ISSN 0873-4275

INSTITUTO NACIONAL DE ESTATÍSTICA

PORTUGAL

REVISTA DE ESTATÍSTICA

VOLUME II
2º QUADRIMESTRE 1998

VOLUME 2

2º QUADRIMESTRE DE 1998

REVISTA DE ESTATÍSTICA
3ª PAGINA
2º QUADRIMESTRE DE 1998

ÍNDICE

INDEX

REVISTA DE ESTATÍSTICA
5ª PAGINA
2º QUADRIMESTRE DE 1988

THE INFORMATION GAP ON THE VERY SMALL ENTERPRISES; CONCEPTUAL AND MEASUREMENT ISSUES OF THE DATAMED APPROACH

Autor:
Daniel Santos

VOLUME 2

2º QUADRIMESTRE DE 1998

REVISTA DE ESTATÍSTICA
7ª PAGINA
2º QUADRIMESTRE DE 1998

THE INFORMATION GAP ON THE VERY SMALL ENTERPRISES; CONCEPTUAL AND MEASUREMENT ISSUES OF THE DATAMED APPROACH

LACUNAS DE INFORMAÇÃO NO DOMÍNIO DAS MICROEMPRESAS; ALGUNS PROBLEMAS CONCEPTUAIS E DE AVALIAÇÃO NA ABORDAGEM ESTATÍSTICA DESTE UNIVERSO

Autor: Daniel Santos
Director de Departamento, Instituto Nacional de Estatística

ABSTRACT

- This paper was presented in the Workshop "*Data Capturing on the overlapping area between enterprises and households in the Mediterranean countries*", held in Rome. In five points, it reviews some features of DATAMED first report.

The paper begins, in the first point, by analysing the role of micro enterprises in the Mediterranean economies, specially in the labour market dynamics. This seems to be the reason for new information needs. Second point examines the capacity of traditional information systems to give an answer to these emerging needs, concluding that the present situation does not allow this kind of adjustment. Third point presents the new statistical approach, developing by DATAMED which is characterised for being a transversal approach of two different realities existing in the same unit - the Household with production activity. Fourth point raises three important problems related with data collection and dissemination of results, to which DATAMED could give a positive answer: quality improvement, reducing burden on respondents and costs for collectors and increasing statistical coverage. Finally, the author finishes, discussing some conceptual and measurement issues, namely: the reason for restricting the scope of DATAMED to recorded units; the main domains for indicators; problems associated to sources' linkage; sample design and consistency with other existent sources and measurement of non monetary flows.

KEY-WORDS

- *Data Capturing, Very Small Enterprise, Household*

RESUMO

- O artigo corresponde à comunicação apresentada no Workshop "*Data Capturing on the overlapping area between enterprises and households in the Mediterranean countries*", realizado em Roma, passando em revista os principais traços do primeiro relatório do projecto DATAMED.

O artigo inicia-se por uma análise do papel das micro-empresas nas economias dos países Mediterrânicos, fundamentalmente no que respeita à dinâmica do mercado de trabalho, determinando novas necessidades de informação para as quais os

sistemas de informação existentes não têm capacidade de resposta. É apresentado o modelo de abordagem estatística proposto pelo projecto DATAMED que se caracteriza por possibilitar uma visão transversal de duas diferentes realidades coexistentes na mesma unidade institucional - Família desenvolvendo uma actividade de produção. São abordados os problemas que, no campo da recolha de informação estatística e difusão de resultados, habitualmente se colocam aos sistema estatísticos – qualidade dos resultados, sobrecarga sobre os inquiridos, custos da recolha de informação e cobertura estatística. Por último, o artigo aflora questões conceptuais e de quantificação no âmbito do projecto: domínio de observação, principais indicadores, ligação entre bases de dados, consistência entre diferentes fontes e avaliação de fluxos não monetários.

PALAVRAS-CHAVE

- *Recolha de informação estatística, Micro-empresa, Família*

INTRODUCTION

DATAMED[1] is a research project with the purpose of developing a new statistical approach to the small production units belonging to the Household sector and a prototype system applying advanced technologies to data capturing.

Expected results of this research are the following:

- to increase statistical coverage in the complex reality of micro-enterprises,
- to reduce burden on respondents and costs for collecting data,
- to provide the tools for gathering integrated statistical data on Household units with economic production activity, for the social and economic domains as a whole.

In order to achieve these objectives, DATAMED will explore all possibilities of using information sources such as administrative registers, data derived from existing surveys, databases of intermediaries[2], combined and complemented with information derived from a specific survey applied to households with production activity.

ROLE OF SMALL ECONOMIC PRODUCING UNITS IN THE CONTEXT OF THE HOUSEHOLD SECTOR

The boundary between production and consumption does not imply a correspondent frontier between producers and consumers as different units. This happens in Households, where the two realities may coexist in the same institutional unit. Production in the household sector takes place in *Unincorporated enterprises* – Very Small Enterprises (VSE) – directly owned and controlled by members of a household.

This kind of units has an important weight in the economic activity, specially in the Mediterranean countries, where very small business plays a significant role in the production structure, namely in terms of number of economic units and employment. Table 1 shows the importance of VSEs in total business for the three partners involved in DATAMED, comparing with European Union.

Table 1 - Share (%) of VSEs in total business (1996)

	Enterprises	Employment	Turnover
EU	93	33	23
Greece	97	37	41
Italy	90	46	25
Portugal	94	35	24

[1] DATA capturing in MEDiterranean countries. National Statistical Institutes of Italy, Greece and Portugal, Olivetti Research and University of Patras, make up the consortium that develops this research project financed by the European Community within the Fourth Framework Programme in the information technologies sector. Further information is available on the web site: www.istat.it/datamed/datamed..htm

[2] Professional associations, suppliers of accountancy services to micro-enterprises,...

VSEs are characterised for being labour-intensive; in order to produce the same output, they use more quantity of labour as input. There are mainly three reasons for this: activity location, level of qualification of employment and capacity of investment.

VSEs locate mainly in traditional and labour intensive activities, such as agriculture, textile industries, construction, retail trade, repair of transport equipment and other durable goods, accommodation services and restaurants and transport services. VSEs hire less educated and trained employees which reflect the lower level of labour productivity. Finally, VSEs have more difficulties in attracting financial means and to access credits; this results on lower rates of investment.

Some studies carried out on small business show that the share of small units in total employment has increased in the seventies and eighties, although rates of unemployment remained at high level. Some reasons can be pointed to justify this movement:

- technological development have decreased the economies of scale,
- downsizing of large enterprises linked to objectives of increasing flexibility and efficiency.

Small business reveals potential conditions for job creation and accommodation of job destruction in large firms. Micro-enterprises play an important role in labour market; normally VSEs create more jobs than large enterprises do. This does not mean that all VSEs grow faster than all other firms -- the growth rate of an enterprise depends on several factors, among which, size and activity sector are important. Although VSEs may be characterised as volatile units[5] -- some of them will close, others will survive and grow to a size that ensures efficiency -- if we consider a net rate of job creation, micro-enterprises show higher rates than other firms which may be related to the fact that they are relatively young.

Different reasons may justify the engagement of a Household or one of its members in a productive activity:

- as a reaction to a crisis in the economy increasing unemployment,
- as a way of increasing household income,
- as a result of the entrepreneurial capacity.

The first corresponds to a passive role, reflecting the change in behaviour induced by slowing-down of the economic activity; the second is generally connected to low income (low wage level, underemployment,...). Both these are not relevant as generating employment growth in the economy.

Finally, the third reason requiring initiative, enterprise and capital equipment, corresponds to an active role in the economy -- creation of jobs in the long term.

In conclusion, performance and behaviour of Households-VSEs play an important role in the economy: they are decisive in some economic activities; they could support the dynamics of employment growth in the economy and be an instrument to reduce regional differences in levels of economic development.

[5] They generally present the lowest survival rates.

THE PRESENT SITUATION OF INFORMATION SYSTEMS IN SOCIAL AND ECONOMIC STATISTICS

In general, statistical systems have based their approach for collecting data and produce statistics, in the traditional separation between consumers and producers. This approach reflects the theoretical division between production and consumption, which in the case of a Household-VSE does not correspond to two different institutional units in reality.

In this sense, National Statistical Institutes (NSIs) conduct separate surveys to collect data on social conditions of Households -- Labour Force Survey (LFS), Household Budget Survey (HBS), European Community Household Panel (ECHP),... -- and on production activity (Enterprise Surveys).

This situation lead to macrodata on different domains such as employment, unemployment, income, consumption and production that, most of the times, are not integrated and consistent and became too restrictive, since do not allow further analysis, specially considering the relationship between social and economic in the core of the same unit – the Household.

The present statistic model is characterised by systematic undercoverage in what respects the evaluation of the output for small business; enterprise surveys normally present a reasonable coverage of corporations, but fail on seizing micro-enterprises. Data on demographic events for small business (births, deaths, survival rates, job creation and job destruction,...); on the effects of economic policy in the behaviour of VSEs (removing barriers to entrepreneurship, training programs, tax reductions, interest bonus,...) and reactions of VSEs to the economic cycle, is not normally available.

The situation of undercoverage of small production units is presently one of the weaknesses of statistical systems as the basis for producing National Accounts for the Household institutional sector.

The existing statistic model is inflexible in the sense that it does not allow, based on the existing surveys, to draw an integrated and comprehensive picture of this reality which could give an answer to users' needs (policy makers, national accountants, researchers, professional associations,...).

THE NEED FOR A DIFFERENT APPROACH

As pointed above, there is a considerable information gap on the Household-VSE domain, for which the present structure of the statistical model does not give a satisfactory answer.

DATAMED project was designed to fill the lack of integrated data on this sector and to satisfy new demands of information on the household sector, developing simultaneous:

- new technologies for capturing data,
- a comprehensive questionnaire.

To fulfil these objectives, DATAMED developed in the first workpackage:

- the inventory of potential new technologies that could be used in the field,
- the analysis of the environment - legal and statistical constraints, existing sources and intermediaries as potential providers of data,
- an evaluation of users needs in order to built the instruments for surveying.

The content for the database is derived from users' proposals concerning the statistical output, that is to say, indicators that were pointed by users as important needs, fixed the target variables to be surveyed. These variables can be collected by capturing data in each institutional unit, the Household-VSE, referred to different observation units:

- micro-enterprise,
- household,
- individuals (members of the household or employees of the micro-enterprise).

Table 2 provides a summary of indicators outlined by potential users.

Table 2 - Main users and related indicators

USERS	OBJECTIVE	SCOPE	GROUPING OF INDICATORS
Ministry of Economics	optimise economic growth and productivity	Country social-economic aspects)	• economic results by sector of industry • employment by sector of industry • unemployment by: age, sex, education level, profession, ...
Ministry of Labour	optimise employment	Country and Region (social-economic aspects)	• employment by sector of industry • unemployment by: age, sex, education level, profession, ... • same indicators by region • changes over time
Ministry of Transport	optimise infrastructure	Country and Region (social-economic aspects)	• supply and use of infrastructure
Ministry of Agriculture	optimise employment, economic growth and productivity in agriculture	Country and Region(social-economical aspects)	• economic results in agriculture • employment in agriculture • agricultural infrastructures
Ministry of Finance	optimise tax collection	Country (social-economic aspects)	• income from enterprises (tax) • income from households (tax)
Social Security	optimise welfare of citizens	Country (social-economic aspects)	• average income of people according to: age, sex, employed/unemployed,

			• distribution of income and expenses between categories of employed and unemployed people
Regional Government	optimise social-economic aspects in the region	Region (social-economic aspects)	• all above indicators by regions
Local Government	optimise social-economic aspects in the municipalities	Municipality	• all above indicators by municipalities
Chamber of Commerce	optimise economic activity in the region	Region (social-economic aspects)	• demography of enterprises • economic results by detailed sector of industry
NSI • national accounts • specific statistical departments	providing data on behalf of the users (specially those addressed above)	Country and Region (social-economic aspects)	• derived from information needs mentioned above

Three aspects are innovative in this development for the project DATAMED:

- the subsystem of information on Households-VSEs is built strictly responding to users needs - new demands for data structured the model for collecting and producing data,
- for the first time users will accede to data linking two correlated domains -- social and economic -- referred to the same statistical unit; this can produce more consistent results at macro level,
- finally, this kind of correlated data will be available at micro level, allowing modelling.

This new approach does not intend to substitute the traditional household surveys (LFS, HBS, ECHP and other household surveys). First, it only applies to a subset of Households -- those who act in the economy simultaneous as producers and as consumers. Second, it does not provide a detailed picture on specific thematics for which traditional surveys are more adequate. DATAMED approach is complementary to existing traditional statistics on Households -- for instance, this approach has a potential capacity for explaining dynamics in labour market complementing the detailed description provided by Labour Force Surveys.

SPECIFIC PROBLEMS RELATED TO DATA CAPTURING ON HOUSEHOLDS-VSES UNITS

Data collection and dissemination of results on small economic units, deals with specific problems, mainly:

- quality improvement,
- burden on respondents and costs for NSIs,
- coverage.

Development of DATAMED project should give some progress to these aspects of statistical nature.

Quality of data should be envisaged in a broad sense considering timeliness, methodological harmonisation and accuracy. The new and innovative technologies for data capturing are the backbone to attain quality. Methodological harmonisation is settled for all partners in the context of the statistical approach developed by DATAMED providing a single definition of domain and the use of the same socio-economic concepts in coherence with the National Accounts framework. Accuracy is a consequence of methodological harmonisation and a result of applying new data capturing technics that can support sophisticated control rules for microdata. Finally, new technologies can also reduce the existent lag between reference period and availability of data.

Burden on respondents is normally pointed as one of the reasons for undercoverage (significant non response rates) and reflects also inefficiency of the statistical process.

Two reasons may be in the origin of this inefficiency. First, instruments for collecting data are not adapted to the respondents or accommodate very detailed breakdown for variables, some of them unuseful. To avoid this problem a statistical project, and this is the case of DATAMED, should pay attention to the definition of questionnaire contents, taking account, as main objective for collecting data, of what is relevant for users and of what is possible for respondents. Second, every possibility of using information from intermediaries is not being explored. Of course there are some problems that must be solved in order to allow this practice, mainly: the existing legal constraints and the necessary correspondence with statistical concepts. DATAMED model for data capture is being designed to benefit of potential databases detained by information brokers, with obvious advantages for respondents (reduced burden) and for NSIs (reduced costs).

Statistical coverage is a weakness in the VSEs field. We have seen that traditional statistic process does not give an acceptable answer to new needs in the socio-economic domain. Also in the production side traditional statistics (Enterprise Surveys) do not pay special attention to micro enterprises. DATAMED can provide, in this transversal approach, the tools to answer specific users needs, to reach significant coverage in VSEs statistics and to combine social and economic perspectives for this kind of units.

CONCEPTUAL AND MEASUREMENT ISSUES

DATAMED project sets some conceptual and measurement problems namely:

- the definition of domain,
- the target variables defined according to the indicators,

- linkage between different existing sources,
- sampling, extrapolation of data and consistency with other statistical sources,
- evaluation of non monetary flows and other measurement problems.

Definition of domain is crucial for DATAMED framework. In order to determine the grey area between Enterprises (production units) and Households, some combined criteria can be used: size related to employment and/or turnover, sector of economic activity and legal status. Normally, the kind of economic activity is not decisive to define a VSE linked to the institutional unit household. Naturally, the density of VSE is more relevant certain activities! Analyses done in the first workpackage as conclude that legal status (defining or not an independent economic unit from the household or individual) and size defined by employment are the fundamental criteria to consider.

The definition of Household-VSE proposed in DATAMED context is the following:

"*a group of persons living together in the same dwelling unit, pooling income and resources, managing a join budget, collectively consuming goods and services, one of whom owns an enterprise producing goods and services for the market which is not a corporation or quasi-corporation.* "

DATAMED domain is defined as:

"*all enterprises and households producing not only for own consumption to be addressed as Very Small Enterprises with the following characteristics: unincorporated (non-legal status); with no more than 9 employees and not bound to keep a complete set of accounts*"

Although this definition of domain places the scope of the project in a broad sense including informal units, it was decided, in this phase of development, to cover only the official economy, that is to say, only recorded units in administrative (fiscal, social security, or other files) or business registers. No doubt, this decision does not comply totally with the aim of statistical coverage for all productive activity based in the Household sector -- unrecorded small economic units are not in scope. The inclusion of these units – informal units – could lead the project to difficulties, specially in data capturing, presenting disadvantages greater than the benefit of an ideal total coverage.

In the first workpackage partners sound out the potential users in order to establish the core of indicators in DATAMED. Target variables are identified and defined according to this core. Indicators cover essential domains such as: the demography of the units; characteristics of the production process; characteristics of the household and insertion in the overall economy (Productive system and Household sector). These indicators provide a transversal overview for VSE/Households.

One of the possibilities for reducing NSIs costs and burden on respondents is to appeal intermediaries for getting data on VSEs that could be linked and complemented by other domains with origin in specific surveys conducted by NSIs. For instance, one

could use, referred to the same unit – household or individual member of the household -- fiscal records (tax declarations) and social security records combined and complemented with data from existing surveys (LFS, ECHP) or with data from a new survey specifically designed for this objective. This method maximises the use of existent sources, internal and external to the statistical system. Special attention, in this case, should be drawn to a common identification key that allows the linkage of data for the same unit with different origins. A solution should be found in order to go beyond existing legal constraints related to privacy protection.

Administrative registers and Business registers are the sampling frame for surveying VSEs/Households. Sample design should allow, for reasonable detail levels, significant results by kind of activity, regions, size of economic units, type of households. Since we are in presence of a transversal overview, extrapolation of data should produce consistent results with other specific sources (LFS, ECHP, Enterprise statistics,...). In order to achieve this objective, DATAMED should adopt a complex set of weights so that macrodata should reflect the same structures presented in those specific sources. Here, the experience developed in ECHP could give some help.

Scope of DATAMED excludes a great part of non monetary flows – households only producing for its own consumption are excluded. Nevertheless, non market flows may exist in potential surveyed units, namely goods and services produced for own final use --excluding domestic and personal services – that are included in the production boundary: partial consumption of the output by household members and self-investment.

Methods used by National Accounts for evaluating this kind of non market flows can be adopted. The best procedure to value these goods and services, is to use the same price at which they could be sold if offered for sale in the market. This procedure will require reliable information on market prices for this kind of goods and services. If these data cannot be obtained, then a second best procedure -- the production costs -- should be applied, so that value of these goods and services could correspond to the value of inputs used in their production:

- intermediate consumption,
- compensation of employees,
- fixed capital consumption,
- taxes on production.

Also, special attention should be paid to some particular flows that could present evaluation problems, namely: the separation of household acquisitions by purpose (intermediate consumption/gross fixed capital formation or private consumption) and financing flows, external to the banking system. The questionnaire design is essential to provide the means for a correct evaluation of this kind of flows. As pointed above, the institutional unit should be split into three observation units helping in the separation between acquisitions as an input for producing and as final consumption of the household. Acquisitions should be affected to each observation unit (Household, Household member and VSE) taking account its main purpose – some could be clearly affected, others could be partially shared between final consumption and intermediate consumption.

BIBLIOGRAPHY

The European Observatory for SMEs, Fifth Annual Report, 1997, European Network for SME Research..

Report on Users' requirements, February 1998, DATAMED.

VOLUME 2

2º QUADRIMESTRE DE 1998

REVISTA DE ESTATÍSTICA
19ª PAGINA
2° QUADRIMESTRE DE 1998

EQUIVALÊNCIA ENTRE OS COEFICIENTES DE GINI E DE GOODMAN-KRUSKAL EM ÁRVORES DE DECISÃO

Autor:
Joaquim Fernando P. Costa

VOLUME 2
2º QUADRIMESTRE DE 1998

REVISTA DE ESTATÍSTICA
21ª PÁGINA
2º QUADRIMESTRE DE 1998

EQUIVALÊNCIA ENTRE OS COEFICIENTES DE GINI E DE GOODMAN-KRUSKAL EM ÁRVORES DE DECISÃO

IN DECISION TREES, THE SPLITTING CRITERIA OF GINI AND GOODMAN-KRUSKAL ARE EQUIVALENT

Autor: Joaquim Fernando P. Costa
- Professor Auxiliar no Departamento de Matemática Aplicada da Universidade do Porto
e
- Investigador do LIACC

RESUMO

- As árvores de decisão são uma técnica não paramétrica de análise discriminante que nas últimas décadas se tornou muito popular em ambas as comunidades de estatística e de inteligência artificial. Geralmente, o principio subjacente à sua construção, consiste em dividir recursivamente o espaço de descrição em regiões e depois, dentro de cada região, escolher a classe cuja frequência é máxima. A divisão recursiva do espaço é obtida comparando as variáveis descritivas com a variável a predizer, e escolhendo a que optimiza um certo critério, ou coeficiente. Neste trabalho demonstramos que o uso do critério de Gini, que é utilizado no célebre método CART (Breiman & al., 1984), é equivalente ao uso do coeficiente de (Goodman & Kruskal, 1954).

PALAVRAS CHAVE:

- *Análise discriminante, árvores de decisão, coeficientes de associação entre variáveis qualitativas.*

ABSTRACT:

- Decision trees are a non-parametric technique of discriminant analysis, which has become very popular in the last decades in both communities of statistics and artificial intelligence. In general, the principle underlying their construction, consists in splitting recursively the description space in regions, and then, inside each region, choosing the more frequent class. The recursive division of the space is obtained by comparing the descriptive variables (attributes) with the variable (or concept) to be predicted; then one chooses the descriptive variable which maximises a certain splitting criterion. In this work we demonstrate that using the Gini splitting criterion, which is employed in the celebrated program CART (Breiman & al., 1984), is equivalent to using the coefficient of (Goodman-Kruskal, 1954).

KEYWORDS:

- *Discriminant analysis, decision trees, measures of association between qualitative variables*

REVISTA DE ESTATÍSTICA
23ª PAGINA
2º QUADRIMESTRE DE 1998

1. INTRODUÇÃO

Neste trabalho tratamos o problema da discriminação através do uso de Árvores de Decisão. Dado um conjunto de aprendizagem t_r de n indivíduos descritos por $K+1$ variáveis, $v^1, v^2, \ldots, v^K, c$ ($v^1, v^2, \ldots, v^K$ são as variáveis descritivas e c é a variável classe a discriminar), pretende-se construir um conjunto de regras que permitam identificar, para um novo indivíduo o, a que classe ele pertence, supondo conhecidos os valores $v^1(o), v^2(o), \ldots, v^K(o)$. As Árvores de Decisão tem sido usadas com sucesso como ajuda à tomada de decisão em problemas muito diversos. Esta técnica, desenvolvida sobretudo a partir dos trabalhos de (Quinlan, 1986) e (Breiman & al., 1984), originou uma forte interacção entre a Estatística e a Inteligência Artificial, mais concretamente, na área da Aprendizagem ((Nakhaeizadeh, 1994), (Taylor, 1994)). A decisão final sobre a classe a atribuir a um novo indivíduo é uma consequência de uma série de decisões parciais e pode eventualmente ser tomada sem ter necessidade de conhecer todos os valores $v^1(o), v^2(o), \ldots, v^K(o)$, o que reduz o custo; por exemplo, em diagnóstico médico, as variáveis $v^1, v^2, \ldots, v^K$ podem ser o resultado de um conjunto de testes médicos dispendiosos, prejudiciais e eventualmente fatais para o paciente e portanto só devem ser realizados todos se forem realmente necessários.

As árvores de decisão podem ser binárias (em que cada nó t é dividido em exactamente dois nós descendentes t_l e t_r) ou não binárias. A construção de uma Árvore de Decisão processa-se de uma forma recursiva. Começa-se por "colocar" todos os indivíduos do conjunto de aprendizagem na raiz da árvore. De entre o conjunto de variáveis descritivas $V = \{v^1, v^2, \ldots, v^K\}$ escolhe-se aquela que é a mais discriminante em relação à variável classe, c. Por exemplo, se quisermos construir uma árvore binária, temos de começar por binarisar todas as variáveis descritivas. Para isso, se por exemplo a variável v^j for qualitativa nominal, tomando L valores diferentes, então v^j dá origem a $2^{L-1}-1$ variáveis binárias, correspondentes a todas as bipartições das L modalidades. No caso de uma variável numérica v^j, tomando valores $x_1, x_2, \ldots, x_n$, teremos de considerar $n-1$ variáveis binárias do tipo $I(v^j \le c_j)$, sendo por exemplo $c_j = \frac{x_j + x_{j+1}}{2}$, $j = 1,2,\ldots,n-1$. Cada uma destas variáveis binárias terá de ser comparada com a variável c, para determinar qual a mais discriminante. Procede-se então à divisão da raiz da árvore, criando dois novos nós. De seguida faz-se o mesmo para cada um destes novos nós, e assim sucessivamente.

Como ilustração, consideremos a árvore de decisão da figura 1, onde $V = \{v^m / 1 \le m \le 4\}$ designa o conjunto dos atributos já binarisados e $C = \{C_j / 1 \le j \le 3\}$ o conjunto das classes (ou grupos) a discriminar.

Um dos problemas principais em árvores de decisão tem a ver com a questão de saber quando parar a construção da árvore. Evidentemente, se continuarmos a

segmentar o espaço descritivo até que cada folha da árvore contenha elementos de uma só classe, essa árvore classifica correctamente todos os exemplos do conjunto de aprendizagem. No entanto, não é de esperar que essa árvore dê bons resultados num outro conjunto de exemplos independente do conjunto de aprendizagem. Uma das razões tem a ver com o facto de nos últimos nós da árvore as estimativas das probabilidades de cada classe serem pouco fiáveis, devido ao pequeno número de exemplos nesses nós terminais. Para resolver este problema, as estratégias iniciais consistiam em parar a segmentação de um nó sempre que o decréscimo no erro de classificação ao usar a variável (atributo) escolhida para dividir esse nó, era inferior a uma certa constante. De uma maneira geral, isto não deu bons resultados, pois era difícil de saber que constante utilizar. Uma constante que seria boa para um nó, poderia não o ser para outros. Por outro lado, embora num dado nó possa não existir nenhum atributo suficientemente predictivo, poderá ser que num descendente desse nó um desses atributos se revele predictivo.

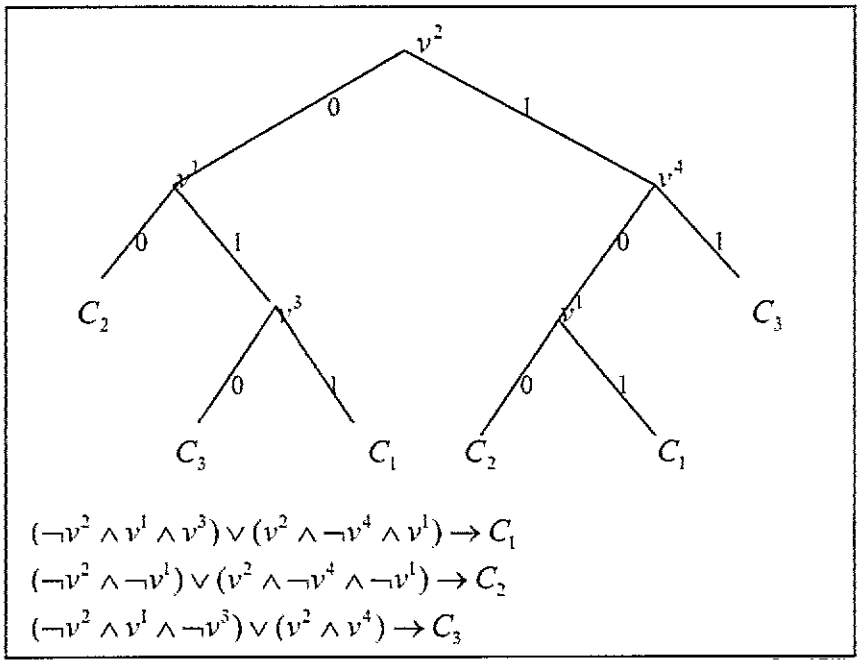


FIG. 1 - Exemplo ilustrativo de uma árvore de decisão binária e respectivas regras de decisão.

(Breiman & al., 1984) propuseram um método radicalmente diferente. A ideia consiste em fazer crescer uma árvore muito "grande" em que cada folha contém elementos de uma só classe, ou então um número muito pequeno de elementos. De seguida "poda-se" essa árvore, obtendo uma sucessão de árvores em que cada uma delas resulta da anterior por "poda" de um ou mais dos seus nós. Para escolher a árvore óptima da sequência, pode-se por exemplo usar um conjunto de exemplos independente do conjunto de aprendizagem, estimar o erro (custo) de cada árvore, e escolher a melhor. Este método tem dado bons resultados.

O problema que nós abordamos neste trabalho refere-se à escolha do atributo binário a utilizar na divisão de um nó. Vamos considerar apenas o caso de árvores binárias, embora a extensão para as árvores não binárias seja imediata. Suponhamos então que estamos num nó t da árvore de decisão e queremos saber qual dos atributos binários do conjunto V é mais predictivo em relação à variável classe, c. Temos portanto de comparar cada um deles com a variável c. Para isso, estabelecemos a tabela de contingência que cruza cada atributo v^i com o atributo c. De seguida,

medimos a dependência entre v^i e c através de um coeficiente, como por exemplo o do qui-quadrado, os de (Lerman, 1992a, 1992b), o de Gini (Breiman & al., 1984), o de (Goodman-Kruskal, 1954) ou os critérios derivados deste (Light et Margolin, 1971 ; Haldane, 1940 (citado em Kendall & Stuart, 1961)). O objectivo principal deste trabalho é o de demonstrar que os coeficientes de Gini e de Goodman-Kruskal, quando usados em árvores de decisão, são na realidade o mesmo coeficiente.

2. O COEFICIENTE DE GINI

Seja t um nó da árvore de decisão e O_t o subconjunto de objectos correspondente a esse nó. A partição induzida pela variável classe c sobre O_t é designada por

$$\{O_{tj} / 1 \leq j \leq K\}.$$

Assim, a probabilidade de um objecto de O_t pertencer à classe j de c é estimada pela proporção relativa

$$p_j^t = \frac{card(O_{tj})}{card(O_t)}.$$

Como vimos na introdução, vamos apenas considerar árvores de decisão binárias. Designemos então por t_l e t_r os dois nós descendentes de t determinados pela variável binária w. Da tabela de contingência de c com w, podemos formar a seguinte tabela:

c / w	1	...	k	...	K	
t_l	$p_1^{t_l}$	...	$p_k^{t_l}$	...	$p_K^{t_l}$	1
t_r	$p_1^{t_r}$	...	$p_k^{t_r}$	...	$p_K^{t_r}$	1
t	p_1^t	...	p_k^t	...	p_K^t	1

onde as proporções condicionais são dadas por

$$p_l^t = \frac{card(O_{t_l})}{card(O_t)};\ p_r^t = \frac{card(O_{t_r})}{card(O_t)};$$

$$p_k^{t_l} = \frac{card(O_k \cap O_{t_l})}{card(O_{t_l})};\ p_k^{t_r} = \frac{card(O_k \cap O_{t_r})}{card(O_{t_r})};\ p_k^t = \frac{card(O_k \cap O_t)}{card(O_t)}.$$

A ideia fundamental na construção de uma árvore de decisão binária é a de selecionar, em cada nó t, um atributo binário de forma a que os nós descendentes, t_l e t_r, sejam mais "puros" que o nó "pai", t. (Breiman & al., 1984) sugeriram então que se começasse por definir uma medida de impureza de um nó, como por exemplo o indice de Gini:

$$\varphi(p_1^t,\ldots,p_K^t) = \sum_{1\le i\ne j\le K} p_i^t p_j^t = 1 - \sum_{1\le i\le K}(p_i^t)^2 \qquad (1)$$

φ satisfaz as condições:

- $\varphi(\frac{1}{K},\frac{1}{K},\ldots,\frac{1}{K})$ é máximo;
- $\varphi(1,0,\ldots,0) = \varphi(0,1,0,\ldots,0) = \ldots = \varphi(0,\ldots,0,1) = 0$;
- φ é uma função simétrica das quantidades $p_1^t, p_2^t,\ldots,p_K^t$.

