International Standard Book Number*);
- Implementação do módulo, anexo ao Inquérito ao Emprego, relativo ao "Emprego de pessoas com incapacidades" (colaboração INE/SNRIPD);
- Reformulação da metodologia das "Estimativas da população" e cálculo de novas séries para o período intercensitário 1991 / 2001;
- Obtenção do apoio jurídico da Comissão Nacional de Protecção de Dados para a legalização da constituição de um ficheiro exaustivo de endereços postais de alojamento;
- Implementação da infraestrutura de reconhecimento dos endereços postais;

- Concepção e desenvolvimento do "Índice de Custo Médio do Trabalho", em articulação com o Regulamento respectivo e o Índice de Custo do Trabalho, já disponibilizado a nível nacional;
- Desenvolvimento dos "Indicadores do mercado de trabalho", em articulação com as metodologias utilizadas nas Contas Nacionais;
- Concepção e desenvolvimento de módulos, no âmbito do Sistema Europeu de Estatísticas Integradas de Protecção Social (SEEPROS), sobre "Despesas líquidas da protecção social" e "Beneficiários da protecção social";
- Reestruturação do processo de recolha de dados das estatísticas vitais, de modo a suportá-lo num circuito completamente digitalizado (médio prazo) e implementação progressiva da leitura óptica dos actuais verbetes demográficos registrais;
- Inclusão da informação estatística sobre imigração temporária de estrangeiros (concessões e renovações de autorizações de permanência); expansão da cobertura estatística da população estrangeira, ao nível territorial (concelho) e das características sócio-demográficas dos estrangeiros legalmente residentes;
- Estudo de fontes alternativas para a recolha da informação estatística sobre a Emigração;
- Revisão da Tipologia das áreas urbanas, à luz do Recenseamento da População de 2001;
- Elaboração de um Atlas Estatístico das cidades de Portugal, com destaque visual de indicadores básicos e valorização da componente cartográfica e gráfica.;
- Realização da operação estatística "Instituições de Reabilitação", com o objectivo de enriquecer e actualizar o conhecimento das instituições, programas e recursos colocados ao serviço da Reabilitação de Deficientes no nosso país;
- Elaboração de indicadores que permitam avaliar a qualidade do emprego;
- Concepção de um sistema regular de informação estatística sobre o mercado de formação profissional, na óptica da oferta e da procura, abrangendo a nível nacional a totalidade da formação efectuada;
- Caracterização através de indicadores apropriados, dos diferentes domínios de observação do mercado de trabalho – demografia, emprego, desemprego, educação e formação, estrutura empresarial, remunerações – quer do ponto de vista nacional, quer regional;
- Conhecimento das diferentes formas de emprego que estruturam o mercado de trabalho português;
- Reformulação da metodologia do Inquérito de Ganhos;
- Acompanhamento e apoio na elaboração do Plano Nacional de Emprego;

- Alteração do método de recolha das estatísticas da justiça, passando a uma recolha automatizada, quando possível, ou através de formulários web, nos restantes casos;
- Desenvolvimento de um repositório de dados estatísticos sobre a justiça (Data Warehouse), e da respectiva ferramenta de análise multidimensional.

1.3. Estatísticas e Estudos sobre Ciência e Tecnologia

Principais objectivos:

- Implementar um sistema integrado de informação de ciência e tecnologia;
- Actualizar o inventário dos recursos humanos e financeiros disponíveis para actividades de investigação;
- Melhorar o conhecimento do sistema de inovação das empresas portuguesas.

Acções a desenvolver:

- Continuação dos trabalhos de implementação do sistema integrado de informação de ciência e tecnologia através da actualização e divulgação da informação constante das bases de dados relativas a:
  - Doutoramentos por universidades portuguesas;
  - Programas de formação avançada;
  - Projectos de investigação em curso;
  - Produção científica portuguesa referenciada internacionalmente;
  - Dotações orçamentais para a Ciência e Tecnologia;
  - Imagens da Ciência e Tecnologia em Portugal;
  - Instituições de Ciência e Tecnologia.
- Lançamento, recolha, tratamento e divulgação dos resultados do Inquérito ao Potencial Cientifico e Tecnológico Nacional, com o objectivo de actualizar o inventário dos recursos humanos e financeiros disponíveis para actividades de investigação em 2001 nos sectores das Empresas, Estado, Ensino Superior e Instituições Privadas sem Fins Lucrativos e poder assegurar os compromissos nacionais e internacionais na área das estatísticas de Ciência e Tecnologia;
- Tratamento, análise e divulgação de dados do Inquérito Comunitário à Inovação (CIS III) na Indústria e nos Serviços;
- Aprofundamento do conhecimento sobre o sistema de C&T através do desenvolvimento dos estudos relativos a:
  - Inserção profissional de doutorados em Portugal;
  - Instituições de Ciência e Tecnologia.

1.4. Estatísticas e Estudos sobre a Sociedade da Informação

Principal objectivo:

- Desenvolver a produção de indicadores relacionados com a sociedade e a economia da informação.

Acções a desenvolver:

- Realização e apresentação dos resultados dos inquéritos relativos à utilização das tecnologias de informação em Portugal nas famílias, empresas, Administração Pública e escolas;
- Recolha, tratamento e análise dos dados relativos a oferta e programa da formação para a Sociedade da Informação, empresas e emprego na Sociedade de Informação, tecnologias de informação e comunicação: infra-estruturas e acessos;
- Divulgação dos resultados dos estudos relativos a:
  - Conteúdos na Internet;
  - Desenvolvimento do comércio electrónico;
  - Factores facilitadores e de bloqueio na massificação das tecnologias de informação.

1.5. Estatísticas do Ambiente

Principal objectivo:

- Desenvolver um sistema integrado de estatísticas do ambiente.

Acções a desenvolver:

- Desenvolvimento de projectos relativos à conta satélite do ambiente e aos fluxos de materiais;
- Reforço do nível de articulação com o Ministério do Ambiente, designadamente com o Instituto dos Resíduos, o Instituto da Água e a Direcção-Geral do Ambiente, para preparação da aplicação do novo regulamento comunitário nesta matéria.

2. Investigação, Estudos, Análise Económica/Social e Previsão

Principais objectivos:

- Proporcionar informação (de curto prazo) actualizada e fiável, análises da evolução e tendências da economia portuguesa e do seu enquadramento internacional;
- Desenvolver e coordenar processos de investigação aplicada à produção, conjugando a disponibilidade de informação de base, inerente ao SEN, com as oportunidades de desenvolvimento de estudos diversificados;
- Identificar e promover processos de inovação nos métodos e práticas da produção estatística, tendentes a aumentar a eficiência desta no que respeita quer à utilização de recursos, quer à qualidade dos dados de base;
- Integrar progressivamente a questão do desenvol-vimento sustentável em todos os estudos de carácter demográfico e social;

- Fortalecer o estudo das tendências demográficas, a identificação das respectivas causas e a avaliação das implicações no domínio económico e social;
- Desenvolver estudos demográficos e sociais, permitindo comparações internacionais com base em inquéritos e indicadores harmonizados.

