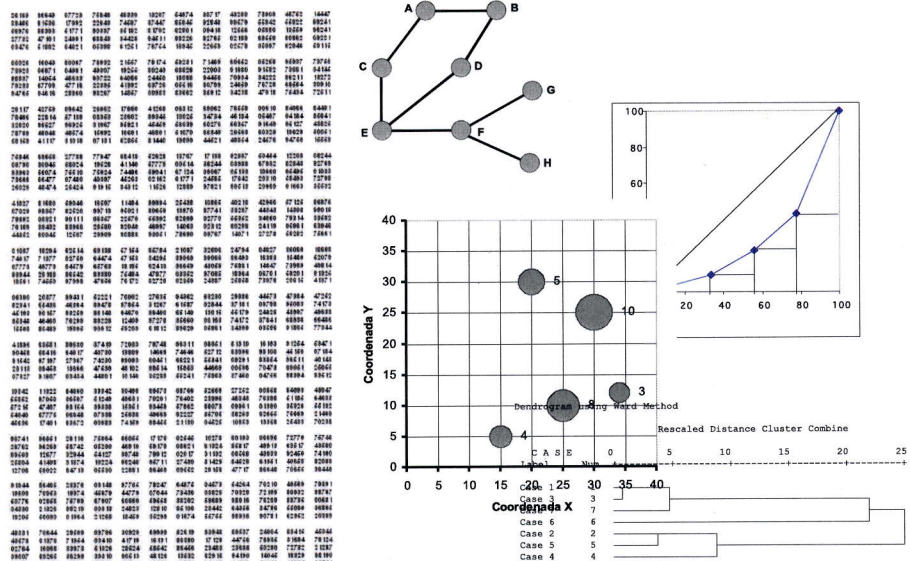


DIOGO DE ABREU



UNIVERSIDADE DE LISBOA

Análise de Dados II

PROGRAMA

DIOGO JOSÉ BROCHADO DE ABREU

Estudos para o Planeamento Regional e Urbano nº 69

Centro Estudos Geográficos

Universidade de Lisboa

2006

FICHA TÉCNICA

Título: Análise de Dados II. Programa

Autor: Diogo de Abreu

Unidade de Investigação: Centro de Estudos Geográficos da Universidade de Lisboa

Edição e Distribuição: Centro de Estudos Geográficos da Universidade de Lisboa

Impressão: David Barreira

Tiragem: 150 exemplares

ISBN: 978-972-636-174-9

UNIVERSIDADE DE LISBOA

**FACULDADE DE LETRAS
DEPARTAMENTO DE GEOGRAFIA**

Análise de Dados II

(PROGRAMA PARA UMA DISCIPLINA OBRIGATÓRIA DA
LICENCIATURA EM GEOGRAFIA)

Relatório sobre “o programa, os conteúdos e os métodos de ensino das matérias da disciplina de ANÁLISE DE DADOS II, no cumprimento do Art.º 9º -1, alínea a) do Decreto n.º 301/72 de 14 de Agosto, relativo às provas para obtenção do título de Professor Agregado.

ÍNDICE

NOTA PRÉVIA	7
I - APRESENTAÇÃO DO PROGRAMA E DOS MÉTODOS DE ENSINO	13
I.1 INTRODUÇÃO	15
1. Antecedentes	19
2. A situação actual	21
3. Perspectivas futuras	23
I.2 OBJECTIVOS	25
1. Objectivos gerais da disciplina	25
2. Objectivos específicos	26
I.3 PROGRAMA	29
1. Considerações preliminares	29
2. A estrutura do programa	30
3. O enunciado do programa	33
I.4 MÉTODOS DE ENSINO	41
I.5 AVALIAÇÃO	45
I.6 MATERIAIS DE APOIO	49
I.7 BIBLIOGRAFIA	51
I.8 PLANIFICAÇÃO DAS AULAS	65

II - DESENVOLVIMENTO DO PROGRAMA	69
II.1 CONCEITOS BÁSICOS E DE ENQUADRAMENTO	71
1. Objectivos a atingir	71
2. Conteúdos a desenvolver	71
3. Meios e estratégias a utilizar	72
4. Bibliografia	72
II.2 ALGUMA MATEMÁTICA PARA GEÓGRAFOS	73
1. Objectivos a atingir	73
2. Conteúdos a desenvolver	73
3. Meios e estratégias a utilizar	74
4. Bibliografia	74
II.3 RECOLHA DA INFORMAÇÃO	75
1. Objectivos a atingir	75
2. Conteúdos a desenvolver	75
3. Meios e estratégias a utilizar	76
4. Bibliografia	76
II.4 TABELAS DE CONTINGÊNCIAS	79
1. Objectivos a atingir	79
2. Conteúdos a desenvolver	80
3. Meios e estratégias a utilizar	80
4. Bibliografia	80
II.5 ANÁLISE DE VARIÂNCIA	81
1. Objectivos a atingir	81
2. Conteúdos a desenvolver	81
3. Meios e estratégias a utilizar	82
4. Bibliografia	82
II.6 REGRESSÃO MÚLTIPLA	85
1. Objectivos a atingir	85
2. Conteúdos a desenvolver	86
3. Meios e estratégias a utilizar	86
4. Bibliografia	87
II.7 ANÁLISE FACTORIAL	89
1. Objectivos a atingir	89
2. Conteúdos a desenvolver	90
3. Meios e estratégias a utilizar	90
4. Bibliografia	91
II.8 CLASSIFICAÇÃO NUMÉRICA	93
1. Objectivos a atingir	93
2. Conteúdos a desenvolver	94

3. Meios e estratégias a utilizar	94
4. Bibliografia	95
II.9 ANÁLISE DISCRIMINANTE	97
1. Objectivos a atingir	97
2. Conteúdos a desenvolver	97
3. Meios e estratégias a utilizar	98
4. Bibliografia	98
II.10 REDES E GRAFOS	99
1. Objectivos a atingir	99
2. Conteúdos a desenvolver	100
3. Meios e estratégias a utilizar	100
4. Bibliografia	101
II.11 GEOESTATÍSTICAS	103
1. Objectivos a atingir	103
2. Conteúdos a desenvolver	103
3. Meios e estratégias a utilizar	104
4. Bibliografia	104
II.12 MODELOS INTEGRADOS EM GEOGRAFIA	107
1. Objectivos a atingir	107
2. Conteúdos a desenvolver	108
3. Meios e estratégias a utilizar	109
4. Bibliografia	109
ANEXO – Conteúdos a desenvolver (Capítulos 1 a 11)	113

I. - NOTA PRÉVIA

O meu interesse pela utilização dos métodos quantitativos em Geografia tem sido uma das constantes da minha vida académica, desde que ingressei no Curso de Geografia da Faculdade de Letras da Universidade de Lisboa, nos já distantes idos de Março de 1975.