Isto significa que a impureza é máxima quando todas as classes estão igualmente presentes no nó e mínima quando o nó contém apenas uma classe; ela é tanto maior quanto menor for o poder discriminante desse nó. Para um nó t, a regra ou critério de Gini consiste em escolher a variável binária w - que segmenta t em t_l e t_r - que maximiza a diminuição de impureza, que é dada por:

$$G(t,w) = \varphi(t) - p_l^t\varphi(t_l) - p_r^t\varphi(t_r) \qquad (2)$$

Este coeficiente pode-se escrever sobre a forma:

$$G(t,w) = p_l^t\sum_{j=1}^{K}(p_j^{t_l})^2 + p_r^t\sum_{j=1}^{K}(p_j^{t_r})^2 - \sum_{j=1}^{K}(p_j^t)^2 \qquad (3)$$

$$G(t,w) = p_l^t\sum_{j=1}^{K}(p_j^{t_l} - p_j^t)^2 + p_r^t\sum_{j=1}^{K}(p_j^{t_r} - p_j^t)^2 - p_l^t\sum_{j=1}^{K}(p_j^t)^2 -$$
$$p_r^t\sum_{j=1}^{K}(p_j^t)^2 + 2p_l^t\sum_{j=1}^{K}p_j^t p_j^{t_l} + 2p_r^t\sum_{j=1}^{K}p_j^t p_j^{t_r} - \sum_{j=1}^{K}(p_j^t)^2 =$$
$$p_l^t\sum_{j=1}^{K}(p_j^{t_l} - p_j^t)^2 + p_r^t\sum_{j=1}^{K}(p_j^{t_r} - p_j^t)^2 - (p_l^t + p_r^t + 1)\sum_{j=1}^{K}(p_j^t)^2 +$$
$$2(\sum_{j=1}^{K}(p_l^t p_j^{t_l} + p_r^t p_j^{t_r})\cdot p_j^t) = p_l^t\sum_{j=1}^{K}(p_j^{t_l} - p_j^t)^2 + p_r^t\sum_{j=1}^{K}(p_j^{t_r} - p_j^t)^2 -$$
$$2\sum_{j=1}^{K}(p_j^t)^2 + 2(\sum_{j=1}^{K}(p_l^t p_j^{t_l} + p_r^t p_j^{t_r})\cdot p_j^t)$$

Ora

$$p_j^t = p_l^t p_j^{t_l} + p_r^t p_j^{t_r} \tag{4}$$

e por conseguinte os dois últimos termos de $G(t,w)$ anulam-se:

$$G(t,w) = p_l^t \sum_{j=1}^{K} (p_j^{t_l} - p_j^t)^2 + p_r^t \sum_{j=1}^{K} (p_j^{t_r} - p_j^t)^2 . \tag{5}$$

Relativamente à tabela da página anterior, se usarmos a representação geométrica da análise das correspondências de t_l e t_r pelos seus perfis respectivos, obtemos uma nuvem formada por dois pontos pesados:

$$\{(p_J^{t_l}, p_l^t), (p_J^{t_r}, p_r^t)\} \tag{6}$$

em que $p_J^{t_l} = (p_1^{t_l}, p_2^{t_l}, \ldots, p_K^{t_l})$, $p_J^{t_r} = (p_1^{t_r}, p_2^{t_r}, \ldots, p_K^{t_r})$ e p_l^t (resp. p_r^t) representa o peso do nó t_l (resp. t_r). O centro de gravidade desta nuvem é o ponto $p_J^t = (p_1^t, p_2^t, \ldots, p_K^t)$ (cf. (4)). O coeficiente de Gini (5) é precisamente o momento total de inércia desta nuvem em relação à métrica euclidiana ordinária. Daqui se vê logo que muitos outros coeficientes deste tipo podem ser formados, usando métricas diferentes. Por exemplo, (Lerman & Costa, 1996) e (Costa, 1996) demonstram que se usarmos a métrica do qui-quadrado no cálculo da inércia daquela nuvem de pontos, obtemos precisamente o coeficiente do qui-quadrado habitual para a tabela de contingência acima.

Vimos na introdução que a utilização de variáveis qualitativas com L modalidades (categorias) em árvores de decisão binárias leva-nos a considerar um número exponencial de atributos binários em cada nó da árvore de decisão $(2^{L-1}-1)$. O grande interesse do coeficiente de Gini tem precisamente a ver com a binarização de variáveis qualitativas. Com efeito, os autores do célebre método CART (Breiman & al., 1984) demonstram que, se se utilizar o coeficiente de Gini e o número de classes, K, for 2, então a procura do melhor atributo binário, de entre todos os $2^{L-1}-1$, é de complexidade $O(L \cdot \log(L))$. (Lerman & Costa, 1996) e (Costa, 1996) demonstram este resultado de uma forma geométrica:

2.1. O CASO L = 2

Vamos comecar por considerar o caso $L = 2$, que permite uma interpretação geométrica imediata, e de seguida generalisaremos. Neste caso a variável qualitativa é binária e portanto não levanta problemas de complexidade. Designemos então por l e r os dois valores possiveis da variável. Por uma questão de simplificação, vamos omitir o indice t do coeficiente de Gini (5). Assim, $L = \{l, r\}$ e como $K = 2$,

$$G(w) = p_l \sum_{j=1}^{2} (p_j^l - pj)^2 + p_r \sum_{j=1}^{2} (p_j^r - pj)^2 = p_l \left\| p_J^l - p_J \right\|^2 + p_r \left\| p_J^r - p_J \right\|^2$$

e como $p_J = p_l p_J^l + p_r p_J^r$ (4), então

$$G(w) = p_l \left\| p_J^l \right\|^2 + p_l \left\| p_J \right\|^2 - 2p_l \left\langle p_J^l, p_J \right\rangle + p_r \left\| p_J^r \right\|^2 + p_r \left\| p_J \right\|^2 -$$
$$2p_r \left\langle p_J^r, p_J \right\rangle = p_l \left\| p_J^l \right\|^2 + p_r \left\| p_J^r \right\|^2 + \left\| p_J \right\|^2 - 2\left\langle p_l p_J^l + p_r p_J^r, p_J \right\rangle$$

O último termo desta expressão vale , e por isso o coeficiente de Gini reduz-se a $p_l \left\| p_J^l \right\|^2 + p_r \left\| p_J^r \right\|^2 - \left\| p_J \right\|^2$

Em relação ao critério do χ^2, a expressão é a mesma; mas usando a métrica do χ^2.

Como vimos atrás, o coeficiente de Gini é um momento de inércia de uma nuvem de dois pontos, num espaço de dimensão 2, a priori. Na realidade, essa nuvem de pontos encontra-se num espaço de dimensão 1; mais exactamente sobre o segmento de recta definido por: $\{(p,q) / p > 0, q > 0, p + q = 1\}$. A nuvem de dois pontos é

$p_{\underline{2}}^l = (p_1^l, p_2^l); p_{\underline{L}}$ ($p_{\underline{L}}$ é o peso afectado ao nó esquerdo)
$p_{\underline{2}}^r = (p_1^r, p_2^r); p_{\underline{r}}$ ($p_{\underline{r}}$ é o peso afectado ao nó direito)

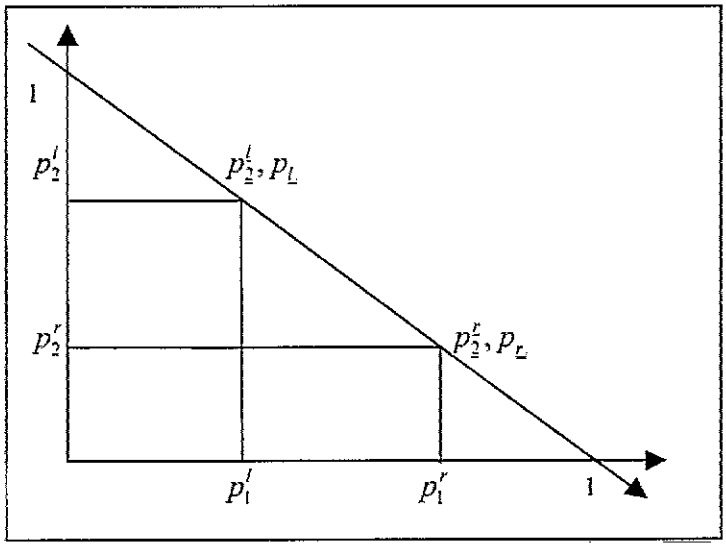


FIG. 2 - Segmento de recta contendo os dois pontos $p_{\underline{2}}^l$ et $p_{\underline{2}}^r$.

O que nos interessa é portanto a inércia desta nuvem de pontos em relação à métrica euclidiana (no caso do indice de Gini). Sobre o eixo transversal do desenho, e que é paralelo à segunda bissectriz, podemos determinar a abcissa de $p_{\underline{2}}^l$ (resp. $p_{\underline{2}}^r$) em função (crescente) de p_1^l (resp. p_1^r).

2.2. O CASO L > 2

Consideremos agora para o caso geral, a tabela de contingência que cruza a variável qualitativa t, com L modalidades, com a variável classe, que neste caso tem 2 modalidades:

j l	1	2
1	n_{11}	n_{12}
⋮	⋮	⋮
L	n_{L1}	n_{L2}

$$p_j^l = p(j/x=l) = \frac{p(j,l)}{p(l)} = \frac{\pi(j)p(l/j)}{\sum_j \pi(j)p(l/j)}$$

Seja $p_1^{l_1} \leq p_1^{l_2} \leq \ldots \leq p_1^{l_L}$ a sequência ordenada das probabilidades $p_1^l, l = 1,\ldots,L$. O coeficiente de Gini neste caso,

$$\sum_{i=1}^{L} p_{l_i} \sum_{j=1}^{2} (p_j^{l_i} - p_j)^2 ;$$

representa a inércia de uma nuvem de pontos situada sobre o segmento de recta ((0,1),(1,0)) (cf. Fig. 2). O nosso objectivo é o de agrupar os L valores da variável v em duas classes de forma a maximizar a inércia da nuvem de dois pontos constituída pelos centros de gravidade dessas classes:

$$\{(p_{\underline{l}}^{l}, p_{\underline{l}.}), (p_{\underline{l}}^{r}, p_{\underline{r}.})\},$$

$\underline{l}$ representa um subconjunto das L modalidades e $\underline{r}$, representa o subconjunto complementar.

A priori, existem $2^{L-1}-1$ bipartições das L modalidades a considerar; no entanto (Fisher, 1958) demonstrou que a partição optimal é em intervalos conexos. No nosso caso temos então de considerar apenas $L-1$ bipartições, pois existem apenas $L-1$ divisões em 2 intervalos conexos.

Como no nosso caso $K = 2$, o indice de Gini é igual a

$$\sum_{l=1}^{L} p_l (p_1^l - p_1)^2 + \sum_{l=1}^{L} p_l (p_2^l - p_2)^2$$

Por outro lado, $p_2^l = 1 - p_1^l$ et $p_2 = 1 - p_1$, de forma que $(p_2^l - p_2)^2 = (p_1^l - p_1)^2$, e por isso a segunda soma da última equação é idêntica à

primeira soma; isto é, a inércia da nuvem projectada sobre o eixo horizontal é idêntica à inércia da nuvem projectada sobre o eixo vertical. Nestas condições, é suficiente maximizar a inércia da nuvem projectada por exemplo sobre o eixo horizontal. Por outro lado, as $L-1$ fusões a considerar são da forma:

$$l_1 \vee l_2 \vee \ldots \vee l_k = \underline{l} \qquad \text{e} \qquad l_{k+1} \vee l_{k+2} \vee \ldots \vee l_L = \underline{r}$$

e por isso,

$$p_1^{\underline{l}} = \frac{p(l_1)p_1^{l_1} + \ldots p(l_k)p_1^{l_k}}{p(l_1) + \ldots + p(l_k)} \text{ e } p_1^{\underline{r}} = \frac{p(l_{k+1})p_1^{l_{k+1}} + \ldots p(l_L)p_1^{l_L}}{p(l_{k+1}) + \ldots + p(l_L)}$$

A maximização da inércia da nuvem $\{(p_1^{\underline{l}}, p(\underline{l})),((p_1^{\underline{r}}, p(\underline{r}))\}$, tem portanto uma complexidade $O(L \cdot \log(L))$, como se queria demonstrar. O mesmo acontece no caso de utilizarmos o critério do qui-quadrado, pois isso corresponde a considerar a métrica do χ^2, em vez da euclideana. (Costa, 1996) e (Lerman & Costa, 1996) demonstram ainda que, no caso de termos apenas duas classes (K=2), os dois critérios, Gini e qui-quadrado, são equivalentes, isto é, dão exactamente os mesmos resultados, pois escolhem sempre a mesma bipartição das L modalidades.

3. O COEFICIENTE DE GOODMAN-KRUSKAL E O CRITÉRIO DE HALDANE, LIGHT E MARGOLIN

O indice proposto por (Goodman & Kruskal, 1954), para comparar duas variáveis qualitativas u e v, dissimetriza a relação entre essas variáveis. Com efeito, o objectivo subjacente à concepção deste coeficiente é o de predizer, para um indivíduo escolhido aleatóriamente da população, que modalidade m_v da variável v lhe corresponde, supondo conhecida m_u.

Se quisermos estimar m_v sem ter em conta m_u, então a proporção de predicções correctas é dada por

$$\sum_{v=1}^{q} (\frac{n_v}{n})^2 .$$

Se tomarmos em consideração a modalidade de u, m_u, correspondente a esse indivíduo, então a proporção de predicções correctas é dada por

$$\sum_{u=1}^{p} \sum_{v=1}^{q} \frac{n_{uv}^2}{nn_u} .$$

A diferença relativa entre as proporções de predicções correctas nos dois casos, que designamos por τ_b, constitui o indice de associação de Goodman-Kruskal entre as duas variáveis u e v:

$$\tau_b = \frac{\sum_{u=1}^{p}\sum_{v=1}^{q}\frac{n_{uv}^2}{nn_{u.}} - \sum_{v=1}^{q}(\frac{n_{.v}}{n})^2}{1-\sum_{v=1}^{q}(\frac{n_{.v}}{n})^2}.$$

(consultar Goodman & Kruskal, 1979 ou Marcotorchino, 1984)

- $0 \le \tau_b \le 1$.
- $\tau_b = 0$ em caso de independência entre as duas variáveis, e
- $\tau_b = 1$ no caso de associação completa, isto é, quando as variáveis u e v são idênticas.

(Lauro e D'Ambra, 1984) mostram as vantagens de τ_b em relação ao coeficiente $\frac{\chi^2}{n}$:

- τ_b, contrariamente a $\frac{\chi^2}{n}$, está limitado superiormente;
- τ_b é menos sensível do que $\frac{\chi^2}{n}$ quando as margens de v são muito assimétricas;
- τ_b aumenta, contrariamente a $\frac{\chi^2}{n}$,quando a variabilidade de v diminui.
- $\frac{\chi^2}{n}$ é muito sensível quando as frequências teóricas são pequenas.

(Lauro et D'Ambra, 1984) consideram portanto que τ_b é um bom coeficiente de predicção. Isto interessa-nos pois, no nosso trabalho, o objectivo é o de predizer a variável classe

$$c : O \rightarrow \{c_1, c_2, \ldots, c_k, \ldots, c_K\}.$$

Na realidade o que nos interessa é o numerador de τ_b, uma vez que o denominador é constante. Com efeito, o nosso objectivo é o de escolher uma variável

binária para segmentar um nó da árvore de decisão. Para isso, comparamos cada uma das variáveis binárias com a variável c. Como o denominador de τ_b só depende de c, ele permanece constante. O numerador de τ_b, multiplicado por 2, constitui o critério de Haldane, Light e Margolin (Light & Margolin, 1971; Haldane, 1940 (citado em (Kendall & Stuart, 1961)).

A expressão do numerador de τ_b no nosso caso ($p=2$ e $q=K$), num nó t da árvore de decisão, reduz-se a:

$$\sum_{i=1}^{2}\sum_{j=1}^{K}\frac{\left(n_{ij}^{t}\right)^2}{n^t n_{i.}^{t}}-\sum_{j=1}^{K}\left(\frac{n_{.j}^{t}}{n^t}\right)^2=\sum_{i=1}^{2}\sum_{j=1}^{K}\left(\frac{n_{ij}^{t}}{n_{i.}^{t}}\right)^2\left(\frac{n_{i.}^{t}}{n^t}\right)-\sum_{j=1}^{K}\left(\frac{n_{.j}^{t}}{n^t}\right)^2=$$

$$\sum_{j=1}^{K}\left(\frac{n_{1j}^{t}}{n_{1.}^{t}}\right)^2\left(\frac{n_{1.}^{t}}{n^t}\right)+\sum_{j=1}^{K}\left(\frac{n_{2j}^{t}}{n_{2.}^{t}}\right)^2\left(\frac{n_{2.}^{t}}{n^t}\right)-\sum_{j=1}^{K}\left(\frac{n_{.j}^{t}}{n^t}\right)^2=$$

$$p_l^t\sum_{j=1}^{K}\left(p_j^{t_l}\right)^2+p_r^t\sum_{j=1}^{K}\left(p_j^{t_r}\right)^2-\sum_{j=1}^{K}\left(p_j^{t}\right)^2 \qquad (7)$$

Este coeficiente é exactamente o coeficiente de Gini (cf. 3). Demonstramos assim que, para árvores de decisão binárias, os coeficientes de Gini e de Goodman-Kruskal, Haldane, Light & Margolin são equivalentes. A demonstração generaliza-se facilmente ao caso de árvores não binárias de decisão.

REFERÊNCIAS BIBLIOGRÁFICAS

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, A., and STONE, C.J. (1984). "*Classification and Regression Trees*". Wadsworth, Belmont.

COSTA, J.F.P., (1996) "*Coefficients d'association et binarisation par la classification hiérarchique dans les arbres de décision. Application à l'identification de la structure secondaire des protéines*", (1996). Thèse de doctorat, Université de Rennes I, Rennes, France.

FISHER, W.D., (1958). "On grouping for maximum homogeneity". *J. Am. Statist. Assoc.* 53, pp.789-798.

GOODMAN, L.A., and KRUSKAL, W.H., (1954). "Measures of association for cross classifications". *Journal of the American Statistical Association*, vol. 49, pp. 732-764.

GOODMAN, L.A., and KRUSKAL, W.H., (1963). "Measures of association for cross classifications III: Approximate sampling theory". *Journal of the American Statistical Association*, vol. 58, pp. 310-364.

GOODMAN, L.A., and KRUSKAL, W.H., (1979) "*Measures of Association for Cross Classification*". Springer Verlag, Berlin, New-York.

KENDALL, M.G., and STUART, A., (1961). " *The Advanced Theory of Statistics*", Vol 2, Griffin, Londres.

LAURO, N. and D'AMBRA, L., (1984). "Analyse non Symétrique des correspondances". *Proceedings of the third Congress "International Data Analysis and Informatics"*, North Holland, Amsterdam.

LERMAN, I.C., (1992a). "Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles I". *Rev. Math. Infor. & Sci. Hum.*, 30e année, Paris, n. 118, pp. 35-52.

LERMAN, I.C., (1992b). "Conception et analyse de la forme limite d'une famille de coefficients statistiques d'association entre variables relationnelles II". *Rev. Math. Infor. & Sci. Hum.*, 30e année, Paris, n. 119, pp. 75-100.

LERMAN, I. C. and COSTA, J. F., (1996). "Coefficients d'association et variables à très grand nombre de catégories dans les arbres de décision. Application à l'identification de la structure secondaire d'une protéine". *Publication Interne Irisa n° 982 et Rapport de Recherche Inria n° 2803, 46 pages.*

LIGHT, J.R., and MARGOLIN, B.H., (1971). "An Analysis of Variance for Categorical Data". *Journal of the American Statistical Association*, vol. 66, No 331.

MARCOTORCHINO, F., (1984). "Utilisation des Comparaisons par Paires en Statistique des Contingences. Partie I". *Etude du centre Scientifique IBM-France* No F 069, Février 1984.

NAKHAEIZADEH, G., (1994). "Interaction Between Machine Learning and Statistics, An Overview". *Proc. of the ECML 94*, 7th European Conference on ML, Catania, Sicily.

QUINLAN, J.R., (1986). "Induction of Decision Trees". *Machine Learning*, pp.81-106.

TAYLOR, C.C., (1994). "Distance-based Decision Trees". *Proc. of the ECML 94*, 7th European Conference on ML, Catania, Sicily.

VOLUME 2

2º QUADRIMESTRE DE 1998

REVISTA DE ESTATÍSTICA
35ª PAGINA
2º QUADRIMESTRE DE 1998

VARIABILIDADE ESPACIAL NO MODELO LINEAR GERAL MISTO

Autores:
José António Pinheiro
Pedro Coelho

VOLUME 2
2º QUADRIMESTRE DE 1998

REVISTA DE ESTATÍSTICA
37ª PAGINA
2º QUADRIMESTRE DE 1998

VARIABILIDADE ESPACIAL NO MODELO LINEAR GERAL MISTO

SPATIAL VARIABILITY IN THE GENERAL LINEAR MIXED MODEL

Autores: José António Pinheiro
- Professor Auxiliar Convidado na Faculdade de Economia da Universidade Nova de Lisboa

e

Pedro Coelho
- Professor Auxiliar Convidado no Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa

RESUMO:

- A variabilidade espacial, refere-se à variação entre observações no espaço, cujos padrões poderão ser objecto de modelização e estimação estatística. São, neste artigo, analisados vários modelos de covariância espacial, bem como ferramentas de utilização habitual no âmbito da estatística espacial, como o variograma e o semivariograma. Propõe-se a representação da variabilidade espacial no quadro do modelo linear geral misto, através de estruturas de covariância adequadas. Aborda-se ainda teoria necessária à estimação dos parâmetros e à realização de inferências em vários espaços. É apresentada uma aplicação ao estudo de uma função consumo para um conjunto de dados que exibem correlação espacial entre conjuntos de observações. Conclui-se que a abordagem seguida fornece não só uma forma de caracterizar as estruturas de correlação espacial, mas também de incorporar mais informação em modelos de ajustamento com o propósito melhorar a eficiência da estimação ou previsão.

PALAVRAS CHAVE:

- *modelo misto, BLUP, espaço de inferência, correlação espacial, semivariograma, covariograma*

ABSTRACT:

- Spatial variability refers to the variation between observations in the space, hose patterns may be statistically modelled and estimated. In this article, we analyse several models for spatial covariance, as well as some well none tools for spatial statistics, namely the covariogram and the semivariogram. We propose representing spatial variability in the frame of the general linear mixed model, using adequate covariance structures. We also approach theory concerning parameter estimation and inference under several spaces. We present a case concerning the study of a consumption function using a data set exhibiting spatial correlation between groups of observations. We conclude that the followed

approach offers not only a way of characterising spatial correlation structures, but also to bring more information into an adjustment model in order to improve the estimation or prediction efficiency.

KEY-WORDS:

- *mixed model, BLUP, predictable function, spatial correlation, semivariogram, covariogram*

REVISTA DE ESTATÍSTICA
39ª PAGINA
2º QUADRIMESTRE DE 1998

1 INTRODUÇÃO

A variabilidade espacial, refere-se à variação entre observações no espaço. Esta variação pode exibir determinados padrões passíveis de serem modelizados e estimados estatisticamente. Há dois grandes campos de aplicação: os métodos aplicados a localizações aleatórias e os métodos que se aplicam a localizações fixas, sobre as quais se faz, por exemplo, amostragem.

Para estes últimos é de particular interesse a correlação espacial positiva ou dependência espacial positiva em pequena escala, i.e., a tendência que as observações espacialmente próximas têm para serem mais "parecidas" do que as espacialmente distantes.

A variabilidade espacial pode ser conceptualizada num quadro de extensão multidimensional de medidas repetidas (múltiplas medidas de uma variável resposta tomadas sobre uma mesma unidade).

Os métodos estatísticos que abordam a correlação espacial podem ser divididos em duas classes: caracterização e ajustamento. A caracterização envolve essencialmente a estimação dos parâmetros de covariância. O ajustamento tem por objectivo a acomodação dos efeitos da correlação espacial a fim de obter estimativas mais precisas.

2 UM MODELO ESPACIAL: DEFINIÇÃO E GENERALIDADES

Seja $y(l)$ uma variável de interesse cujo argumento é um ponto de um domínio $D \subseteq \Re^n$. Seja $y(l_i)$ o valor que y toma no local l_i ; sem perda de generalidade usaremos a notação $y(l_i) = y_i$.

Considere-se o seguinte modelo de correlação espacial:

$$y_i = \mu + \varepsilon_i \qquad (1)$$

y_i é a i-ésima observaçãoe e ε_i o erro respectivo cujo valor esperado é nulo. De (1) vem

$$E(y_{l_i}) = E(y_{l_j}) = \mu \quad \forall l_i, l_j \in D \tag{2}$$

ou, numa notação que nos interessa para o futuro,

$$E\left(y(l_i + \vec{h})\right) = E\left(y(l_i)\right) \tag{3}$$

onde $\vec{h}$ é um vector de norma h, sendo pois h a distância euclideana entre l_i e l_j. Sendo, por exemplo em $\Re^2$, l_i o local físico onde se encontra y_i, este pode ser especificado por duas coordenadas , e.g. latitude e longitude.

Sejam as características aleatórias do modelo (1)

$$\begin{aligned} &Var(\varepsilon_{l_i}) = Var(y_{l_i}) = \sigma^2 = C(\vec{0}) \\ &Cov(\varepsilon_{l_i}, \varepsilon_{l_j}) = Cov(y_{l_i}, y_{l_j}) = \sigma_{ij} = \sigma^2 f(l_i - l_j) = C(l_i - l_j) \end{aligned} \tag{4}$$

A fórmula (4) determina que a covariância entre ε_{l_i} e ε_{l_j} depende apenas do vector $l_i - l_j$. A função $C(.)$ designa-se por covariograma ou função de covariância estacionária.

Uma variável de interesse que verifique (3) e (4) diz-se *estacionária de segunda ordem, fracamente estacionária* ou simplesmente *estacionária.* Se além disso $C(l_i - l_j)$ for função apenas de $\|l_i - l_j\|$, a variável de interesse diz-se *isotrópica*, designação que se pode aplicar também à função $C(.)$. Neste caso a função depende apenas da distância entre dois locais e não da posição relativa desses locais. Sendo a covariância uma função da distância entre dois locais l_i e l_j, pode ser escrita como

$$Cov(\varepsilon_{l_i}, \varepsilon_{l_j}) = \sigma^2 [f(h)] = C\left(\|l_i - l_j\|\right) \tag{5}$$

sendo h a distância referida.

Com esta notação os modelos **estacionários** são aqueles onde o valor $f(.)$ é idêntico para todos os pares l_i, l_j tais que $l_i - l_j = \vec{h}$. Os modelos **isotrópicos** são aqueles onde o valor $f(.)$ é idêntico para todos os pares l_i, l_j tais que $\|l_i - l_j\| = h$.

Definem-se ainda os modelos **intrinsecamente estacionários**. São aqueles que verificam (3) e onde ainda

$$Var(y_i - y_j) = 2\gamma(l_i - l_j) \tag{6}$$

ou seja, onde a variância no membro esquerdo pode ser escrita apenas como função do vector $l_i - l_j$.

Note-se que a classe de processos estacionários está estritamente contida na classe de todos os processos estritamente estacionários. Existem, no entanto, modelos onde $2\gamma(.)$ está definido mas $C(.)$ não está. Um exemplo é apresentado em Cressie (1991, pag 68).

Estabelece-se assim a relação

Isotrópicos $\subset$ *Estacionários* $\subset$ *Intrinsemente estacionários*

3 O SEMIVARIOGRAMA

3.1 DEFINIÇÃO

Uma ferramenta importante para a análise dos modelos intrinsecamente estacionários é o semivariograma. É uma medida estatística da variabilidade espacial baseada na distância entre as observações.

Define-se

$$\begin{aligned} &Var\left[y\left(l+\vec{h}\right)-y(l)\right]=2\gamma(\vec{h}) \\ &ou \\ &Var\left[y(l_i)-y(l_j)\right]=2\gamma(l_i-l_j) \end{aligned} \tag{7}$$

sendo a função $2\gamma(\vec{h})$ designada por variograma (Matheron, 1962). A função $\gamma(\vec{h})$ é designada por semivariograma.

Da definição verifica-se que $\gamma(\vec{h}) = \gamma(-\vec{h})$ (imediato a partir da segunda fórmula em 7) e $\gamma(\vec{0}) = 0$. Alarguemos esta definição permitindo que $\gamma(\vec{h}) \to c_0 > 0$, quando $\vec{h} \to \vec{0}$. O valor c_0 costuma designar-se por **efeito de pepita** (*nugget effect*). Formalmente isto significa que se admite a existência de variabilidade à microescala que induz uma descontinuidade na origem do semivariograma. Tal não pode formalmente acontecer em processos estacionários nos quais se verifica a propriedade

$$E\left(y(l+\vec{h}) - y(l)\right)^2 \underset{h\to 0}{\to} 0. \tag{8}$$

Caso o fenómeno em observação seja contínuo à micro-escala, a única razão pela qual $c_0 > 0$ será a existência de erros de observação. A introdução deste efeito de pepita no modelo pode ser matematicamente consubstanciado por um ruído branco que se adiciona ao processo (1):

$$y_t = \mu + \varepsilon_t + \xi_t = \mu + e_t \tag{9}$$

onde

$$\begin{aligned} &Var(\varepsilon_t) = \sigma^2 \\ &Var(\xi_t) = c_0 \\ &Var(e_t) = Var(y_t) = \sigma^2 + c_0 = C(\vec{0}) \\ &Cov(e_t, e_j) = Cov(\varepsilon_t, \varepsilon_j) = Cov(y_t, y_j) = \sigma_{ij} = \sigma^2 f(l_t - l_j) = C(l_t - l_j) \end{aligned} \tag{10}$$

A função f é tal que $\lim_{h\to 0} f = 1$. De facto, quando $\vec{h} \to \vec{0}$, i.e., se l_t e l_j estão muito próximos, tem-se que a covariância tende para o seu valor máximo $C(l_t - l_j) \to \sigma^2$.

Sendo
$2\gamma(\vec{h}) = 2\gamma(l_i - l_j) = Var(y_i - y_j) = Var(y_i) + Var(y_j) - 2Cov(y_i, y_j)$,
verifica-se que

$$\lim_{\vec{h}\to\vec{0}} 2\gamma(\vec{h}) = \lim_{l_i \to l_j} Var(y_i - y_j) = \sigma^2 + c_0 + \sigma^2 + c_0 - 2\sigma^2 = 2c_0 \quad (11)$$

Donde, enfim

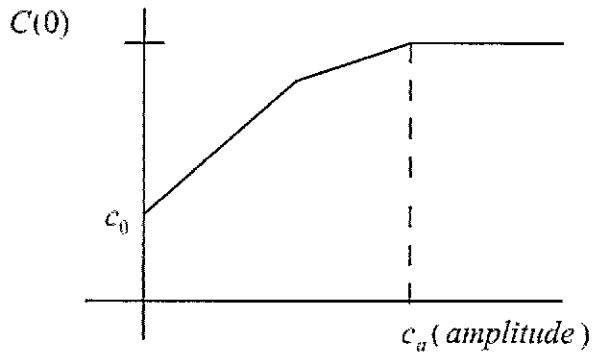
$$\lim_{\vec{h}\to\vec{0}} \gamma(\vec{h}) = c_0 \quad (12)$$

A determinação de c_0 é difícil a partir de dados cuja distância seja excessivamente grande para darem informação à microescala. Costuma ser empiricamente aproximada a partir da extrapolação no semivariograma estimado para distâncias próximas de zero.

O mais pequeno valor de $\left\|\vec{h}_0\right\|$ tal que $\gamma(\vec{h}_0.(1+\delta)) = C(\vec{0})$, com δ uma qualquer constante positiva, designa-se por **amplitude** (c_a) do semivariograma na direcção dada pelo versor $\dfrac{\vec{h}_0}{\left\|\vec{h}_0\right\|}$. Nos termos de (10), c_a pode ainda ser visto como o mais pequeno valor de $\left\|\vec{h}_0\right\|$, na direcção dada pelo versor $\dfrac{\vec{h}_0}{\left\|\vec{h}_0\right\|}$, para o qual $f(l_i - l_j) = f(\vec{h}_0) = 0$ ou, de forma equivalente, $C(\vec{h}_0) = 0$.

Em modelos isotrópicos toda a notação acima utilizada pode-se declinar para argumentos não vectoriais. Falar-se-á de $\gamma(h)$, de $C(h)$, etc. Todos os conceitos permanecem ainda que se tornem necessárias adaptações de notação cuja compreensão é imediata. Por exemplo $C(\vec{h})$ é uma função real de variável vectorial, enquanto $C(h)$ será uma função real de variável real, devendo em boa verdade escrever-se $C^*(h)$ para salientar essa diferença.

Estas referências permitem esboçar o seguinte andamento geral de um semivariograma:



3.3 ALGUMAS RELAÇÕES ENTRE O COVARIOGRAMA E O VARIOGRAMA

Recorde-se a relação

$$2\gamma(\bar{h}) = 2\gamma(l_i - l_j) = Var(y_i - y_j) = Var(y_i) + Var(y_j) - 2Cov(y_i, y_j) \quad (13)$$

Constata-se que:

- um modelo que possua variograma e variâncias finitas induz um modelo estacionário
- um modelo estacionário com variâncias finitas induz um modelo que possui variograma, com a forma

$$Var(y_i - y_j) = 2\{C(\bar{0}) - C(l_i - l_j)\}, \quad (14)$$

ou, de forma equivalente

$$2\gamma(\bar{h}) = 2\{C(\bar{0}) - C(\bar{h})\}. \quad (15)$$

Rescrevendo (15) da seguinte forma

$$\gamma(\bar{h}) = C(\bar{0}) - C(\bar{h}) \quad (16)$$

constata-se que, quando $\|\vec{h}\| \geq c_a$, $C(\vec{h}) = 0$ (o que em correlação espacial quer dizer que à medida que os locais se afastam a ligação entre os valores da variável de interesse diminui) e $\gamma(\vec{h}) = C(\vec{0}) = \sigma^2 + c_0$. Por este motivo a quantidade $C(\vec{0})$ designa-se por **limiar** do semivariograma. Na mesma linha de raciocínio a quantidade $c_s = \sigma^2 = C(\vec{0}) - c_0$ é o **limiar parcial**, sendo c_0 o já referido efeito de pepita.