Acções a desenvolver:

- Reactivação de uma publicação de Síntese de Conjuntura (previsão: 2º semestre de 2002);
- Avaliação das condições ao nível da produção estatística necessárias à produção regular da Matriz de Contabilidade Social;
- Desenvolvimento de estudos com aplicação específica à melhoria dos processos de produção no subsistema de indicadores económicos de curto prazo;
- Contribuição para o estabelecimento de modelos sistemáticos de análise da coerência da informação económica;
- Intensificação do processo de investigação de indicadores económicos, coincidentes e avançados;
- Criação de condições que permitam o início do estudo da produção de novos indicadores de curto prazo com recurso à informação de carácter administrativo;
- Início da publicação da Revista de Estudos Económicos e Sociais, estabelecendo e coordenando o plano de contribuições internas e externas para a publicação;
- Aplicação e monitorização das recomendações das conferências Desenvolvimento de estudos temáticos e indicadores no campo demográfico e das condições de vida dos indivíduos e famílias nomeadamente natalidade, fecundidade, nupcialidade, mortalidade e migrações. As problemáticas do envelhecimento demográfico e solidariedade intergeracional, a dimensão "género" ligada à igualdade de oportunidades, bem como as questões da exclusão social e pobreza, especialmente em grupos populacionais mais vulneráveis, merecem igualmente uma atenção particular;
- Desenvolvimento de todos os trabalhos inerentes à preparação da Conferência - 2ª Assembleia Mundial sobre o Envelhecimento - a realizar em Abril de 2002, e participação na mesma. A esta actividade será atribuída uma relevância especial nos primeiros meses do ano, dado que está atribuída a Portugal uma das vice-presidências do Bureau, em representação do *Grupo Ocidental* (Europa, excepto países com economia em transição, Canadá e Estados Unidos).

3. Metodologia Estatística

Principais objectivos:

- Conceber modelos gerais de apoio à definição da arquitectura de sistemas de informação estatística, e acompanhar a sua implementação;
- Harmonizar e integrar o Sistema de Metainformação do SEN;
- Manter, de forma eficaz, os ficheiros dos universos de referência que são em geral usados para a extracção dos planos de amostragem;
- Reforçar a capacidade de desempenho na área da amostragem, das técnicas de estimação, do tratamento da não resposta, do controlo estatístico da qualidade e da protecção dos dados individuais;
- Criar e manter os instrumentos cartográficos de suporte à implementação e desenvolvimento do Sistema de Informação Geográfica do INE (INESIG);
- Actualizar a Base de Amostragem Agrícola.

Acções a desenvolver:

- Identificação e caracterização dos subsistemas estatísticos do Sistema de Informação do INE;
- Desenvolvimento de modelos gerais de suporte à concepção de sistemas de informação;
- Harmonização do processo "Captura de Dados" para a recolha directa e desenvolvimento de procedimentos de controlo da qualidade associados aos erros de não-amostragem;
- Reforço e integração das componentes Conceitos, Nomenclaturas e Fontes Estatísticas com vista à construção de um sistema integrado de Metainformação;
- Intervenção no processo de revisão internacional das grandes nomenclaturas económicas;
- Prossecução dos esforços visando assegurar o normal acesso às fontes administrativas adequadas para permanente actualização do Ficheiro de Empresas e Estabelecimentos;
- Reforço do controlo da qualidade dos ficheiros de Empresas e Estabelecimentos, Administração Pública e Instituições sem Fins Lucrativos;
- Normalização dos sistemas de ficheiros de dados dos universos de referência, que servem de bases de amostragem às operações estatísticas oficiais;
- Definição de metodologias das operações estatísticas solicitadas pelos diferentes Departamentos e Direcções Regionais do INE e entidades com delegação de competências;

- Harmonização dos métodos de dimensionamento de amostras utilizados nas diferentes operações estatísticas;
- Revisão das amostras definidas para cada um dos Indicadores de curto prazo, no âmbito do processo de mudança de base;
- Aprofundamento da metodologia estatística dos inquéritos da área de produção industrial;
- Continuação do desenvolvimento e implementação do INESIG nas vertentes de manutenção e actualização da Base Geográfica (BG);
- Integração da Base Geográfica nos subsistemas de informação;
- Desenvolvimento de aplicações de acesso à BG e dados dos Censos 2001;
- Realização de um estudo conducente à georeferenciação de ficheiros de endereços postais;
- Reforço da utilização de fontes administrativas para actualização do ficheiro de explorações agrícolas;
- Construção e actualização da Base Geográfica de Edifícios (lançamento do projecto piloto de construção da BGE).

4. Tecnologias de Informação

Principais objectivos:

- Consolidar a arquitectura do Sistema de Informação do INE;
- Favorecer a inter-operabilidade dos sistemas de informação das entidades co-produtoras de estatísticas oficiais;
- Reformular toda a rede de comunicação de dados que assegura a interligação entre os vários edifícios do INE, incluindo as Direcções Regionais;
- Adoptar tecnologias mais recentes de comutação e segurança da comunicação de dados;
- Implementar uma solução para recepção electrónica dos dados obtidos por recolha directa;
- Implementar novas soluções no processo de captura electrónica de dados;
- Criar um subsistema de difusão que contenha motores genéricos geradores dos vários tipos de publicações;
- Reformular a intranet;
- Potenciar a utilização do DataWarehouse.

Acções a desenvolver:

- Identificação e caracterização das infra-estruturas computacionais e de comunicação das entidades co-produtoras de estatísticas oficiais;

- Identificação e caracterização dos principais fluxos de informação entre as referidas entidades;
- Estabelecimento de normas e procedimentos para o acesso a bases de dados de interesse comum;
- Implementação de uma solução de "Balanced Scorecard" para apoio à decisão;
- Consolidação do actual parque de servidores Unix e NT;
- Desenvolvimento de soluções de difusão pela Web utilizando tecnologias "On-line Analytical Processing (OLAP);
- Implementação de um sistema de "workflow" e arquivo que suporte toda a componente administrativa;
- Criação de um portal interno que substituirá a actual intranet, de modo a rentabilizar-se a partilha de conhecimento interno bem como a comunicação formal;
- Continuação da inclusão progressiva de projectos no Data Warehouse do INE.

5. Âmbito Regional

Principais objectivos:

- Promover o desenvolvimento do sistema de informação regional compatível com as necessidades de planeamento e avaliação de programas de âmbito regional;
- Desenvolver estudos de âmbito regional que tenham por base informação estatística do SEN;
- Favorecer a construção de um interface único para transferência de informação a nível local, processo que deverá evoluir para uma comunicação por via electrónica e distribuída;
- Desenvolver as estatísticas transfronteiriças, no contexto das preocupações comunitárias com a revitalização das regiões de fronteira e dos novos impulsos assegurados pelo Interreg III;
- Difundir e promover novos estudos, publicações e outras iniciativas regionais;
- Promover a cooperação com as Universidades e Centros de Investigação.