Nesses tempos revolucionários – revolucionários até mesmo para a pacífica Geografia portuguesa – viveram-se tempos de permanente crítica, de alguma contestação, mas principalmente de muita reflexão sobre o que era, o que tinha sido e o que deveria ser a Geografia em Portugal. Foi então que docentes e discentes, conjuntamente, se lançaram abertamente na discussão sobre muitos dos conceitos fundamentais da Geografia, do modo como ela deveria ser praticada e, principalmente, de como se poderia fazer a ligação dos saberes e competências geográficas ministrados na Faculdade com as práticas de intensa transformação do território nacional que então se vivia. Foi a partir desse trabalho conjunto que se veio a prefigurar grande parte do que viria a ser o paradigma dominante no trabalho académico e nas práticas de intervenção no território que se vieram a desenvolver na Geografia portuguesa nos trinta anos que se seguiram.

Como já disse, data de então o meu interesse pela utilização dos métodos quantitativos em Geografia e também noutras ciências sociais e

naturais. Como razão encontro talvez o desejo de encontrar maior rigor no conhecimento dos territórios, até então baseados principalmente nas impressões subjectivas das observações individuais impressionistas, sempre contingentes e muitas vezes insuficientes. E também a consciência de que a urgência em encontrar soluções para os diversos problemas do país (nas suas diversas escalas) não se compadecia com um conhecimento que se baseava fundamentalmente na vivência dos espaços e nos espaços vividos. Considerava então que as novas necessidades de intervenção territorial exigiam competências mais abrangentes e a capacidade de tratar múltiplos e diversificados espaços.

De entre as múltiplas influências que formaram a minha maneira de ser e de estar na Geografia e a minha personalidade de investigador, salientam-se as influências duradouras de três mestres, ao mesmo tempo três amigos, que, de algum modo, formaram a minha personalidade de investigador e nortearam toda a minha actividade académica.

O Prof. Chris Jensen-Butler, que me mostrou que os complexos métodos que se liam nos livros tinham aplicação prática à realidade concreta e podiam ser soluções para os problemas do dia a dia com que as populações se defrontam. Tenho de referir especialmente o incentivo para o desenvolvimento dos grandes modelos regionais, como foi o Rex e as avaliações de impactes territoriais.

Tenho também de lembrar a memória do Prof. Peter Gould, um gentil homem, cuja excepcional imaginação e capacidade criadora, aliada à permanente actualização dos seus conhecimentos técnicos, mesmo das mais recentes complexidades, me mostrou as aventuras da descoberta e da inovação científicas e como se podem associar à alegria de viver.

Mas foi e é do Prof. Jorge Gaspar que tenho recebido a maior lição. Entusiasmo na permanente aprendizagem, capacidade científica e humildade na aplicação do conhecimento são características que percorrem toda a sua carreira científica, académica e profissional. O grande conhecimento que foi adquirindo dos territórios do País e do resto do mundo, associado a uma intensa preocupação no entendimento das razões para os factos e processos

geográficos que neles se desenvolvem, fez dele a personalidade mais importante da Geografia portuguesa dos últimos tempos. É a ele e ao seu trabalho que se deve o prestígio de que a Geografia actualmente goza em Portugal.

Sempre motivador, é com o Prof. Jorge Gaspar, e desde os tempos distantes em que fui seu aluno em Geografia Humana Geral, que tenho mantido as mais interessantes discussões técnicas e científicas, sempre estimulantes, exigentes e motivadoras. E é também ao Prof. Jorge Gaspar que eu, como tantos outros, devo a oportunidade de ter podido participar em muitos trabalhos fundamentais para a estruturação do território nacional, assim como também é a ele que devo os desafios de investigação mais estimulantes do meu percurso académico.

Este programa também muito deve aos colegas com quem, ao longo de quase três décadas de trabalho conjunto, fui fazendo a descoberta das técnicas de análise de dados e dos diversos modos de os aplicar.

Lembro especialmente Lucinda Fonseca e João Lemos, e o desafio que constituiu a introdução no curso de Geografia da disciplina de “Elementos de Estatística aplicados à Geografia”, em Outubro de 1978. E que mais posso dizer sobre a colaboração com Lucinda Fonseca, que iniciada então, continua ininterruptamente até hoje, em variados trabalhos e projectos de investigação comuns, nas disciplinas leccionadas em conjunto e nos diversos trabalhos de administração académica em que tanto nos temos empenhado?

No campo técnico mais específico dos métodos de análise e da modelização, mais recentemente, muito do trabalho que executo é feito com a colaboração de Nuno Costa, ontem um dos meus primeiros e mais brilhantes alunos, hoje um apoio seguro nas técnicas e modelos que tanto gosto temos em desenvolver.

Quero referir também a colaboração proveitosa que tenho tido da parte de Alina Esteves, José Rodriguez, Carlos Silva e Luís Moreno com quem tenho leccionado disciplinas de Estatística, Métodos Quantitativos e de Análise de Dados, e, mais recentemente, de Jorge Rocha e Paulo Morgado, novos assistentes, companheiros interessados nos Sistemas de Informação

Geográfica, com quem tenho mantido uma relação científica fecunda, procurando o esclarecimento mútuo de muitos aspectos que respeitam, simultaneamente, à Análise de Dados e às técnicas analíticas dos Sistemas de Informação Geográfica.

Não posso também deixar de referir os mestres, colegas e amigos com quem partilho o Centro de Estudos Geográficos e o Departamento de Geografia da Faculdade de Letras da Universidade de Lisboa, e com quem discuto e aprendo todos os dias. Saliento especialmente a minha coordenadora na Área de Investigação de Geografia Humana, a Prof^a. Carminda Cavaco, e os colegas com quem partilho a Área de Investigação.