Note-se ainda que

- se $2\gamma(.)$ é uma função contínua na origem, então $y(l)$ é uma função contínua.
- se $2\gamma(.)$ não tende para zero quando $\vec{h} \to \vec{0}$, então $y(l)$ não é par nem contínua. Esta descontinuidade de γ na origem constitui o referido efeito de pepita.
- se $2\gamma(.)$ é uma constante positiva (excepto na origem onde será zero) então $y(l_i)$ e $y(l_j)$ são não correlacionados para todos os pares $l_i \neq l_j$, independentemente da sua proximidade; neste caso $y(l)$ é um ruído branco.

De (16) pode escrever-se

$$C(\vec{h}) = C(\vec{0}) - \gamma(\vec{h}) \tag{17}$$

Desta fórmula e de (4) teremos sucessivamente

$$\gamma(\vec{h}) = C(\vec{0}) - C(\vec{h}) = C(\vec{0}) - \sigma^2\left[f(\vec{h})\right] = c_0 + \sigma^2\left[1 - f(\vec{h})\right] \tag{18}$$

$$C(\vec{h}) = C(\vec{0}) - \gamma(\vec{h}) = C(\vec{0}) - \left\{c_0 + \sigma^2\left[1 - f(\vec{h})\right]\right\} = \sigma^2 f(\vec{h}) \tag{19}$$

Esta fórmula é particularmente útil no que relaciona a covariância espacial com a função de distância.

Para o caso em que não exista efeito de pepita, isto é $\gamma(\vec{0}) = 0$ e $C(\vec{0}) = \sigma^2$, as fórmulas (18) e (19) vêm

$$\gamma(\vec{h}) = C(\vec{0}) - C(\vec{h}) = C(\vec{0}) - \sigma^2\left[f(\vec{h})\right] = \sigma^2\left[1 - f(\vec{h})\right] \qquad (20)$$

e

$$C(\vec{h}) = C(\vec{0}) - \gamma(\vec{h}) = C(\vec{0}) - \left\{\sigma^2\left[1 - f(\vec{h})\right]\right\} = \sigma^2 f(\vec{h}). \qquad (21)$$

Como se disse atrás, a $C(\vec{h})$, covariancia entre duas observações separadas por $\vec{h}$, também se costuma chamar covariograma. Desde que $C(\vec{0}) > 0$, define-se

$$\rho(\vec{h}) = \frac{C(\vec{h})}{C(\vec{0})} \qquad (22)$$

que se designa por correlograma (equivalente do que em sucessões cronológicas se chama de função de autocorrelação).

As propriedades seguintes são imediatas:

$$\begin{aligned} C(\vec{h}) &= C(-\vec{h}) \\ \rho(\vec{h}) &= \rho(-\vec{h}) \\ \rho(\vec{0}) &= 1 \end{aligned} \qquad (23)$$

3.4 ALGUNS MODELOS PARA SEMIVARIOGRAMAS

3.4.1 SEMIVARIOGRAMAS ISOTRÓPICOS

São agora apresentados alguns semivariogramas correntemente utilizados e as respectivas funções de distância, válidos no espaço $\Re^n$. As parametrizações acomodam-se aos casos de presença e ausência de efeito de pepita, integrando também os restantes parâmetros.

3.4.1.1 MODELO LINEAR

A função de distância é

$$f(h)=\left(1-\frac{h}{c_a}\right)I(h<c_a) \tag{24}$$

onde $I(h<c_a)$ é uma função indicatriz.

Desta função vem o semivariograma

$$\begin{aligned}\gamma(h;c_o,c_a,\sigma^2) &= \sigma^2+c_0+\sigma^2(h/c_a-1)I(h<c_a)\\ &= c_0+\sigma^2\left[1+(h/c_a-1)I(h<c_a)\right]\end{aligned} \tag{25}$$

$c_0\geq 0,\sigma^2\geq 0,c_a\geq 0$.

3.4.1.2 MODELO ESFÉRICO

A função de distância é

$$f(h)=\left[1-\frac{3h}{2c_a}+\frac{1}{2}\left(\frac{h}{c_a}\right)^3\right]I(h<c_a) \tag{26}$$

Desta função vem o seguinte semivariograma

$$\begin{aligned}\gamma(h;c_o,c_a,\sigma^2) &= \sigma^2+c_0-\sigma^2\left[1-\frac{3h}{2c_a}+\frac{1}{2}\left(\frac{h}{c_a}\right)^3\right]I(h<c_a)\\ &= c_0+\sigma^2\left\{1-\left[1-\frac{3h}{2c_a}+\frac{1}{2}\left(\frac{h}{c_a}\right)^3\right]I(h<c_a)\right\}\end{aligned} \tag{27}$$

$c_0\geq 0,\sigma^2\geq 0,c_a\geq 0$.

3.4.1.3 MODELO EXPONENCIAL (FORMA 1)

A função de distância é

$$f(h)=\left[exp\left(-\frac{h}{c_e}\right)\right] \tag{28}$$

Desta função vem o seguinte semivariograma

$$\gamma(h;c_o,c_a,\sigma^2)=c_0+\sigma^2\left[1-exp\left(-\frac{h}{c_e}\right)\right] \tag{29}$$

$c_0 \geq 0, \sigma^2 \geq 0, c_e \geq 0$.

Neste caso, c_e é um parâmetro que controla a forma da função. A amplitude nos moldes anteriores é infinita já que $\lim_{h\to\infty}\gamma(h)=C(0)=\sigma^2+c_0$.

3.4.1.4 MODELO EXPONENCIAL (FORMA 2)

A função distância é

$$f(h)=\rho^h \tag{30}$$

Desta função vem o seguinte semivariograma

$$\gamma(h;c_0,\sigma^2,\rho)=c_0+\sigma^2\left[1-\rho^h\right] \tag{31}$$

$c_0 \geq 0, \sigma^2 \geq 0, 0 \leq \rho \leq 1$.

As duas formas da função exponencial são equivalentes pois pode escrever-se $\left[\exp\left(-\frac{h}{c_e}\right)\right] = \left[\exp\left(-\frac{1}{c_e}\right)\right]^h = \rho^h$, com $\exp\left(-\frac{1}{c_e}\right) = \rho$.

3.4.1.5 MODELO GAUSSIANO

A função distância é

$$f(h) = exp\left(-\frac{h^2}{c_g^2}\right) \tag{32}$$

Desta função vem o seguinte semivariograma

$$\gamma(h; c_0, \sigma^2, c_g) = c_0 + \sigma^2\left[1 - \exp\left(-\frac{h^2}{c_g^2}\right)\right] \tag{33}$$

$c_0 \geq 0, \sigma^2 \geq 0, c_g \geq 0$.

Mantém-se válida a observação sobre a amplitude que foi efectuada para o modelo exponencial.

3.4.1.6 MODELO LOG-LINEAR

A função distância é

$$f(h) = \left(1 - \frac{1}{c_a} log\, h\right) I(log\, h < c_a) \tag{34}$$

Desta função vem o seguinte semivariograma

$$\gamma(h;\sigma^2,c_0,c_a)=c_0+\sigma^2\left[1+\left(\frac{1}{c_a}\log h-1\right)I(\log h<c_a)\right] \quad (35)$$

$c_0\geq 0,\sigma^2\geq 0,c_a\geq 0$.

3.4.2 SEMIVARIOGRAMAS ESTACIONÁRIOS ANISOTRÓPICOS

Quando um processo é não isotrópico, adiante designado por **anisotrópico**, a covariância entre $y(l_i)$ e $y(l_i+\vec{h})$ é uma função quer da magnitude quer da direcção de $\vec{h}$. Neste caso, o variograma deixa igualmente de ser uma função que depende unicamente da distância entre as duas localizações espaciais.

Os fenómenos que exibem estas características são normalmente causados por um processo físico que envolva a diferenciação no espaço. Por exemplo, um processo na direcção vertical (eg. altitude) será tipicamente diferente de outro que se estabeleça nas direcções horizontais (eg. latitude e longitude). Outros fenómenos poderão igualmente exibir diferentes processos segundo várias direcções horizontais.

Em alguns casos, a anisotropia pode ser corrigida através de transformações lineares do vector $\vec{h}$, vindo

$$\gamma(\vec{h})=\gamma^*\left(\left\|A\vec{h}\right\|\right) \quad \vec{h}\in\Re^n \quad (36)$$

onde A é uma matriz $n\times n$ e γ^* é uma função de variável real.

Neste caso, o espaço Euclidiano não é considerado apropriado para medir a distância entre localizações, mas uma sua transformação linear é.

3.4.2.1 MODELO EXPONENCIAL ANISOTRÓPICO

Para representar um processo anisotrópico pode ser usada uma generalização do modelo exponencial (que inclui como caso particular o modelo gaussiano).

Para esse efeito, recorde-se o modelo exponencial isotrópico cuja função distância é

$$f(h)=\left[exp\left(-\frac{h}{c_a}\right)\right]. \quad (37)$$

A sua generalização a processos anisotrópicos poder-se-á fazer através de um processo multiplicativo de um conjunto de funções distância exponenciais, uma para cada uma das n direcções definidas em $\Re^n$.

A função f vem então

$$f(\vec{h}) = \prod_{k=1}^{n} \exp\left(-\theta_k d(i,j,k)^{p_k}\right) = \exp\left(-\sum_{k=1}^{n} \theta_k d(i,j,k)^{p_k}\right) \quad (38)$$

onde n é o número de coordenadas e $d(i,j,k)$ é a distância absoluta entre a k-ésima coordenada da i-ésima e da j-ésima observações.

Quando se tem $p_k = 2; \quad k = 1,\ldots,n$ vem

$$f(\vec{h}) = f^*\left(\left\|A\vec{h}\right\|\right) = \exp\left(-\sum_{k=1}^{n} \theta_k d(i,j,k)^2\right) \quad (39)$$

correspondendo a um modelo gaussiano isotrópico para a transformação $A\vec{h}$, onde $A = diag_{1\le k\le n}\left\{\sqrt{\theta_k}\right\}$.

Quando se impõe $p_k = 2, \theta_k = \theta; \quad k = 1,\ldots,n$ vem

$$f(\vec{h}) = f^*(h) = \exp\left(-\theta\sum_{k=1}^{n} d(i,j,k)^2\right) = \exp\left(-\theta\, d(i,j)^2\right) \quad (40)$$

reduzindo-se a um modelo isotrópico gaussiano.

Desta função vem o seguinte semivariograma

$$\gamma(\vec{h}; c_o, \sigma^2, \theta, p) = c_0 + \sigma^2\left[1 - \exp\left(-\sum_{k=1}^{n} \theta_k d(i,j,k)^{p_k}\right)\right] \quad (41)$$

$c_0 \geq 0, \sigma^2 \geq 0, \theta_k \geq 0, p_k \geq 0 \quad k = 1, \ldots, n$.

3.5 ESTIMAÇÃO DO SEMIVARIOGRAMA

O estimador natural do semivariograma baseado no método dos momentos e devido a Matheron (1962) é

$$\hat{\gamma}(\vec{h}) = \frac{1}{2\left|N(\vec{h})\right|} \sum_{N(\vec{h})} \left[y_i - y_j\right]^2 \tag{42}$$

onde $N(\vec{h}) \equiv \left\{(i,j) : l_i - l_j = \vec{h}\right\}$ e $\left|N(\vec{h})\right|$ é o número de elementos distintos do conjunto $N(\vec{h})$.

Note-se que $N(-\vec{h}) \neq N(\vec{h})$ apesar de que $\left|N(h)\right| = \left|N(-h)\right|$ e $\hat{\gamma}(-h) = \hat{\gamma}(h)$, propriedade amostral equivalente a $\gamma(\vec{h}) = \gamma(-\vec{h})$.

Como se verá mais adiante, no âmbito deste estudo, o semivariograma estimado tem como principal utilidade fornecer uma primeira aproximação aos parâmetros de covariância de modelos lineares mistos. Tem-se verificado empiricamente que a utilização destes valores iniciais poderá refinar e acelerar o processo numérico de optimização das respectivas funções de verosimilhança, evitando, nomeadamente, a convergência a extremos locais.

4 INTEGRAÇÃO DA VARIABILIDADE ESPACIAL NO MODELO LINEAR MISTO

Uma vez analisada a formalização da variabilidade espacial, vê-se agora como esta pode ser considerada no âmbito do modelo linear geral misto.

No modelo misto

$$y = X\beta + Zu + \varepsilon. \tag{43}$$

onde

$$E\begin{bmatrix}\varepsilon \\ u\end{bmatrix} = \begin{bmatrix}0 \\ 0\end{bmatrix}$$
$$Var\begin{bmatrix}\varepsilon \\ u\end{bmatrix} = \begin{bmatrix}R & 0 \\ 0 & G\end{bmatrix} \qquad (44)$$

a correlação espacial pode reflectir-se em G (matriz de variâncias e covariâncias do vector aleatório u) ou em R (matriz de variâncias e covariâncias do vector de erros ε).

A estimação do modelo (43) integra-se na estimação geral dos modelos mistos, contemplando alguns aspectos particulares na parametrização. Os parâmetros de covariância especificam aqui as características dos modelos espaciais.

Através da selecção de uma das funções de distância apresentadas (cf. 3.4), ou outras, especificam-se matrizes

$$Var(\varepsilon) = R = \sigma^2 F_\varepsilon$$
$$Var(u) = G = \sigma_u^2 F_u \qquad (45)$$

onde F_ε e F_u são matrizes de dimensões nxn e gxg, respectivamente, sendo g o número de efeitos aleatórios. Os seus elementos genéricos (ij) são calculados por funções $f(\vec{h})$, como as apresentadas, sendo $\vec{h} = l_i - l_j$. Se o modelo for isotrópico os elementos das matrizes F_ε e F_u serão formados a partir de funções $f(h)$.

Caso haja efeito de pepita, as matrizes terão a forma

$$Var(\varepsilon) = R = c_{0\varepsilon} I + \sigma_\varepsilon^2 F$$
$$Var(u) = G = c_{0u} I + \sigma_u^2 F \qquad (46)$$

Este procedimento corresponde à formulação de uma estrutura de variâncias e covariâncias parametrizada nos parâmetros de covariâncias clássicos (neste caso σ^2 e c_0) e ainda nos parâmetros das funções de distância.

Uma vez obtidas estimativas para estes parâmetros de covariância, por exemplo através de métodos de verosimilhança, torna-se então possível estimar/prever os

efeitos fixos e aleatórios β e u, que resultam da solução das equações do modelo misto (ver Coelho e Pinheiro, 1997).

Saliente-se que se pode estabelecer uma relação funcional entre as matrizes R e G e o semivariograma. Tome-se por exemplo R. Para uma estrutura de covariância espacial sem efeito de pepita, onde $Var(\varepsilon) = R = \sigma_\varepsilon^2 F_\varepsilon$ e $\gamma(\vec{h}) = \sigma_\varepsilon^2 - C(\vec{h})$, tem-se $Var(\varepsilon) = \left[C(\vec{h}) + \gamma(\vec{h})\right]F$. Para uma estrutura com efeito de pepita, onde $Var(\varepsilon) = R = c_{0\varepsilon} I + \sigma_\varepsilon^2 F$ e $\gamma(\vec{h}) = c_0 + \sigma_\varepsilon^2 - C(\vec{h})$, tem-se $Var(\varepsilon) = \left[C(\vec{h}) + \gamma(\vec{h})\right]F + c_0(I - F)$.

4.1 ESTIMADORES E PREVISORES

A estrutura do modelo misto (43) e (44), na qual o modelo espacial se integra como caso particular, obriga à introdução de um novo conceito que não cabe no tradicional conceito de estimador. Consagrado com o acrónimo BLUP, é o ***Best Linear Unbiased Predictor***, Melhor Previsor Linear Centrado. De facto nesta nova estrutura há que ***prever*** valores do vector aleatório u além de ***estimar*** valores do vector β.

As expressões seguintes esclarecem as diferenças na conceptualização:

$$\begin{aligned} E(\hat{\beta}) &= \beta \\ E(\hat{u}) &= E(u) = 0 \end{aligned} \qquad (47)$$

A média não condicionada de y é $E(y) = X\beta$ e pode ser compreendida como a média de y em toda a população. A média de y condicionada por u é $E(y \mid u) = X\beta + Zu$ e pode ser compreendida como a média de y respeitante a um **conjunto de efeitos aleatórios na amostra**. Como este média condicionada depende dos valores realizados por u, as duas médias, em geral, diferem.

Seja $K'\beta$ uma combinação linear de efeitos fixos. Diz-se que $K'\beta$ é uma **função estimável** se puder ser obtida como combinação linear **de médias não condicionadas** de observações: ou seja se $K'\beta = T'\left[E(y)\right] = T'X\beta$ para algum T'. Estas funções estimáveis não dependem dos efeitos aleatórios.

Uma generalização do conceito de funções estimáveis surge nas combinações lineares que englobam β e u, seja $K'\beta + M'u$. Trata-se de uma combinação linear de médias condicionadas, designada por **função predictível**. Uma função como $K'\beta + M'u$ é **predictível** se $K'\beta$ for **estimável**.

A teoria dos modelos mistos dá resposta às duas situações: estimando valores para β permite construir um estimador BLUE para $K'\beta$; *estimando/prevendo* valores para β e para u permite construir o BLUP para $K'\beta + M'u$.

Uma nova terminologia estatística tem vindo a ser necessária para o estudo de estimadores e previsores, de acordo com os alvos de inferência em estudo. Essa terminologia estrutura-se em torno de dois pólos que tendem a cristalizar diversas propostas:

- **espaço de inferência largo** e **espaço de inferência restrito** (McLean, Sanders e Stroup, 1991)
- **inferência na população** e **inferência especifica no indivíduo** (Zeger et al., 1988)

De uma forma muito simplificada, **espaço de inferência largo** e **inferência na população** referem-se à inferência sobre efeitos fixos e funções estimáveis; **espaço de inferência restrito** e **inferência especifica no indivíduo** referem-se a inferência sobre funções predictíveis.

Os estimadores de $\hat{\beta}$ e de $\hat{u}$ (por facilidade de linguagem falaremos com frequência de estimadores, sabendo que para u é válido o conceito de previsor) são dados pelas fórmulas

$$\begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} (X'V^{-1}X)^{-1}X'V^{-1}y \\ GZ'V^{-1}(y - X\hat{\beta}) \end{bmatrix}. \tag{48}$$

onde

$$V = Var(y) = ZGZ' + R. \tag{49}$$

A matriz de variâncias e covariâncias dos erros de previsão de β e de u é

$$C = E\left\{ \begin{bmatrix} \hat{\beta} - \beta \\ \hat{u} - u \end{bmatrix} \begin{bmatrix} \hat{\beta} - \beta \\ \hat{u} - u \end{bmatrix}' \right\} = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ X'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix}^{-1} \tag{50}$$

4.2 ESTIMAÇÃO DAS COMPONENTES DE VARIÂNCIA

O conhecimento de V é indispensável para a obtenção dos estimadores (48). Por sua vez a estimação dos elementos de V (componentes de variância que são simbolicamente representadas num vector θ de dimensão apropriada) deve ser realizada em dois passos: o primeiro consiste em **formular** *uma estrutura de covariâncias* adequada à descrição dos fenómenos que se estudam; o segundo diz respeito à **estimação** dos elementos da matriz.

Uma possível abordagem para esta estimação recorre a métodos de verosimilhança que se baseiam nos pressupostos da normalidade de u e ε.

Dois métodos de verosimilhança estão consagrados: um designado por **máxima verosimilhança** e outro de **máxima verosimilhança restrita** ou **residual**, que se designam abreviadamente por ML e REML, respectivamente, de acordo com a tradição anglo-saxónica. Uma importante propriedade destes métodos consiste na sua capacidade de lidar com dados não equilibrados e acomodar dados omissos (Wolfinger, 1993). De facto, em abordagens tradicionais a existência de dados omissos obriga ao abandono de observações ou ao recurso a métodos de imputação por vezes complexos. Referências a estes métodos podem ser encontradas em Hartley e Rao (1967), Patterson e Thompson (1974), Harville (1977), Laird e Ware (1982) e Jennrich e Schluchter (1986).

4.3 INFERÊNCIA NO MODELO MISTO

4.3.1 INFERÊNCIA SOBRE OS PARÂMETROS DO MODELO

O ensaio de hipóteses sobre os parâmetros fixos e aleatórios pode ser formulado em termos matriciais

$$H_0: L \begin{bmatrix} \beta \\ u \end{bmatrix} = r \qquad (51)$$

onde L é uma matriz de dimensões compatíveis e elementos coerentes com as hipóteses formuladas.

Se L for uma matriz linha, consubstanciando uma combinação linear dos parâmetros, a estatística

$$\frac{L\begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} - r}{\sqrt{L\hat{C}L'}} \qquad (52)$$

tem, sob a normalidade de ε e u, distribuição aproximada t de Student (McLean e Sanders, 1988 e Stroup, 1989), sendo que os graus de liberdade (ν) devem ser estimados, usando, por exemplo, o método de Satterwaite (Giesbrecht e Burnes, 1985; McLean e Sanders, 1988 e Fay e Cornelius, 1993). O seu cálculo permite fazer testes de hipóteses e construir intervalos de confiança. Naturalmente que $\hat{C}$ é a matriz de variâncias e covariâncias estimadas de $\left[(\hat{\beta}-\beta),(\hat{u}-u)\right]'$.

Quando L tem característica superior a 1, a estatística

$$\frac{\left\{L\begin{bmatrix}\hat{\beta}\\ \hat{u}\end{bmatrix}-r\right\}'\left[(L\hat{C}L')^{-1}\right]\left\{L\begin{bmatrix}\hat{\beta}\\ \hat{u}\end{bmatrix}-r\right\}}{\rho(L)} \qquad (53)$$

segue aproximadamente uma distribuição $F(\rho(L),\nu)$, onde $\rho(L)$ representa a característica de L. Esta estatística, baseia-se na estatística de Wald, sendo $\hat{C}$ a matriz a matriz de variâncias-covariâncias estimadas dos erros de previsão dos parâmetros $\left[(\hat{\beta}-\beta)\quad(\hat{u}-u)\right]'$.

4.3.2 INFERÊNCIA SOBRE AS COMPONENTES DE VARIÂNCIA

Os estimadores das componentes de variância, sendo obtidos por métodos de verosimilhança, gozam das propriedades genéricas associadas: consistentes, assimptoticamente normais e assimptoticamente eficientes; as suas matrizes de variâncias-covariâncias são as inversas das respectivas matrizes de informação de Fisher, $I(\theta)=-E\left[\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}\right]$ que por sua vez coincidem com o limite inferior de Cramer-Rao.

A estes estimadores podem-se ainda aplicar os três testes clássicos, assimptoticamente equivalentes: o de Wald, o de Lagrange e o da razão de verosimilhanças. Os três testes podem apresentar resultados bastante diferentes em amostras pequenas, situação em que as suas propriedades são ainda mal conhecidas (Greene, 1997).

Estes procedimentos permitem testar situações gerais de hipóteses nulas como

$$H_0 : c(\theta)=q,$$

onde $c(\theta)$ é uma função dos parâmetros que estão a ser estimados.

4.4 SELECÇÃO DE MODELOS

Do que se expôs entende-se que existe grande flexibilidade na especificação do modelo misto, quer na forma de X e Z, quer na estrutura de G e de R.

O diagnóstico e a selecção de modelos, no âmbito de modelos lineares gerais mistos, podem ser levados a cabo através de um procedimento em 3 etapas, tal como o proposto por Diggle (1988) e estendido por Wolfinger (1993).

Primeira etapa

Compreende a realização de transformações apropriadas aos dados e selecção de efeitos fixos. Nesta fase são preferidos modelos sobreajustados relativamente a modelos subajustados a fim de evitar a possível introdução de correlações espúrias.

Segunda etapa

Consiste na selecção de estruturas de covariância iniciais para G e R, usando teoria científica relevante para o âmbito do estudo, bem como a análise de gráficos e semivariogramas de resíduos.

Uma vez efectuada uma primeira selecção de estruturas para G e R, os valores obtidos para os BLUP's $\hat{u}$ e para os resíduos $\hat{\varepsilon} = y - X\hat{\beta} - Z\hat{u}$ podem ser analisados a fim de a confirmar ou infirmar empiricamente.

A normalidade quer dos previsores $\hat{u}$, quer dos resíduos $\hat{\varepsilon}$ pode ainda ser formalmente ensaiada através de testes adequados, como o de Shapiro-Wilk ou de Kolmogorov.

Terceira etapa

Na última etapa utilizam-se técnicas estatísticas formais para comparar várias estruturas de covariâncias alternativas. Estas técnicas incluem:

- Construção de testes de Wald ou do multiplicador de Lagrange para ensaiar a significância dos parâmetros de covariância. Tendo, no entanto, em conta que as propriedades destas estatísticas para pequenas amostras são pouco conhecidas, será preferível, na comparação de modelos encaixados, a construção de testes de razão de verosimilhanças (a este respeito ver Jenrich e Schluchter, 1986; Schaalje et al., 1991).
- Maximização de critérios de informação REML, tais como

$$AIC_{REML} = l_{REML} - q.$$
$$HQC_{REML} = l_{REML} - q \log\log(n - p).$$

$$SBC_{REML} = l_{REML} - \frac{1}{2} q \log(n-p). \tag{54}$$

$$CAIC_{REML} = l_{REML} - \frac{1}{2} q(\log(n-p)+1).$$

onde q é o número de parâmetros de covariância e p a ordem de X.

Estes testes são particularmente úteis na comparação de modelos alternativos não encaixados, i.e. quando nenhum dos modelos pode ser encarado como um subconjunto do outro.

- Ensaiar a redução da especificação dos efeitos fixos. Para este fim, poder-se-ão construir versões ML da razão de verosimilhanças e dos critérios de informação, os quais virão

$$AIC_{ML} = l_{ML} - (p+q).$$
$$HQC_{ML} = l_{ML} - (p+q)\log\log(n).$$

$$SBC_{ML} = l_{ML} - \frac{1}{2}(p+q)\log(n). \tag{55}$$

$$CAIC_{ML} = l_{ML} - \frac{1}{2}(p+q)(\log(n-p)+1).$$

A metodologia apresentada para a selecção de modelos é de carácter geral. Nos modelos de variabilidade espacial há, no entanto, situações específicas a abordar. A primeira delas é a própria existência da variabilidade espacial representada por (45) que aqui se reescreve

$$\begin{aligned} Var(\varepsilon) &= R = \sigma^2 F_\varepsilon \\ Var(u) &= G = \sigma_u^2 F_u \end{aligned} \tag{45}$$

onde cada F é uma matriz cujo elemento genérico (ij) é o valor de $f(\vec{h})$ ou de $f(h)$. A estimação de um modelo misto com variabilidade espacial produzirá estimativas para σ^2, bem como para os restantes parâmetros das funções de distância. Caso a distância entre l_i e l_j não intervenha na formação da estrutura de covariâncias que geram (45), o facto de que $f(0)=1$ leva a que $Var(.)=\sigma^2 I$. Assim sendo, a presença da variabilidade espacial pode ser testada através dos ensaios apresentados, em particular o da razão de verosimilhanças. Para tal, basta que sejam produzidos valores das verosimilhanças para especificações alternativas do modelo: uma incluindo os parâmetros de covariância espacial e outra correspondente a uma formulação mais simples em que esse fenómeno seja ignorado.

Por outro lado, as diferentes especificações para a estrutura de covariância espacial podem ser testadas comparativamente através do recurso aos critérios de informação já apresentados.

5 ESTUDO DE UM CASO

Procede-se ao estudo dos níveis de consumo de famílias de um determinado espaço económico. São analisadas 248 famílias, que estão dispersas por 64 regiões (bairro, cidade, distrito ou outra divisão espacial). As regiões são referenciadas como se se tratasse de uma matriz de 8x8 com a notação (L.C) para linha e coluna. Existe um duplo objectivo: (1) indagar se os consumos têm uma relação geográfica que poderá ter a ver com características urbanas ou económicas da localidade e em caso afirmativo conhecer a sua estrutura; (2) estimar os parâmetros de uma função consumo e com base nestes efectuar inferência sobre os consumos de conjuntos de famílias ou mesmo de famílias especificas. Os dados que suportam o exemplo encontram-se no anexo 1.

O primeiro objectivo passa por ensaiar a existência de covariância espacial e por encontrar uma estrutura adequada para a sua modelização; o segundo passa pela estimação/previsão dos efeitos fixos e/ou aleatórios do modelo especificado.

5.1 MODELOS ENSAIADOS

São especificados 4 tipos de modelos.

Modelo (i): é um modelo de erros independentes, onde se ignora a existência de qualquer tipo de variabilidade espacial:

$$C_{ij} = \alpha + \beta Y_{ij} + \varepsilon_{ij} \tag{56}$$

onde C_{ij} representa o consumo da família j da região i, Y_{ij} o rendimento disponível da família j da região i, e ε_{ij} é tal que

$$\begin{aligned} &E(\varepsilon_{ij}) = 0 \\ &Var(\varepsilon_{ij}) = c_0 \quad \forall i, j \\ &Cov(\varepsilon_{ij}, \varepsilon_{ij'}) = 0 \quad j \neq j' \end{aligned} \tag{57}$$

Modelo (ii)

$$C_{ij} = \alpha + \beta Y_{ij} + u_i + \varepsilon_{ij} \tag{58}$$

onde C_{ij}, Y_{ij} e ε_{ij} são definidos tal como anteriormente e os u_i são não correlacionados com os ε_{ij}, tais que

$$\begin{aligned} &E(u_i) = 0 \\ &Var(u_i) = \sigma_u^2 \quad \forall i \\ &Cov(u_i, u_{i'}) = 0 \quad i \neq i' \end{aligned} \tag{59}$$

Neste caso, u_i representa o efeito de região e ε_{ij} o termo residual associado à família. Trata-se de um modelo de erro encaixado, que permite considerar a existência de uma covariância comum entre as famílias de uma mesma região. A existência de variabilidade espacial, está assim implícita neste modelo, o qual consiste na abordagem tradicionalmente utilizada para a representação deste tipo de fenómenos.

Modelo (iii)

$$C_{ij} = \alpha + \beta Y_{ij} + u_i + \varepsilon_{ij} \tag{60}$$

onde C_{ij}, Y_{ij} e ε_{ij} são definidos tal como anteriormente e u_i é tal que

$$\begin{aligned} &E(u_i) = 0 \\ &Var(u_i) = \sigma_u^2 \quad \forall i \\ &Cov(u_i, u_{i'}) = \sigma_u^2 f\left(\left\| l_i - l_{i'} \right\|\right) \end{aligned} \tag{61}$$

Trata-se de um modelo de erro encaixado, onde se considera explicitamente a existência de correlação espacial como uma função da distância entre as regiões. A

consideração de diferentes estruturas de covariância pode ser conseguida pela formulação de funções distância $f(\|l_i - l_{i'}\|)$ alternativas.

Modelo (iv)

$$C_{ij} = \alpha + \beta Y_{ij} + \varepsilon_{ij} \tag{62}$$

onde C_{ij}, Y_{ij} são definidos tal como anteriormente e ε_{ij} é tal que

$$\begin{aligned} &E(\varepsilon_{ij}) = 0 \\ &Var(\varepsilon_{ij}) = c_0 + \sigma^2 \quad \forall i, j \\ &Cov(\varepsilon_{ij}, \varepsilon_{i'j'}) = \sigma^2 f(\|l_i - l_{i'}\|) \end{aligned} \tag{63}$$

Esta última especificação corresponde a um modelo sem efeitos aleatórios, onde a variabilidade espacial é parametrizada ao nível da variável residual ε. Note-se que o argumento da função $f(\|l_i - l_{i'}\|)$ continua a ser a distância entre as regiões, pelo que a covariância entre as variáveis residuais ε associadas a duas famílias da mesma região é $Cov(\varepsilon_{ij}, \varepsilon_{ij'}) = \sigma^2$.