Acções a desenvolver:

- Aprofundamento da articulação inter-institucional ao nível regional e local;
- Início do processo de implementação de um Sistema de Informação Regional suportado em tecnologia de Sistemas de Informação Geográfica (SIG);

- Desenvolvimento de iniciativas de produção, estudos e difusão no âmbito da cooperação transfronteiriça;
- Continuação do Estudo de caracterização sócio-económica da área de influência do empreendimento para fins múltiplos do Alqueva;
- Desenvolvimento de um sistema de difusão de dados de âmbito regional integrado no projecto Alentejo Digital;
- Realização de estudos de âmbito regional, nomeadamente baseados nos dados censitários;
- Participação na realização de estudos de âmbito nacional;
- Desenvolvimento das iniciativas de organização e análise de informação estatística com vista à elaboração dos Anuários Estatísticos Regionais, Boletim Trimestral de Estatística e Revista Estatísticas & Estudos Regionais;
- Participação no processo de elaboração das Contas Regionais Portuguesas;
- Concepção e realização do Inventário Municipal de 2002;
- Preparação de um Seminário internacional sobre Estatísticas Regionais.

ACÇÕES DESENVOLVIDAS PELO INE NO ÂMBITO DA COOPERAÇÃO

ACTIONS ACHIEVED BY NSI IN THE SCOPE OF COOPERATION

(DE 1 DE MAIO A 31 DE DEZEMBRO DE 2001)

- **COOPERAÇÃO DESENVOLVIDA COM OS PALOP:**

Teve lugar na cidade da Praia, em Cabo Verde, nos dias 10 e 11 de Outubro de 2001, a XIª Reunião dos Directores-Gerais de Estatística de Portugal, dos Países Africanos de Língua Portuguesa e de Macau, tendo sido seguida, no dia 12 de Outubro, de um seminário subordinado ao tema "Censos da População e Habitação". Os eventos reuniram os Directores-Gerais e Presidentes dos Institutos Nacionais de Estatística de Angola, Cabo Verde, Moçambique, Portugal e S. Tomé e Príncipe, e respectivas delegações, tendo também contado com a participação da Subdirectora do Gabinete de Assuntos Europeus e Relações Externas do Ministério do Planeamento de Portugal e de representantes da Comissão Europeia. A reunião dos Directores-Gerais foi precedida de uma sessão solene de abertura, presidida pelo Ministro das Finanças de Cabo Verde, tendo os trabalhos sido conduzidos pelo Presidente do Instituto Nacional de Estatística daquele país. No decurso da reunião foram discutidos aspectos relativos ao balanço da cooperação estatística nos últimos dois anos, quer ao nível bilateral, quer no que respeita a projectos de interesse comum a vários países, tendo também sido abordadas as perspectivas de cooperação no futuro. Na área da formação estatística, foi analisada a possibilidade do recurso às novas técnicas de ensino à distância, tendo sido também avaliada a possibilidade de utilização das novas tecnologias da informação e comunicação para o estabelecimento de uma Intranet entre os países. Ficou acordada a realização de encontro entre os Directores-Gerais, no primeiro trimestre de 2002, para uma reflexão sobre o estabelecimento de um novo paradigma da cooperação. Ficou também acordado que a próxima reunião dos Directores-Gerais de Estatística terá lugar em S. Tomé, no ano de 2002, em data a definir.

No quadro do Programa Estatístico da CPLP, deu-se início às missões de diagnóstico do projecto sobre "Estatísticas da Educação" com a realização da primeira acção que teve lugar junto do INE de Angola. A missão foi efectuada pelo Dr. José Alexandre Paredes, Chefe da Divisão de Estatística do Ministério da Educação, entre 18 e 17 de Novembro, tendo incidido sobre a análise e reformulação dos instrumentos de notação neste domínio. No decurso da missão, para além das reuniões efectuadas com a Direcção do INE de Angola, foram realizados encontros com o Vice Ministro da Educação e Cultura e com o Director do Gabinete de Estudos e Planeamento deste Ministério, bem como com a respectiva Direcção de Estatística, tendo ainda sido efectuadas visitas a uma Secção Municipal da Educação e a uma Direcção Provincial da Educação.

No âmbito do projecto comum aos PALOP sobre "Classificações, Conceitos e Nomenclaturas", tiveram lugar no INE, entre Setembro e Dezembro, dois estágios

sobre a Classificação Nacional de Bens e Serviços destinados, respectivamente, ao Dr. Jorge Utui, do INE de Moçambique, com a duração de três semanas, e à Dra Elsa Cassandra, do INE de S. Tomé e Príncipe, com a duração de um mês. Foi ainda realizada uma missão de formação à Classificação de Actividades Económicas, junto do INE de S. Tomé e Príncipe, no mês de Outubro, com a duração de uma semana. A missão foi efectuada pelo coordenador do projecto, Dr. Saraiva Aguiar, do Departamento de Metodologia Estatística do INE.

Foi organizada no INE, a quarta reunião do Grupo de Trabalho de projecto comum aos PALOP sobre Ficheiros de Unidades Estatísticas, no período compreendido entre 28 de Maio e 1 de Junho, com a participação de dois delegados oriundos de cada país beneficiário: Angola, Cabo Verde, Guiné-Bissau e São Tomé e Príncipe, para além de representantes do Instituto da Cooperação Portuguesa (ICP), do Gabinete de Assuntos Europeus e Relações Externas (GAERE) do Ministério do Planeamento e da Agência Portuguesa de Apoio ao Desenvolvimento (APAD), bem como de técnicos do INE afectos ao Departamento de Coordenação e Contas Nacionais e ao Gabinete de Planeamento, Relações Internacionais e Qualidade. Considerando que esta reunião marcou a conclusão deste projecto comum, o Grupo de Trabalho aprovou um conjunto de recomendações no sentido de cada país vir a submeter ao INE de Portugal, propostas de desenvolvimento de acções de cooperação futuras, neste domínio.

No período em apreço, e no âmbito da **Cooperação Bilateral com os PALOP**, foram realizadas as seguintes acções:

ANGOLA

Foram realizadas duas acções no âmbito do projecto "Coordenação, Controlo e Avaliação".

Foi realizada uma visita de trabalho do Dr. Flávio Couto, Director Geral do INE de Angola, durante a qual foram efectuadas diversas reuniões sobre as perspectivas de desenvolvimento de novos projectos. Na sequência da visita, realizou-se a 8ª reunião da Comissão Coordenadora da Gestão do Acordo de Cooperação Estatística Luso-Angolano, durante a qual foi actualizado o programa de cooperação até ao final de 2002.

CABO VERDE

Foram realizadas sete acções no âmbito dos projectos "Coordenação, Controlo e Avaliação", "Infra-estruturas Estatísticas" e "Produção Estatística" e preparados dois novos projectos no âmbito da Reforma do Sector Público de Cabo Verde.