Uma palavra ainda para os colegas de outras Universidades. Um agradecimento especial é devido aos colegas dos Departamentos de Geografia da Universidade de Coimbra, que me convidaram por duas vezes para colaborar com eles no ensino da aplicação de métodos quantitativos à Geografia, no seu mestrado, e do Departamento de Ambiente e Ordenamento da Universidade de Aveiro, com quem mantenho nesta área do conhecimento e de docência uma estreita colaboração desde há mais de uma década.

Mas têm sido sempre os alunos os auxiliares fundamentais do meu processo de ensino / aprendizagem. A activa participação dos melhores, e as dificuldades que todos experimentam para a compreensão do que ensino, as dificuldades que mostram nas provas de avaliação que fazem, são a grande avaliação do que ensino e das estratégias de como o faço.

Em última análise é com eles e para eles que faço o meu trabalho.

No concurso para Professor Associado apresentei o programa da disciplina de Análise Regional e Urbana. Poderia agora ter apresentado um programa actualizado dessa disciplina, que continuo a leccionar, ou então desenvolver as questões teóricas da organização do espaço regional e urbano ou as importantes questões do desenvolvimento geral, regional e local, que desde sempre estiveram no cerne dos meus interesses científicos e académicos.

Mas decidi tratar da problemática dos métodos de análise de dados em Geografia. Considero que tenho já uma experiência razoável na utilização e

no ensino destas metodologias de análise, que penso ser meu dever transmitir ao ensino que se pratica no Departamento. Com efeito, tenho leccionado métodos quantitativos nas licenciaturas e mestrados do Departamento de Geografia na Universidade de Lisboa, nos cursos de mestrado em Geografia da Universidade de Coimbra e no mestrado em Inovação e Desenvolvimento do Departamento de Ambiente e Ordenamento da Universidade de Aveiro, e em cursos mais curtos para psicólogos, psiquiatras, jornalistas e bibliotecários, arquivistas e documentalistas.

Penso portanto que esta é uma boa oportunidade para apresentar os resultados da minha experiência.

Por outro lado, acho, cada vez mais, que os licenciados que formamos necessitam de uma sólida formação nestas matérias da análise de dados, assim como nos SIGs - Sistemas de Informação Geográfica, que são indispensáveis nas suas práticas profissionais e constituem uma mais valia de competência, que contribui para aumentar a sua empregabilidade.

Estas as razões que a meu ver justificam a opção que tomei de tratar nestas provas o ensino dos métodos quantitativos em Geografia. Mesmo encarando eventuais opiniões de que esta matéria não é exactamente Geografia, a maneira como a vou apresentar e a importância que creio que tem para que os actuais alunos e futuros geógrafos possam fazer boa Geografia, justifica amplamente, em meu entender, a escolha que faço.

O entendimento que tenho sobre o que deve ser a preparação dos alunos e a melhor maneira de suprir às suas necessidades veio a prevalecer sobre quaisquer outras considerações.

O programa que aqui apresento corresponde a um segundo nível do ensino de métodos quantitativos em Geografia, e pressupõe que as matérias básicas foram apresentadas na disciplina de *Análise de Dados I*. Penso que no conjunto destas duas disciplinas semestrais devem ser apresentadas aos alunos as técnicas quantitativas e qualitativas basilares para o exercício da profissão de geógrafo, e isto logo desde o início da sua formação académica. Na disciplina de *Análise de Dados II* devem ser abordados os temas

analíticos mais avançados, de um modo que privilegie sempre a sua aplicação à Geografia. Por isso, o que se segue trata um pouco de métodos quantitativos e muito de Geografia, segundo o entendimento que tenho de que os métodos em Geografia, como aliás em qualquer outro ramo do conhecimento, são os meios que ilustram e simplificam o entendimento das temáticas disciplinares.

Definitivamente não aceito que as disciplinas de análises de dados sejam em, Geografia, uma dificuldade mais que é preciso ultrapassar e para pouco ou nada serve, tanto na prática da investigação durante o curso, como após a sua conclusão.

Por último, uma palavra sobre o Departamento de Geografia da minha Universidade de Lisboa. Nele tenho desenvolvido a minha carreira, nele fui aluno, assistente estagiário, assistente, professor auxiliar e associado, nele me licenciiei e me doutorei.

Tudo isto teria sido impossível sem o ambiente de excelente trabalho, de muito companheirismo e cumplicidade e de grande amizade que desde há muito se vive no Departamento.

Todos temos conseguido manter um extraordinário ambiente de trabalho, fecundo e entusiasmante, que tem sido a causa dos nossos pequenos e grandes sucessos.

Que assim continue são os meus votos mais sinceros.

Diogo de Abreu

Lisboa, Outubro de 2005

Primeira Parte

**Apresentação do Programa
e dos Métodos de Ensino**

I.1 INTRODUÇÃO

O programa da disciplina de “**Análise de Dados II**”, que aqui se apresenta, de acordo com o disposto no nº 9 do Decreto-Lei 3012/72, de 14 de Agosto, reflecte a importância que as metodologias de análise de dados têm no ensino da Geografia e na prática dos geógrafos.

Os problemas geográficos são muito diversificados. Vão desde o ambiente natural ao comportamento dos indivíduos. A sua complexidade e abrangência obrigam, actualmente, ao conhecimento diversificado de técnicas rigorosas de recolha e de tratamento da informação disponível.

A delapidação e o esgotamento de recursos, a poluição do meio e a degradação da qualidade de vida das populações, a protecção contra riscos naturais, económicos e sociais, a evolução da população e as migrações, a organização espacial dos indivíduos, das instituições e das suas actividades, o desenvolvimento nacional, regional ou local e a urbanização, os sistemas de transporte e comunicação são aspectos da realidade geográfica que exigem ao geógrafo que os descreve, que os compreende e que neles tem de intervir, competências alargadas, em que se devem aliar ao conhecimento teórico dos problemas as capacidades técnicas de interpretação da informação disponível ou que é necessário recolher.