Sublinhe-se ainda que os modelos (iii) e (iv) conduzem à mesma estrutura de covariâncias. Tem-se para (iii)

$$\begin{aligned} &Var(Y_{ij}) = Var(u_i) + Var(\varepsilon_{ij}) = \sigma_u^2 + c_0 \\ &Cov(Y_{ij}, Y_{i'j'}) = Cov(u_i, u_{i'}) = \sigma_u^2 f(\|l_i - l_{i'}\|) \quad (j \neq j') \end{aligned} \tag{64}$$

e para (iv)

$$\begin{aligned} &Var(Y_{ij}) = Var(\varepsilon_{ij}) = \sigma^2 + c_0 \\ &Cov(Y_{ij}, Y_{i'j'}) = Cov(\varepsilon_{ij}, \varepsilon_{i'j'}) = \sigma^2 f(\|l_i - l_{i'}\|) \quad (j \neq j') \end{aligned} \tag{65}$$

O valor esperado e a estrutura de variâncias e covariâncias da variável de interesse é a mesma para ambos os modelos, sendo que para (iii) a variabilidade espacial é considerada ao nível da matriz G (matriz de covariâncias do vector de

parâmetros u), enquanto que para (iv) esta é considerada ao nível de R (matriz de covariâncias do vector de parâmetros ε).

A especificação (iv) apresenta a vantagem de não envolver efeitos aleatórios, pelo que a sua estimação se torna mais simples, deixando, nomeadamente, de ser necessário o recurso à teoria dos modelos lineares mistos. No entanto, esta é computacionalmente mais complexa devido à maior dimensão da matriz R. Utilizando um computador pessoal Pentium a 120Mhz, os tempos de cálculo, na produção de estimativas REML para os parâmetros dos dois modelos, com uma função de distância esférica e utilizando o conjunto de dados em anexo foram de cerca de 32 s para o modelo (iii) e de 9,47 m para o modelo (iv).

Existe, por outro lado, uma diferença conceptual de base entre os 2 modelos. De facto, o modelo envolvendo efeitos aleatórios (iii), abre espaço à inferência sobre funções predictíveis do tipo $K'\beta + M'u$, enquanto que para (iv) a inferência possível se resume aos efeitos fixos do modelo.

Imagine-se que se pretende conhecer o consumo esperado das famílias desta população com um dado rendimento y_0. A função a estimar é então $E(C \mid y_0) = \alpha + \beta y_0$, que é uma função dos efeitos fixos do modelo α e β. Ambos os modelos (iii) e (iv) poderão responder cabalmente a este objectivo.

Suponha-se, no entanto, que o objectivo é o de conhecer o consumo esperado das famílias de uma dada região i com o mesmo nível de rendimento y_0. A função a prever é agora $E(C \mid u_i, y_0) = \alpha + \beta y_0 + u_i$, que é função dos efeitos fixos e aleatórios do modelo.

A estimação de $E(C \mid y_0)$ corresponde a inferência no espaço largo, enquanto que a previsão de $E(C \mid u_i, y_0)$ corresponde a inferência no espaço restrito (cf. Secção 4.1). O modelo (iii), mostra-se mais flexível no que diz respeito a este tipo de inferência. Este é particularmente adequado quando há o objectivo de inferir em espaços mais restritos, caso onde a inferência sob o modelo (iv) se torna mais complexa, só podendo ser conseguida por recurso à previsão pontual $E(C \mid \varepsilon_{ij}, y_0) = \alpha + \beta y_0 + \varepsilon_{ij}$.

5.2 PRINCIPAIS RESULTADOS DA ESTIMAÇÃO[4]

Modelo (i)

$$\hat{C}_{ij} = \underset{(1.17)}{8.32} + \underset{(0.07)}{0.65} Y_{ij}$$

$$\hat{c}_0 = 4.63$$

[4] Indicam-se entre parêntesis as estimativas dos erros padrão dos estimadores dos efeitos fixos dos modelos.

REML=-543.8
AIC = -544.8
SBC=-546.5

Modelo (ii)[5]

$$\hat{E}(C_{ij}) = \underset{(0.79)}{9.07} + \underset{(0.05)}{0.61} Y_{ij}$$

$$\hat{c}_0 = 1.44$$

$$\hat{\sigma}_u^2 = 3.13$$

REML=-470.0
AIC = -472.0
SBC=-475.5

Modelos (iii) e *(iv)*

Os resultados são obtidos com a função de distância esférica.

$$\hat{E}(C_{ij}) = \underset{(0.88)}{9.10} + \underset{(0.05)}{0.61} Y_{ij}$$

$$\hat{c}_0 = 1.43$$

$$\hat{\sigma}_{(u)}^2 = 3.29$$

$$\hat{c}_a = 2.91$$

REML=-457.9
AIC = -460.9
SBC=-466.1

5.3 DIAGNÓSTICO DO MODELO

5.3.1 TESTE À EXISTÊNCIA DE CORRELAÇÃO ESPACIAL

Testam-se as hipóteses $\sigma_u^2 = 0$ e $c_a = 0$ recorrendo ao teste da razão de verosimilhanças e comparando sucessivamente os modelo (i) e (ii) e os modelos (ii) e (iii). Em ambos os casos a estatística teste segue um χ^2 com um grau de liberdade.

Seja $H_0 : \sigma_u^2 = 0 ; -2 log \frac{f_R^*}{f_U^*} = 147.6$.

Seja $H_0 : c_a = 0 ; -2 log \frac{f_R^*}{f_U^*} = 24.1$

[5] Para os modelos (ii) e (iii) omitem-se as previsões dos efeitos aleatórios u_i (i = 1,...,64).

Há evidência amostral, clara em ambos os casos, para rejeitar a nulidade dos parâmetros em teste. Conclui-se pela existência de covariância espacial neste conjunto de dados expressa na função de distância considerada.

5.3.2 TESTE À FORMA DA CORRELAÇÃO ESPACIAL

Podemos comparar os diferentes modelos através de critérios de informação. Os respectivos valores foram calculados a partir da forma $l(\theta) - F(n,q,p)$ onde $l(\theta)$ é uma verosimilhança e $F(n,q,p)$ é uma função do número de parâmetros estimados, da dimensão da amostra e da dimensão da matriz X. A lógica destes critérios é a de adoptar o modelo com *maior verosimilhança penalizada*. Há outros critérios e todos eles apresentam formas diversas que podem introduzir alguma imprecisão na linguagem e confusão ao decisor. A esse propósito ver Akaike (1974), Schwarz (1978), Hannan e Quinn(1979) e Bozdogan (1987). A forma que apresentamos é por vezes designada *quanto maior melhor*.

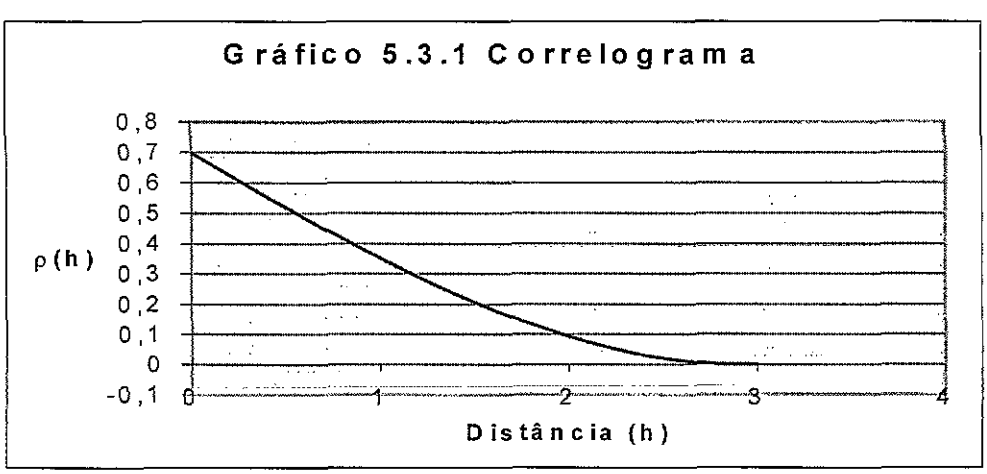
Ao modelo (iii) foram ajustadas diversas estruturas de covariância espacial associadas a diferentes funções de distância isotrópicas. No quadro 5.3.1 apresentam-se os valores de verosimilhança e dos critérios de informação AIC e SBC relativos a cada uma das especificações.

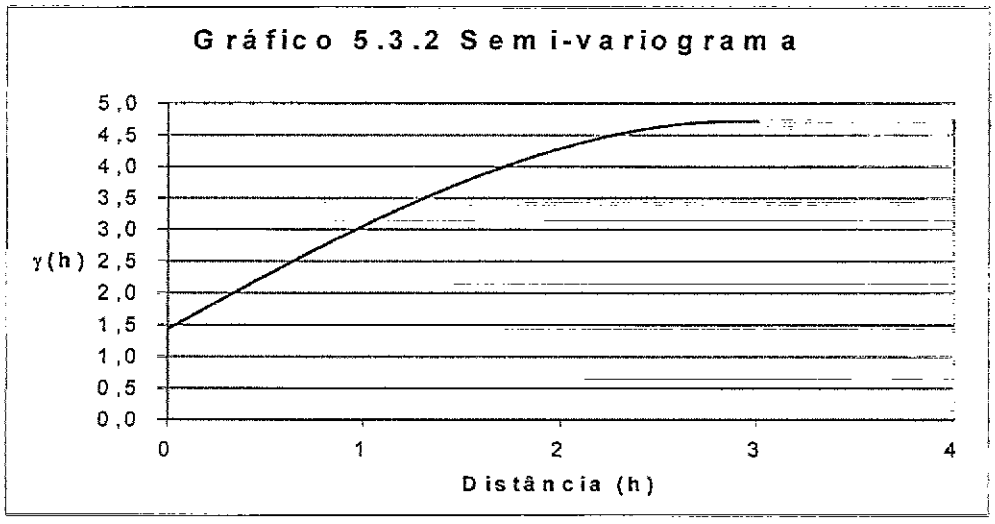
Quadro 5.3.1 Logaritmo da verosimilhança e critérios de informação

Modelo	Log. da Verosimilhança restrita	AIC	SBC
Esférico	-457,887	-460,887	-466,145
Potência	-459,215	-462,215	-467,473
Gaussiano	-458.547	-461.547	-466.805
Linear	-469.953	-472.953	-478.211
Log linear	-469.953	-472.953	-478.211

Há uma indiciação clara para o modelo esférico, que maximiza quer os critérios de Akaike e Schwarz quer o próprio logaritmo da verosimilhança.

Os gráficos 5.3.1 e 5.3.2 apresentam respectivamente o correlograma e o semivariograma resultantes da estimação dos parâmetros do modelo adoptado.





Note-se que a existência de covariância espacial entre as observações é, no modelo (iii) consubstanciada nos 64 efeitos aleatórios de região. As previsões para estes parâmetros são apresentadas no anexo 2. É particularmente notória a concentração quer de efeitos positivos, quer de efeitos negativos em grupos de regiões, i.e., na vizinhança de uma região é mais provável encontrar efeitos com o mesmo sinal que o da região considerada. Tal facto, evidencia a existência de correlação espacial, particularmente notória para as regiões mais próximas.

5.3.3 TESTE À EXISTÊNCIA DO EFEITO DE PEPITA

Estimar o modelo sem efeito de pepita consiste em especificar uma matriz $V = \sigma^2 F$ em vez de $V = \sigma^2 F + Ic_0$.

Para tal, testa-se no modelo (iii) a hipótese $c_0 = 0$ recorrendo ao teste de Wald. Note-se que o teste da razão de verosimilhanças torna-se inviável pois que a estimação do modelo (iv) com $c_0 = 0$ conduz a uma matriz R estimada não definida positiva. De facto, há que ter em conta que a discriminação espacial das famílias da amostra se faz unicamente através da sua região. Não sendo possível, com a informação disponível, distinguir espacialmente as famílias de uma mesma região, a estimação de um modelo sem efeito de pepita seria equivalente a especificar o modelo (iii) sem uma variável residual que pudesse acomodar as especificidades das famílias individuais.

Seja $H_0 : c_0 = 0$. A estatística teste é $W = \dfrac{\hat{c}_0^2}{\hat{V}(\hat{c}_0)}$ que sob a hipótese nula segue aproximadamente uma distribuição do χ^2 com um grau de liberdade. De forma equivalente, o teste pode ser realizado com base na estatística $\sqrt{W} = \dfrac{\hat{c}_0}{\sqrt{\hat{V}(\hat{c}_0)}}$ que segue aproximadamente uma distribuição normal standartizada.

Tem-se $\sqrt{W} = \frac{1.43}{0.149} = 9.6$, pelo que se rejeita claramente a hipótese nula, concluindo-se pela existência de variabilidade espacial à micro-escala.

5.3.4 ENSAIO À NORMALIDADE DOS RESÍDUOS

A fim de analisar a normalidade dos previsores $\hat{u}$ e dos residuos $\hat{\varepsilon}$ procede-se primeiramente à sua standartização através das transformações

$$\hat{u}^* = \hat{G}^{-1/2}\hat{u}$$

$$\hat{\varepsilon}^* = \hat{R}^{-1/2}\hat{\varepsilon}$$

onde $\hat{G}^{-1/2} = \sum_{i=1}^{r} \lambda_i^{-1/2} A_i$, os λ_i $(i = 1,\ldots r)$ são os r valores próprios distintos de $\hat{G}$ e $A_i = B_i B_i'$, sendo B_i a matriz dos m_i vectores próprios ortonormados associados a λ_i. $\hat{R}^{-1/2}$ tem o mesmo significado relativamente à matriz $\hat{R}$.

Os gráficos 9.3.3 e 9.3.4 apresentam os 64 valores standartizados $\hat{u}^*$ e os 248 valores de $\hat{\varepsilon}^*$. Note-se que em ambos dos casos não parecem verificar-se quaisquer padrões que violem os pressupostos de normalidade.

Procedeu-se ainda ao ensaio formal da normalidade de $\hat{u}^*$ e de $\hat{\varepsilon}^*$ através do teste de Shapiro-Wilk.

O valor da estatística teste calculada sobre os 64 valores de $\hat{u}^*$ é W=0,972. Para os 248 $\hat{\varepsilon}^*$ o valor da estatística é W=0,982.

Para ambos os casos verifica-se não haver evidência amostral para rejeitar a hipótese nula da normalidade de $\hat{u}^*$ e de $\hat{\varepsilon}^*$.

Gráfico 5.3.3 Resíduos standartizados ($\hat{\varepsilon}^*$)

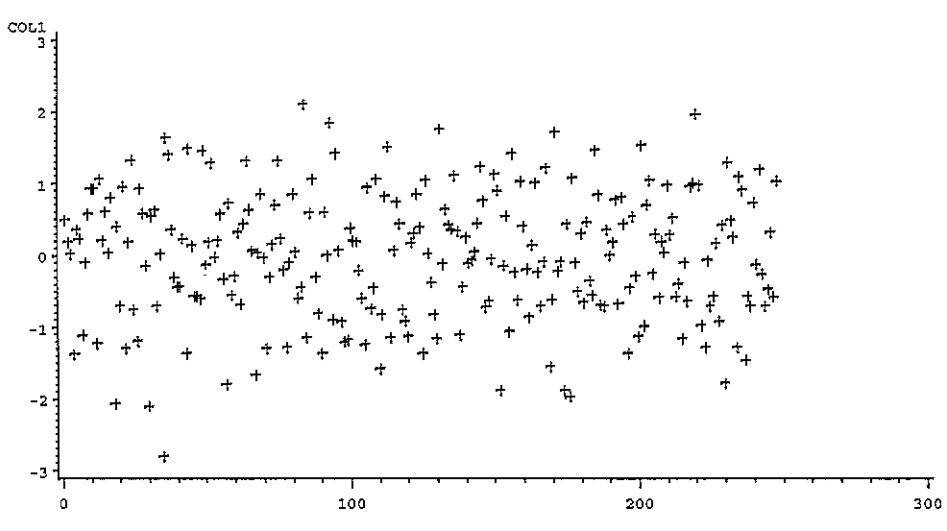
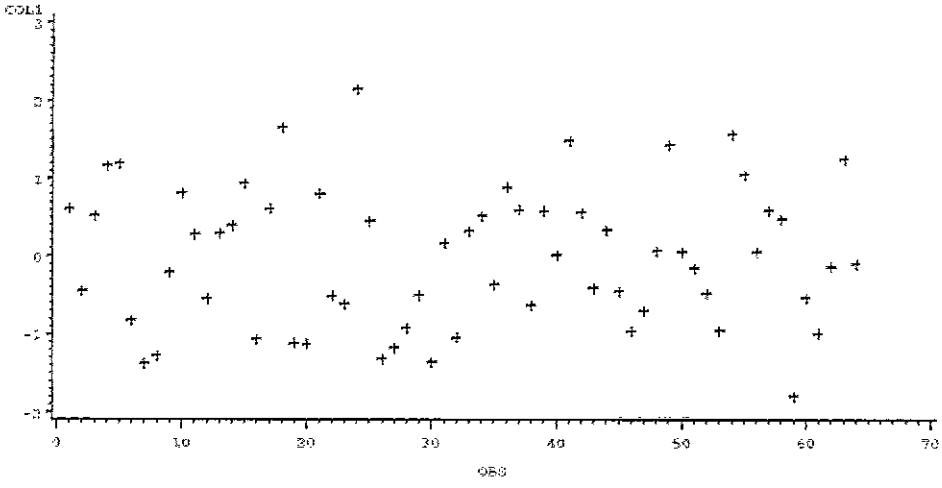


Gráfico 5.3.4 Efeitos aleatórios standartizados ($\hat{u}^*$)



5.4 ESTIMAÇÃO DAS COMPONENTES DE VARIÂNCIA

A estimação das componentes de variância σ^2, σ_u^2 e c_a foi efectuada pelo método de verosimilhança restrita. A optimização da função de verosimilhança é obtida por métodos numéricos. Alguns dos algoritmos apropriados para este efeito são referidos em Coelho e Pinheiro (1997). Tem-se verificado empiricamente que quando estão envolvidas estruturas de covariância espacial, a complexidade da função de verosimilhança faz com que esta optimização se torne particularmente sensível aos valores iniciais para arranque do algoritmo.

Foi ensaiada a estimação das componentes de variância pelo método de verosimilhança restrita. Para a optimização da função objectivo é utilizado o algoritmo de Fisher com o valor inicial de $\sigma_u^2 = 1$ e partindo de 6 diferentes valores iniciais para c_a. O quadro 5.4.1 apresenta as estimativas obtidas para os parâmetros σ_u^2 e c_a.

Quadro 5.4.1 Estimativas de σ_u^2 e c_a para diversos valores iniciais de c_a

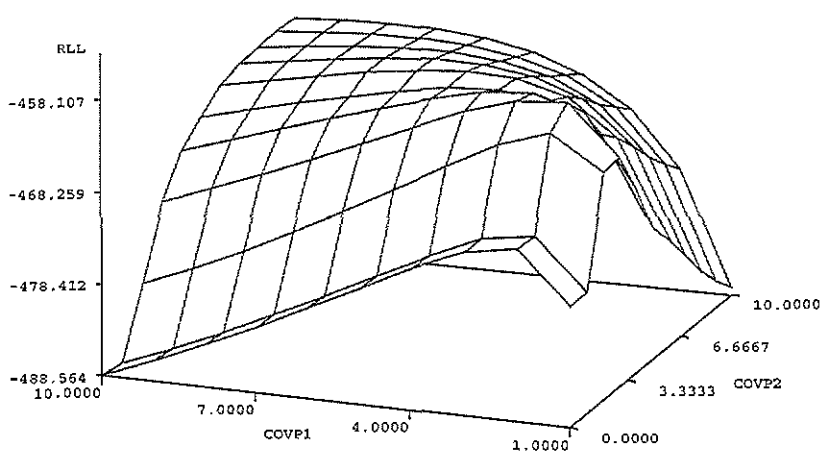
Valor inicial de c_a	$\hat{\sigma}_u^2$	$\hat{c}_a$
0	3.13	0.00
2	3.65	3.27
4	3.29	2.91
6	7.15	6.16
8	7.15	6.16
10	10.73	9.10

Repare-se que o algoritmo de Fisher tende a ser mais robusto à escolha dos valores iniciais do que outros algoritmos e em particular o de Newton-Raphson (Jennerich e Sampson, 1976). Mesmo assim, verifica-se, no caso vertente, que a escolha dos valores iniciais de c_a condiciona fortemente as estimativas obtidas.

Assim, e a fim de minimizar o risco de convergência a óptimos locais que forneçam estimativas pouco razoáveis para as componentes de variância, torna-se fundamental que o algoritmo seja iniciado com valores plausíveis para esses parâmetros. Nesse sentido, procedeu-se a uma primeira pesquisa dos valores da função de verosimilhança resultantes de diversas combinações plausíveis de valores das componentes de variância. O algoritmo foi então iniciado com a combinação de valores que resultou numa maior verosimilhança.

A pesquisa foi efectuada sobre 110 pontos definidos pelo cruzamento dos valores $\sigma_u^2 = 1,2,\ldots,10$ e $c_a = 0,1,\ldots,10$. O gráfico 5.4.1 apresenta a superfície de verosimilhança obtida a partir dos valores de verosimilhança calculados para os pontos da grelha.

Gráfico 5.4.1 Superfície de verosimilhança calculada para 110 pontos



O maior valor da função de verosimilhança foi -458.107, obtido para o ponto definido por $\sigma_u^2 = 2$ e $c_a = 3$. Estes valores inciais foram assim utilizados para arranque do algoritmo, tendo-se obtido as estimativas referidas na secção 5.2.

5.5 OUTROS RESULTADOS DA ESTIMAÇÃO

O conjunto de dados que serviu de base a este estudo empírico foi gerado com um valor $\beta = 0.6$. A análise dos resultados apresentados na secção 5.2 mostra claramente que a não consideração da existência de correlação espacial entre as observações resulta numa distorção das estimativas dos efeitos fixos do modelo (normalmente associada a uma perda de eficiência do estimador). Repare-se que o ajustamento do modelo (i) conduz a uma estimativa da propensão marginal ao consumo $\hat{\beta} = 0.65$, enquanto que os restantes modelos conduzem a estimativas próximas do seu verdadeiro valor $\hat{\beta} = 0.61$. De igual forma, se verifica que os erros padrão estimados, dos estimadores dos efeitos fixos são, para o modelo (i),

apreciavelmente maiores do que para os restantes modelos. Facilmente se imaginam conjuntos de dados onde as conclusões sobre os efeitos do modelo sejam seriamente afectadas.

Repare-se, no entanto, que as estimativas dos efeitos do modelo e respectivos erros-padrão, não são senão ligeiramente afectados quando se consideram outras estruturas de covariância espacial (cf. Anexo 3). De facto, os estudos empíricos mostram que o maior impacto nas conclusões da inferência resulta da consideração de uma estrutura de covariância razoável, sendo que a opção por diversos modelos plausíveis não afecta significativamente essas conclusões.

5.6 INFERÊNCIA SOBRE FUNÇÕES PREDICTÍVEIS

Retomem-se os exemplos referidos em 5.1 relativamente aos propósitos de inferência sobre funções do tipo $K'\beta + M'u$.
Tomem-se as seguintes funções

$$E(C \mid y_0) = \alpha + \beta y_0 \tag{66}$$

$$E(C \mid u_i, y_0) = \alpha + \beta y_0 + u_i \tag{67}$$

$$E(C \mid u_i, \varepsilon_{ij}, y_0) = \alpha + \beta y_0 + u_i + \varepsilon_{ij} \tag{68}$$

A função (66) representa um consumo médio das famílias com o mesmo nível de rendimento y_0. (67) refere-se ao consumo esperado para as famílias de rendimento y_0 residentes na região *i*. A função (68) consubstancia a previsão do consumo de uma dada família de rendimento y_0 e residente na região *i*.

Os BLUP das 3 funções sob o modelo (iii) são, respectivamente, dados por

$$\hat{E}(C \mid y_0) = \hat{\alpha} + \hat{\beta} y_0 \tag{69}$$

$$\hat{E}(C \mid u_i, y_0) = \hat{\alpha} + \hat{\beta} y_0 + \hat{u}_i \tag{70}$$

$$\hat{E}\left(C \mid u_i, \varepsilon_{ij}, y_0\right) = \hat{E}\left(C \mid u_i, \varepsilon_{ij}, y_0\right) = \hat{\alpha} + \hat{\beta} y_0 + \hat{u}_i \qquad (71)$$

As respectivas variâncias dos erros de previsão são

$$\begin{aligned} E\left[\hat{E}(C \mid y_0) - E(C \mid y_0)\right]^2 &= E\left[(\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) y_0\right]^2 \\ &= E(\hat{\alpha} - \alpha)^2 + y_0^2 E\left(\hat{\beta} - \beta\right)^2 + 2y_0 E\left[(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)\right] \end{aligned} \qquad (72)$$

$$\begin{aligned} &E\left[\hat{E}(C \mid u_i, y_0) - E(C \mid u_i, y_0)\right]^2 = E\left[(\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) y_0 + (\hat{u}_i - u_i)\right]^2 \\ &= E(\hat{\alpha} - \alpha)^2 + y_0^2 E\left(\hat{\beta} - \beta\right)^2 + E(\hat{u}_i - u_i)^2 \\ &+ 2y_0 E\left[(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)\right] + 2E\left[(\hat{\alpha} - \alpha)(\hat{u}_i - u_i)\right] + 2y_0 E\left[(\hat{u}_i - u_i)(\hat{\beta} - \beta)\right] \end{aligned} \qquad (73)$$

$$\begin{aligned} &E\left[\hat{E}\left(C \mid u_i, \varepsilon_{ij}, y_0\right) - E\left(C \mid u_i, \varepsilon_{ij}, y_0\right)\right]^2 = E\left[(\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) y_0 + (\hat{u}_i - u_i) - \varepsilon_{ij}\right]^2 \\ &= E(\hat{\alpha} - \alpha)^2 + y_0^2 E\left(\hat{\beta} - \beta\right)^2 + E(\hat{u}_i - u_i)^2 + E\left(\varepsilon_{ij}\right)^2 \\ &+ 2y_0 E\left[(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)\right] + 2E\left[(\hat{\alpha} - \alpha)(\hat{u}_i - u_i)\right] + 2y_0 E\left[(\hat{u}_i - u_i)(\hat{\beta} - \beta)\right] \end{aligned} \qquad (74)$$

O quadro 5.6.1 apresenta os previsores das 3 funções e respectivos erros-padrão, considerando um nível de rendimento $y_0 = 16$.

Quadro 5.6.1 Funções predictíveis e erros de previsão

Função predictível	Previsor	Erro-padrão
(66)	18.80	0.47
(67)	19.81	0.63
(68)	19.81	1.35

Os estimadores das variâncias e covariâncias dos erros de previsão dos parâmetros do modelo $\alpha, \beta, u_1, \ldots, u_{64}$ são obtidos substituindo as componentes de variância σ_u^2, c_0 .e c_a pelas suas estimativas REML e ignorando a variabilidade adicional introduzida por esta substituição.

Os valores apresentados sugerem os seguintes comentários:

- A função (66), que representa o consumo médio das famílias com rendimento $y_0 = 16$, pode ser entendida como uma função estimável, conduzindo à estimação no espaço de inferência largo; as restantes enquadram-se no conceito de previsão no espaço de inferência restrito. Em particular, a função (68) diz respeito à inferência especifica no indivíduo nos termos de Zeger et al.(1988).
- As previsões para as funções (67) e (68) são idênticas. Há, no entanto, diferenças no erro-padrão da previsão que se apresenta maior para o caso (68). De facto é de esperar que a incerteza na previsão do consumo de uma família especifica seja maior do que a incerteza associada à previsão do consumo médio das famílias de uma dada região. A diferença entre as variâncias dos erros de previsão de (67) e (68) é exactamente igual ao valor estimado da variância das variáveis residuais ε_{ij}. Foi já visto que esta variância $V(\varepsilon_{ij}) = c_0$ corresponde ao chamado efeito de pepita e que corresponde à existência de descontinuidade na variabilidade espacial à micro-escala.
- A função (67) representa o consumo médio das famílias da região 1 com rendimento $y_0 = 16$. O valor previsto é maior que o apresentado para a função (66). Tal facto, evidencia que se trata de uma região que contribui positivamente para o consumo das familias, i.e., onde o consumo tende a ser, em média, superior ao consumo médio do conjunto da população. O respectivo erro-padrão é igualmente superior ao da função (66) o que resulta da existência de uma variância apreciável do erro de previsão do efeito aleatório da região. Tal facto, não é de estranhar, dada a escassez de observações disponíveis em cada região, cujo número não excede em caso algum as 5 famílias.

6 CONCLUSÕES

Os modelos de variabilidade espacial têm lugar, quer em análises de regressão clássica, quer em análises de regressão em modelos mistos. Sendo estes uma visão mais geral daqueles, a variabilidade espacial encontrará neles uma posição mais fecunda.

Trata-se de utilizar conceitos espaciais ou físicos para melhor modelizar a estrutura de variâncias e covariâncias e, ao mesmo tempo, encontrar ferramentas de explicação e previsão de fenómenos que exibam variabilidade espacial. O primeiro passo, a modelização, pode também designar-se pelo objectivo de **caracterizar**. O segundo passo, explicar e prever, vem na sequência directa da obtenção de melhores estimativas ou previsões e pode designar-se pelo objectivo de bem **ajustar**.

Para ambos os objectivos a utilização formal do semivariograma constitui um valor acrescentado, já que constitui um recurso relativamente novo para detecção e representação de características espaciais, que se manifesta particularmente útil para a iniciação de métodos iterativos.

A utilização de estruturas de correlação espacial permite incorporar mais informação no modelo tornando a estimação mais eficiente. Quando esta estimação é

feita no quadro dos modelos mistos, as diversas formas de indução nestes definidas são naturalmente utilizáveis nos espaços de inferência larga ou restrita. Estas diferentes formas de inferência, juntas com a percepção da realidade espacial, dão origem a um instrumental estatístico particularmente potente e flexível .

Particular interesse desta teoria tem lugar quando a correlação espacial se manifesta entre grupos de observações, como é o caso do exemplo apresentado. De facto, o modelo misto permite que esta seja contemplada, quer ao nível das variáveis residuais, quer ao nível dos efeitos aleatórios, os quais poderão constituir factores de agregação de observações com diferentes níveis de granularidade.

Alguns campos de aplicação mais imediata dos modelos com variabilidade espacial são as sondagens, o *marketing*, a agricultura e a economia (em particular a dos recursos naturais). Em economia, por exemplo, há a possibilidade de estudar funções consumo com efeitos aleatórios e correlação espacial por regiões procurando acomodar, entre muitos outros, efeitos de contágio ou moda. Em economia de recursos naturais há todo um conjunto de possibilidades de estudo de relações geográficas em poluição, fluxos hídricos, propagação de culturas, etc. As aplicações em marketing podem cobrir casos variados como a penetração de campanhas promocionais ou estudos de audiência. As sondagens terão aqui uma fonte quase inesgotável para recomposição de estimadores.

REFERÊNCIAS BIBLIOGRÁFICAS

COELHO, P. (1996). Estimadores combinados para pequenos domínios. *Revista de Estatística*, 2, 23-43.

COELHO, P. e PINHEIRO J. (1997). Modelo linear geral misto-Teoria e estudo de um caso. . *Revista de Estatística*, 3, 63-86.

CORBEIL, R. e SEARLE, S. (1976). Restricted Maximum Likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, 18, 31-38.

CRESSIE, N. (1991). *Statistics for spatial data.* John Wiley & Sons, Inc.

DEMPSTER, A., LAIRD, N. e RUBIN, D. (1977). Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Serie B, 39, 1-38.

DIGGLE, P. (1988) . An approach to the analysis of repeated measurements. *Biometrics* 44, 959-971.

FULLER, W. A. e BATTES e, G. E. (1973). Transformations for estination of linear models with nested-error structures. *Journal of American Statistical Association*, 68, 625-632

GERALD, C. (1970). *Applied numerical analysis.* Adison-Wesley.

GIESBRECHT, F e BURNES, J (1985). Two-stage analysis based on a mixed model: large-sample asymptotic theory and small-sample simulation results. *Biometrics* 41, 477-486.

GREENE, W. (1997). *Econometric Analysis*, 3ª ed. Prentice-Hall.

HARTLEY, H. e RAO, J. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54, 93-108.

HARVILLE, D. (1977). Maximum Likelihood Approaches to Variance Components Estimation and to Related Problems. *Journal of American Statistical Association*, 72, 320-337

HARVILLE, D. (1985). Decomposition of prediction error. *Journal of American Statistical Association*, 80, 132-3138.

HARVILLE, D. (1988). Mixed-model methodology: theoretical justifications and future directions. *Proceedings of the statistical computing section*. American Statistical Association, New Orleans, 41-49.

HARVILLE, D. (1990). BLUP (Best linear unbiased prediction), and beyond, in *Advances in statistical methods for genetic improvement of lifestock*. Springer-Verlag, 239-276.

HENDERSON, C. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics*, 21, 309-310.

HENDERSON, C. (1973). Sire evaluation and genetic trends. *Proceedings of the animal breeding and genetics symposiom in honour of Dr. Jay L. Lush*. American Society Animal Science-American Dairy Science Association-Poultry Science Association, Champaign. III, 10-41.