No âmbito do projecto "Coordenação, Controlo e Avaliação" foram realizadas duas visitas de trabalho de curta duração pelo Eng. Francisco Fernandes Tavares, Presidente do INE de Cabo Verde, durante as quais tiveram lugar diversas reuniões para actualização do programa de cooperação, bem como sobre as perspectivas de desenvolvimento de novos projectos. Foi ainda realizada em Cabo Verde, por ocasião da Reunião dos DGINE de Portugal e PALOP, a 7ª Reunião da Comissão

coordenadora da Gestão do Acordo de Cooperação Luso-Caboverdiano no domínio da Estatística, a qual permitiu a actualização do programa de cooperação até ao final de 2002.

No domínio do projecto "Infra-estruturas Estatísticas" foi realizada, pela Eng.ª Júlia Cravo, uma missão de assistência técnica cujos principais objectivos visaram a adaptação do programa de imputação de não respostas e a determinação do coeficiente de extrapolação, aos dados do inquérito Anual às Empresas, bem como a análise sócio-económica relativa aos anos de 1998 e 1999.

Relativamente ao projecto "Produção Estatística" foram realizados: um estágio para uma técnica do INE de Cabo Verde e uma missão de cooperação e assistência técnica.

O estágio, realizado pela Dra. Alice Monteiro, decorreu na Direcção Regional do Centro e teve como objectivo a análise do processo de concepção e realização do Inventário Municipal.

A missão de assistência técnica foi realizada pela Dra. Cristina Fernandes Lopes e incidiu sobre os trabalhos de actualização e alargamento de cobertura do IPC de Cabo Verde.

Foram preparados dois novos projectos, na sequência de concursos internacionais, no âmbito da "Reforma das Contas Nacionais de Cabo Verde" e da "Criação da Base de Dados de Estatísticas Oficiais".

Foi ainda realizada, pelo Dr. Sérgio Bacelar e pelo Eng. Carlos Dias, uma missão para análise e adaptação do projecto "Base de Dados de Estatísticas Oficiais de Cabo Verde".

MOÇAMBIQUE

Foram realizadas seis acções no âmbito dos projectos "Coordenação, Controlo e Avaliação" e "Apoio Institucional".

No âmbito do projecto "Coordenação, Controlo e Avaliação" foi realizada em Cabo Verde, por ocasião da Reunião dos DGINE de Portugal e PALOP, a 9ª Reunião da Comissão Coordenadora da Gestão do Acordo de Cooperação Luso-Moçambicano no domínio da Estatística, a qual permitiu efectuar a actualização do programa de cooperação até ao final de 2002.

Relativamente ao projecto "Apoio Institucional" foram realizados: uma visita de trabalho e quatro estágios.

A visita de trabalho, realizada pela Dra. Manuela Xavier e pelo Dr. Domingos Maingue, teve como principal objectivo a análise da organização e funcionamento das áreas Administrativas, Financeiras e de Recursos Humanos.

Um dos estágios, realizado junto da Direcção Regional do Algarve pela D. Lúcia Moiane, oriunda da Delegação Provincial de Maputo do INE de Moçambique,

incidiu na organização e funcionamento de uma Direcção Regional, no âmbito do processo de descentralização e regionalização do INE de Moçambique.

Os outros três estágios, realizados junto das Direcções Regionais do INE: de Lisboa e Vale do Tejo, do Centro e do Alentejo, bem como junto dos Serviços Centrais, incidiram sobre a análise da organização e funcionamento do sector da Difusão nas Delegações Provinciais, no âmbito do processo de descentralização e regionalização do INE de Moçambique e foram efectuados por técnicos das Delegações Provinciais de Maputo (Dra. Manuela Beca), Manica (Sr. Boaventura Williamo) e Nampula (D. Isaura Mamad).

S. Tomé e Príncipe

Foram realizadas onze acções no âmbito dos projectos "Coordenação, Controlo e Avaliação", "Apoio Institucional", "Operações Estatísticas de Base", "Produção Estatística", "Infra-estruturas Estatísticas" e "Difusão Estatística".

No âmbito do projecto "Coordenação, Controlo e Avaliação" foi realizada uma visita de trabalho pelo Dr. Albano Germano de Deus, novo Director Geral do INE de São Tomé e Príncipe, durante a qual foi apresentada a organização e funcionamento dos principais sectores do INE e analisadas as perspectivas de desenvolvimento dos projectos estatísticos. Na sequência da visita, realizou-se a 13ª reunião da Comissão Coordenadora da Gestão do Acordo de Cooperação Estatística Luso-Santomense, durante a qual foi actualizado o programa de cooperação até ao final de 2002.

Relativamente ao projecto "Apoio Institucional" foram realizados uma missão de assistência técnica, um estágio, e fornecidos alguns equipamentos informáticos.

A missão de assistência técnica, realizada pelo Dr. Adrião Ferreira da Cunha, teve como principais objectivos a realização de um curso de formação sobre "O Novo Sistema Estatístico Nacional", dirigido aos quadros do INE e dos Serviços de Estatística do Banco Central e dos Ministérios que serão Órgãos Delegados do INE, bem como a organização de um Seminário sobre "O Novo Sistema Estatístico Nacional", cujo objectivo foi de dar informação sobre as principais linhas de força do Sistema, proporcionando o consequente debate com os participantes convidados.

O estágio, realizado na Direcção Regional do Alentejo, pelo Sr. Anastácio Ramos, técnico responsável pela Delegação Regional do Príncipe, incidiu no processo de organização e funcionamento de uma Delegação Provincial. Na sequência deste estágio, foi ainda doado ao INE de São Tomé e Príncipe um computador portátil para reforço dos recursos afectos à referida Delegação Regional.

Foi ainda fornecida uma placa gráfica para reparação do equipamento afecto ao Ficheiro de Unidades Estatísticas.

No quadro do projecto "Operações Estatísticas de Base" foi realizada uma missão de assistência técnica e fornecidos diversos materiais para apoio ao Recenseamento Geral da População e Habitação-2001.

A missão de assistência técnica, realizada pelo Dr. Fernando Casimiro, visou avaliar os resultados do inquérito-piloto e todos os suportes de apoio à realização da operação de recolha de dados e a análise com os quadros nacionais do plano operacional para a recolha de dados no terreno. Foi ainda efectuada a definição das regras de validação para registo e tratamento dos questionários, o inquérito de qualidade, a análise da cartografia para ensaio de digitalização e a sensibilização dos intervenientes nacionais para a operação.

Relativamente ao apoio material, foram editados diversos materiais para as acções de promoção e sensibilização da operação e de questionários para recolha de informações, bem como, fornecidos diversos materiais de apoio às operações, em conformidade com o projecto aprovado pela Cooperação Portuguesa.

No que concerne ao projecto "Produção Estatística", foi realizado um estágio pela D. Elsa Rosário na área das estatísticas do comércio externo, para melhoria do processo de produção estatística neste domínio.