Cada escala de análise ou de intervenção no trabalho que se faz em Geografia tem associada uma tipologia de formas de recolha da informação, que tem a ver com a perspectiva teórica do investigador, a natureza intrínseca dos dados obtidos e com as técnicas de tratamento a utilizar.

Ao nível local, privilegia-se a observação directa, a observação participante, as entrevistas pessoais e os encontros de grupo, os dados recolhidos são muitas vezes de natureza nominal, e as técnicas de tratamento qualitativas. No âmbito regional, nos procedimentos da recolha de informação dominam os inquéritos e o acesso a informação estatística

estruturada, muitas vezes recolhida e disponibilizada por organismos especializados, onde a natureza dos dados é quase sempre paramétrica e os tratamentos são fundamentalmente do tipo multivariado. E nas macro-análises às pequenas escalas, tem de se usar os dados que existem geralmente apresentados em escalas de razão enquanto os tratamentos são do tipo quantitativo multivariado.

A Geografia apresenta ainda uma grande riqueza e diversidade no posicionamento filosófico dos seus praticantes, que vai desde o possibilismo e o positivismo até ao neo-modernismo e ao pós-estruturalismo.

Nos seus estudos, uns geógrafos utilizam métodos quantitativos e outros métodos qualitativos, uns usam a matemática e a modelização, outros exploram valores subjectivos, significados ou emoções, numa escolha que sendo sempre pessoal e em relação com os valores e a ideologia do investigador, se encontra também indubitavelmente relacionada com as condições materiais em que se efectua, a escala, o âmbito da investigação e a natureza da informação de que se dispõe.

O mundo de hoje gera quantidades enormes de informação, não apenas a que é recolhida, de forma tradicional, através das observações dos investigadores, pelos procedimentos administrativos e pelos organismos especializados (estes últimos cada vez mais abrangentes), mas principalmente a informação recolhida continua ou intermitentemente pelos mais variados sensores remotos – satélites, câmaras de vídeo vigilância, gravadores áudio directos e de telecomunicações. Toda esta informação, em regra georreferenciada, que é necessário analisar e que muitas vezes está já disponível para efeitos de pesquisa e investigação, coloca novos problemas e levanta novos desafios, nos aspectos de recolha, armazenamento, tratamento e exploração.

As diversas disciplinas científicas acordam para esta verdadeira “inundação de informação”, muita da qual é informação espacialmente referenciada, e desenvolvem modos de incorporar o território nos seus saberes e técnicas sofisticadas que permitam o seu tratamento e interpretação.

Ao mesmo tempo, não deixa de ser curioso verificar como a Geografia académica e muita da profissional reage a esta situação. Numa perspectiva de recusa da oportunidade, negando a importância das análises quantitativas de dados espaciais, muitos dos geógrafos actuais assumem um posicionamento teórico não-quantitativo, quando não verdadeiramente anti-quantitativo (Fotheringham, 2000:xi).

Cremos poder identificar as duas razões principais que, em nosso entender, explicam esta situação paradoxal.

A primeira terá a ver com a desilusão que a abordagem positivista dos anos setenta e oitenta provocou em muitos de nós, em que parte considerável do trabalho efectuado com a utilização de métodos estatísticos e de técnicas de análise de dados se mostrou insuficiente, quando não errado, na identificação dos padrões geográficos e na compreensão dos processos que lhe estavam associados.

A segunda, menos referida, mas talvez mais importante, é que a utilização adequada destas análises é difícil, exige uma longa aprendizagem e aprofundados conhecimentos técnicos, na óptica do utilizador que deve interpretar os resultados. Esta dificuldade de muitos recém licenciados, e também de alguns académicos com formação quantitativa deficiente, é talvez a maior causa da pouca atracção que o conhecimento das técnicas de análise estatística e espacial de dados tem para muitos praticantes de Geografia.

Importa ainda referir que, um pouco contraditoriamente, tem sido a grande divulgação de programas para computadores pessoais de análise estatística, multivariada e espacial que tem agravado a situação, pois o uso indevido e incorrecto dos dados e as interpretações pouco ou nada fundamentadas tiveram e estão a ter um crescimento exponencial. Não admira, pois, que muitos geógrafos mais rigorosos vejam no seu uso um grave perigo para a credibilidade dos trabalhos de investigação em Geografia.

Porém, seja qual for a posição filosófica ou a perspectiva que os geógrafos tenham sobre a utilização de métodos quantitativos, a realidade é

que a diversidade dos temas e de abordagens metodológicas que os investigadores e os profissionais têm de encarar nos seus trabalhos, obrigam a que o geógrafo tenha de ter uma preparação técnica alargada, que lhe dê a capacidade de recolher, analisar e interpretar eficazmente a informação disponível, seja ela de que natureza for.

Tem sido esta a forma de pensar dominante no Departamento de Geografia da Universidade de Lisboa, e é esta a que se adopta no contexto deste programa: a de que o nosso ensino da Geografia deve incluir um ensino abrangente dos métodos de recolha, tratamento e interpretação da informação geográfica, que trate das técnicas mais clássicas mas em que se motivem os alunos para os novos desenvolvimentos que estão a ocorrer.

Existe um grande consenso no Departamento de Geografia sobre a forma como deve ser efectuado o ensino e aprendizagem das técnicas de análise de dados.

É geralmente aceite que devem existir dois níveis, correspondentes a duas disciplinas semestrais, leccionadas sequencialmente em dois semestres seguidos.

Na primeira disciplina de análise de dados considera-se que deverão ser apresentados os conceitos básicos da estatística elementar. Não cabe aqui apresentar o programa, o conteúdo e os métodos de ensino teórico e prático das matérias dessa disciplina, mas não poderão deixar de ser referidos alguns dos seus aspectos fundamentais, que justificam as opções que tomámos neste nosso programa.

Nessa disciplina inicial leccionam-se os conceitos fundamentais, os métodos mais elementares de recolha de informação, o conceito teórico e empírico de probabilidade, os tratamentos a uma variável, com destaque para a determinação dos parâmetros estatísticos das distribuições, os testes de hipótese e a correlação simples entre duas variáveis.