HENDERSON, C. (1984). *Applications of linear models in animal breeding*. University of Guelph.

JENNRICH R. e SAMPSON, P. (1976). Newton-Raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics*, 18, 11-17.

JENNRICH R. e SCHLUCHTER, M. (1986). Unbalanced Repeated-measures models with structured covariance matrices. *Biometrics*, 42, 805-820.

JUDGE et AL. (1985). The theory and practice of econometries, 2ª ed. John Wiley & Sons, Inc.

KACKAR, R.N. e HARVILLE, D.A. (1984). Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models. *Journal of American Statistical Association*, 79, 853-862

LAIRD, N., LANGE, N e STRAM, D (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of American Statistical Association*, 82, 97-105.

LAIRD. N., WARE, J. (1982). Random-effects models for longitudinal data. *Biometrika*, 73, 13-22.

LIDSTROM. M. e BATES, D.(1988). Newton-Raphson amd EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of American Statistical Association*, 83, 1014-1022

PATTERSON, H. e THOMPSON, R. (1974). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545-554.

PRASAD, N e RAO, J (1990). The estimation of mean squared error of small-area estimators, *Journal of American Statistical Association*, 85, 163-171

RAO, C. (1972). Estimation of variance and covariance components in linear models. *Journal of American Statistical Association*, 67, 112-115.

RAO, C. (1973). *Linear statistical inference and its applications*, 2ª ed., John Wiley &Sons, Inc.

ROBINSON, G. K. (1991). That BLUP is a good thing; the estimation of random effects (with discussion). *Statistical Science*, 6, 15-51.

SEARLE, S. (1971). *Linear models*. John Wiley & Sons, Inc.

SEARLE, S. (1982). *Matrix algebra usefull for statisticians*. John Wiley & Sons, Inc.

SERFLING, R. (1980). *Approximation theorems of mathematical statistics*. John Wiley & Sons, Inc.

STROUP, W. (1989). Predictable functions and prediction space in the mixed model procedure, in *Applications of mixed models in agricultural and related disciplines*, Southern cooperative series bulletin, 343, Baton Rouge: Louisiana agricultural experiment station, 39-48.

WOLFINGER, R (1993). Covariance structures selection in general mixed models. *Comunications in statistics, simulation and computing*, 22(4), 1079-1106.

ZEGER, S. et AL.(1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049-1060.

ZIMMERMAN, D. e HALVILLE, D. (1991). A random fiels approachn to the analysis of field-plot experiments and other spatial experiments. *Biometrics*, 47, 223-239.

ANEXO 1 CONJUNTO DE DADOS

Região	Linha	Coluna	Rendimento	Consumo
1	1	1	16,26	20,57
1	1	1	16,53	20,37
1	1	1	14,43	18,89
2	1	2	14,87	16,78
2	1	2	14,73	18,76
2	1	2	15,71	19,20
3	1	3	17,80	19,98
3	1	3	12,81	18,16
3	1	3	15,97	20,89
3	1	3	18,98	23,14
4	1	4	18,85	24,08
4	1	4	13,64	18,34
4	1	4	13,57	21,04
5	1	5	13,73	19,96
5	1	5	16,93	22,38
5	1	5	16,59	21,48
6	1	6	17,38	19,57
6	1	6	17,02	15,93
6	1	6	18,05	19,50
7	1	7	16,69	15,51
7	1	7	15,27	16,63
7	1	7	14,13	13,26
7	1	7	16,77	16,62
8	1	8	17,58	18,43
8	1	8	14,81	14,27
8	1	8	16,84	14,98
9	2	1	16,57	21,12
9	2	1	19,98	22,77
9	2	1	18,15	20,80
9	2	1	16,44	17,41
10	2	2	18,33	22,69
10	2	2	13,69	19,98
10	2	2	12,54	17,68
11	2	3	15,82	19,27
11	2	3	20,09	18,48
11	2	3	14,54	20,43
11	2	3	18,12	22,31
12	2	4	15,15	18,49
12	2	4	17,33	19,01
12	2	4	15,37	17,66
12	2	4	16,99	18,65
13	2	5	13,72	18,59
13	2	5	16,04	18,10
13	2	5	16,18	21,59
13	2	5	15,38	19,49
13	2	5	10,56	15,71
14	2	6	15,33	18,21
14	2	6	13,80	17,25
14	2	6	15,38	20,67
15	2	7	16,68	19,44
15	2	7	20,63	22,20
15	2	7	16,67	21,11
15	2	7	16,78	19,62
16	2	8	10,09	14,03
16	2	8	15,57	17,79
16	2	8	15,77	16,83
16	2	8	17,08	15,87
17	3	1	15,67	21,13
17	3	1	16,15	19,89
17	3	1	14,35	19,12
17	3	1	17,48	21,74
18	3	2	13,68	18,51
18	3	2	16,85	21,76
18	3	2	14,87	21,62

18	3	2	15,59	21,24
18	3	2	12,30	18,57
19	3	3	17,05	15,20
19	3	3	11.43	13,83
19	3	3	13,92	16,32
19	3	3	16,23	16,66
20	3	4	17,14	15,46
20	3	4	16,30	16,14
20	3	4	14.43	15,55
20	3	4	15,34	16,75
21	3	5	18,60	21,92
21	3	5	15,09	18,51
21	3	5	16,91	19,08
22	3	6	20,69	19,22
22	3	6	15,01	17,18
22	3	6	16,20	19,04
23	3	7	17,41	19,18
23	3	7	15,60	17,31
23	3	7	18,28	19,11
24	3	8	19,39	25,86
24	3	8	16,58	20,27
24	3	8	15,55	21,72
24	3	8	20.16	25,08
24	3	8	16.81	21,43
25	4	1	14,62	18,13
25	4	1	15,07	17,73
25	4	1	13.77	19,29
25	4	1	17,40	20,78
25	4	1	15,96	22,11
26	4	2	16,31	16,36
26	4	2	17,50	19.86
26	4	2	17,65	18,34
26	4	2	12.12	13,80
26	4	2	18.57	17,38
27	4	3	19.11	16,47
27	4	3	16.65	16,83
27	4	3	18,46	17,71
28	4	4	16,55	16,98
28	4	4	13.11	14,41
28	4	4	12,38	13,51
29	4	5	17,66	16,79
29	4	5	14,62	17,57
29	4	5	15,35	16,00
29	4	5	18,04	17,97
29	4	5	17,90	19,70
30	4	6	16,45	14,74
30	4	6	17,01	15,99
30	4	6	20,97	20,36
30	4	6	12,77	16,20
30	4	6	19,45	17,09
31	4	7	17,41	19,28
31	4	7	17.55	20,18
31	4	7	16,70	19.30
31	4	7	18,17	18,75
32	4	8	17,03	17.56
32	4	8	16,05	16,72
32	4	8	18.71	19,90
32	4	8	17,05	19,04
33	5	1	15,26	20,90
33	5	1	17,84	21,91
33	5	1	19,35	20,74
34	5	2	14,42	19,84
34	5	2	16,54	19,89
34	5	2	17,30	19,87
35	5	3	17,65	17,90

35	5	3	16,35	16,72
35	5	3	14,50	19,08
36	5	4	16,26	19,35
36	5	4	12,53	17,98
36	5	4	14,24	18,75
37	5	5	19,40	21,36
37	5	5	17,75	21,26
37	5	5	17,78	20,35
37	5	5	13,39	15,96
38	5	6	15,37	16,22
38	5	6	14,58	16,56
38	5	6	16,07	17,03
38	5	6	13,86	15,74
38	5	6	14,48	16,27
39	5	7	17,88	20,24
39	5	7	15,76	19,89
39	5	7	15,97	19,45
39	5	7	15,50	17,40
39	5	7	14,52	16,91
40	5	8	15,17	18,04
40	5	8	14,33	18,93
40	5	8	17,27	20,45
40	5	8	16,61	16,71
41	6	1	15,39	21,86
41	6	1	15,93	23,02
41	6	1	18,52	22,67
41	6	1	17,44	24,97
42	6	2	17,73	21,36
42	6	2	17,88	20,99
42	6	2	14,29	20,79
43	6	3	16,93	19,53
43	6	3	15,25	17,80
43	6	3	16,17	17,56
44	6	4	16,38	19,31
44	6	4	17,36	20,94
44	6	4	14,88	17,96
44	6	4	17,00	18,67
45	6	5	15,77	17,66
45	6	5	14,88	18,66
45	6	5	18,31	17,44
46	6	6	15,38	16,44
46	6	6	16,99	20,22
46	6	6	14,76	16,55
46	6	6	14,09	16,30
46	6	6	14,88	14,62
47	6	7	16,36	19,05
47	6	7	10,55	12,64
47	6	7	13,39	18,01
47	6	7	14,33	17,17
47	6	7	17,35	18,52
48	6	8	14,32	18,30
48	6	8	14,59	17,32
48	6	8	16,25	19,67
49	7	1	17,15	22,60
49	7	1	12,90	19,78
49	7	1	15,75	23,92
49	7	1	15,46	22,99
49	7	1	16,10	21,55
50	7	2	10,73	15,79
50	7	2	18,40	21,72
50	7	2	17,32	20,64
51	7	3	15,19	17,49
51	7	3	13,88	17,41
51	7	3	15,44	16,61
52	7	4	17,22	18,57
52	7	4	18,39	18,84
52	7	4	13,09	13,47
52	7	4	18,56	17,88

52	7	4	17,20	18,23
53	7	5	17,61	17,63
53	7	5	15,51	15,35
53	7	5	16,72	19,26
53	7	5	15,49	15,51
54	7	6	17,10	22,16
54	7	6	15,12	21,37
54	7	6	16,39	20,60
54	7	6	18,60	22,59
55	7	7	17,10	21,03
55	7	7	15,65	21,08
55	7	7	17,12	21,79
55	7	7	15,80	22,11
56	7	8	16,57	20,26
56	7	8	16,42	20,45
56	7	8	17,67	19,89
56	7	8	17,34	19,91
57	8	1	19,74	21,53
57	8	1	13,87	19,23
57	8	1	17,44	20,76
57	8	1	15,23	21,33
57	8	1	16,54	22,17
58	8	2	14,37	20,74
58	8	2	18,47	22,04
58	8	2	14,23	17,14
58	8	2	16,24	18,00
58	8	2	17,58	20,26
59	8	3	18,14	16,12
59	8	3	14,94	14,34
59	8	3	15,41	15,52
59	8	3	12,52	12,46
59	8	3	16,00	16,17
60	8	4	15,03	13,58
60	8	4	14,20	16,75
60	8	4	14,28	15,84
60	8	4	17,44	17,47
61	8	5	15,93	15,18
61	8	5	17,94	19,24
61	8	5	14,59	16,99
61	8	5	16,41	15,26
62	8	6	17,47	19,49
62	8	6	19,86	20,79
62	8	6	14,74	19,40
63	8	7	16,86	21,80
63	8	7	18,17	24,17
63	8	7	18,61	22,69
64	8	8	15,41	18,28
64	8	8	17,52	19,85
64	8	8	15,96	19,85
64	8	8	15,13	18,27
64	8	8	15,22	20,25

ANEXO 2 – PREVISÕES PARA OS 64 EFEITOS ALEATÓRIOS DE REGIÃO SOB O MODELO (III) COM FUNÇÃO DE DISTÂNCIA ESFÉRICA

1,010109	0,299059	1,414754	2,433166	2,278087	-1,03225	-2,87823	-2,91783
0,855755	1,821074	0,54355	-0,23909	0,889056	0,491989	0,365187	-1,44629
1,654076	1,917636	-2,25211	-2,49933	-0,03354	-0,90999	-0,54851	2,478363
1,116579	-1,55795	-2,83046	-2,3966	-1,53303	-2,45289	-0,45936	-0,77264
1,525299	0,727639	-0,92059	0,506515	0,054871	-1,69056	-0,24863	-0,21599
3,594896	1,779855	-0,32465	0,098695	-0,91533	-1,2529	-0,50469	0,152644
3,508247	1,025711	-1,04872	-1,94796	-1,81429	1,848011	2,249518	0,760543
1,836325	0,570269	-3,14094	-2,51409	-2,05696	0,47898	2,612359	0,677687

ANEXO 3 – RESULTADOS DA ESTIMAÇÃO DO MODELO (III) COM DIFERENTES FUNÇÕES DE DISTÂNCIA

Função de distância exponencial

$$\hat{E}(C_{ij}) = \underset{(1.14)}{9.31} + \underset{(0.05)}{0.61} Y_{ij}$$

$$\hat{c}_0 = 1.43$$
$$\hat{\sigma}^2_{(u)} = 4.35$$
$$\hat{c}_e = 2.07$$

Função de distância gaussiana

$$\hat{E}(C_{ij}) = \underset{(0.85)}{8.97} + \underset{(0.05)}{0.61} Y_{ij}$$

$$\hat{c}_0 = 1.44$$
$$\hat{\sigma}^2_{(u)} = 3.18$$
$$\hat{c}_g = 1.15$$

Função de distância linear

$$\hat{E}(C_{ij}) = \underset{(0.79)}{9.07} + \underset{(0.05)}{0.61} Y_{ij}$$

$$\hat{c}_0 = 1.44$$
$$\hat{\sigma}^2_{(u)} = 3.13$$
$$\hat{c}_a = 0.00$$

Função de distância log-linear

$$\hat{E}(C_{ij}) = \underset{(0.79)}{9.07} + \underset{(0.05)}{0.61} Y_{ij}$$

$$\hat{c}_0 = 1.44$$
$$\hat{\sigma}^2_{(u)} = 3.13$$
$$\hat{c}_a = 0.00$$

REVISTA DE ESTATÍSTICA
81ª PAGINA
2º QUADRIMESTRE DE 1998

DIMENSÕES GEOGRÁFICAS NA SEGMENTAÇÃO DOS CLIENTES PORTUGUESES DAS POUSADAS DE PORTUGAL

Autores:
Margarida G. M. S. Cardoso,
Isabel Hall Themido
Luís Chambel Cardoso

VOLUME 2

2º QUADRIMESTRE DE 1998

REVISTA DE ESTATÍSTICA
83ª PAGINA
2º QUADRIMESTRE DE 1998

DIMENSÕES GEOGRÁFICAS NA SEGMENTAÇÃO DOS CLIENTES PORTUGUESES DAS POUSADAS DE PORTUGAL

THE CONTRIBUTION OF GEOGRAPHICAL DIMENSIONS TO THE SEGMENTATION OF THE PORTUGUESE CLIENTS OF *POUSADAS DE PORTUGAL*

Autores: Margarida G. M. S. Cardoso
- Estudante de doutoramento em Engenharia de Sistemas no CESUR- IST, bolseira do Programa Praxis XXI
Isabel Hall Themido
- Prof.ª Associada no Departamento de Engenharia Civil, CESUR-IST
e
Luís Chambel Cardoso
- Estudante de doutoramento em Engenharia de Minas e Georrecursos no IST, bolseiro do Programa Praxis XXI

RESUMO:

- Neste artigo procura-se complementar um estudo de segmentação dos clientes portugueses das Pousadas de Portugal, considerando a contribuição de dimensões geográficas, nomeadamente a região de residência dos clientes e a região destino onde se situa a Pousada.

  Constata-se que os respondentes da amostra em questão são, maioritariamente, oriundos dos distritos com maior poder de compra *per capita*: Lisboa, Porto e Setúbal.

  Verifica-se que alguns dos hábitos e preferências destes clientes estão associados à região de residência.

  Estuda-se a distribuição geográfica dos segmentos de *Clientes frequentes*, *Clientes comuns* e *Clientes pela primeira vez*, identificando as regiões *fornecedoras* de cada tipo de clientes.

  Por outro lado, a atractividade das Pousadas é medida por meio de um índice de penetração nos distritos e como função da distância média aproximada que percorrem os seus clientes.

PALAVRAS CHAVE:

- *Análise estatística multivariada; marketing; segmentação; turismo; geografia.*

ABSTRACT:

- This paper complements an earlier segmentation study of the Portuguese clients of *Pousadas de Portugal*, a network of differentiated hotels, considering the contribution of geographical dimensions, namely the area of residence of clients and the destination region, where the *Pousada* is located.

We conclude that the majority of clients comes from the districts with the highest purchasing power *per capita*- Lisbon, Oporto and Setúbal - and that this geographical origin influences some of their habits and preferences.

The geographic origin of the heavy users, light users and first time users is also studied and the regions which generate more clients of each type are identified.

The attractiveness of the Pousadas is measured by means of a penetration index by district and as a function of the average distance travelled by its clients.

KEY-WORDS:

- *Multivariate data analysis, marketing, segmentation, tourism, geography.*

1 INTRODUÇÃO

1.1 SOBRE A AMOSTRAGEM E RECOLHA DE DADOS

Os dados aqui considerados referem-se a respostas a inquéritos dirigidos aos clientes portugueses das Pousadas de Portugal. Estes inquéritos foram elaborados e distribuídos no âmbito de um projecto de Segmentação dos Clientes Portugueses das Pousadas de Portugal.

A recolha de dados para a realização do referido projecto foi efectuada em todas as unidades hoteleiras da ENATUR, tendo sido organizada por trimestres e por grupos de Pousadas entre Outubro de 96 e Outubro de 97. Foram devolvidos preenchidos mais de 2500 questionários o que correspondeu a uma taxa média de resposta de 76 % (nº de inquéritos preenchidos/nº de inquéritos distribuídos).

O questionário integra questões sobre as Pousadas (desde a motivação e escolha até à caracterização e avaliação da estadia e região onde se situa a Pousada) assim como sobre as férias e o perfil demográfico e psicográfico dos respondentes, num total de 39 questões que se traduzem (após codificação) em 142 variáveis básicas[6].

Neste trabalho analisam-se os dados correspondentes a 238 respondentes ao questionário, clientes de quatro Pousadas de diversa *tipologia*/classificação comercial da própria ENATUR. Esta classificação das Pousadas destaca os grupos CH, CSUP, C e B aos quais correspondem, simplificadamente, níveis decrescentes de preços. Por outro lado, as quatro Pousadas em causa têm diversas situações geográficas no continente. Concretamente esta análise refere-se às Pousadas: Quinta da Ortiga (tipo B), em Santiago do Cacém; Stª Cristina (tipo C), em Condeixa-a-Nova; S. Bento (tipo CSUP) em Vieira do Minho; e Lóios (tipo CH), em Évora.

1.2 ENQUADRAMENTO

Este estudo sucede-se a um outro (Cardoso e Themido, 1998) no qual se procedeu a uma dupla segmentação dos respondentes usando critérios observáveis e de relação com os produtos-Pousada. Segundo um critério relacionado com o tipo e frequência de uso constituiram-se segmentos de *Clientes frequentes*, *Clientes comuns* e *Clientes pela primeira vez*, aqui denominados *Segmentos-frequência*. Por outro lado, foram constituídas classes de respondentes consoante a sua escolha actual/preferência revelada i.e. os respondentes foram agrupados segundo a Pousada em que se encontravam alojados.

Os critérios de segmentação referidos que são, em geral, considerados eficientes (Wedel e Kamakura, 1998, p.16) , conduziram a bons resultados com esta amostra em particular, tendo-se obtido diferenças significativas ao nível de

[6] Mediante contacto com uma das autoras o questionário poderá ser disponibilizado para consulta.

preferências manifestadas pelos respondentes e ainda intenção de voltar a uma Pousada, por exemplo.

1.3 DESCRIÇÃO SUMÁRIA DA AMOSTRA

Na amostra em questão os respondentes são na sua maioria homens (77%), casados (81%) cuja idade média é de 45 anos com desvio padrão de 14 anos. A formação académica dos respondentes é maioritáriamente *superior*: 52% são licenciados e 9% mestres ou doutorados. Segundo a distribuição das classes de rendimento do agregado familiar (líquido, mensal) na amostra, 50% dos respondentes auferem rendimentos superiores a 500 contos sendo que, para 10% da amostra este rendimento ultrapassa os 1000 contos.

A razão da estadia na Pousada é *lazer* para 88% dos respondentes, sendo a estadia, no geral, de uma noite (58%) ou duas (25%). O desejo de conhecer e o gosto pela região onde se situa uma Pousada são dos motivos mais frequentemente invocados no processo de escolha da Pousada como destino de férias. A qualidade global da estadia - independentemente da unidade hoteleira particular a que se refere - é considerada *Boa* por 57% dos respondentes e *Muito Boa* por 37%. Finalmente, a intenção de voltar a uma (qualquer) Pousada de Portugal é expressa por *Volto com certeza* por 80% dos inquiridos.

1.4 ESTRUTURA DE ANÁLISE

O objectivo particular deste trabalho é conjugar as análises de segmentação já efectuadas com o estudo sobre as dimensões geográficas disponíveis: situação geográfica das unidades hoteleiras e região de origem dos clientes. Estas dimensões são consideradas particularmente relevantes tendo em conta que o presente estudo se enquadra na área do turismo (Opperman,1997). Além do mais, considera-se que o estudo destas dimensões, quer do ponto de vista da oferta quer da procura, permitirá obter uma informação mais adequada para suportar futuras decisões em marketing (ponto de vista sustentado por Spotts,1997) .

Como estrutura inicial, base da análise, adopta-se a que é ilustrada na Figura 1. Nesta estrutura, evidencia-se, para além dos critérios de segmentação já referidos, o sistema de distritos, divisão geográfica por que se opta tendo em conta a dimensão da amostra.

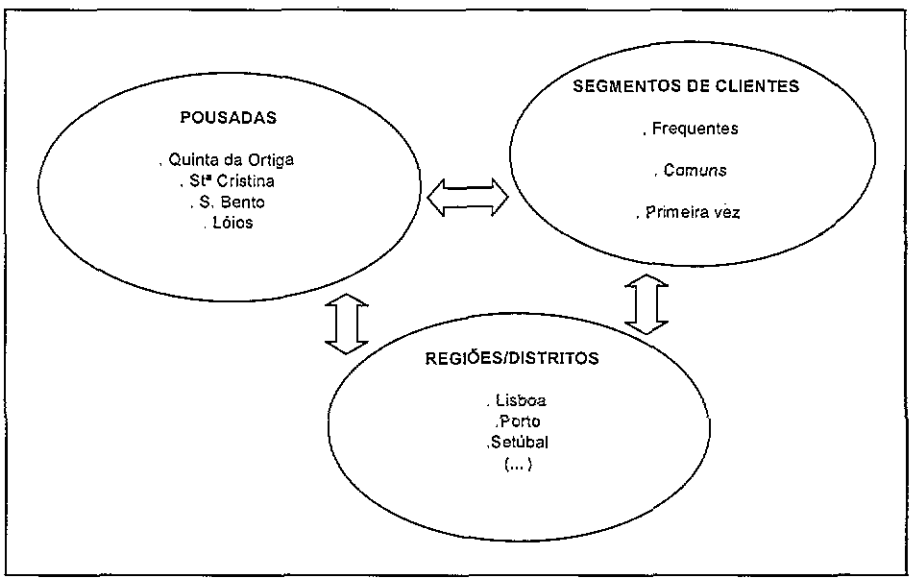


Figura 1-Estrutura de suporte da análise dos dados

Segundo a estrutura de análise proposta para os dados podem observar-se, na Tabela 1 os valores correspondentes à distribuição dos respondentes em estudo.

Distrito	Pousadas	Segmentos de clientes			Totais
		Frequentes	Comuns	1ª vez	
Lisboa	S. Bento	4	25	3	32
	Stª Cristina	6	25	3	34
	Lóios	8	26	10	44
	Quinta da Ortiga	3	20	4	27
	Total	21	96	20	137
		15,33%	70,07%	14,60%	
Porto	S. Bento	1	7	5	13
	Stª Cristina	3	3	0	6
	Lóios	0	2	2	4
	Quinta da Ortiga	1	4	1	6
	Total	5	16	8	29
		17,24%	55,17%	27,59%	
Setúbal	S. Bento	0	7	0	7
	Stª Cristina	1	5	0	6
	Lóios	0	1	1	2
	Quinta da Ortiga	0	6	0	6
	Total	1	19	1	21
		4,76%	90,48%	4,76%	
	Total	**27**	**131**	**29**	**187**

Tabela 1- Distribuição da amostra pelos distritos de Lisboa, Porto e Setúbal e segundo as Pousadas e os *Segmentos-frequência*

Note-se que estão ausentes desta tabela os respondentes que não residem nos três distritos indicados (os mais representados na amostra) assim como a pequena percentagem de não respostas referidas ao código postal de residência.

2 A AMOSTRA POR REGIÕES

Tendo em conta o código postal de residência apontado pelos respondentes a amostra é agrupada por distritos observando-se que os respondentes provêm maioritariamente de Lisboa (64%) e ainda do Porto (14%) e Setúbal (10%) (v. números de respondentes na Tabela 1). Quanto aos restantes distritos do País eles estão pouco representados na amostra.

Na análise que a seguir se apresenta sobre a amostra são utilizados os dados disponíveis sobre a População (Fonte: INE, Estimativas de População Residente, nº24)

e sobre um Índice de poder de compra *per capita* (Fonte: INE, Estudo sobre o Poder de Compra Concelhio, Número III, 1997).

2.1 ANÁLISE DO PODER DE COMPRA

Tendo em conta os valores de um Índice do poder de compra *per capita* dos distritos (valores presentes, na Tabela 2) – valores calculados como uma média ponderada pela população do Índice por Concelho disponível no INE - nota-se que os distritos mais representados na amostra são precisamente os que têm os maiores Índices de poder de compra: 173 para Lisboa, 107 e 108 para o Porto e Setúbal (respectivamente), sendo 100 o índice base para o País.

Viseu	Vila Real	Bragança	Guarda	Viana do Castelo	Beja	Açores	Castelo Branco	Madeira	Portalegre	Braga	Santarém	Aveiro	Évora	Leiria	Coimbra	Faro	Porto	Setúbal	Lisboa
46	53	55	57	58	60	61	63	64	64	65	73	75	75	78	79	104	107	108	173

Tabela 2- Índice do Poder de Compra *per Capita*

Curiosamente, analisando os dados disponíveis sobre o rendimento mensal líquido dos clientes da amostra (v. Tabela 3) não há diferenças significativas na sua distribuição entre os distritos mais representados na amostra. Mais ainda, não há qualquer associação entre os rendimentos declarados pelos respondentes de uma região e o Índice de poder de compra dessa região facto que se justifica já que, em cada região, são clientes das Pousadas, os indivíduos de rendimento relativamente mais elevado.

	< 200 contos	200-500 contos	500-1000 contos	> 1000 contos	Número de respostas
Lisboa	5	48	55	16	124
	4,03%	38,71%	44,35%	12,90%	
Porto	1	9	12	4	26
	3,85%	34,62%	46,15%	15,38%	
Setúbal	1	5	10	2	18
	5,56%	27,78%	55,56%	11,11%	
Totais	7	62	77	22	168
	4,17%	36,90%	45,83%	13,10%	

Tabela 3- Classes de rendimento mensal líquido dos respondentes (Lisboa, Porto e Setúbal)

2.2 INDICADOR DE DISTRIBUIÇÃO GEOGRÁFICA

No sentido de avaliar a importância de cada região como fornecedora de cliente para a ENATUR determinou-se um Indicador de distribuição geográfica: número de respondentes residentes em cada distrito/população do distrito (valor corrigido por um factor de 100000 e posteriormente reescalado de 0 a 1). Segundo os resultados obtidos (v.**Tabela 4**) regista-se o maior valor em Lisboa (1), seguido por Setúbal (0,429) e Porto (0,257). Note-se ainda que Coimbra, Portalegre e Madeira têm um valor do indicador da ordem de grandeza do apontado para o Porto, embora a sua presença na amostra, em termos absolutos, tenha pouca expressão.

Aveiro	0
Beja	0
Braga	0,058
Bragança	0,099
Castelo Branco	0
Coimbra	**0,213**
Évora	0
Faro	0,130
Guarda	0,083
Leiria	0,069
Lisboa	**1**
Portalegre	**0,236**
Porto	**0,257**
Santarém	0,102
Setúbal	**0,429**
Viana do Castelo	0
Vila Real	0,065
Viseu	0
Madeira	**0,232**
Açores	0

Tabela 4- Indicador de distribuição geográfica dos clientes

2.3 MODOS DE ACTUAR

A partir de alguns dados de resposta ao questionário procuraram-se padrões geográficos de comportamento.

Apenas na forma particular de efectuar a reserva da estadia se constatam (v Tabela 5), diferenças significativas associadas aos distrito de origem dos clientes. Evidencia-se, concretamente, o diferente modo de proceder dos clientes do Porto que, pouco utilizando a Central de Reservas recorrem, muito mais que os clientes de Lisboa e os Setúbal, à utilização da intermediação das Agências de Viagens.

	Forma de efectuar a reserva			
	Pousada	**Central de Reservas**	**Agência**	**Totais**
Lisboa	62	55	12	129
%	48.06%	42.64%	9.30%	129
Porto	19	4	6	29
%	65.52%	13.79%	20.69%	29
Setúbal	10	8	2	20
%	50.00%	40.00%	10.00%	20
Total	91	67	20	72.47%
% Total	51.12%	37.64%	11.24%	100.00%
Valor de probabilidade associado a teste de independência do Qui-Quadrado de Pearson				**0,051**

Tabela 5- Modo de efectuar a reserva da estadia/distrito

2.4 PREFERÊNCIAS

A análise das associações entre as preferências manifestadas pelos respondentes e os seus distritos de origem centrou-se, sobretudo, nos conceitos propostos no questionário, de *Pousada ideal* e *férias ideais*. Outras análises que exigiam ainda a associação das preferências às Pousadas ou às regiões onde estas se situam, foram adiadas para o estudo da amostra global, tendo em conta a reduzida dimensão da amostra presente.

No que diz respeito à análise dos conceitos de *Pousada ideal* segundo os distritos de origem dos respondentes, registam-se (v. Tabela 6) algumas diferenças, embora pouco significativas. A este propósito note-se, apenas, a mais destacada preferência dos respondentes do Porto por Pousadas de Montanha e situadas na região Norte, preferências associadas, por outro lado a uma menor reserva relativa à vida social do que a manifestada por Lisboa ou Setúbal.

	Lisboa	Porto	Setúbal	Valor de probabilidade associado a teste de qui-quadrado
Edifício				0,346
Monumento	68,80%	54,17%	61,11%	
Regional	31,20%	45,83%	38,89%	
Paisagem				0,098*
Montanha	59,41%	73,91%	46,67%	
Cidade	9,90%	0,00%	0,00%	
Campo	22,77%	21,74%	53,33%	
Praia	7,92%	4,35%	0,00%	
Zona do País				0,120
Norte	55,56%	68,18%	61,54%	
Centro	15,15%	27,27%	7,69%	
Sul	29,29%	4,55%	30,77%	
Ambiente				0,233
Requintado	66,95%	84,00%	66,67%	
Simples	33,05%	16,00%	33,33%	
Decoração				0,751**
Moderna	6,31%	13,64%	0,00%	
Clássica	36,04%	36,36%	36,84%	
Rústica	31,53%	27,27%	47,37%	
Antiga	26,13%	22,73%	15,79%	
Local				0,353
Litoral	30,28%	16,00%	26,67%	
Interior	69,72%	84,00%	73,33%	
Vida Social				0,059
Reservada	34,96%	11,54%	33,33%	
Moderada	60,98%	84,62%	66,67%	
Animada	4,07%	3,85%	0,00%	

*excluindo"cidade" e "praia" c/ insuficiente nº de observações

**excluindo "moderna" c/ insuficiente nº de observações

Tabela 6 – Caracterização da *Pousada ideal* por distrito de origem do respondente

Comparando algumas diferenças na apreciação do conceito de *Férias ideais* sobressai a menor preferência por paisagem de praia manifestada pelos respondentes do Porto em favor das paisagens de campo e montanha. Outras diferenças e correspondentes graus de significância podem ser observadas na tabela seguinte:

	Lisboa	Porto	Setúbal	Valor de probabilidade associado a teste de qui-quadrado
Local				0,264
País	63,44%	55,00%	83,33%	
Estrangeiro	36,56%	45,00%	16,67%	
Estação				0,119*
Primavera	17,24%	11,76%	23,08%	
Verão	66,67%	58,82%	46,15%	
Outono	8,05%	23,53%	30,77%	
Inverno	8,05%	5,88%	0,00%	
Paisagem				0,080**
Montanha	16,00%	37,50%	22,22%	
Cidade	8,00%	0,00%	0,00%	
Campo	20,00%	31,25%	22,22%	
Praia	56,00%	31,25%	55,56%	
Zona do País				0,186
Norte	25,45%	50,00%	36,36%	
Centro	23,64%	8,33%	0,00%	
Sul	50,91%	41,67%	63,64%	

*excluindo "Inverno" c/ insuficiente nº de observações

**excluindo "Cidade" e "Setúbal" c/ insuficiente nº de observações

Tabela 7 – Caracterização das *férias ideais* segundo distrito de origem dos respondentes

2.5 *POUSADAS EM QUE OS RESPONDENTES JÁ ESTIVERAM ALOJADOS*

Analisando a distribuição, em percentagem, das Pousadas em que os respondentes de Lisboa e Porto já estiveram alojados há um padrão comum de escolha referido às Pousadas que se podem considerar mais emblemáticas - Rainha Stª Isabel, Stª Maria, Castelo, Stª Marinha, S. Gonçalo e S. Lourenço. No entanto os respondentes do Porto destacam-se por conhecerem mais Pousadas que os de Lisboa, Pousadas que ficam, na grande maioria, situadas nas regiões Cento e Norte do Continente. O número de Pousadas conhecidas pelos clientes provenientes de Setúbal é muito mais reduzido, sobressaindo a escolha de Palmela.