Relativamente ao projecto "Infra-estruturas Estatísticas" foi realizada uma missão de identificação, pela Eng.ª Júlia Cravo, com o objectivo de preparar o próximo Recenseamento Empresarial, que englobou a definição da metodologia e preparação/adaptação de um questionário para recolha exaustiva da informação relativa às empresas com vista à actualização do Ficheiro de Unidades Estatísticas.

No domínio do projecto "Difusão Estatística" foram preparadas capas para a edição das Folhas de Informação Rápida do INE de São Tomé e Príncipe.

- **COOPERAÇÃO DESENVOLVIDA COM OS PAÍSES DA EUROPA CENTRAL E ORIENTAL, NO QUADRO DO PROGRAMA PHARE:**

Ao abrigo do Programa PHARE (Assistência técnica aos Países da Europa Central e Oriental) da União Europeia, realizaram-se no período em apreço, doze acções de cooperação, em diferentes áreas, nomeadamente sobre Estatísticas Agrícolas e das Pescas, Estatísticas Ambientais, Contas Regionais e Difusão Estatística.

- **Cooperação com o Instituto Nacional de Estatística (INS) da Roménia:**

Projecto "Difusão e Disseminação Estatística":

- No âmbito do projecto em epígrafe, o INE participou, através do Dr. Pedro Campos da DRN, no Seminário "*Dissemination of Statistical Information*", que ocorreu de 28 de Agosto a 1 de Setembro, na Roménia.
- Também no quadro deste Projecto, foi realizada pelo Eng. Pinto Martins, técnico do DDP, uma missão ao INS, de 3 a 9 de Setembro, sobre "Finalização da elaboração do orçamento sobre equipamentos de tipografia".

Projecto "Contas Regionais":

- Com o intuito de finalizar o projecto em epígrafe, a colaboradora do DCN, Dr.ª Raquel Paulino, efectuou, de 3 a 7 de Setembro, uma missão de assistência técnica à Roménia.

Projecto Piloto "*Supply Balance Sheets*":

- No domínio das Estatísticas Agrícolas, a Eng.ª Carla Marques, colaboradora do DEAP, realizou uma missão de assistência técnica à Roménia, a 14 e 15 de Junho.

Cooperação com o Instituto Nacional de Estatística da Bulgária (NSI):

Projectos no âmbito das Estatísticas Agrícolas:

- Na área das Estatísticas Agrícolas, o INE tem vindo a participar em diversos projectos de cooperação com o NSI, nomeadamente no domínio do "*Acquis Communautaire*". Esta assistência desenvolve-se através de missões de assistência técnica, tendo no período referido ocorrido 2 acções, nomeadamente:
  - de 5 a 11 de Maio, missão realizada pelo Director do DEAP, Eng. António Macedo;
  - de 14 a 18 de Maio, missão realizada pela Eng.ª Anabela Delgado, colaboradora do DEAP.
- No âmbito do Projecto Piloto "*Supply Balance Sheets*", a técnica do DEAP, Eng.ª Carla Marcos, realizou uma missão a este país, a 11 e 12 de Junho.

Projectos no âmbito das Estatísticas Financeiras:

- O INE recebeu, de 14 a 16 de Maio, a visita de uma delegação búlgara, composta por 3 elementos. Esta visita sobre "*RTU and NUTS*" foi coordenada pelo DME.

Cooperação com o Instituto Nacional de Estatística da Hungria (NSI):

Projectos no âmbito das Estatísticas Agrícolas:

- O INE recebeu, a 24 e 25 de Outubro, a visita de uma delegação húngara, composta por elementos do Research Institute for Agricultural Economics (AKII) sobre o tema "Contas Económicas Agrícolas". Esta visita foi coordenada pelo DEAP.

Projecto Piloto "*Environment*":

- No âmbito deste projecto, o INE participou, através da presença do Dr. Nuno Romão, colaborador da DRLVT, no Workshop "*Environment - Expenditure Statistics*", o qual teve lugar nos dias 12, 13 e 14 de Novembro na Hungria.

Cooperação com o Instituto Nacional de Estatística da Lituânia

Projecto Piloto "*Supply Balance Sheets*":

- Foi realizada, pela Eng.ª Carla Marques, colaboradora do DEAP, uma missão de assistência técnica à Lituânia, a 20 e 21 de Setembro.

Outras actividades no âmbito do PHARE - Workshops:

- No domínio das Estatísticas Agrícolas, a Eng.ª Carla Marques, colaboradora do DEAP, participou no Workshop "*Supply Balance Sheets*", que ocorreu lugar no Luxemburgo, a 4 e 5 de Outubro último.

- **COOPERAÇÃO DESENVOLVIDA NO ÂMBITO DO PROGRAMA MEDSTAT:**

Ao abrigo do Programa MEDSTAT (Assistência técnica aos Países do Mediterrâneo) da União Europeia, o INE colabora no Projecto "MED-IS" sobre Sistemas de Informação Estatística.

Durante o período de referência, e com o objectivo de preparar a 2ª fase deste projecto, os colaboradores do INE, Eng. Carlos Dias, Chefe de Serviço do DDP, e Eng. Paulo Saraiva, Chefe de Serviço do DEIS, estiveram presentes na Reunião de Coordenadores Nacionais, em 28 e 29 de Maio, no Luxemburgo.

O INE encontra-se envolvido igualmente neste projecto, através da participação do Eng. José Pinto Martins, colaborador do DDP, na qualidade de *Project Manager*.

- **COOPERAÇÃO DESENVOLVIDA NO ÂMBITO DO PROGRAMA UNIÃO EUROPEIA — MERCOSUL E CHILE:**

Ao abrigo do Programa MERCOSUL (Assistência técnica aos Países do MERCOSUL e ao Chile) da União Europeia, realizaram-se durante o período em referência, 6 acções de cooperação.

Grupo de Trabalho n. 7 - "Nomenclaturas e Classificações":

- Realizou-se em Lisboa a 2ª Acção de Formação deste Grupo de Trabalho. A participação do INE, que acolheu, no seu edifício Sede, os técnicos dos países beneficiários, foi feita através da contribuição dos seus colaboradores Drs. Adrião Ferreira da Cunha, António Portugal, Hermínio Saraiva Aguiar, José Castro Pinto, Pedro Dias, Arminda Brites, Isabel Farinha, Luísa Pereira e Engª Julia Cravo.
- No âmbito deste Grupo de Trabalho, o técnico do DME, Dr. Saraiva Aguiar, participou, na qualidade de perito europeu, na 5ª reunião que decorreu no Uruguai, de 12 a 14 de Novembro.

Grupo de Trabalho n. 5 - "Produtividade e Competitividade nas Empresas":

- A participação do INE na 5ª reunião deste Grupo de Trabalho foi assegurada pela colaboradora do DEE, Dr.ª Amélia Paisana. Esta reunião realizou-se no Brasil, a 20 e 21 de Setembro.