Neste programa, em que se apresenta a segunda parte de um curso de análise de dados em Geografia, considera-se que já foi ultrapassada, com sucesso, a formação básica em estatística monodimensional adquirida no âmbito dessa disciplina introdutória de análise de dados.

1. Antecedentes

A vontade de tratar a informação disponível tem sido uma constante na investigação e no ensino da Geografia.

Desde sempre que se procurou analisar a informação disponível através da elaboração de mapas, de gráficos e de alguns tratamentos executados de forma dispersa pelas diversas disciplinas do curso e pelos trabalhos de investigação.

É a Amorim Girão e a Orlando Ribeiro que se devem as primeiras tentativas, como se vê nos mapas, atlas e trabalhos que publicaram. Consta até que o Prof. Orlando Ribeiro, por volta dos anos cinquenta, quando se debatia com análises de grande dimensão, como foram os mapas de evolução da população por freguesias, terá procurado, numa perspectiva verdadeiramente inovadora, o apoio do Prof. Tiago de Oliveira, especialista em probabilidades e estatística na Secção de Matemática da Faculdade de Ciências de Lisboa, para melhorar as competências nessa área dos alunos do Curso de Geografia¹.

Perdida a oportunidade, a grande transformação veio a ocorrer já na década de setenta, através de dois pioneiros, o Prof. Jorge Gaspar e a Prof.^a Paula Lema e devido às ideias novas que então se desenvolviam nas universidades em que efectuaram estágios de formação, a de Lund e a de Nottingham, respectivamente.

O Prof. Jorge Gaspar, que contactara com um excepcional conjunto de jovens geógrafos que desenvolvia as novas correntes positivistas na Universidade de Lund, sob a direcção de Thorsten Hagerstrand, decidiu apresentar como prova ao concurso de Professor Extraordinário um programa de Geografia Humana² em que dava um grande relevo à técnica de

¹ - E consta também que ao ser informado que para calcular melhor as médias, os alunos teriam de frequentar duas ou três disciplinas de Matemáticas Gerais, de Probabilidades e de Estatística, terá afirmado que então era melhor que aprendessem mais Geografia, mesmo errando no cálculo da média (informação pessoal).

² - GASPAR, Jorge (1973) – Programa de Geografia Humana, Universidade de Lisboa, Lisboa.

correlação linear, na procura de relações entre distintas variáveis. Em relação à investigação, já anteriormente desenvolvera ou fomentara o desenvolvimento de outros métodos analíticos, especialmente de taxionomias numéricas e de análise factorial, para o Gabinete da Área de Sines.

Entretanto, a Prof.^a Paula Lema frequentava um estágio em Nottingham, na altura um dos mais dinâmicos pólos de desenvolvimento das técnicas de análise de dados no Reino Unido. Aí contactara com os Prof.s Cole e King, que estavam a editar um dos primeiros e mais completos manuais de métodos quantitativos em Geografia (Cole & King, 1970) e com o Prof. Paul Mather, também um grande especialista de metodologias em Geografia. Publicou alguns artigos, quer durante a estada em Inglaterra quer posteriormente, e incorporou alguns deles na sua dissertação de doutoramento. A sua influência não foi porém tão importante como a do Prof. Jorge Gaspar, por não se ter conseguido tornar autónoma, mantendo sempre a dependência do Departamento de Nottingham na execução dos programas. O seu percurso pode servir de ilustração para a necessidade dos alunos e dos licenciados em Geografia garantirem a autonomia em todas as fases dos seus procedimentos de análise de dados, desde o estabelecimento do plano de trabalho, recolha da informação e execução dos tratamentos analíticos até à interpretação dos resultados.

Mas as grandes transformações vieram a ocorrer com o 25 de Abril de 1974. O regresso de Paula Lema e a entrada para o Departamento de Geografia da Faculdade de Letras do Prof. Jorge Gaspar foram marcos importantes na difusão das técnicas de métodos quantitativos. O Prof. Jorge Gaspar introduziu o programa que defendera nas provas para professor extraordinário na disciplina de Geografia Humana, logo nas “experiências pedagógicas” de 1974 e 75, e tanto nessa como nas outras disciplinas que ministrou ou nos trabalhos de investigação que dirigiu, foi um dos promotores da utilização adequada de muitas técnicas de análise de dados na Geografia portuguesa.

1977 é um ano marcante, pois foi então que a tutela, com a Reforma dos Cursos de Geografia nas universidades portuguesas, estabeleceu pela primeira vez e formalmente uma disciplina de análise de dados chamada

“Estatística Aplicada à Geografia”, que começou a ser ensinada nas licenciaturas de Geografia do País em 1977/78 e em Lisboa apenas um ano depois, em 1978/77. O atraso de Lisboa deveu-se à consciência de que esta disciplina deveria ser ensinada, não por matemáticos, como nas outras faculdades se fez, mas por especialistas em Geografia, que se tiveram de preparar. O futuro veio a dar razão à opção da Faculdade de Letras da Universidade de Lisboa. Com efeito, a opção de preparar geógrafos para ministrar esta disciplina, permitiu um maior entrosamento com as potenciais aplicações e um maior rigor teórico na utilização.

Desde então, em todas as reformas dos cursos de Geografia e em todas as suas alterações curriculares, esteve sempre presente uma disciplina de elementos de estatística ou de análise de dados, que eu próprio leccionei ou de que fui coordenador.

Com a entrada do regime semestral e a reestruturação das licenciaturas de Geografia em 1998³, a disciplina anual em que se tinha vindo a ensinar técnicas de análise de dados foi desdobrada em duas disciplinas semestrais, uma mais elementar e outra mais avançada, situação que se mantém até ao momento presente.

2. A Situação Actual

A reforma de 1998, ainda em grande parte em vigor, considerava uma disciplina anual de introdução metodológica básica muito geral, logo no primeiro semestre – ***Métodos e Técnicas em Geografia*** – em que deveriam ser ministrados conhecimentos elementares de informática, de cartografia e de recolha e tratamento de dados, a par de uma breve e ligeira introdução à teoria do pensamento geográfico⁴.

A análise estatística ficaria a cargo de duas outras disciplinas, semestrais, a Análise de Dados I, a decorrer no segundo semestre do primeiro ano, e a Análise de Dados II, no primeiro semestre do segundo ano.