3 SEGMENTOS-FREQUÊNCIA

3.1 PROCESSO DE SEGMENTAÇÃO

Segundo o de segmentação anteriormente efectuado foi constituído, *a priori,* um primeiro grupo/segmento dos respondentes que estão numa Pousada de Portugal pela primeira vez. Este grupo é constituído por 41 respondentes, representando cerca de 17% da amostra.

Em seguida foram seleccionadas variáveis de segmentação que traduzem frequência e tipos de Pousadas em que os respondentes já estiveram alojados, mais concretamente:

CH - número de Pousadas tipo CH em que o respondente já esteve alojado

CSUP - número de Pousadas tipo CSUP em que o respondente já esteve alojado

C - número de Pousadas tipo C em que o respondente já esteve alojado

B - número de Pousadas tipo B em que o respondente já esteve alojado

O número de segmentos a constituir com base nestas variáveis foi previamente fixado em dois. Esta opção ficou a dever-se à dimensão da amostra e ao interesse, estabelecido *a priori*, em constituir dois grupos de clientes das Pousadas: um grupo de clientes comuns e um grupo de clientes diferenciados.

A partir da realização de uma análise classificatória *K-Médias* foram, então, constituídos dois segmentos – designados *Clientes frequentes* e *Clientes comuns* das Pousadas – tendo a maioria dos respondentes (70%) sido classificada neste último segmento. Os *Clientes frequentes* já estiveram alojados, em média, em 20 Pousadas e os *Clientes comuns* em cinco, sendo a distribuição destes valores pelos diversos tipos de Pousadas idêntica nos dois segmentos.

3.2 CARACTERIZAÇÃO

No estudo já efectuado sobre este sistema de segmentos e tal como se pode observar na Tabela 8, foram identificadas diversas dimensões para as quais se registaram diferenças significativas entre os segmentos. Em geral, da análise efectuada, destaca-se o facto de, tendo-se segmentado com base em indicadores comportamentais da relação cliente-produto/Pousada, sobressairem, entre os segmentos, as diferenças de preferências, nomeadamente as que dizem respeito a *férias ideais* e *Pousada ideal*. Note-se que, entre os *Segmentos-Frequência*, é nos aspectos mais subjectivos das componentes do conceito de *Pousada ideal*, (como *ambiente* e *vida social*) que se manifestam mais as diferenças.

Note-se ainda que a única variável demográfica que distingue os segmentos formados é a idade.

Finalmente constate-se também a diferença significativa registada na intenção de voltar a uma (qualquer) Pousada de Portugal que, tal como seria de esperar, está relacionada com o facto dos *Clientes pela primeira vez* não estarem tão certos de voltar como os restantes.

Numa análise acerca das escolhas dos produtos-Pousadas revelou-se que todas as Pousadas foram já destino de uma percentagem elevada dos *Clientes frequentes*, percentagem que chega a mais de 70% quando referida às Pousadas S. Gonçalo, S. Lourenço, Stª Bárbara, Stª Luzia, e Rainha Stª Isabel. A escolha dos *Clientes comuns* recai sobre algumas das Pousadas mais emblemáticas da ENATUR, nomeadamente S. Lourenço na Serra da Estrela, Stª Maria no Marvão, Rainha Stª Isabel em Estremoz e ainda Stª Luzia, perto da fronteira.

	Segmentos-Frequência	Segmentos-Pousada
Tipo de destino		x
Quem escolheu	x	
Modo de efectuar a reserva	x	x
Avaliação da qualidade da estadia	n.a.	x
Avaliação do preço da estadia	n.a.	x
Intenção de voltar à Pousada	n.a.	x
Intenção de voltar a uma (qualquer) Pousada	x	
Alojamento nas últimas férias	x	
Férias ideais		
Local		x
Paisagem		x
Zona do País		x
Estação do ano	x	
Pousada ideal		
Localização	x	x
Zona do País		x
Paisagem		x
Ambiente	x	
Vida social	x	
Idade do respondente	x	

n.a.-não analisada por insuficiente dimensão da amostra

Tabela 8–Registos de diferenças significativas entre os segmentos

3.3 SOBRE A ORIGEM GEOGRÁFICA DOS CLIENTES DE CADA SEGMENTO

Reportando aos correspondentes dados presentes na Tabela 1 revelam-se diferenças significativas ao nível 0,073 de significância nas distribuições dos clientes de Lisboa, Porto e Setúbal pelos Segmentos de *Clientes frequentes*, *comuns* e *pela primeira vez*. É no Porto que se regista não só a maior percentagem relativa de Clientes frequentes (17%) como também a de Clientes pela primeira vez (26%).

Numa análise complementar foi calculado um Indicador distribuição geográfica dos *Segmentos-frequência* (cálculo semelhante ao efectuado em 2.2). De acordo com os valores observados na **Figura 2** evidencia-se a presença em Lisboa de qualquer dos segmentos e a presença em Setúbal do grupo de *Clientes comuns.*

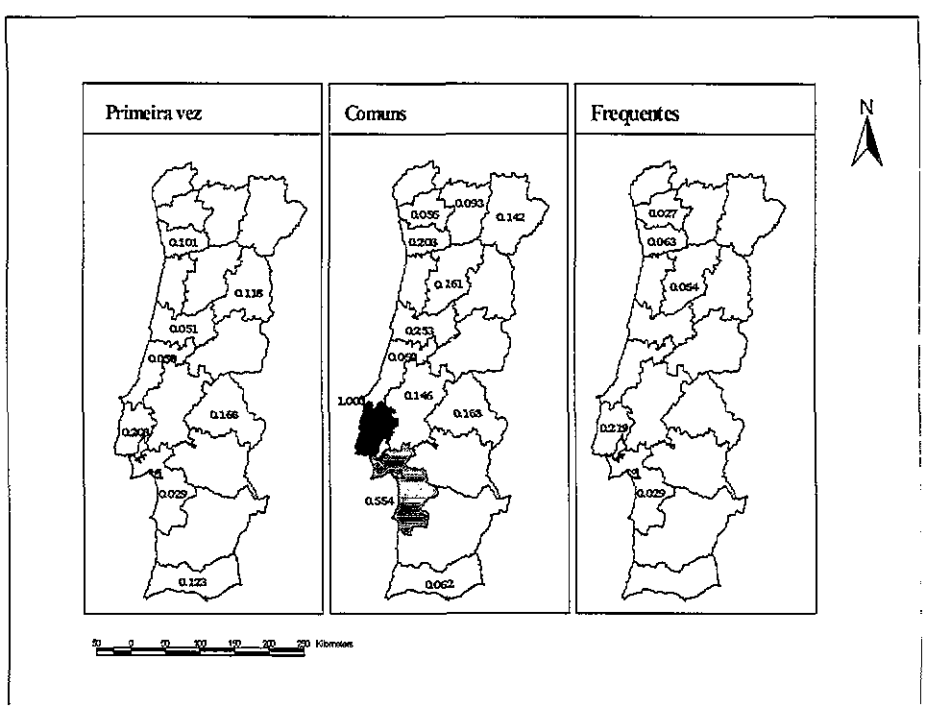


Figura 2 - **Indicador de distribuição geográfica dos Segmentos/dis**

4 SEGMENTOS-POUSADA

4.1 CARACTERIZAÇÃO

Tal como já foi referido os dados em análise correspondem a clientes de quatro unidades hoteleiras da ENATUR. As principais características das unidades em questão podem ser observadas na tabela seguinte:

	Quinta da Ortiga	Stª Cristina	S. Bento	Lóios
Classificação ENATUR (1997)	B	C	CSUP	CH
Monumento Nacional				x
Zona Histórica				
Regional	x	x	x	
Montanha			x	
Zona Arborizada	x		x	
Nº Quartos (1996)	10	45	29	30
Nº suites (1996)	3			2
Capacidade Total	13	45	29	32
Seminários e conferências		x		x
Piscina	x	x	x	x
Ténis		x	x	
Pesca	x		x	x
Desportos Náuticos	x		x	
Golfe				
Caça	x		x	x
Cavalos	x		x	x
Código Postal	7540	3150	4850	7000
Localidade	Santiago do Cacém	Condeixa-a-Nova	Vieira do Minho	Évora
Zona Turístico Promocional	Planícies	Costa de Prata	Costa Verde	Planícies

Tabela 9 - Caracterização das Pousadas da amostra.

De acordo com o estudo já efectuado sobre os respondentes alojados em cada Pousada foi possível identificar algumas diferenças significativas entre estes grupos/*Segmentos-Pousada*, diferenças cujos registos se resumem na Tabela 8.

A grande maioria dos respondentes, como já foi referido, está na Pousada a passar uns dias de férias. No entanto os *Segmentos-Pousada* diferem na apreciação do *tipo de destino* que cada Pousada representa. Assim, comparativamente, S. Bento (Gerês) é a Pousada mais citada como *um de vários destinos* (59%); Lóios (Évora) e Quinta da Ortiga (Santiago do Cacém) são as mais citadas como *destino único* da viagem de férias (58% e 55%, respectivamente); Stª Cristina (Condeixa-a-Nova) é a mais frequentemente citada como *ponto de passagem* (25%, contra percentagens de 7 a 16% para os grupos alojados nas outras Pousadas).

A forma de efectuar a reserva para os Segmentos-Pousada é também distinta: Lóios (tipo CH) é a Pousada onde se regista maior percentagem de reservas efectuadas

através de Agência (20%) ou Central de Reservas da ENATUR (49%). Os grupos nas restantes Pousadas efectuam, maioritariamente, a reserva na própria Pousada.

Sublinha-se, nos resultados obtidos, que os clientes das diferentes Pousadas têm opções significativamente diferentes no que respeita ao conceito de *Pousada ideal*, nomeadamente nas suas componentes geográficas. São ainda significativamente diferentes as avaliações no que diz respeito à qualidade e preço da estadia e ainda intenção de voltar à Pousada onde se encontram alojados os respondentes. Por outro lado, verificam-se ainda diferenças no que diz respeito às preferências dos diversos segmentos relativamente a *férias ideais*, diferenças que se concentram (tal como para a *Pousada ideal*), nas variáveis geográficas associadas a este conceito.

4.2 SOBRE A ORIGEM GEOGRÁFICA DOS CLIENTES DE CADA POUSADA

A distribuição dos clientes de Lisboa e Porto pelas Pousadas apresenta diferenças significativas ao nível 0,08 de significância: os respondentes de Lisboa distribuem-se pelas quatro Pousadas em percentagens que vão de 19% em Quinta da Ortiga a 32% em Lóios; os respondentes do Porto estão mais concentrados em S. Bento (44%) e menos em Lóios (13%).

Associados às regiões de origem dos clientes dispõem-se de valores para um Índice do Poder de Compra (já mencionado em 2.1) que permitiram estudar, nesta perspectiva, mais uma associação entre as regiões de origem e os destinos/Pousadas dos clientes na amostra. Assim, calculando uma média do Índice do Poder de Compra para os clientes de cada Pousada, conclui-se que é em Lóios (pousada CH, de nível de preço mais elevado) que este índice atinge o máximo (164). Quanto ao seu valor mínimo ele é atingido em S. Bento, uma Pousada emblemática da ENATUR. Note-se contudo que, para o nível de significância destas diferenças (0,00002) contribui, a maior ou menor concentração dos clientes oriundos de Lisboa.

S. Bento	129
Stª Cristina	146
Lóios	164
Quinta da Ortiga	150

Tabela 10 - Média do Índice do Poder de Compra dos clientes de cada Pousada tendo em conta a sua região de origem

Da análise do Indicador de distribuição geográfica dos clientes de cada Pousada por distrito (calculado como em 2.2) resultam os valores apresentados na Tabela 11.

Estes valores apresentam uma relação com o Índice de poder de compra *per capita* das regiões expressa do modo seguinte: o coeficiente de determinação associado à relação entre o Indicador e o Índice do poder de compra atinge o seu máximo de 85% quando referido aos clientes de Quinta da Ortiga (a mais barata), o

seu mínimo de 27% para a Pousada mais emblemática (S. Bento) e valores semelhantes (cerca de 75%) para as outras duas Pousadas da amostra.

	S. Bento	Stª Cristina	Lóios	Quinta da Ortiga
Aveiro	0	0	0	0
Beja	0	0	0	0
Braga	0,120	0,060	0	0
Bragança	0,309	0	0	0
Castelo Branco	0	0	0	0
Coimbra	0,331	0,331	0	0
Évora	0	0	0	0
Faro	0,135	0,135	0	0,135
Guarda	0,258	0	0	0
Leiria	0,108	0	0,108	0
Lisboa	0,727	0,773	1	0,614
Portalegre	0,734	0	0	0
Porto	0,359	0,166	0,110	0,166
Santarém	0,106	0,106	0	0,106
Setúbal	0,446	0,382	0,127	0,382
Viana do Castelo	0	0	0	0
Vila Real	0,202	0	0	0
Viseu	0	0	0	0
Madeira	0,542	0,181	0	0
Açores	0	0	0	0

Tabela 11 - Indicador de distribuição geográfica dos clientes de cada Pousada/distrito de origem

5 ATRACTIVIDADE DAS POUSADAS

A análise anteriormente efectuada sobre o Indicador de distribuição geográfica dos clientes de cada Pousada relativamente aos distritos de origem pode traduzir-se, numa perspectiva do lado da oferta, num Indicador do grau de penetração de cada Pousada nas regiões/distritos. Realça-se (v. Figura 3) deste ponto de vista, a capacidade de S. Bento cativar clientes de regiões dispersas por todo o continente e ainda da Madeira. Por outro lado, Stª Cristina mobiliza um número relativamente grande de clientes nos distritos de Coimbra e Setúbal, tendo em conta a sua população comparativamente à de Lisboa.

Como indicador resumido da capacidade de cada Pousada atrair clientes dos vários distritos foi utilizada uma medida aproximada das distâncias médias percorridas pelos clientes nessa Pousada. Este cálculo baseou-se numa matriz de distâncias entre as capitais de distrito. Em conclusão realça-se a capacidade de S. Bento atrair clientes que percorrem, em média a maior distância para visitar esta Pousada (v. Tabela 12). Esta conclusão, condizente com a que foi retirada a partir do grau de penetração das Pousadas por distrito, (indicador de atractividade usado acima), é reafirmada na análise de uma subamostra que se refere apenas aos clientes que apontam a Pousada

em que estão alojados com sendo o único destino da sua viagem de lazer (note-se que o facto de uma Pousada ser um de vários destinos ou apenas ponto de passagem poderia pôr em causa as conclusões retiradas sobre a atractividade das Pousadas).

S. Bento (Braga)	287
Stª Cristina (Coimbra)	185
Lóios (Évora)	166
Quinta da Ortiga (Évora)	176

Tabela 12- Média (aproximada) de Km percorridos pelos clientes das Pousadas

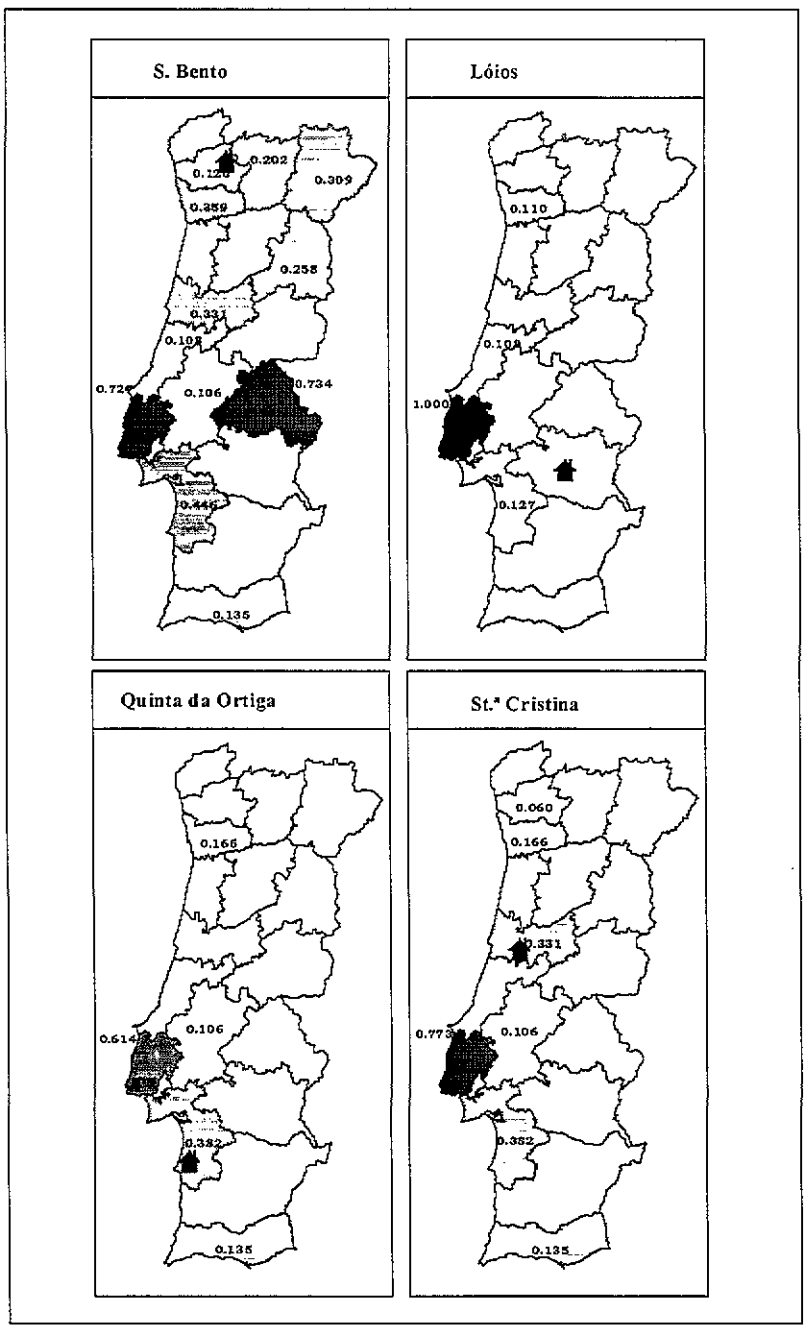


Figura 3 – Grau de penetração das Pousadas nos distritos

6 CONCLUSÕES E PERSPECTIVAS

Neste estudo apresenta-se uma análise geográfica como complemento de um processo de segmentação efectuado com base numa amostra de clientes portugueses das Pousadas de Portugal.

Conclui-se que estes clientes são maioritariamente oriundos dos distritos que apresentam um Índice de poder de compra *per capita* mais elevado: Lisboa (64%), Porto (14%) e Setúbal (10%). No entanto não se encontrou qualquer associação entre os rendimentos declarados pelos respondentes de uma região e o correspondente índice de poder de compra tendo-se chegado, por outro lado, à conclusão que, em cada região, são clientes das Pousadas as famílias de rendimentos elevados.

Verifica-se que a forma de efectuar a reserva está associada à região onde residem os clientes: os residentes em Lisboa pouco recorrem às Agências de Viagens optando igualmente entre a reserva na própria Pousada ou através da central de reservas; já os clientes do Porto usam mais a intermediação das Agências de Viagens, praticamente ignorando o serviço da Central de Reservas da ENATUR.

As preferências dos respondentes e, em particular os seus conceitos de Pousada ideal foram também analisados tendo-se constatado, a título de exemplo, que a paisagem de montanha é o cenário escolhido para a Pousada ideal da larga maioria da amostra, embora mais destacadamente pelos respondentes do Porto (diferenças estatisticamente significativas). Ainda o tipo de *vida social* associado à Pousada ideal varia também com a origem geográfica dos clientes: a maioria das preferências concentram-se nas opções *reservada* (31 %) e *moderada* (65 %), mas os clientes oriundos do Porto mostram preferir uma actividade social moderada (85%) em detrimento da *reservada* (12%).

O cruzamento da origem geográfica dos clientes com uma classificação desses mesmos clientes em *Segmentos-frequência - Clientes pela primeira vez, Clientes comuns e Clientes frequentes-* revelou uma forte presença de Lisboetas em todos os segmentos. O Porto detém uma representação acima da média nos segmentos dos *Clientes frequentes* e *primeira vez* sendo Setúbal um distrito origem de *Clientes comuns* (90%).

Comparando as origens dos clientes das quatro Pousadas incluidas na amostra – S. Bento, na região do Parque Natural do Gerês, Santa Cristina em Condeixa-a Nova, Lóios em Évora e Quinta da Ortiga, uma Pousada regional situada no Alentejo, em Santiago do Cacém, - é possível concluir que S. Bento, provavelmente pelas características ímpares da sua localização, atrai clientes provindo de quase todo o País, oriundos de distritos com uma maior diversidade do Índice do Poder de Compra *per Capita* e residindo em regiões mais distantes da Pousada.

O trabalho aqui presente, de análise das dimensões geográficas, será alargado à amostra de 2500 respondentes do inquérito efectuado junto dos clientes portugueses das Pousadas de Portugal. Nesse contexto tornar-se-ão viáveis análises mais desagregadas ao nível de cada região, permitindo estudar padrões de avaliação e satisfação dos clientes das Pousadas em relação às regiões onde estas se situam e tornando possível a análise da atractividade de todas as unidades hoteleiras. Este estudo permitirá certamente tirar ilações que serão úteis para apoiar a tomada de decisões estratégicas ao nível do marketing da ENATUR, abrangendo a análise da

configuração geográfica desta rede de unidades hoteleiras e, em geral, a configuração do marketing-mix dos produtos-Pousada.

Será ainda possível efectuar um estudo acerca de padrões de facto e padrões ideais de férias dos portugueses que se caracterizam por perfis socio-demográficos semelhantes ao da presente amostra, procurando retirar algumas conclusões para o País ao nível deste estrato social

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERECK, KATHLEEN L; CALDWELL, LINDA L. (1994). Variable selection in tourism market segmentation models; Journal of Travel Research, Fall, p. 40-46

CARDOSO,M.G.M.S.; THEMIDO I. (1998). Clientes Portugueses das Pousadas de Portugal: um estudo de segmentação; Publicação CESUR Nº 902

BAKER, KENNETH G.; HOZIER JR., GEORGE C.; ROGERS, ROBERT D. (1994). Marketing Research Theory and Methodology and the Tourism Industry: a Nontechnical Discussion; Journal of Travel Research, Winter, p.3-7

JOHNSON, DOUGLAS (1996). Some observations on, and developments in the analysis of multivariate survey data; Journal of the Market Research Society, Vol. 38-Nº 4, October, pp. 449-471

McELROY, JEROME L.; ALBUQUERQUE, KLAUS DE (1998). Tourism penetration index in small caribbean islands; Annals of tourism research, January, Vol. 25 – Nº1 , pp. 145-168

OPPERMAN, MARTIN (1997). Geography's Changing Role in Tourism Marketing; in Opperman, Martin. *Geography and Tourism Marketing*. The Haworth Press, Inc.New York, London, p. 1-3

SPOTTS, DANIEL M. (1997). Regional analysis of tourism resources for marketing purposes; Journal of Travel Research, Winter, pp. 3-15

STEMERDING, MARCUS P.; OPPEWAL, HARMEN; BECKERS, THEO A. M.; TIMMERMANS, HARRY J. P. (1996). Leisure market segmentation: an integrated preferences/constraints-based approach; Journal of Travel and Tourism Marketing. Vol.5, Nº 3, pp. 161-185

TAYLOR, GORDON D. (1996). A new direction; Journal of Travel and Tourism Marketing. Vol.5, Nº 3, pp. 253-263

WEDEL, M.; KAMAKURA,W. (1998). Market Segmentation: Conceptual and Methodological Foundations; Kluwer Academic Publishers

VOLUME 2
2º QUADRIMESTRE DE 1998

REVISTA DE ESTATÍSTICA
103ª PAGINA
2º QUADRIMESTRE DE 1998

A AVALIAÇÃO DA QUALIDADE NOS RECENSEAMENTOS DA POPULAÇÃO E HABITAÇÃO DE 2001 EM PORTUGAL

Autor:
Fernando Simões Casimiro

VOLUME 2
2º QUADRIMESTRE DE 1998

REVISTA DE ESTATÍSTICA
105ª PAGINA
2º QUADRIMESTRE DE 1998

A AVALIAÇÃO DA QUALIDADE NOS RECENSEAMENTOS DA POPULAÇÃO E HABITAÇÃO DE 2001 EM PORTUGAL[7]

QUALITY EVALUATION IN THE 2001 POPULATION AND HOUSING CENSUSES IN PORTUGAL[8]

Autor: Fernando Simões Casimiro
Director do Gabinete dos Censos 2001 - Instituto Nacional de Estatística

RESUMO:

- Os dados dos Recenseamentos da População e Habitação de 1991 (Censos 91) constituíram uma verdadeira surpresa para a maioria dos utilizadores. As principais causas desta surpresa foram as estimativas da população feitas pelo INE durante a década de oitenta e a forte expectativa dos municípios quanto aos valores da respectiva população residente, baseadas nos dados do recenseamento eleitoral.

  As estimativas da população são calculadas, fundamentalmente, com os dados do último censo da população e com os saldos natural e migratório de cada ano.

  Por outro lado, a base do recenseamento eleitoral existente era de 1979 e os procedimentos de actualização não eram suficientemente fiáveis devido a mortos e mudanças de residência não abatidas, quer fosse por nova inscrição ou emigração.

  A suspeita de um elevado erro de cobertura (cerca de 5% para a população residente) nos Censos 91 tem subsistido, embora o respectivo inquérito de qualidade tenha dado um erro líquido de cobertura da população residente de cerca de 1%.

  Assim, para os Recenseamentos da População e Habitação de 2001 (Censos 2001), estamos a preparar um Programa de Controlo e Avaliação dos respectivos dados cujo principal objectivo consiste na produção de indicadores sobre a qualidade dos Censos 2001 que sejam inquestionáveis para os utilizadores.

PALAVRAS-CHAVE:

- *Recenseamento da população, Censos 91, Censos 2001, estimativa, recenseamento eleitoral, inquérito de qualidade, cobertura, conteúdo.*

ABSTRACT:

- Data coming from the 1991Population and Housing Censuses have been very surprising for the majority of the users. The key elements for that kind of surprise were the population estimates made by National Statistical Institute (NSI) during the eighty decade and a strong expectation of municipalities based on the electoral roll.

[7] Documento apresentado no Seminário Internacional sobre Metodologias dos Recenseamentos da População, realizado em Portsmouth (R. U.), de 30 de Abril a 1 de Maio de 1998.

[8] Paper submitted to the International Seminar on Census Methodology, April 30-1 May 1998, Portsmouth (U. K.).

The population estimates are based on the balance of net migrations and natural increase.

On the other way, the electoral roll basis was built in 1979 and updating procedures are not enough reliable due to deaths not erased, external migration flow without residence change and differences between the two residence concepts (population census and electoral roll).

The suspicion of significant underenumeration (about 5%) in the 1991 Population Census remains up to now although the respective post enumeration survey pointed out a net underenumeration of 1% for resident population.

So for 2001 Population and Housing Censuses we are designing an exhaustive control and evaluating programme for which the main objective is to get quality indicators for 2001 Population and Housing Censuses, out of question to users.

KEY-WORDS:

- *Population census, 91 Censuses, 2001 Censuses, estimate, electoral roll, post enumeration survey, coverage, content.*

1- INTRODUÇÃO

Os dados dos Recenseamentos da População e Habitação de 1991 (Censos 91) constituíram uma verdadeira surpresa para a maioria dos utilizadores. As principais causas desta surpresa foram as estimativas da população feitas pelo INE durante a década de oitenta e a forte expectativa dos municípios quanto aos valores da respectiva população residente, baseadas nos dados do recenseamento eleitoral.

As estimativas da população são calculadas, fundamentalmente, com os dados do último censo da população e com os saldos natural e migratório de cada ano. Enquanto o saldo natural era calculado com base nos dados de óbitos e nascimentos provenientes dos registos administrativos e pode considerar-se como tendo um erro praticamente nulo, o saldo migratório era o resultado da equação da concordância e de outros dados provenientes de várias fontes como sejam os registos administrativos de emigrantes, que deixaram de existir a partir de 1987, e do Inquérito ao Emprego; de facto, não existia uma fonte estatística suficientemente fiável para a produção de dados sobre as migrações internas e externas.

Por outro lado, a base do recenseamento eleitoral existente era de 1979 e os procedimentos de actualização não eram suficientemente fiáveis devido a mortos e mudanças de residência não abatidas, quer fosse por nova inscrição ou emigração. De facto, enquanto no recenseamento da população a residência que é observada é aquela onde as pessoas passaram a maior parte do ano, no recenseamento eleitoral é possível ter pessoas inscritas como residentes em Portugal e que trabalham e vivem no estrangeiro. Devido a razões políticas, o recenseamento eleitoral foi objecto de um procedimento especial de actualização (nova lei e novo sistema de controlo administrativo) durante o primeiro semestre de 1998.

A suspeita de um elevado erro de cobertura (cerca de 5% para a população residente) nos Censos 91 tem subsistido, embora o respectivo inquérito de qualidade tenha dado um erro líquido de cobertura da população residente de cerca de 1%.

Assim, para os Recenseamentos da População e Habitação de 2001 (Censos 2001), estamos a preparar um Programa de Controlo e Avaliação dos respectivos dados, o qual se baseia nos seguintes elementos:

- Um "Sistema de Alerta", baseado num conjunto de indicadores relacionados com os saldos demográficos, o recenseamento eleitoral, a informação geográfica suportada por mapas digitalizados e contagem de células pretas referentes a edifícios, população escolar, moradas postais e consumidores de electricidade, para cada nível de desagregação geográfica máximo possível; este sistema permitir-nos-á um valor esperado para cada unidade estatística coberta no respectivo nível de desagregação geográfica; se o valor esperado não for atingido ou ultrapassado para além de um determinado desvio, deverá ser investigada a razão desse desvio;
- Procedimentos de controlo para a selecção e formação dos intervenientes locais, recolha, registo, codificação e imputação de dados;
- Inquérito de Qualidade, com o objectivo de produzir dados sobre a qualidade da cobertura e do conteúdo e com uma amostra significativa para cada região

NUTS II; os dados deste inquérito deverão constituir o elemento-base de avaliação da qualidade dos Censos 2001; para assegurar uma maior garantia de independência, aos utilizadores, sobre os resultados finais deste inquérito de qualidade, prevemos ter a participação de uma entidade externa, independente, prestigiada e proveniente do meio científico;

- Uma publicação com a análise comparativa dos resultados dos Censos 2001 com os dados de outras fontes utilizadas habitualmente pelos utilizadores; este procedimento permitir-nos-á antecipar e analisar eventuais incompatibilidades dos respectivos dados e facilitar a utilização adequada de comparações de dados entre várias fontes.

Com este Programa de Controlo e Avaliação da Qualidade pretendemos atingir o seguinte objectivo: ter indicadores sobre a qualidade dos Censos 2001 que sejam inquestionáveis para os utilizadores.

2 - O PROCESSO DE AVALIAÇÃO DA QUALIDADE DO CENSO EM 1991

A avaliação da qualidade dos Censos 91 baseou-se, sobretudo, na análise comparativa com outras fontes de dados e nos resultados obtidos com o Inquérito de Qualidade.

Uma vez que não existiam outras fontes de dados suficientemente fiáveis para analisar a parte da habitação, a análise comparativa com outras fontes centrou-se nos dados referentes à população, com especial destaque para os saldos naturais e migratórios.

De acordo com os dados apresentados no anexo 1, em 1981, Portugal tinha um saldo natural anual de 56.330 pessoas, enquanto em 1991 esse saldo tinha baixado para 12.440 pessoas, o que equivale a 22% do saldo que existia 10 anos antes; este decréscimo é contínuo ao longo da década e apenas a suposição de que se estava a verificar um forte saldo migratório positivo, provocado pelo retorno de emigrantes portugueses e pela entrada de imigrantes vindos das ex-colónias portuguesas, levava a estimar uma população que já teria ultrapassado os 10.300.000 habitantes em 1989. Os "picos" imigratórios calculados antes da correcção do censo posterior situaram-se em 1983, 1984 e 1985 e contrastam com os valores dos anos imediatamente anterior (1982) e posterior (1986); assim, enquanto em 1982 este saldo migratório correspondia a 18100 pessoas, em 1983 e 1984 passou para cerca de 33000 pessoas; em 1985 é de 22900 e em 1986 passa para 13700 pessoas, o que denota variações relativas bastante fortes. Após a correcção deste saldo com os dados dos Censos 91, os respectivos valores passam a ser crescentemente negativos e as maiores diferenças entre o estimado antes e depois da correcção situam-se nos anos de 1987 a 1989.