Grupo de Trabalho n. 3 - "Estatísticas dos Serviços":

- Realizou-se em Lisboa, de 28 a 30 de Maio, a 1ª Acção de Formação deste Grupo de Trabalho. Esta acção foi ministrada em conjunto com o INE-Espanha e o Banco de Portugal, tendo o INE acolhido os técnicos dos países beneficiários. A participação do INE foi feita através dos seus colaboradores Drs. Adrião Ferreira da Cunha, Rogério Reis e Fernando Reis.

Cooperação Bilateral com o Paraguai - "Estatísticas do Sector Industrial":

- Em Novembro último deu-se início à intervenção do INE neste projecto, com a 1ª missão de assistência técnica, realizada pelo colaborador do INE, Dr. Humberto Pereira.

Outras actividades no âmbito do MERCOSUL:

- A 25 e 26 de Junho, decorreu em Madrid a 3ª Reunião de Peritos Europeus, tendo o INE sido representado pelos Drs. Hermínio Saraiva Aguiar e Amélia Paisana.

- **COOPERAÇÃO DESENVOLVIDA NO ÂMBITO DO PROGRAMA DE COOPERAÇÃO UNIÃO EUROPEIA – PAÍSES TERCEIROS:**

Projecto de Cooperação "Comunidade Andina":

- No âmbito do projecto em epígrafe, o INE participou na Acção de Formação sobre o "Documento Administrativo Único", que ocorreu em Lima, no Peru, de 26 a 29 de Novembro. A representação do INE esteve a cargo do Dr. António Marques, colaborador do DEIS.

CONGRESSOS, SEMINÁRIOS, COLÓQUIOS E CONFERÊNCIAS

CONGRESS, SEMINARS AND CONFERENCES

2002

- 15-19 January
  First International ICSC Congress on Neuro-Fuzzy NF'2002 to be held at The Capitolio de la Habana, Cuba.
  Informações: INTERNATIONAL COMPUTER SCIENCE CONVENTIONS
  Head Office: 5101C-50 Street, Wetaskiwin AB, T9A 1K1, Canada
  (Phone: +1-780-352-1912 / Fax: +1-780-352-1913)
  Email: operating@icsc.ab.ca
  or
  planning@icsc.ab.ca
  URL: http://www.icsc.ab.ca/NF2002.htm
  or
  http://www.icsc.ab.ca/

- 16-18 January
  Food-Industry and Statistics, to be held in Villeneuve d'Ascq (LILLE), France. Bât. EUDIL IAAL - Cité Scientifique, F 59655.
  Information: E-mail: agrostat2002@eudil.fr
  URL: http://www.eudil.fr/~agrostat.

- 24-25 January
  International Biometric Society (French region) conference on mixed models in biometry, to be held in Paris, France.
  Information: URL: bioserv.univ-lyon1.fr/SFB/SFB.html

- 4-8 February
  ProbaStat 2002, the 4th International Conference on Mathematical Statistics, to be held at Smolenice Castle, Smolenice, Slovak Republic.
  Information: E-mail: probastat@savba.sk
  URL: http://www.um.savba.sk/lab_15/probastat.html.

- 12-15 February
  First International ICSC-NAISO Congress on Autonomous Intelligent Systems ICAIS 2002 to be held at Deakin University, Geelong, Australia.
  Information: E-mail: icais02@itstransnational.com
  URL: http://www.icsc-naiso.org/conferences/icais2002/index.html

- 18 February-8 March
  Scaling and Cluster Analysis 31st Spring Seminar, to be held at the Zentralarchiv, Cologne, Germany.
  Information: E-mail: maria.rohlinger@za.uni-koeln.de
  URL: www.gesis.org

- 14-27 February
  IEA South East Asia congress of epidemiology, to be held in Khajuraho, India.
  Informações: Organising secretariat: Tel. +91 517 320 858.
  E-mail: epidcong@rediffmail.com
  URL: www.epidcong.8m.com

- 15-21 March
  ENAR/IMS Eastern Regional to be held in Washington, DC, USA.
  Informações: Program Chair: Jiayang Sun, Case Western Reserve University
  Local Arrangements Chair: Colin Wu, John Hopkins University
  Contributed Papers Chair: Nidhan Choudhuri;
  E-mail: jiayang@sun.STAT.cwru.edu
  colin@mts.jhu.edu
  nidhan@nidhan.cwru.edu
  URL: http://sun.cwru.edu/ims

- 19-22 March
  German Open Conference on Probability and Statistics, to be held at the University of Magdeburg, Germany.
  Informações: Magdeburger Stochastik-Tage 2002, c/o Prof. Dr. Gerd Christoph, Otto-von-Guericke-Universität Magdeburg, Institut für Mathematische Stochastik, Universitätsplatz 2, D-39106 Magdeburg, Germany. Fax: 0049-(0)391-67 11172 Phone: 0049-(0)391-67 18652.
  E-mail: stoch2002@uni-magdeburg.de
  URL: www.math.uni-magdeburg.de/stoch2002/

- 25-26 March
  Young Statisticians' Meeting, to be beld at the University of Edinburgh, UK.
  Information: E-mail: ysm2002@bioss.ac.uk
  URL: www.bioss.ac.uk/ysm2002

- 17 April
  Sampling, Weighting, Profiling, Segmentation and Modelling. "Targeting Mr X - but is he Mr Right" to be held at Imperial College, London, UK.
  Informações: Contact: Diana Elder, Administrator ASC, PO box, Chesham, Bucks HP5 3QH, UK. Tel/fax 01494 793033.
  E-mail: admin@asc.org.uk
  URL: www.asc.org.uk

- 28 April - 3 May
  Workshop on nonparametric smoothing in complex statistical models, to be held in Ascona, Switzerland.
  Informações: Organizer: Hans-Ruedi Kuensch, Enno Mammen, Theo Gasse.
  URL: www.unizh.ch/biostat/smoothing2002/

- 13-17 May
  34th Journées de Statistique of the Société Française de Statistique, jointly organized by the Institutes of Statistics of the Université libre de Bruxelles and of the Université catholique de Louvain, will be held in Brussels and Louvain-la-Neuve, Belgium.
  Informações: JSBL 2002, Institute of Statistics, Université catholique de Louvain, 20 Voie du Roman Pays, B-1348 Louvain-la-Neuve, Belgium. Telephone : +32 10 47 43 54, Fax : +32 10 47 30 32.
  URL: www.stat.ucl.ac.be/jsbl2002

- 26-29 May
  Annual Meeting of the Statistical Society of Canada, Hamilton, Ontario, Canada.
  Informações: Local Arrangements Chair: Peter Macdonald Department of Mathematics and Statistics, McMaster University, 1280 Main Street West, Hamilton, Ontario, L8S 4K1, Canada; Phone: (905) 525-9140 x 23423 Fax: (905) 522-0935. Program Chair: Bruce Smith Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, B3H 3J5, Canada; Phone: (902) 494-2257 Fax: (902) 494-5130.
  E-mail: pdmmac@mcmail.cis.mcmaster.ca
  bsmith@mathstat.dal.ca