³ -. Deliberação nº. 283/1998 de 18 de Maio, alterada posteriormente pela Deliberação nº 1115/2003 de 6 de Agosto.

⁴ - Esta disciplina desde 2003 que passou a semestral.

Na primeira, a *Análise de Dados I*, tratam-se os teste estatísticos, as distribuições e os erros, os tratamentos estatísticos elementares de uma variável e faz-se uma introdução à recolha de informação e à correlação e regressão entre duas variáveis. E na outra, a *Análise de Dados II*, assegura-se a leccionação das técnicas estatísticas multidimensionais mais frequentes, da análise de redes e das geoestatísticas.

O ensino dos métodos quantitativos em Geografia reveste-se agora de uma excepcional importância, nesta mesma ocasião em que muitos geógrafos, de forma paradoxal, recusam as abordagens mais quantitativas (consideradas depreciativamente como demasiado positivistas), mas ao mesmo tempo, como já se referiu, o desenvolvimento dos Sistemas de Informação Geográfica e a sua rápida difusão por todos os sectores e níveis da nossa sociedade, exigem aos seus operadores sólidos conhecimentos de análise quantitativa espacial e a capacidade teórica e técnica de os aplicar.

E, simultaneamente, as capacidades de utilização banalizam-se. Existe mais informação, melhor documentada e mais actualizada, e meios mais fáceis para a tratar. Os programas comerciais de análise de dados, de análise gráfica e cartográfica, assim como os pequenos sistemas integrados com potencialidades de tratamento de informação de base territorial são hoje muito mais baratos e mais fáceis de operar.

Mudaram-se os tempos. E esta mudança, a facilidade com que se podem fazer as análises tem multiplicado erros, incoerências e más interpretações. Como exemplo, basta referir os índices de compra e de venda, as ordenações e os “rankings” de concelhos, de cidades, de escolas de universidades, de hospitais e de tudo o mais, que ultimamente têm inundado a comunicação social, e cuja qualidade, para dizer o mínimo, deixa bastante a desejar. E estou a referir-me ao topo da escala da investigação !

Foram todas estas circunstâncias que me conduziram à opção por este programa. Espero que venha a constituir um contributo, modesto embora, para a melhoria das capacidades e das competências dos futuros geógrafos que estamos a formar.

3. Perspectivas futuras

O futuro está já connosco.

O desafio que está a constituir a adaptação dos cursos de Geografia ao processo desencadeado em Bolonha, é mais do que uma simples operação de cosmética a mascarar o “emagrecimento” do sector público de ensino universitário.

É o desafio de responder às necessidades e inquietações que a sociedade portuguesa coloca ao futuro e é uma oportunidade para que se faça mais e melhor geografia, e principalmente que se torne o ensino mais eficaz, isto é, que se consiga capacitar mais os alunos com menos esforço de docentes e discentes.

À Geografia portuguesa e muito especialmente aos seus representantes nas universidades incumbe a responsabilidade de preparar a transformação. O futuro irá demonstrar se a Geografia, enquanto ramo autónomo do saber e da sua aplicação, continuará a ter na sociedade portuguesa o sucesso que conseguiu nos últimos trinta anos.

Para tal, o que agora importa, para além de discutir e preparar, detalhadamente, as desejáveis competências dos geógrafos, os planos curriculares, as listas de disciplinas obrigatórias e de opção, os conteúdos programáticos e os métodos de ensino, é ter uma ideia clara das tarefas passíveis de ser desempenhadas pelos geógrafos.

Parece-me um bom sinal a opção que se configura na reestruturação dos nossos cursos, de conciliar uma formação básica de largo espectro, com uma sólida base de pré-especialização teórica e o desenvolvimento de competências alargadas dos alunos nos diversos aspectos da auto-formação.

Neste contexto, toda a capacidade metodológica que se desenvolva no campo da análise da informação assume especial relevância, pois garante aos geógrafos a capacidade de entender facilmente as realidades e os contextos concretos em que irão trabalhar.

Tenho de salientar ainda que é minha opinião que o ensino dos

métodos de análise da informação em Geografia deve ser preparado de modo a dar para dar uma resposta mais cabal às necessidades actuais e futuras do país, formando geógrafos mais competentes na compreensão dos problemas e dos conflitos territoriais a todas as escalas e com a diversidade de perspectivas e abordagens que só a Geografia permite.

Por isso creio que, sem descurar a formação teórica básica, que tem de continuar a ser a estruturação fundamental dos nossos saberes, a Geografia tem de desenvolver bastante mais as capacidades metodológicas, especialmente na área do tratamento de dados qualitativos, fundamentais na intervenção à escala local e municipal, e as técnicas de interpretação em sistemas de informação geográfica, desde as multivariadas mais clássicas, em teledetecção, até às mais sofisticadas de análise de vizinhança, fractais, e redes neuronais, entre muitas outras que se aplicam em todos os ramos da geografia física e humana. É para mim evidente a necessidade de um ensino de segundo ciclo que inclua todas estas matérias aplicadas à generalidade dos temas geográficos, razão pela qual apoio sem reservas a continuação da especialização em Cartografia e Sistemas de Informação em Geografia, que iniciámos em 1998.

I.2 OBJECTIVOS

1. Objectivos gerais da disciplina

Como objectivos gerais desta disciplina enunciam-se os seguintes:

- 1)- Dar a conhecer os métodos de análise estatística multivariada e as técnicas de análise de dados que são mais utilizados na investigação geográfica;
- 2)- Saber em que circunstâncias e em que condições se deve proceder à sua aplicação;
- 3)- Compreender os aspectos teóricos fundamentais e os procedimentos de utilização dos métodos de análise de dados;
- 4)- Conhecer as potencialidades e os limites das suas aplicações;
- 5)- Ter a capacidade de executar as análises estatística multivariadas e as técnicas espaciais de análise de dados mais utilizadas na investigação geográfica, nomeadamente quanto à selecção adequada das variáveis e à escolha das opções de funcionamento dos métodos;
- 6)- Saber interpretar os resultados, avaliando criticamente a aplicação dos métodos analíticos enquanto instrumentos de compreensão das realidades e dos processos geográficos;
- 7)- Fornecer uma base que permita a compreender o utilizar técnicas de análise mais avançada, de forma autónoma.