Por outro lado, o recenseamento eleitoral (inscrição obrigatória de todos os cidadãos portugueses com 18 ou + anos de idade) que é actualizado anualmente e cuja base se reporta a 1979, tinha 7.098.492 inscritos em 1981, enquanto em 1991 tinha passado para 8.341.192 eleitores o que corresponde a um crescimento de 17,5%. Convém assinalar que o recenseamento da população de 1981 tinha -5% de população portuguesa com 18 ou + anos do que o recenseamento eleitoral do mesmo ano, o que

já permitia perceber que existiam diferenças importantes entre as duas fontes e apesar de apenas terem passado dois anos sobre a base do recenseamento eleitoral. Em 1991, a diferença entre as duas fontes, para a população equivalente, é de -12,5% no recenseamento da população, o que deixou os utilizadores com suspeitas sobre a qualidade da cobertura dos Censos 91. Contudo, recentemente a maior parte dos partidos políticos reconheceu a necessidade de fazer uma actualização extraordinária do recenseamento eleitoral por se verificar que não estavam efectuados os abates de muitos cidadãos que morreram ou mudaram de residência.

O Inquérito de Qualidade dos Censos 91 foi realizado em 262 secções estatísticas, as quais foram objecto de uma segunda observação exaustiva em todas as respectivas áreas; estas 262 secções correspondem a uma taxa de 2% de todas as unidades estatísticas observadas nos Censos 91. Apesar do esforço feito com a realização deste inquérito de qualidade, verificou-se que os dados sobre o conteúdo tinham falhas importantes de resposta que inviabilizaram o apuramento dos erros de conteúdo. Assim, apenas foram apurados os erros líquidos de cobertura de cada unidade estatística observada; as taxas líquidas de cobertura para cada unidade estatística são as seguintes:

- Edifício 99,60 %
- Alojamento 99,42 %
- Família 99,24 %
- Indivíduo 99,04 %

Face às estimativas existentes, estes dados foram de alguma forma surpreendentes; contudo, em operações estatísticas de grande dimensão, realizadas em 1994 e 1996 e com uma metodologia de observação baseada na amostragem areolar e com secções estatísticas iguais às utilizadas nos Censos 91, verificou-se que nestas secções havia menos população do que a encontrada nos Censos 91.

3 - QUE MODELO DE AVALIAÇÃO PARA 2001?

Devido ao importante efeito de surpresa gerado pelos dados censitários de 1991, estamos convencidos de que os Censos 2001 irão ser objecto de uma observação bastante atenta por parte dos principais utilizadores.

De certa maneira os dados censitários de 2001 acabarão por constituir, ainda que passados dez anos, um importante factor de avaliação dos Censos 91 uma vez que não houve nem se prevê qualquer "acidente" demográfico significativo no processo de evolução demográfica do país.

Assim, o controlo e a avaliação da qualidade vão ser assumidos como tarefas fundamentais na realização dos Censos 2001, de modo a garantir que o processo produtivo detecte, em tempo oportuno, todos as falhas existentes e que o resultados definitivos sejam tendencialmente inquestionáveis quanto à sua qualidade.

As Autarquias Locais são entidades quase imprescindíveis na execução destas operações estatísticas, tanto pelo conhecimento pormenorizado que têm do terreno como pelo facto de serem utilizadores crescentemente interessados nos respectivos resultados; os aspectos mais marcantes neste envolvimento são os seguintes:

- Muitos dos indicadores utilizados no cálculo das transferências financeiras da Administração Central para a Local são estimados com base em dados provenientes destes recenseamentos;
- O controlo e planeamento da ocupação do território é regulado pelos Planos Directores Municipais (PDMs) e constitui uma das mais importantes funções da Administração Local, pelo impacto actual e futuro na qualidade de vida dos respectivos cidadãos;
- A maioria dos municípios está preocupada com a evolução da respectiva população, ou porque a estão a perder, ou porque os afluxos populacionais descontrolados geram efeitos marginais na ocupação do território.

Deste modo, o envolvimento das autarquias locais nestas operações estatísticas, para além de constituir um elemento facilitador da sua execução, pelo conhecimento que possuem, também permite introduzir um factor de avaliação e corresponsabilização dos respectivos resultados.

Assim, paralelamente ao envolvimento das Autarquias Locais, enquanto parceiros a privilegiar em todo o processo executivo, estamos a desenhar um Programa de Controlo e Avaliação da Qualidade que assenta em três pilares:

- Controlo do processo executivo;
- Inquérito de qualidade;
- Análise de índices e de fontes alternativas de dados.

Para garantir a total utilização dos indicadores de qualidade, prevemos elaborar uma publicação com a análise exaustiva de todos estes indicadores, de forma que os utilizadores possam estar completamente informados da qualidade e das eventuais limitações dos dados que lhes forem disponibilizados.

3.1 - CONTROLO DO PROCESSO PRODUTIVO

O controlo do processo produtivo assume características algo semelhantes ao de qualquer outra cadeia de produção, através da implementação de instrumentos de controlo da qualidade dos vários elementos passíveis de cometerem erros.

As actividades previamente seleccionadas para a implementação destes controlos são as seguintes:

- Selecção, formação e avaliação técnica dos intervenientes regionais e locais;
- Distribuição e recolha dos questionários;
- Tratamento dos dados.

Quanto à selecção, formação e avaliação técnica dos intervenientes regionais e locais, para além dos instrumentos de avaliação dos conhecimentos possuídos e adquiridos e da implementação de níveis mínimos e admissíveis de conhecimentos,

procurar-se-á implementar um sistema remuneratório que permita tornar suficientemente atractiva a execução deste trabalho a nível local e regional. Além disso, serão estruturados modelos de formação que garantam a aquisição de conhecimentos em condições de tempo tão reduzido quanto possível, mas com duração intrínseca e standardizada para garantir um padrão de formação adaptado aos conhecimentos necessários, em termos médios. Assim, os instrumentos de controlo a considerar para esta fase são os seguintes:

- Teste de selecção para candidatos à formação;
- Controlos de duração da formação;
- Modelos de avaliação técnica dos conhecimentos adquiridos;
- Modelos de acompanhamento da formação prática.

A distribuição e recolha dos questionários constitui a tarefa central e mais fortemente influenciadora da qualidade destes recenseamentos; todos os instrumentos de controlo da qualidade são fundamentais para esta fase, mas não devem ceder à tentação fácil de, a pretexto de se obter um controlo "férreo", acabar por inviabilizar o fluir normal dos trabalhos, introduzindo uma carga de trabalho excessivo no processo produtivo. Assim, para além do controlo individual do trabalho realizado, aposta-se na construção de um "sistema de alerta" baseado em "valores esperados" ao nível máximo possível da desagregação geográfica; este último sistema permitirá detectar e reverificar situações com fortes divergências entre dados provenientes de fontes diferentes. Os instrumentos de controlo a considerar para esta fase são os seguintes:

- Nos controlos individuais, é feita a verificação, pelo coordenador, de indicadores de cobertura e de conteúdo de uma amostra dos questionários distribuídos e recolhidos por cada agente recenseador; o instrumento a utilizar nesta verificação será validado pela hierarquia superior do coordenador e constituirá elemento necessário para pagamento;
- Para o sistema de indicadores de verificação e alerta, será construída uma base de dados provenientes de várias fontes (BGRI, CTT, EDP, Recenseamento Eleitoral, população escolar, saldos naturais da década, e outras que oportunamente estejam disponíveis), para estabelecer um valor esperado para a população e alojamentos.

No tratamento dos dados, prevê-se apostar no aumento da automatização em relação a 1991, nomeadamente na utilização da leitura óptica, no alargamento da codificação automática dos campos alfabéticos (melhorando o sistema C91, implementado com os Censos 91) e no reforço das correcções automáticas. Contudo, é previsível que em todos estes sistemas existam limiares de decisão que coloquem problemas de qualidade. Alguns dos instrumentos de controlo a implementar são os seguintes:

- Contagem de "campos" lidos sem dúvidas;
- Contagem de "campos" lidos com dúvidas;
- Contagem de "campos" brancos (respostas imputadas automaticamente);
- Contagens de correcções efectuadas;
- Controlo, por amostra, dos registos criados.

3.2 - INQUÉRITO DE QUALIDADE

O inquérito de qualidade é o verdadeiro instrumento de medida definitiva da qualidade dos Censos 2001 e, por isso, deverá ser rodeado de cuidados muito especiais, tanto na sua preparação como na sua execução.

Este inquérito é desenhado e executado tendo em conta os seguintes objectivos:

- Avaliar os erros de cobertura, bruto e líquido, de cada uma das unidades estatísticas observadas;
- Avaliar os erros de conteúdo das variáveis e respectivas modalidades, referentes a cada unidade estatística.

Habitualmente, este segundo objectivo não abrange todas as variáveis em causa, havendo necessidade de seleccionar as que são mais importantes e representativas para a análise de conteúdo.

A execução deste inquérito baseia-se no princípio de que a segunda observação (inquérito de qualidade) funciona como modelo de "informação correcta", sendo as diferenças em relação à primeira observação (recolha censitária) consideradas como erros de cobertura ou conteúdo; assim a qualidade e isenção da segunda observação obriga a um conjunto importante de cuidados que vão desde a qualidade da preparação dos respectivos entrevistadores até ao processo de execução, tratamento e mesmo auditoria externa que reforce a credibilidade deste inquérito.

Deste modo, prevê-se associar, a este processo do inquérito de qualidade, uma entidade externa ao INE, de reconhecido prestígio e proveniente da área científica, para garantir que os resultados sejam assumidos como devidamente auditorados e suficientemente fiáveis para os utilizadores.

Por outro lado, está a desenhar-se uma tendência crescente de correcção das estimativas intercensitárias da população, incorporando os erros de cobertura e conteúdo verificados, pelo que se deve garantir o adequado rigor dos resultados do inquérito de qualidade.

3.2.1 - PRINCIPAIS CARACTERÍSTICAS METODOLÓGICAS DO INQUÉRITO DE QUALIDADE

Embora o Programa de Controlo e Avaliação da Qualidade ainda não esteja totalmente concluído, alguns dos procedimentos técnicos do Inquérito de Qualidade (I.Q.) são relativamente standardizados e estamos a prever os seguintes:

- A amostra deste inquérito será de tipo areolar e representativa de todas as unidades estatísticas e das principais características de cada uma delas, ao nível de NUTS II (7 regiões);

- A recolha de dados da segunda observação será feita após a conclusão da recolha da primeira observação e por uma pessoa diferente e desconhecida da que fez a primeira observação;
- Enquanto a avaliação da cobertura é feita com todas as unidades estatísticas observadas na mesma área, a avaliação do conteúdo é feita, apenas, com as mesmas unidades estatísticas que foram observadas em ambas as observações.

3.2.1.1 - AVALIAÇÃO DA COBERTURA

Os indicadores da avaliação da cobertura dos dados censitários constituem os elementos centrais de formação da opinião pública definitiva sobre a qualidade censitária. A avaliação deste tipo de indicadores deverá obedecer a um modelo de apuramento que permita conhecer os erros de cobertura brutos e líquidos, para se fazer a distinção entre as unidades estatísticas efectivamente não recenseadas e a estimativa das unidades recenseáveis.

Assim, a avaliação dos erros de cobertura deverá apoiar-se no seguinte modelo de quadro:

Quadro 3.1

U. E. (unidade estatística)	TOTAL	%
U. E. recenseadas	X1	100
U. E. correctamente recenseadas	X2	X2/X1*100
U. E. indevidamente recenseadas	X3	X3/X1*100
U. E. observadas no inquérito	X4	X4/X1*100
U. E. omitidas no recenseamento	X5	X5/X1*100
Diferença líquida entre o recenseamento e o inquérito	X1-X4	(X1-X4)/X1*100
Diferença bruta entre o recenseamento e o inquérito	X3+X5	(X3+X5)/X1*100

No modelo de cálculo das células deste quadro assume-se que não são feitas correcções aos dados definitivos do recenseamento, razão pela qual o denominador é sempre X1; se a opção for a correcção daqueles dados, então o denominador deverá ser X4.

3.2.1.2 - AVALIAÇÃO DO CONTEÚDO

A avaliação dos erros de conteúdo constitui um elemento fundamental de apreciação da qualidade técnica "intrínseca" destes recenseamentos; de facto podemos ter todas as unidades recenseadas, mas com características diferentes das que realmente possuem, o que conduz a erros de conteúdo na informação que se pretende recolher com estas operações estatísticas; por vezes estes erros de conteúdo podem ser interpretados como sendo de cobertura. Um dos problemas típicos desta situação é o "conflito" entre o auto-preenchimento de um questionário e a recolha de dados por entrevista directa na variável "situação perante a actividade económica", para um reformado que apenas trabalhou uma hora na semana de referência; na primeira situação tenderá a considerar-se como reformado, enquanto na segunda, através da capacidade "crítica" do entrevistador, poderá ser considerado como população activa.

Assim, a avaliação dos erros de conteúdo deverá apoiar-se nos seguintes modelos de quadro:

Quadro 3.2

Censo / IQ	a)	b)	c)	unidades correctamente incluídas
a)	n11	n12	n13	n11
b)	n21	n22	n23	n22
c)	n31	n32	n33	n33
unidades correctamente incluídas	n11	n22	n33	n

- Reside e está presente (a);
- Reside mas está ausente (b);
- Temporariamente presente e residente noutro local (c).

Quadro 3.3

Censo / IQ	Com a modalidade "RESIDENTE MAS AUSENTE"	Com a modalidade "RESIDENTE MAS AUSENTE"	TOTAL
Com a modalidade "residente mas ausente"	a	b	a+b
Sem a modalidade "residente mas ausente"	c	d	c+d
TOTAL	a+c	b+d	n

Este quadro permite analisar a qualidade do conteúdo da resposta a cada uma das modalidades de cada variável; assim, seleccionámos a modalidade "residente, mas está ausente" (residente ausente) para verificar a consistência dos respectivos dados, através da comparação das respostas das mesmas pessoas ao censo e ao inquérito de qualidade.

Com os dados provenientes destes dois tipos de quadros é possível construir um conjunto alargado de indicadores que permitem avaliar a qualidade da resposta a cada variável e a cada tipo de população observada.

3.3 - ANÁLISE DE ÍNDICES E DE FONTES ALTERNATIVAS DE DADOS

Estes recenseamentos são habitualmente objecto de estudos e da produção de um conjunto de índices que permitem avaliar a coerência de uma parte do seu conteúdo, em termos qualitativos; por outro lado, a disponibilização dos dados definitivos proporciona a comparação com os dados provenientes de outras fontes, o que leva os utilizadores a estabelecer comparações que poderão ser melhor interpretadas e analisadas se lhes forem proporcionados indicadores de comparabilidade, através de um documento ou publicação onde se faça, à partida, essa análise. Assim, os indicadores de análise da qualidade a disponibilizar no âmbito desta parte são os seguintes:

- Produção e análise comparativa dos Índices de Whipple, de Irregularidade, Combinado das Nações Unidas, e de Myers, e da Equação da Concordância, para os três últimos recenseamentos realizados; esta análise deverá ser regionalizada, pelo menos, até NUTS II;
- Análise comparativa dos resultados dos Censos 2001 com outras fontes de dados disponíveis e passíveis de serem comparados com os censitários (população escolar, nados-vivos dos últimos anos, recenseamento eleitoral, etc.).

4 - POR UM MODELO HARMONIZADO DE INDICADORES DE QUALIDADE PARA A PRÓXIMA SÉRIE DE RECENSEAMENTOS NA U.E.

A produção de indicadores de qualidade é uma exigência crescente dos principais utilizadores estatísticos e tenderá a generalizar-se aos organismos internacionais que utilizam dados provenientes de vários países e de fontes equivalentes.

No âmbito da União Europeia, os dados sobre nacionalidade e migrações assumem uma importância bastante grande tanto para os países de destino, como para os de origem; as estimativas intercensitárias da população e todos os projectos estatísticos que elas suportam, como o Inquérito ao Emprego, poderão ser crescentemente "influenciados" pelo Acordo de Schengen, na medida em que este estabelece a liberdade de circulação no espaço comunitário. Deste modo, o saldo

migratório anual, além de ser o elemento de cálculo mais difícil para as estimativas, é também o mais influenciado pelos erros de cobertura e de conteúdo destes recenseamentos.

Por outro lado, assiste-se a uma tendência crescente para a correcção dos dados censitários em função dos respectivos erros, sobretudo os de cobertura. Esta opção poderá representar uma visão algo diferente da tradicional, pois enquanto se torna relativamente fácil fazer correcções nos níveis agregados, o mesmo não se passa para as pequenas áreas estatísticas, para as quais estes recenseamentos constituem a principal fonte de dados estatísticos.

Deste modo, parece-nos oportuna a sugestão de um conjunto de medidas que, para além de permitirem comparar a qualidade dos recenseamentos nos vários países que os realizam, permitem uma melhor avaliação de algumas variáveis demográficas que influenciam e podem ajudar a explicar algumas "surpresas censitárias". O essencial destas medidas decorre da execução de um inquérito de qualidade com procedimentos de tratamento e apresentação dos respectivos dados de forma harmonizada. Destacamos as seguintes propostas:

- Realização de um inquérito de qualidade que permita apresentar dados sobre a qualidade da cobertura e do conteúdo dos respectivos recenseamentos;
- Apresentação de indicadores sobre a cobertura bruta e liquida das unidades estatísticas observadas;
- Avaliação da cobertura para as nacionalidades, pelo menos da União Europeia;
- Apresentação de indicadores da qualidade do conteúdo, pelo menos, para as seguintes variáveis e respectivas modalidades:
  - Situação perante a residência;
  - Residência anterior, para todos os períodos observados, com destaque para as residências anteriores no estrangeiro e com frequências superiores a 2000 pessoas;
  - Nacionalidade, para todas as modalidades que apresentem frequências superiores a 2000 pessoas;
  - Naturalidade, com tratamento idêntico ao da nacionalidade.

Se chegarmos a uma dimensão harmonizada dos erros de cobertura e de conteúdo, deverá ou não fazer-se a correcção dos respectivos dados censitários?

As correcções de dados censitários sempre serviram de base a discussões metodológicas sem que se tivessem efectuado alterações nos dados efectivamente apurados nos respectivos recenseamentos. Em Portugal, as razões destas contestações têm sido dominadas por motivos políticos e tenderão a agravar-se com a previsível estabilização geral da população, associada a alguns decréscimos acentuados em determinadas regiões. Por outro lado, se as correcções podem não levantar problemas significativos ao nível nacional e até mesmo de NUTS II, já para desagregações mais finas podem conduzir a imputações inconsistentes sobretudo para pequenas populações.

Daí que uma eventual solução harmonizada seja a "incorporação" dos principais erros da qualidade destes recenseamentos nas estimativas demográficas anuais, de modo a obter-se uma população de referência já devidamente corrigida.

REFERÊNCIAS BIBLIOGRÁFICAS

INE (1991) – Metodologia para o Inquérito de Qualidade dos Censos-91 – Lisboa – Portugal

ANEXO 1 – EVOLUÇÃO DE ALGUNS INDICADORES DA POPULAÇÃO ENTRE 1981 E 1997[1]

	1981	1982	1983	1984	1985
1	2	3	4	5	6
População Residente (censo)	9 833 014				
População Residente (estimada antes da correcção com o censo posterior)	9 891 900	9 968 600	10 049 700	10 128 900	10 185 100
População Residente (estimada após a correcção com o censo posterior)	9 883 940	9 939 110	9 969 940	10 008 530	10 014 300
Saldo natural	56 330	58 620	48 110	45 800	33 350
Saldo migratório externo (estimado antes da correcção com o censo posterior)	16 600	18 100	33 000	33 400	22 900
Saldo migratório externo (estimado após a correcção com o censo posterior)	8 630	-3 450	-17 280	-7 210	-27 580
População eleitoral portuguesa (≥18 anos)	7 098 492	7 183 817	7 261 633	7 458 966	7 606 087
População portuguesa no censo (≥18 anos)	6 745 117				

ANEXO 1 – EVOLUÇÃO DE ALGUNS INDICADORES DA POPULAÇÃO ENTRE 1981 E 1997[1]

(continuação)

	1986	1987	1988	1989	1990	1991
1	7	8	9	10	11	12
População Residente (censo)						9 867 147
População Residente (estimada antes da correcção com o censo posterior)	10 230 000	10 270 000	10 304 700	10 337 000	[2]9 858 600	9 864 890
População Residente (estimada após a correcção com o censo posterior)	10 007 050	9 981 360	9 955 050	9 919 690	9 877 480	
Saldo natural	31 180	28 060	24 240	22 730	13 520	12 404
Saldo migratório externo (estimado antes da correcção com o censo posterior)	13 700	11 900	10 500	9 500	-33 100	-25 001
Saldo migratório externo (estimado após a correcção com o censo posterior)	-38 430	-53 750	-50 550	-58 090	-55 730	
População eleitoral portuguesa (≥18 anos)	7 773 132	7 915 566	8 010 857	8 134 976	8 255 726	8 341 192
População portuguesa no censo (≥18 anos)						7 300 476

ANEXO 1 – EVOLUÇÃO DE ALGUNS INDICADORES DA POPULAÇÃO ENTRE 1981 E 1997[1]

(continuação)

	1992	1993	1994	1995	1996	1997
1	13	14	15	16	17	18
População Residente (censo)						
População Residente (estimada antes da correcção com o censo posterior)	9 869 170	9 892 160	9 912 140	9 920 760	9 934 110	
População Residente (estimada após a correcção com o censo posterior)						
Saldo natural	14 270	7 999	9 981	3 620	3 362	
Saldo migratório externo (estimado antes da correcção com o censo posterior)	-10 000	15 000	9 999	5 000	10 000	
Saldo migratório externo (estimado após a correcção com o censo posterior)						
População eleitoral portuguesa (≥18 anos)	8 406 194	8 555 733	8 669 014	8 743 287	8 815 520	8 926 129
População portuguesa no censo (≥18 anos)						

[1] **Os dados aqui apresentados foram retirados das publicações dos Censos 81 e 91 (INE), das séries sobre Estimativas da População (INE) e das publicações anuais do Recenseamento Eleitoral (STAPE); qualquer eventual diferença destes dados em relação aos disponíveis é da exclusiva responsabilidade do autor.**

[2] **Esta estimativa já foi calculada com os dados preliminares dos Censos 91**

REVISTA DE ESTATÍSTICA
119ª PAGINA
2º QUADRIMESTRE DE 1998

INFORMAÇÕES

VOLUME 2
2º QUADRIMESTRE DE 1998

ACTIVIDADES E PROJECTOS IMPORTANTES NO ÂMBITO DO SISTEMA ESTATÍSTICO NACIONAL

IMPORTANTS ACTIVITIES AND PROJECTS IN THE SCOPE OF THE NATIONAL STATISTICAL SYSTEM

ÍNDICES DE VOLUME DE NEGÓCIOS, EMPREGO, REMUNERAÇÕES E HORAS TRABALHADAS NA INDÚSTRIA

SÍNTESE METODOLÓGICA

O presente artigo tem como objectivo, exclusivo, contribuir para a divulgação da metodologia de cálculo dos Índices de Volume de Negócios, Emprego, Remunerações e Horas Trabalhadas na Indústria (base 1995), indicadores estes que, no seu conjunto, formam o projecto estatístico adiante designado por IVNEI.

Na prossecução deste objectivo será apresentado, de seguida, um breve enquadramento do contexto no âmbito do qual este projecto foi implementado, ao que se seguirá a descrição do campo sectorial coberto, dos resultados apurados, das amostras consideradas, dos conceitos empregues e das fórmulas utilizadas.

1. ENQUADRAMENTO

A implementação do IVNEI (base 1995) pretendeu dar continuidade à versão anterior deste projecto, em base 1990, mantendo como objectivo central a medição da evolução mensal, no âmbito da indústria, de quatro variáveis: o volume de negócios (total, realizado com o mercado nacional e realizado com o mercado externo), o emprego, as remunerações e as horas trabalhadas.

Um outro objectivo deste projecto que convirá mencionar, se bem que secundário, é o de produzir informação relevante para efeitos de cálculo do Índice de Produção Industrial (IPI), base 1995, de entre a qual se destaca o valor mensal das vendas realizadas de produtos acabados e intermédios (valor este que, conjuntamente com o valor das vendas de mercadorias e o valor da prestação de serviços, integram o conceito volume de negócios).

A adopção da nova base pelo IVNEI ocorreu paralelamente a um vasto esforço levado a cabo, não só a nível nacional, mas também a nível da União Europeia, no sentido de se aumentar a quantidade e a qualidade dos indicadores quantitativos de conjuntura disponíveis, e também de se melhorar a harmonização existente entre os resultados produzidos pelos vários Estados-Membros.

Consequentemente, os trabalhos realizados procuraram levar em linha de conta as exigências estatísticas decorrentes da entrada em vigor do Regulamento (CE) n.º 1165/98 do Conselho, de 19 de Maio, relativo às estatísticas conjunturais.

2. CAMPO SECTORIAL

O IVNEI abrange a globalidade dos sectores integrantes das secções "C - Indústrias Extractivas", "D - Indústrias Transformadoras" e "E - Produção e Distribuição de Electricidade, de Gás e de Água", da Revisão 2 da Classificação Portuguesa das Actividades Económicas (CAE-Rev.2).

3. RESULTADOS

Os resultados apurados relativamente a cada variável objecto de disponibilização (o que exclui as variáveis tratadas, exclusivamente, no intuito de se produzir informação relevante para o cálculo do IPI) são constituídos não só pelo seu índice geral respectivo, mas também por índices desagregados, quer em termos sectoriais (em função de posições definidas pela CAE-Rev.2 desde o nível de secção - uma letra de desagregação - até ao nível de grupo - três dígitos de desagregação), quer segundo a finalidade económica dos produtos produzidos (bens de consumo final, bens de consumo final duradouro, bens de consumo final não duradouro, bens de consumo intermédio e bens de investimento).

Os resultados relativos às variáveis volume de negócios (total), emprego, remunerações e horas trabalhadas, desagregados até ao nível de divisões da CAE-Rev.2 (dois dígitos de desagregação) e segundo a finalidade económica dos produtos, são objecto de publicação regular com uma periodicidade mensal. Os restantes resultados, referentes a estas ou a outras variáveis objecto de disponibilização, bem como a estes ou a outros níveis de desagregação, são passíveis de serem fornecidos, em determinadas condições, desde que tal seja solicitado ao INE.

4. AMOSTRAS

As empresas inquiridas pelo Inquérito Anual à Produção Industrial (IAPI) constituem, no seu conjunto, o universo de referência do inquérito que serve de base ao IVNEI: o Inquérito Mensal ao Volume de Negócios e Emprego na Indústria.

Recorde-se que o IAPI inquire, grosso modo, tantas empresas (de entre as mais representativas) quantas as necessárias para que seja alcançada uma cobertura de pelo menos 90% do volume de negócios realizado, em cada sector, pelas empresas

integrantes do Universo de referência do Inquérito (Anual) às Empresas Harmonizado (IEH).

A selecção das empresas integrantes das amostras do IVNEI foi realizada, mediante a constituição de amostras estratificadas no âmbito de cada grupo relevante da CAE-Rev.2.

A estratificação destas amostras teve como base os escalões de pessoal ao serviço 10 a 19, 20 a 49, 50 a 99, 100 a 199, 200 a 499, e 500 ou mais pessoas ao serviço. Foram, pois, excluídas as empresas com menos de 10 pessoas ao serviço e, nalguns casos, excluíram-se ainda as empresas integrantes do escalão 10 a 19 após a constatação, apoiada na análise de resultados calculados com base nas amostras da base 1990, de que aquele estrato não influenciava o andamento dos indicadores a apurar.

São inquiridas exaustivamente as empresas, de maior dimensão, integrantes de tantos escalões quantos os necessários para seja alcançada uma dada cobertura mínima de 25% do volume de negócios realizado por cada sector. Relativamente, aos restantes escalões, foi constituída para cada um deles, uma amostra aleatória específica.

5. CONCEITOS

Os conceitos empregues na recolha da informação de base objecto de tratamento por parte do IVNEI, encontram-se expostos nos parágrafos seguintes.

Volume de negócios

Corresponde ao total da facturação (com exclusão do IVA), relativa às vendas de mercadorias, produtos acabados e intermédios, subprodutos, desperdícios, resíduos e refugos (Contas POC 711,712 e 713) e à prestação de serviços a terceiros (Contas POC 721 a 725). A este valor devem ser deduzidos as devoluções e os descontos e abatimentos (Contas POC 717,718,727 e 728) e, adicionarem-se todas as taxas, encargos ou despesas que recaiam sobre os produtos e que sejam imputados ao cliente, ainda que facturados separadamente. Não devem ser considerados os subsídios de exploração ou quaisquer receitas provenientes da venda de imobilizado.

Total de pessoal (ao serviço)

Refere-se a todos os trabalhadores que fazem parte da "folha de remunerações" da empresa, independentemente do tipo de contrato, de realizarem trabalho a tempo inteiro ou a tempo parcial (os trabalhadores a tempo parcial devem ser contabilizados em número equivalente de trabalhadores a tempo inteiro) ou do local de trabalho, no momento de referência (medido na última semana completa do mês a que se refere a informação).

Inclui:
- os trabalhadores "ao domicílio", desde que constem da folha de remunerações;
- os trabalhadores temporariamente ausentes (por férias, maternidade, conflito de trabalho, formação profissional, doença ou acidentes de trabalho, com duração inferior a um mês).

Não inclui:
- os trabalhadores que se encontrem cedidos para prestar serviço noutras empresas;
- os trabalhadores que se encontrem a prestar serviço militar;
- os trabalhadores de outras empresas a prestar serviços nas instalações da empresa;
- os prestadores de serviços (trabalhadores a recibo verde ou empresários em nome individual).

Remunerações brutas

Referem-se ao montante ilíquido em dinheiro ou em géneros, pagos aos trabalhadores que se incluem no conceito de "pessoal ao serviço", pelo trabalho realizado no período normal e no extraordinário. Inclui ainda o pagamento de horas remuneradas mas não efectuadas (férias, feriados, e outras ausências pagas) e ainda os subsídios que se revistam de carácter regular como sejam os subsídios de alimentação, de função, alojamento ou transporte, diuturnidades ou prémios de antiguidade, produtividade, de assiduidade, isenções de horário de trabalho, subsídio por trabalhos penosos, perigosos ou sujos e subsídios por trabalhos de turnos e nocturnos.

Horas trabalhadas

Refere-se ao número de horas efectivamente trabalhadas pelo "pessoal ao serviço" da empresa, tal como é definido no respectivo conceito.

Vendas de mercadorias

Consiste na facturação, com exclusão do IVA, de produtos adquiridos a outras empresas e vendidos sem transformação (Conta POC 711), após dedução de devoluções e abatimentos (componentes das contas POC 717 e 718 referentes a mercadorias).

Vendas de produtos acabados e intermédios

Consiste na facturação, com exclusão do IVA, de produtos fabricados pela própria empresa (ou mandados fabricar a terceiros com matérias-primas próprias), devendo incluir também subprodutos, desperdícios, resíduos e refugos (Contas POC 712 e 713), após dedução de devoluções, descontos e abatimentos (componentes das contas 717 e 718 referentes a produtos acabados e intermédios).

Prestação de serviços

Respeita aos serviços prestados que sejam próprios dos objectivos ou finalidades da empresa (contas POC 721 a 725). Pode integrar os materiais aplicados no caso de estes não serem facturados separadamente.

6. FÓRMULAS DE CÁLCULO

Uma vez que são utilizadas amostras estratificadas, os primeiros cálculos a realizar passam pela determinação dos valores assumidos por cada variável em cada estrato. A este nível é utilizada a fórmula seguinte:

$$V_{vne} = C_{ne} * \sum_{o=1}^{o=r} V_{vnoe}$$

onde:

V_{vne} = valor assumido pela variável "v" no mês corrente (n), no estrato "e";

C_{ne} = coeficiente de extrapolação no mês corrente (n), do estrato "e";

V_{vnoe} = valor assumido pela variável "v" no mês corrente (n), em cada empresa "o" objecto de inquirição, integrante do estrato "e";

r = número total de empresas com valor V_{vnoe} conhecido ou imputado neste estrato.

O coeficiente de extrapolação relativo a cada estrato é calculado, por seu lado, com base na fórmula:

$$C_{ne} = \frac{u}{r}$$

onde:

C_{ne} = coeficiente de extrapolação no mês corrente (n) do estrato "e";

u = número total empresas integrantes do estrato "e" no universo de referência;

r = número total de empresas com valor V_{vnoe} conhecido ou imputado neste estrato.