- 27-29 May
  Work Session on Statistical Data Editing of the Conference of European Statisticians of the UN, to be held in Helsinki, Finland.
  Informações: Mrs. Jana Meliskova, UNU/ECE Statistical Division, tel. +41 22 917 41 50; fax +41 22 917 00 40.
  E-mail: jana.meliskova@unece.org
  URL: www.unece.org/stats/

- June
  2nd ICMI-EARCOME (East Asia Regional Conference on Mathematics Education) to be in Singapore.
  Informações: EARCOME 2002, Division of Mathematics, National Institute of Education, 469 Bukit Timah Road, Singapore 259756, Republic of Singapore.
  E-mail: earcome2@nie.edu.sg

- 2-5 June
  Annual Meeting of the Statistical Society of Canada, Hamilton, Ontario, Canada.
  Informações: Peter Macdonald, Department of Mathematics and Statistics, McMaster University, 1280 Main Street West, Hamilton, Ontario, L8S 4K1, Canada.
  E-mail: pdmmac@mcmail.cis.mcmaster.ca

- 2-6 June
  Seventh Valencia International Meeting on Bayesian Statistics, to be held at the Playa de las Americas, Tenerife, Canary Islands, Spain.
  Information: URL: www.uv.es/valencia7
  www.stat.duke.edu/valencia7

- 5-9 June
  Hawaii International Conference on Statistics and Related Fields, to be held at the Sheraton Waikiki Hotel, Honolulu, Hawaii, USA.
  Informações: Hawaii International Conference on Statistics, 2440 Campus Road #517, Honolulu, HI, 96822, USA. Tel. (808) 223-1748, fax (808) 947-2420.
  E-mail: statistics@hicstatistics.org
  URL: www.hicstatistics.org

- 9-13 June
  Nordstat 2002 - The 19th Nordic Conference on Mathematical Statistics, to be held in Stockholm, Sweden.
  Information: URL: www.math.kth.se/nordstat/

- 12-15 June
  23rd SCORUS conference "Statistics for the Cities of Tomorrow", to be held in Lisbon, Portugal.
  Informações: Local Organising Committee, tel: +351 21 842 62 90, fax: +351 21 842 63 56.
  E-mail: alberto.pina@ine.pt.
  URL: www.ine.pt/novidades/semin/scorus.html

- 17-20 June
  MMR 2002, Third International Conference on Mathematical Methods in Reliability, to be held at the Norwegian University of Science and Technology, Trondheim, Norway.
  Informações: Professor Bo Lindqvist, Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway. Tel.: +47-73 59 35 20 - Fax: +47-73 59 35 24.
  E-mail: mmr2002@math.ntnu.no
  URL: http://www.math.ntnu.no/mmr2002/

- 18-22 June
  TIES 2002: Annual Conference of The International Environmetrics Society, to be held in the Faculty of Engineering in Genoa, Italy.
  Information: URL: www2.stat.unibo.it/ties2002

- 23-29 June
  The 8th International Vilnius Conference on Probability Theory and Mathematical Statistics, Vilnius, Lithuania.
  Informações: Professor Vytautas Statulevicius, Institute of Mathematics and Informatics, Akademijos str. 4, 2600 Vilnius, Lithuania.
  E-mail: conf@ktl.mii.lt

FUNDAMENTO, OBJECTO E ÂMBITO DA REVISTA

O INE, consciente de como uma cultura estatística é essencial para a compreensão da maioria dos fenómenos do mundo actual, e da sua responsabilidade na divulgação do conhecimento estatístico, fazendo-o chegar ao maior número possível de leitores, tendo reconhecido a necessidade de dar um passo nesse sentido, passou a editar quadrimestralmente a presente *Revista de Estatística* destinada a divulgar:

a) Numa perspectiva científica, artigos originais sobre temas especializados da estatística, tanto pura como aplicada, bem como sobre estudos e análises nos domínios económico, social e demográfico;

b) Informações sobre actividades e projectos importantes do Sistema Estatístico Nacional;

c) Informações sobre acções desenvolvidas pelo INE no âmbito da cooperação.

d) Informações sobre congressos, seminários, colóquios e conferências de interesse estatístico ou afim;

Para tal, são adoptadas as seguintes formas de contribuição para publicação na Revista:

- Quanto aos artigos referidos em a), contribuições da *iniciativa* dos próprios autores e por *convite* do Conselho Editorial, pertencentes ou não ao INE;
- Quanto às informações referidas em *b)*, *c)* e *d)*, contribuições dos departamentos do INE.

As contribuições de artigos por iniciativa dos próprios autores serão objecto de avaliação de mérito científico pelo Conselho Editorial, que decidirá ou não pela sua publicação.

Para a elaboração e envio das contribuições de artigos para publicação na Revista são adoptadas as *Normas de Apresentação de Originais* que figuram na última página.

Os autores dos artigos publicados, a que se refere a alínea a), receberão uma contribuição financeira paga pelo INE, de montante a fixar por despacho da Direcção mediante proposta do Director da Revista.

FOUNDATION, SUBJECT MATTER AND SCOPE OF THE REVIEW

INE is conscious of how statistical awareness is essential to the understanding of the majority of phenomena in the present world and is aware of its responsibility to disseminate statistical knowledge, making it available to the widest possible range of readers. INE has recognised the need to take a step in that direction and will begin publication of this *Statistical Review* three times yearly, designed to provide the following:

a) Within a scientific perspective, original articles on specialised areas of statistics, both pure and applied, as well as studies and analyses within the sphere of economics, social issues and demographics;

b) Information on activities and projects of the National Statistical System;

c) Information on activities developed by INE within the scope of co-operation;

d) Information on congresses, seminars and conferences of a statistical or related nature;

The following approaches for contributing material for publication in the review have been adopted:

- In relation to the articles referred to in section a), contributions are made by the authors themselves and by invitation of the Editorial Committee, whether they are employees of INE or not;
- In relation to the information referred to in section b), c) and d); contributions are from departments of INE.

The Editorial Committee who has sole discretion in deciding whether or not the material will be published will assess the scientific merit of contributions made on the initiative of the authors themselves.

The preparation and delivery of material for publication in the Review are subject to the *Rules for Submitting Originals* presented on the last page.

The authors of the published articles referred to in section a) will receive pecuniary compensation from INE in an amount to be determined by resolution of the Board on the recommendation of the Director of the Review.

NORMAS DE APRESENTAÇÃO DE ORIGINAIS

Nos termos do *Regulamento da Revista de Estatística*, o Conselho Editorial aprovou as seguintes **Normas de Apresentação de Originais**:

1. Os originais dos artigos serão enviados ao Director da Revista pelos respectivos autores, devendo ser escritos em *português* e não terem sido ainda totalmente publicados, ou estar em processo de edição em outra publicação.

2. Poderão também ser apresentados artigos escritos em *inglês*, cabendo ao Director da Revista a decisão sobre a sua aceitação.