Pretende-se que os alunos reconheçam a importância das metodologias de análise de dados, e que adquiram uma visão global das metodologias mais utilizadas em Geografia, as saibam utilizar com correcção, e sejam capazes de interpretar correcta e criticamente os resultados obtidos. Tem de se ter sempre presente que o objecto central da disciplina não é a matemática, nem sequer os métodos de análise, mas sim a geografia,

considerada e assumida como a capacidade de entender os factos e os processos que decorrem no território.

Por isso também se espera que os alunos sejam capazes de reflectir sobre o papel dos métodos de análise de dados na identificação e resolução dos problemas geográficos e saibam avaliar adequadamente a conveniência e a oportunidade da sua utilização,

Por último, deseja-se que os alunos aprendam a aprender. Numa área do saber em que a evolução é intensa e em que estão sempre a ser propostas novas técnicas e perspectivas metodológicas, a capacidade de conseguir desenvolver mais conhecimento de forma autónoma é uma das mais valias essenciais que a disciplina poderá fornecer.

2. Objectivos específicos

Podem concretizar-se estes objectivos pedagógicos e de competência de uma forma mais detalhada.

Consideramos que os alunos que tenham obtido aproveitamento no final desta disciplina de Análise de Dados II, devem estar em condições de aplicar as técnicas e os métodos de análise de dados à realidade, o que implica:

- saber esquematizar uma investigação geográfica, definindo a questão a tratar e escolhendo o tipo de abordagem mais adequado;
- ter a capacidade de esquematizar e preparar todas as fases da investigação, desde a delimitação da área em estudo e a recolha da informação até aos tipos de tratamento analítico;
- saber escolher as variáveis (ou indicadores) mais indicados para medir ou avaliar o efeito dos conceitos teóricos estruturantes da investigação;
- reconhecer os diversos tipos e modelos de dados e saber identificar os tratamentos mais adequados para cada um;

- conhecer as técnicas e metodologias de análise de dados, as respectivas potencialidades e capacidades e as suas especificações, assim como o campo de aplicação e as limitações que apresentam;
- assegurar a aplicação das técnicas e metodologias de análise de dados mais adequadas a cada caso concreto, executando correctamente as necessárias opções de procedimento;
- interpretar os resultados das análises;
- apreciar criticamente os resultados finais e as conclusões dos tratamentos analíticos realizados.

Para cada uma das técnicas específicas de análise de dados espera-se que os alunos:

- compreendam os princípios teóricos básicos que os fundamentam;
- saibam em que situações e circunstâncias podem ser aplicados;
- percebam os limites do seu campo de aplicação;
- conheçam os tipos de dados com que é possível aplicá-los;
- tenham a capacidade de efectuar os procedimentos necessários para a execução de cada método e saibam escolher correctamente as opções do seu funcionamento;
- interpretem os resultados da análise, de acordo com os fundamentos teóricos do problema;
- saibam avaliar criticamente os benefícios decorrentes da utilização de cada metodologia.

e ainda, de um modo mais geral, que:

- saibam apreciar criticamente métodos de análise que lhe venham a ser propostos noutras fontes de informação bibliográfica ou por outros especialistas e ser capaz de fazer a sua avaliação;

- tenham capacidade de efectuar trabalhos de análise de dados geográficos integrado em equipas mono ou pluridisciplinares;
- assegurem a continuidade da sua formação, num processo permanente de auto-aprendizagem.

Para além destes objectivos, os alunos terão ainda, como é habitual, de adquirir o domínio das abstracções da disciplina, revelada pela utilização adequada e rigorosa do vocabulário técnico-científico específico e pela capacidade de associar os conceitos metodológicos das técnicas de investigação aos conceitos e às relações próprias do conhecimento geográfico.

I.3 PROGRAMA

1. Considerações Preliminares

O programa que apresentamos é destinado ao segundo nível do ensino de métodos de análise de dados nos cursos do Departamento de Geografia da Faculdade de Letras da Universidade de Lisboa. Com este programa procura-se garantir o conhecimento das técnicas avançadas de análise de dados e complementar o ensino básico que já foi efectuado no ensino elementar destas matérias.

É também nesta disciplina que serão leccionadas as matérias cuja natureza ou dificuldade, não permitia que fossem ensinadas anteriormente.

O programa corresponde pois ao final do ciclo de ensino obrigatório dos métodos de análise de dados e, por consequência, tem de garantir a preparação suficiente dos alunos na problemática destes métodos.

Optou-se por um programa que permite uma organização com alguma variabilidade, constituído por duas partes: uma primeira, em que são apresentados os conceitos básicos fundamentais, ancorados numa bibliografia básica reduzida, mas bastante para que os alunos possam apreender o fundamental da matéria que se ensina, e uma outra parte, em que o peso relativo dos capítulos é facultativo, constituída pelos temas mais avançados, que poderão ser variáveis de ano para ano, ministrados de forma a apoiar, em cada ano, os interesses diferenciados dos alunos.

Esta segunda parte do programa, também poderá ser encarada como uma forma didáctica de fomento da auto-aprendizagem dos futuros geógrafos, tanto pelo modo como é apresentada, como pelas referências bibliográficas e pelos textos de apoio que se incluem no anexo a este programa.

2. A Estrutura do Programa

Em qualquer disciplina, a elaboração do programa é sempre uma escolha. Uma escolha que decorre entre a multiplicidade de temas possíveis, o tempo disponível para leccionação, o posicionamento da disciplina nos planos de estudos dos cursos, o nível de conhecimentos dos alunos, as necessidade que virão a ter noutras disciplinas e, principalmente, os requisitos das actividades futuras que irão desenvolver na sua actividade profissional.

O programa incide fundamentalmente sobre as técnicas avançadas de análise de dados mais tradicionais e mais difundidas, cuja utilização já pode ser considerada um património consolidado da Geografia, mas também se procurou incluir algumas perspectivas mais recentes.

A opção que aqui se tomou foi a de um duplo equilíbrio.