As não respostas existentes à data do apuramento de resultados são, na generalidade das situações, imputadas para cada empresa com resposta em falta. Essa imputação é efectuada mediante a multiplicação dos valores transmitidos,

relativamente ao mês homólogo do ano anterior, por coeficientes resultantes do quociente (variável a variável) entre os valores globais comunicados no âmbito de cada grupo da CAE-Rev-2, no mês n e no mês n-12, pelas respostas comuns aos dois períodos, convenientemente validadas.

Obtidos os valores de cada variável no âmbito de cada estrato, procede-se ao cálculo dos valores por elas assumidos no âmbito de cada grupo relevante, segundo a fórmula:

$$V_{vgn} = \sum_{e=1}^{e=m} V_{vgne}$$

onde:

V_{vgn} = valor assumido pela variável "v" no grupo "g", no mês corrente (n);

V_{vgne} = valor assumido pela variável "v" no grupo "g", no mês corrente (n), no âmbito do estrato "e";

m = número total de estratos a considerar.

Os índices de grupo são calculados, a partir daqueles valores, de acordo com a fórmula:

$$I_{vgn} = \frac{V_{vgn}}{\frac{\sum_{j=1}^{12} V_{vgj}}{12}} * 100$$

onde:

I_{vgn} = índice da variável "v" referente ao grupo "g", no mês corrente (n);

V_{vgn} = valor assumido pela variável "v" no âmbito do grupo "g", no mês corrente (n);

$\sum_{j=1}^{12} V_{vgj}$ = somatório dos valores assumidos pela variável "v" no âmbito do grupo "g", ao longo dos doze meses integrantes do ano base.

A partir dos valores assumidos mês após mês pelos índices de grupo referentes a cada variável, procede-se ao cálculo dos índices referentes a essa mesma variável a níveis sectorialmente mais agregados, bem como de índices segundo a finalidade

económica dos produtos produzidos. A fórmula utilizada em ambos os casos é a seguinte:

$$I_{vcn} = \sum_{a=1}^{t} P_{vab} * I_{van}$$

onde:

I_{vcn} = índice representativo da evolução do valor assumido pela variável "v" no âmbito do agregado "c", entre o ano base e o mês corrente (n);

P_{vab} = ponderador referente à variável "v", a afectar a cada índice mais detalhado "a" (de grupo ou não), calculado com base em informação relativa ao ano base (b);

I_{van} = índice representativo da evolução do valor assumido pela variável "v" no âmbito de cada posição mais detalhada "a", entre o ano base e o mês corrente (n);

t = número total de índices a agregar.

Os valores assumidos pelos ponderadores foram calculados, por sua vez, com base na fórmula:

$$P_{vab} = \frac{V_{vab}}{\sum_{a=1}^{t} V_{vab}}$$

onde:

P_{vab} = ponderador referente à variável "v", a afectar a cada índice mais detalhado "a", calculado com base em informação relativa ao ano base (b);

V_{vab} = valor assumido por cada variável "v" à escala de cada posição mais detalhada "a", durante o ano base (b);

t = número total de índices a agregar.

VOLUME 2

2° QUADRIMESTRE DE 1998

CONGRESSOS, SEMINÁRIOS, COLÓQUIOS E CONFERÊNCIAS

CONGRESS, SEMINARS AND CONFERENCES

No Estrangeiro:
Abroad:

1998

- 30 de Agosto - 05 de Setembro
  International Colloquium on Mathematics in Gambling, Budapest, Hungary.
  Informações: E - mail: jatek@math-inst.hu
  or
  WWW: http://www.math-inst.hy/~jatek

- 31 de Agosto - 04 de Setembro
  15th IFIP World Computer Congress and SEC '98 – 14th International Information Security Conference, Vienna e Budapest.
  Informações: Oesterreichische Computer Gesellschaft, (Austrian Computer Society OCG), Wollzeile 1-3, A-1010 Vienna, Austria; Telf.: 43 1 5120235; FAX: 43 1 5120235 9.
  E - mail: ifip98@ocg.or.at
  or
  ocg@ocg.or.at
  WWW: http://www.ocg.or.at
  or
  John v. Neumann Computer Society NJSZT, Bathori u. 16, H-1054 Budapest, Hungary; Telf.: 36 1 1329349; FAX: 36 1 1318140.
  E - mail: ifip98@neumann.hu
  or
  h13588zub@ella.hu
  WWW: http://www.njszt.iff.hu

- 01 - 04 de Setembro
  Joint IASS/IAOS Conference, Aguascalientes, Mexico.
  Informações: *Dennis Trewin*, Austrialian Bureau of Statistics, P. O. Box 10, Belconnen, ACT 2616, Australia.
  E - mail: dennis.trewin@abs.gov.au
  or
  Denise Lievesley, ESRC Data Archive, University of Essex, CO4 3SQ, UK.
  E - mail: denise@essex.ac.uk

- 07 - 10 de Setembro
  EUFIT'98, Aachen, Germany.
  Informações: WWW: http://www.mitgmbh.de

- 07 - 11 de Setembro
  1998 Conference of the Royal Statistical Society, to be held at the University of Strathclyde, Glasgow, UK.
  Informações: *Professor Eric Renshaw*, Department of Statistics and Modelling Science, University of Strathclyde, Glasgow G1 1XH, UK; FAX: 44 141 5522079.
  E - mail: rss98@stams.strath.ac.uk
  WWW: http://www.stams.strath.ac.uk/rss98/index.html

- 09 - 11 de Setembro
  PROLAMAT' 98. Tenth International IFIP TC5 WG-5.2 WG-5.3 Conference. The theme is "**The Globalization of Manufacturing in the Digital Communications Era of the 21st Century: Innovation, Agility, and the Virtual Enterprise**", Trento, Italy.
  Informações: *Mara Gruber*, Laboratorio di Ingegneria Informatica, Via f. Zeni, 8, 38068 Rovereto (TN), Italy; Telf: 39 464 443134; FAX: 39 464 443141.
  E - mail: prolamat@lii.unitn.it;
  or
  WWW: http://www.lii.unitn.it/prolamat/

- *14 – 17 de Setembro
  28th International Biometrical Colloquium: in Honour of Tadeusz Calinski, Inowroclaw, Poland.
  Informações: *Prof. Stanislaw Mejza*, Polish Boimetric Society, Agricultural Univ. of Poznan, Wojska Polskiego 28, PL 60-637 Poznan, Poland; Tel.: 48 61 8487140; FAX: 48 61 8487146.
  E - mail: smejza@owl.au.poznan.pl
  or
  WWW: http://www.ii.uni.wroc.pl/~aba/gpol_ibs.html

- 14 - 18 de Setembro
  "German Statistical Week", Lübeck, Germany.
  Informações: Verband Deutscher Städtestatistiker, Bereich Statistik und Wahlen, Schwartzstrasse 73, 46045 Oberhausen, Germany.

- 25 - 26 de Setembro
  Workshop on "Uncertainty in the Risk Assessment of Environmental and Occupational Hazards", Bologna, Italy.
  Informações: *Dr. J. C. Bailar*, Department of Health Statistics, University of Chicago, 5841 S Maryland Ave, Chicago, Illinois 60637, USA; Telf.: 773834 1242; Fax: 773 7021979;
  E - mail: jcbailar@midway.uchicago.edu
  or
  Dr. A. J. Bailer, Dept. of Math and Stats, Maiami University, Oxford, OH 45056-1641, USA; Telf.: 513 5293538; Fax: 513 5291493;
  E - mail: ajbailer@muohio.edu

- *05 - 09 de Outubro
  Journées d'études en statistique: Méthodes bayesiennes en statistique, Marseille, France.
  Informações: *G. Saporta*, Chaire de Statistique Appliquée, CNAM, 292 rue Saint Martin, F 75003 Paris, France; Fax: 33 1 40272746;
  E - mail: jes98@cnam.fr

- 05 - 10 de Outubro
  The Fields Institute Workshop on Hydrodynamic Limits.
  Informações: The Fields Institute for Research in Mathematical Sciences, 222 College Street, Second Floor, Toronto, Ontario M5T 3J1, Canada; Telf.: 416 348 9710; Fax: 416 348 9385;
  E - mail: probability@fields.utoronto,ca
  or
  WWW: http://www.math.yorku.ca/Probability/Fields.html

- *13 - 16 de Outubro
  WAITRO International Seminar on Technology Management, Warsaw, Poland.
  Informações: *Ms. Marianne Jessing*, Waitro Secretariat, Danish Technology Institute, P.O. Box 141, DK-2630 Taastrup, Benmark; Telf.: 45 4350 7047; Fax: 45 4350 7050;
  E - mail: waitro@dti.dk
  or
  Industrial Chemistry Research Institute, Rydygiera 8, PL-01 793 Warsaw, Poland; Telf.: 48 22 6339511; Fax: 48 22 6338295;

- 22 - 24 de Outubro
  International Seminar on Seasonal Adjustment methods (SAM 98), sponsored by Eurostat and ISI, Bucharest, Romania.
  Informações: *Mr. R. Depoutot*; telf.: 352 4301 34926; FAX: 352 4301 34149.

- 25 - 28 de Outubro
  FRACTAL 98, "Complexity and Fractals in the Sciences", 5th International Multidisciplinary Conference, Valletta, Malta.
  Informações: *Dr. Miroslav M. Novak*, School of Physics, Kingston University, Surrey KT1 2EE, UK; Telf.: 44 181 547 7481; FAX: 44 181 547 7562.
  E - mail: novak@kingston.ac.uk;
  or
  WWW: http://www.kingston.ac.uk/fractal.

- 25 - 29 de Outubro
  The Fields Institute Workshop on Monte Carlo Methods.
  Informações: The Fields Institute for Research in Mathematical Sciences, 222 College Street, Second Floor, Toronto, Ontario M5T 3J1, Canada; Telf.: 416 348 9710; Fax: 416 348 9385;
  E - mail: probability@fields.utoronto,ca
  or
  WWW: http://www.math.yorku.ca/Probability/Fields.html

- *27 - 29 de Outubro
  International Conference on "Census'91 and Beyond", Abuja, Nigeria.
  Informações: Lt. Col. Chris Ugokwe (rtd) Chairman, National Population Commission, Office of the Chairman, Lukulu Street Wuse Zone 3, P. M. B. 0281 Garki Abuja, Nigeria.

- *04 - 06 de Novembro
  NTTS'98, International Seminar on New Techniques & Technologies for Statistics, Sorrento, Italy.
  Informações: *Dr. Vicenzo*, Department of Mathematics and Statistics, Faculty of Economics, Univ. of Napoli, Italy; Telf.: 39 81 675114; FAX: 39 81 675113.
  E - mail: ntts98@dms.unina.it;
  or
  WWW: http://www.dms.unina.it/ntts98.html.
  or
  Eurostat website:
  WWW: http://europa.eu.int/en/comm/eurostat/research/conferences/ntts-98

- *21 - 27 de Novembro
  SOFSEM'98. 25th Conference on Current Trends in Theory and Practice of Informatics, Jasna, Slovakia.
  Informações: *Branislav Rovan*, Dept. of Computer Science, Comenius Univ., Mlynska Dolina, 84215 Bratislava, Slovakia.
  E - mail: sofsempc@dcs.fmph.uniba.sk;
  or
  WWW: http://www.dcs.fmph.uniba.sk/~sofsem98.

- 02 - 04 de Dezembro
  ARS' 98 Conference, 3rd Conference on Statistical Computing of the IASC Asian Regional Section, Manila, Philippines.
  Informações: *Dr. Romulo A. Virola*, 2/F Midland Buendia Bldg., 403 Sen. Gil J. Puyat Ave., Makati City, Philippines; Telf.: 632 890 9495 / 632 895 2395; FAX: 632 890 9408.
  E - mail: nscbsg@mozcom.com.

- *07 - 11 de Dezembro
  Envinonmentrics Conference on the theme "Biometry at work towards envinonment 2000", Victoria Falls, Zimbabwe.
  Informações: *Dr. S. Nadarajah*, School of Mathematics and Statistics, Univ. of Plymouth, Plymouth PL4 8AA, UK; Telf.: 44 1752 232723; FAX: 44 1752 232780.
  E - mail: s.nadarajah@plymouth.ac.uk;
  or
  WWW: http://www.tech.plym.ac.uk/maths/conf/zim3.html.

- *08 - 10 de Dezembro
  IV International Symposium on Optimization & Statistics, to be held at Aligarh Muslim University, Aligarh, India.
  Informações: *Professor A. H. Khan*, Chairman, Department of Statistics and Operations Research, Aligarh Muslim University, Aligarh 202002, India.

- *09 - 11 de Dezembro
  Short Course entitled "Bayes and Empirical Bayes methods for data analysis", Leuven, Belgium
  Informações: *Geert Verbeke*, Biostatistical Centre, K.U. Leuven, U.Z. Sint-Rafael, Kapucijnenvoer 35, B-3000 Leuven, Belgium; Telf.: 32 16 336892; FAX: 32 16 336900.
  E - mail: geert.verbeke@med.kuleuven.ac.be.

- *09 - 11 de Dezembro
  2nd International Symposium on Semi-Markov Models: Theory and Applications, Compiegne, France.
  Informações: *N. Limnios*, Division Mathematiques Appliquées, Univ. de Technologie de Compiegne, B.P. 529, 60205 Compiegne, France; FAX: 33 3 44 23 4477.
  E - mail: Nikolaos.Limnios@utc.fr.
  or
  Jacques Janssen, CADESP-ULB, Slovay Business School, av. F. Roosevelt, 50, B.P. 194/7, 1050 Brussels, Belgium; FAX 32 2 6502785.
  E - mail: janssen@ulb.ac.be
  or
  WWW: http://www.ulb.ac.be/soco/cadeps/PSMsympo/

- *12 - 15 de Dezembro
  The 6th Islamic Countries Conference on Statistical Sciences, Dhaka, Bangladesh.
  Informações: *Prof. Pk. Md. Motiur Rahman*, Secretary, Sixth ISOSS Conference, Institute of Statistical Research and Training University of Dhaka – 1000, Bangladesh; Telf.: 880 2 865093; FAX: 880 2 865583.
  E - mail: maislam@citechco.net.

- 13 - 18 de Dezembro
  XIXth International Biometric Conference, Cape Town, South Africa.
  Informações: *Professor Christine McLaren*, Department of Mathematics, Moorhead State University; Telf.: 218 2364004; Fax: 218 2362168;
  E - mail: mclaren@mhd1.moorhead.msus.edu
  or
  Jant Riley, E - mail: janet.riley@bbsrc.ac.uk

- 14 - 16 de Dezembro
  Seventh International Applied Statistics in Industry and Manufacturing Conference, Singapore
  Informações: Seventh IASIM conference, P. O. Box 189, Mulvane, KS67110, USA; FAX: 1 316 7890906;
  E - mail: tracy@isai.org;
  or
  WWW: http://www.isai.org.

1999

- *15 - 20 de Março
  Semstat, Séminar Européen de Statistique – on Complex Stochastic Systems, Eindhoven, The Netherlands.
  Informações: *Claudia Klüppelberg*, Center for Mathematical Sciences, Univ. of Technology, D-80290 Munich, Germany; Telf.: 49 89 289 28211; Fax: 49 89 289 28464;
  E - mail: ck@mathematik.tu-muenchen.de

- *05 - 07 de Maio
  XVI International Symposium on Combing Data from Different Sources, Statistics Canada, Ottawa, Ontario, Canada.
  Informações: *Christian Thibault*, HSMD, Statistics Canada, Tunney`s Pasture, Ottawa, K1A OT6 Canada.
  E - mail: thibchr@statcan.ca

- 17 - 21 de Maio
  Eleventh International Genstat Conference, Poznan, Poland.
  Informações: Genstat'99, *Dr. Pawel Krajewski*, Institute of Plant Genetics, Polish Academy of Sciences, Strzeszynska 34, 60-479 Poznan, Poland; Telf.: 48 61 8233511; Fax: 49 61 8233671;
  E - mail: pkra@igr.poznan.pl

- *25 - 28 de Maio
  International Conference on Large Scale Data Analysis, Cologne, Germany.
  Informações: *Friederika Priemer*, Univ. zu Koeln, Zentralarchiv fuer Empirische Sozialforschung, Bachemer Str. 40, D-50931 Koeln, Germany; Telf.: 49 221 47694 33; 49 221 47031 55; Fax: 49 221 47694 44;
  E - mail: priemer@za.uni-koeln.de

- *26 - 27 de Maio
  IAOS Conference, Official Statistics, Challenges for the Future, The Hague, The Netherlands.
  Informações: *H. D. Dukker*, Secretary of the Programme Committee; Fax: 31 70 337 5991;
  E - mail: hdkr@cbs.nl

- 06 - 09 de Junho
  Annual Meeting of the Statistical Society of Canada, Regina, Saskatchewan, Canada.
  Informações: Local Arrangements Chair, R. J. Tomkins, Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan, SAS OA2, Canada.
  E - mail: jtomkins@max.cc.uregina.ca.

- 14 – 17 de Junho
  The IX International Symposium on Applied Stochastic Models and Data Analysis (ASMDA' 99), Lisboa, Portugal.
  Informações: *Professor Helena Bacelar Nicolau*, Universidade Lisboa;
  E - mail: ulfphelb@cc.fc.ul.pt
  or
  Professor Fernando C. Nicolau, Universidade Lisboa;
  E - mail: fan@laminaria.si.fct.unl.pt.

- *14 - 18 de Junho
  26th Conference on Stochastic Processes and Their Applications, Beijing, China.
  Informações: *Xiaoyu Hu*, Beijing;
  E - mail: xyhu@amath4.amt.ac.cn

- 12 - 16 de Julho
  19th IFIP TC7 Conference on System Modeling and Optimization, Cambridge, UK.
  Informações: E - mail: tc7con@amtp..cam.ac.uk

- *03 - 04 de Agosto
  ISI Satellite Conference on Statistical Publishing, Warsaw, Poland.
  Informações: ISI Permanent Office, 428 Prinses Beatrixlaan, P.O.Box 950, 2270 AZ Voorburg, The Netherlands.

- *05 - 09 de Agosto
  IASS Sponsored Workshop on Recent Trends in the Methodology for Social and Business Studies, Jyväskkylä, Finland.
  Informações: *Mr. Kari Djerf*, Secretary of the Organizing Committee, FIN-00022 STATISTICS FINLAND.

- 08 – 12 de Agosto
  1999 Joint Statistical Meetings, Baltimore, Maryland, USA.
  Informações: ASA, 1429 Duke St., Alexandria, VA 22314-3402, USA; Telf.: 703-6841221; FAX: 703-684 2037.
  E - mail: meetings@asa.mhs.compuserve.com

- 10 - 18 de Agosto
  International Statistical Institute, 52nd Biennal Session, Helsinki, Finland.
  Informações: ISI Permanent Office, 428 Prinses Beatrixlaan, P. O. Box 950, 2270 AZ Voorburg, The Netherlands.

- 19 – 23 de Agosto
  The 6th Tartu Conference on Multivariate Statistics, Satellite meeting to Helsinki ISI Session, Tartu, Estonia.
  Informações: *E.-M. Tiit* or *T. Kollo*, Institute of Mathematical Statistics, University of Tartu, J. Liivi 2, EE2400, Tartu, Estonia; Telf.: 37 27 465488 / 37 27 465486; FAX: 37 27 433509.
  E - mail: etiit@ut.ee or kollo@ut.ee

- *20 - 21 de Agosto
  IASS Satellite Conference on Small Area Estimation, Riga, Latvia.
  Informações: *Professor Jan Kordos*, Central Statistical Office, Al. Niepodleglosci 208, 00-925 Warsaw, Poland; Fax: 48 22 8250395;
  E - mail: j.kordos@stat.gov.pl

- 13 - 16 de Setembro

  44th Annual Conference of the German Society of Medical Informatics, Biometry and Epidemiology (GMDS), Heidelberg, Germany.

  Informações: *Norbert Victor*, Department of Medical Biometry, Institute for Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 305, D-69120 Heidelberg, Germany
  or
  Lutz Edler, Biostatistics Unit, German Cancer Research Center, IM Neuenheimer Feld 280, D-69120 Heidelberg, Germany; Fax: 49 6221 564195
  E - mail: GMDS-ISCB99@dkfz-heidelberg.de
  or
  WWW: http://www.dkfz-heidelberg.de/biostatistics/GMDS-ISCB99

- 13 - 17 de Setembro

  Heidelberg Congress Week: Joint Conference of GMDS – ISCB 99, Heidelberg, Germany.

  Informações: *Norbert Victor*, Department of Medical Biometry, Institute for Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 305, D-69120 Heidelberg, Germany
  or
  Lutz Edler, Biostatistics Unit, German Cancer Research Center, IM Neuenheimer Feld 280, D-69120 Heidelberg, Germany; Fax: 49 6221 564195
  E - mail: GMDS-ISCB99@dkfz-heidelberg.de
  or
  WWW: http://www.dkfz-heidelberg.de/biostatistics/GMDS-ISCB99

- 14 - 17 de Setembro

  20th Annual Conference of the International Society of Clinical Biostatistics (ISCB), Heidelberg, Germany.

  Informações: *Norbert Victor*, Department of Medical Biometry, Institute for Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 305, D-69120 Heidelberg, Germany
  or
  Lutz Edler, Biostatistics Unit, German Cancer Research Center, IM Neuenheimer Feld 280, D-69120 Heidelberg, Germany; Fax: 49 6221 564195
  E - mail: GMDS-ISCB99@dkfz-heidelberg.de
  or
  WWW: http://www.dkfz-heidelberg.de/biostatistics/GMDS-ISCB99

* - Novas entradas – *Denotes new Entries*.

ACÇÕES DESENVOLVIDAS PELO INE NO ÂMBITO DA COOPERAÇÃO BILATERAL E MULTILATERAL

ACTIONS ACHIEVED BY NSI IN THE SCOPE OF BILATERAL AND MULTILATERAL COOPERATION

(DE 1 DE MAIO A 31 DE JULHO DE 1998):

a) Cooperação desenvolvida com os PALOP e Macau:

Com a presença do Secretário de Estado Adjunto do Ministro do Equipamento, Planeamento e Administração do Território, Dr. Consiglieri Pedroso, teve lugar em 19 de Junho, a Sessão Solene de Encerramento da Conferência "Cooperação Estatística no quadro dos Países de Língua Portuguesa (CPLP)". Esta permitiu reunir em Lisboa, durante três dias, as delegações oficiais dos sete países da CPLP, bem como vários observadores convidados, com o principal objectivo de identificar propostas de carácter prático para o estabelecimento de um Programa de Cooperação Estatística entre os países membros. As conclusões da Conferência, entre as quais a institucionalização da mesma como órgão de apoio da CPLP para a preparação e seguimento daquele Programa de Cooperação, deverão ser apresentadas à Cimeira de Chefes de Estado e de Governo da CPLP, a ter lugar em Cabo Verde, em Julho de 1998.

No que se refere ao projecto comum sobre Classificações, Conceitos e Nomenclaturas, foram iniciados os trabalhos para a elaboração da Classificação Nacional de Bens e Serviços da Guiné-Bissau e de S. Tomé e Príncipe, através da realização de dois estágios para técnicos dos INE daqueles países.

No período em apreço e no âmbito da cooperação bilateral com os PALOP e Macau, foram realizadas as seguintes acções :

ANGOLA

Foi realizado um estágio no âmbito do projecto de Digitalização da Base Cartográfica Censitária de Angola.

Além da formação teórica os três técnicos do INE de Angola procederam, com o apoio dos técnicos do INE – Portugal, à digitalização e gravação em CDROM de cartas de algumas cidades de Angola. Inserido no programa e de forma complementar ao trabalho executado no INE, foram realizadas visitas ao Instituto Geográfico do Exército e ao Instituto Português de Cartografia e Cadastro.

Este trabalho insere-se num projecto angolano mais vasto que terá outras acções com participação portuguesa.

Uma das aplicações previstas para esta Base Cartográfica será a próxima realização do Recenseamento Geral da População de Angola.

GUINÉ-BISSAU

Em virtude da situação de guerra na Guiné – Bissau ficaram retidos em Lisboa os membros da delegação Guineense à Conferência "Cooperação Estatística no quadro da CPLP".

Em colaboração com o Instituto da Cooperação Portuguesa e com o CESD – Lisboa foi possível organizar estágios para os três técnicos, nos domínios da formação em Índices de Preços, Contas Nacionais, Estatísticas da Agricultura e do Ambiente.

b) Cooperação desenvolvida com os PECO:

No âmbito do Programa PHARE de assistência técnica aos Países da Europa Central e Oriental (PECO), não se realizou, durante o período acima mencionado, qualquer acção de cooperação entre o INE e os organismos de estatística dos PECO. Esta situação deveu-se ao congelamento das verbas do PHARE para a cooperação estatística, estando previsto o seu inicio em Setembro próximo.

No entanto, realizou-se nos dias 6 e 7 de Maio, no Luxemburgo a reunião anual do Comité Director sobre Cooperação Estatística Comunitária com os PECO. A participação do INE foi assegurada pelo GRIC.

Os principais objectivos foram a apresentação do relatório das actividades de cooperação com aqueles países até 1997 e a nova estrutura de cooperação. Assim, a partir de Setembro próximo os Estados Membros que queiram conduzir acções de cooperação terão que apresentar ao Eurostat propostas anuais por projectos/país, com vista à celebração de contratos entre os EM e o CESD-Comunitário com vista à implementação dos mesmos. A grande diferença entre esta nova estrutura e a anterior, reside no facto de que passará a ser da exclusiva responsabilidade do organismo estatístico doador a gestão financeira dos projectos. Foi dado também especial ênfase aos novos projectos piloto nos PECO, tendo sido referidos os dois projectos que o INE propôs ao Eurostat: Projecto Piloto sobre a Exaustividade das Contas Nacionais e Projecto Piloto Contas Económicas da Agricultura.

FUNDAMENTO, OBJECTO E ÂMBITO

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, passa 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 no âmbito do Sistema Estatístico Nacional;

c) Informações sobre congressos, seminários, colóquios e conferências de interesse estatístico ou afim;

d) Informações sobre acções desenvolvidas pelo INE no âmbito da cooperação bilateral e multilateral.

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 por iniciativa dos próprios autores serão objecto de avaliação de mérito científico pelo Conselho Editorial, que decidirá ou não pela respectiva publicação.

Para a elaboração e envio das contribuições para publicação na Revista são adoptadas as Normas de Apresentação de Manuscritos 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

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 within the scope of the National Statistical System;

c) Information on congresses, seminars and conferences of a statistical or related nature;

d) Information on activities developed by INE within the scope of bilateral or multilateral co-operation;

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 Manuscripts 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 MANUSCRITOS

Nos termos da alínea b) do nº. 3 do Artigo 5º do Regulamento da *Revista de Estatística* do Instituto Nacional de Estatística, o Conselho Editorial aprovou as seguintes **Normas de Apresentação de Manuscritos**:

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 qualquer 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 e cinco 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 artigo apresentado e respectiva *disquette*.
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, e entregues em suporte papel acompanhado da respectiva *disquette*.
6. Na apresentação dos originais, os autores respeitarão ainda as seguintes normas:
   6.1. Quanto à *estrutura*:
      a) O texto deve ser dactilografado em formato A_4, com utilização do tipo de letra *Times New Roman* - 11, e com as seguintes margens: *top*: 2,5 cm, *bottom*: 2 cm, *left*: 2,5 cm, *right*: 5 cm;
      b) A primeira página conterá exclusivamente o título do artigo, bem como o nome, morada e telefone 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 cem palavras, seguido de um parágrafo com indicação de palavras-chave até ao limite de quinze;
      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:

         ANDERSON, C.W., and TURKMAN, K.F, (1995) "Sums and maxima of stationary sequences with heavy tailed distributions", *Sankhya*, Vol. 57, Series A, pp.1-10.

6.3. Quanto à *revisão de provas e publicação*:

a) Uma vez aceite o artigo e antes da sua publicação, receberá o autor dois exemplares de provas para revisão, um dos quais será devolvido 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 Conselho Editorial 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, nº. 5 – 9º.
1 000 Lisbon - Portugal
Tel.: +351 1 842 61 00 (3905) Fax.: +351 1 842 63 66 e-mail: liliana.martins@ine.pt

RULES FOR SUBMITTING MANUSCRIPTS

Within the terms of sub-section a of no. 3 of Article 5 of the regulations of the *Statistical Review* of the National Statistical Institute (INE), the Editorial Committee has approved the following **Rules for Submitting Manuscripts**:

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 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 all articles submitted on the initiative of the respective 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 within a maximum period of thirty-five 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, it will be returned along with the respective diskette.

4. The articles accepted for publication in the *Statistical Review* will also be made public on the INE Internet site.

5. The original articles having no more than thirty pages must be processed in *Word for Windows*, completely at black and white, and they will be delivered in hard copy as well as on diskette.

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 and with the following margins: top: 2.5 cm, bottom: 2 cm, left: 2,5 cm, right: 5 cm;

b) The first page shall contain only the title of the article as well as the name, address and telephone 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 by the *Review* should be forwarded;

c) The second page shall contain only the title and a abstract of the article in Portuguese and English with the maximum of one hundred words followed by a paragraph indicating key words up to the limit of fifteen;

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:

ANDERSON, C.W., and TURKMAN, K.F., (1995) "Sums and maxim of stationary sequences with heavy tailed

distributions", *Sankhya*, Vol. 57, Series A, pp. 1-10.

6.3 Regarding *proof-reading and publication*:

a) Once the article is accepted and prior to its publication, the author will receive two copies for review. One of these copies 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 Editorial Committee 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
Avª. António José de Almeida, nº. 5 – 9º.
1 000 Lisbon - Portugal
Tel.: +351 1 842 61 00 (3905) Fax.: +351 1 842 63 66 e-mail: liliana.martins@ine.pt

BOLETIM DE ENCOMENDA

Nome _____ Data de nascimento: __/__/__

Profissão/Função _____ Instituição/Empresa _____

Telef.:_____ Fax: _____

DESEJO RECEBER OS EXEMPLARES DA REVISTA DE ESTATÍSTICA:

Em casa . ☐ Na Instituição/empresa ☐

Morada para envio: _____

Localidade: _____ Código Postal: _____

Autorizo débito no cartão Visa ☐ ou Mastercard ☐

nº.: ☐☐☐☐ ☐☐☐☐ ☐☐☐☐ ☐☐☐☐

Valor da transacção: 6. 300$00 Validade do cartão __/__/__

☐ Junto cheque nº_____ à ordem do *INSTITUTO NACIONAL DE ESTATÍSTICA* sobre o Banco _____

Data: __/__/__ Assinatura: _____

OS DADOS RECEBIDOS SERÃO PROCESSADOS AUTOMATICAMENTE E DESTINAM-SE AOS ENVIOS RELACIONADOS COM A SUA ASSINATURA, RESPECTIVAS OPERAÇÕES ADMINISTRATIVAS E ESTATÍSTICAS, E À EVENTUAL APRESENTAÇÃO DE OUTROS PRODUTOS E SERVIÇOS DO INSTITUTO NACIONAL DE ESTATÍSTICA.

BOLETIM DE ENCOMENDA

Nome _____ Data de nascimento: __/__/__

Profissão/Função _____ Instituição/Empresa _____

Telef.:_____ Fax: _____

DESEJO RECEBER OS EXEMPLARES DA REVISTA DE ESTATÍSTICA:

Em casa . ☐ Na Instituição/empresa ☐

Morada para envio: _____

Localidade: _____ Código Postal: _____

Autorizo débito no cartão Visa ☐ ou Mastercard ☐

nº.: ☐☐☐☐ ☐☐☐☐ ☐☐☐☐ ☐☐☐☐

Valor da transacção: 6. 300$00 Validade do cartão __/__/__

☐ Junto cheque nº_____ à ordem do *INSTITUTO NACIONAL DE ESTATÍSTICA* sobre o Banco _____

Data: __/__/__ Assinatura: _____

OS DADOS RECEBIDOS SERÃO PROCESSADOS AUTOMATICAMENTE E DESTINAM-SE AOS ENVIOS RELACIONADOS COM A SUA ASSINATURA, RESPECTIVAS OPERAÇÕES ADMINISTRATIVAS E ESTATÍSTICAS, E À EVENTUAL APRESENTAÇÃO DE OUTROS PRODUTOS E SERVIÇOS DO INSTITUTO NACIONAL DE ESTATÍSTICA.

AUTORIZADO PELOS CTT
NO SERVIÇO NACIONAL

RSF

NÃO PRECISA DE SELO

INSTITUTO NACIONAL DE ESTATÍSTICA
SECÇÃO EDITORIAL, PROMOÇÃO, DISTRIBUIÇÃO
E VENDA DE PUBLICAÇÕES

Av. António José de Almeida
1000 LISBOA

AUTORIZADO PELOS CTT
NO SERVIÇO NACIONAL

RSF

NÃO PRECISA DE SELO

INSTITUTO NACIONAL DE ESTATÍSTICA
SECÇÃO EDITORIAL, PROMOÇÃO, DISTRIBUIÇÃO
E VENDA DE PUBLICAÇÕES

Av. António José de Almeida
1000 LISBOA


P1029802