3. Quanto à *avaliação do mérito científico* dos artigos:

   a) Os artigos apresentados por *iniciativa* dos respectivos autores serão submetidos à avaliação do mérito científico pelo Conselho Editorial, com garantia do anonimato tanto do autor como dos avaliadores;

   b) Os autores receberão a informação sobre o resultado da avaliação num prazo máximo de trinta dias, com indicação, nos casos de avaliação positiva, do número da *Revista* em que serão publicados, e nos casos de avaliação negativa com a devolução do original apresentado.

4. Os artigos aceites para publicação na *Revista de Estatística* serão igualmente divulgados no *site* do INE na *Internet*.

5. Os originais, com uma extensão não superior a trinta páginas, serão processados em *Word for Windows*, integralmente a preto e branco, com indicação do(s) software(s) adicional(ais) eventualmente utilizado(s) na produção do documento original, e entregues em suporte papel acompanhado da respectiva *disquette*, ou enviados por E-mail para o seguinte endereço: liliana.martins@ine.pt

6. Na apresentação dos originais, os autores respeitarão ainda as seguintes normas:

   6.1. Quanto à *estrutura*:

      a) O texto deve ser processado em formato A_4, com utilização do tipo de letra *Times New Roman 11*, espacejamento *at least 12*, e com as seguintes margens: *top*: 4 cm, *bottom*: 3 cm, *left*: 2,5 cm, *right*: 5 cm, *header*: 1,25cm, *footer*: 1,25cm;

      b) A primeira página conterá exclusivamente o título do artigo, bem como o nome, morada e telefone, fax e E-mail do autor, com indicação das funções exercidas e da instituição a que pertence, devendo, no caso de vários autores, ser indicado a quem deverá ser dirigida a correspondência da Revista;

      c) A segunda página conterá, em português e inglês, unicamente o título e um *resumo* do artigo, com um máximo de 100 palavras,

seguido de um parágrafo com indicação de *palavras-chave* até ao limite de 15;

d) Na terceira página começará o texto do artigo, sendo as suas eventuais secções ou capítulos numeradas sequencialmente;

6.2. Quanto a *referências bibliográficas*:

a) Os autores eventualmente citados no texto do artigo serão indicados entre parênteses curvos pelo seu nome seguido da data da respectiva publicação e, se for caso disso, do número de página (p. ex.: Malinvaud, 1989, 23);

b) As Referências Bibliográficas serão listadas, por ordem alfabética dos apelidos dos respectivos autores, imediatamente a seguir ao final do texto, de acordo com a fórmula seguinte:

GREENE, W. H., "*Econometric Analysis*", Prentice-Hall, New Jersey, 1993.

6.3. Quanto à *revisão de provas e publicação*:

a) Uma vez aceite o artigo e antes da sua publicação, receberá o autor provas para revisão, as quais serão devolvidas ao Director da Revista no prazo máximo de uma semana contado da data da sua recepção;

b) Serão da responsabilidade dos respectivos autores as consequências de eventuais modificações da versão inicial aceite, bem como de atrasos na revisão das provas, que impossibilitem a publicação no número da Revista previsto, reservando-se o Director o direito de decidir a data da sua publicação futura;

c) Uma vez publicado o artigo, o autor receberá vinte exemplares da sua versão impressa e um exemplar do respectivo número da *Revista*.

7. Para *informações adicionais* contactar o Secretariado de Redacção:

Eduarda Liliana Martins
Instituto Nacional de Estatística
Avª. António José de Almeida, n.º 5 – 9º.
1000-043 Lisboa - Portugal

- Tel.: +351 21 842 62 05
- Fax.: +351 21 842 63 84
- e-mail: liliana.martins@ine.pt

RULES FOR SUBMITTING ORIGINALS

Within the terms of the *Regulation of the Statistical Review*, the Editorial Committee has approved the following **Rules for Submitting Originals**:

1. The original articles will be sent to the Review Director by the respective authors. They should be written in *Portuguese*, they should not have already been published in their entirety nor should they be in the process of being published in any other publication.

2. Articles may also be submitted in *English* to the Review's Director who will decide whether to accept them.

3. In relation to the *evaluation of the scientific merit* of the articles:

   a) The Editorial Committee will assess the articles submitted on the initiative of the authors on the basis of their scientific merit. The identity of both the author and the Committee members will be strictly confidential;

   b) The authors will receive information regarding the results of the evaluation of scientific merit within a maximum period of 30 days. If the article is accepted, the Committee will indicate the issue number of the *Review* in which the article will be published. If the article is not accepted, the original will be returned to the author.

4. The articles accepted for publication in the *Statistical Review* will also be made public on the Internet site of the INE.

5. The original articles having no more than thirty pages must be processed in *Word for Windows*, completely at black and white, with the information on the addicional(s) software(s) eventually used in the production of the original document, and they will be delivered in hard copy as well as on diskette, or sent by E-mail to: liliana.martins@ine.pt

6. With the presentation of the original articles, the authors must also respect the following rules:

   6.1 In relation to the *structure*:

   a) The text shall be printed on A4 format paper utilising the font *Times New Roman* size 11, spacing at least 12, and with the margins: *top* 4cm, *bottom* 3cm, *left* 2,5cm, *right* 5cm, *header* 1,25cm, *footer* 1,25cm;

   b) The first page shall contain only the title of the article as well as the name, address and telephone, fax and E-mail number of the author, indicating the position held and the institution that he/she belongs to. In the case of various authors, it is necessary to indicate the person to whom all correspondence received should be forwarded;

c) The second page shall contain in *Portuguese* and *English* only the *title* and an *abstract* of the article with the maximum of 100 words followed by a paragraph indicating *key words* up to the limit of 15;

d) The third page will begin the text of the article with its respective sections or chapters sequentially numbered;

6.2 Regarding *Bibliographical References*:

a) Authors who are cited in the text of the article shall be indicated in parentheses with their name followed by the date of the respective publication and, if necessary, the page number (ex.: Malinvaud, 1989, 23);

b) All bibliographical references will be listed in alphabetical order by the surnames of the respective authors, immediately following the end of the text, as in the following example:

GREENE, W. H., "*Econometric Analysis*", Prentice-Hall, New Jersey, 1993.

6.3 Regarding *proof-reading and publication*:

a) Once the article is accepted and prior to its publication, the author will receive a copy for review. These copy will be returned to the Director of the Review within a maximum period of one week from the date of its reception;

b) The consequences of subsequent changes to the accepted first version are the responsibility of the respective authors as well as any delays in proof-reading that make its publication in the planned issue of the Review impossible. The Director reserves the right to decide upon the date for future publication;

c) Once the article is published, the author will receive twenty copies of his/her printed version and a copy of the respective issue of the *Review*.

7. For *further information* kindly contact the Editorial Secretary:

Eduarda Liliana Martins
Instituto Nacional de Estatística
Av^{a}. António José de Almeida, n^{o}. 5 – 9^{o}.
1000-043 Lisbon - Portugal

- Tel.: +351 1 21 842 62 05
- Fax.: +351 1 21 842 63 84
- e-mail: liliana.martins@ine.pt