Foi de equilíbrio a opção que se fez no desenvolvimento a dar a cada tema. Entre o tratamento exaustivo de poucas técnicas e a apresentação resumida de muitas, assumiu-se como base para o programa de trabalho a descrição e utilização interpretada das técnicas de análise de dados fundamentais e incontornáveis, tradicionalmente usadas pelos geógrafos.

E foi também de equilíbrio a escolha entre as temáticas tradicionais do passado e as inovações mais recentes. Subjacente a todo o ensino de técnicas quantitativas em Geografia surge, de modo cada vez mais importante, a questão dos SIG, Sistemas de Informação Geográfica, que exigem conhecimentos especializados de complexas análise de dados, e cuja eficiente exploração obriga a uma elevada capacidade na interpretação dos resultados dessas análises.

O ensino proposto consta de uma apresentação e justificação mínimas das formulações matemáticas, de um razoável desenvolvimento dos procedimentos informáticos que conduzem à obtenção de resultados e de um desenvolvimento muito forte dos procedimentos de interpretação.

O programa é extenso. Mas tem de se considerar que os temas da parte final do programa podem, em cada ano, ser desenvolvidos com maior

ou menor profundidade e amplitude, segundo os interesses dos alunos e as condições da actualidade nos cursos.

Os temas do programa estão organizados em doze capítulos, que se podem considerar agrupados em 3 partes.

Na primeira parte encontram-se os três capítulos iniciais.

No primeiro, introdutório, são apresentados os **Conceitos Básicos e de Enquadramento** fundamentais para a análise de dados. Faz-se a revisão dos princípios básicos da metodologia do trabalho científico e discutem-se os tipos de dados geográficos e os tratamentos a que podem ser sujeitos (Cap. 1).

O capítulo seguinte trata de **Alguma Matemática para Geógrafos** (Cap. 2). Aqui é feita a revisão dos conceitos de quocientes e percentagens, como instrumento para compreender a construção e a interpretação de taxas e de índices simples e compostos. Também se apresenta o conceito de função e se definem alguns dos tipos básicos de funções, se desenvolve o conceito o conceito de matriz, e se descrevem as operações matriciais e o seu significado. Tudo isto constitui informação básica necessária para poder executar as análises de que mais adiante se trata.

A **Recolha de Informação** é o assunto do terceiro capítulo. Apresentam-se e discutem-se os tipos de informação geralmente usada nos estudos geográficos, quem a publica ou como pode ser recolhida. Descrevem-se, discutem-se e avaliam-se os diversos métodos de recolha e, por último faz-se a revisão da temática da amostragem, mas agora de uma forma mais dirigida directamente para a investigação geográfica.

A segunda parte, a mais importante da disciplina, consta de seis capítulos e respeita aos tratamentos avançados de análise de dados fundamentais.

Inicia-se com o capítulo quarto sobre a análise de **Tabelas de Contingência**, que são especialmente indicadas para analisar dados qualitativos. São apresentadas formas de transformar o tipo de variáveis, e,

depois, várias medidas de dependência, associação e predição para tabelas de contingência a 2 ou vários estados e para tabelas de Burt.

O capítulo cinco trata da **Análise de Variância** a um e a vários factores e da análise de Covariância, que são métodos auxiliares na busca de relações entre variáveis quantitativas e qualitativas.

O capítulo seguinte, o sexto, é um dos mais importantes, pois é nele que se descreve e se ensina a interpretar a **Regressão Linear Múltipla**, tanto quando é utilizada para descrição, como quando o é para previsão. A regressão linear múltipla é o instrumento ideal para a construção de modelos de previsão e simulação.

Outro tema fundamental é a **Análise Factorial** (Cap. 7). Para além de se descreverem os tipos de análise factorial mais importantes, demonstra-se como se obtêm e interpretam os respectivos resultados. Estas análises são muito utilizadas em análise exploratória de dados, quando ainda não se dispõe de conhecimento estruturado sobre a realidade.

As taxionomias ou **Classificações Numéricas** surgem no capítulo seguinte (o Cap. 8). É um capítulo em que são apresentadas técnicas de classificação, segundo várias variáveis simultaneamente. Para tal apresentam-se várias medidas de distância e de proximidade a utilizar criteriosamente entre as unidades de análise, e diversos métodos de as agregar, agrupar ou classificar.

O capítulo 9 trata da **Análise Discriminante**. É o último capítulo da segunda parte. Neste capítulo, de modo inverso ao anterior das classificações, estudam-se técnicas que permitem compreender quais as variáveis que estão na base da criação de grupos pré-existentes e qual a importância relativa de cada uma.

A terceira parte, que inclui os três capítulos finais, é considerada como informação complementar, em que cada capítulo poderá ser mais ou menos desenvolvido em cada ano, de acordo com as circunstâncias e os interesses dos alunos. Os primeiros dois capítulos são dedicados à análise espacial elementar de dados e o último à modelização em Geografia.

O capítulo 10 trata da análise de **Redes e Grafos**. Nele se apresentam

os conceitos de matriz de conectividade e medidas de ligação ou conectividade, acessibilidade e centralidade. Trata-se das matrizes derivadas da de conectividade e faz-se uma breve introdução à análise de matrizes de fluxo, tanto aditivas como multiplicativas.

O seguinte, o capítulo 11, é sobre **Geoestatísticas**, e nele se estudam as medidas espaciais de centralidade e de dispersão, os potenciais e se analisa a vizinhança.

No último capítulo, o 12, trata-se de uma forma introdutória da construção e utilização de **Modelos Integrados em Geografia**. Referem-se princípios básicos da sua construção e exemplificam-se as possibilidades de utilização de dois tipos de modelos, um demográfico (o *cohort-survival*) e outro de economia regional (*input/output*) isoladamente. Finalmente serão ainda apresentados aos alunos os modelos REX e Espon114--WP5 que se desenvolveram e são utilizados no Centro de Estudos Geográficos.

O objectivo deste último capítulo é o de exemplificar a utilidade da utilização de abordagens mais quantitativas e motivar os alunos para aprofundar estas temáticas, quer no decorrer dos seus cursos, quer mesmo após a conclusão das suas licenciaturas.

3. O Enunciado do Programa

1 - CONCEITOS BÁSICOS E DE ENQUADRAMENTO

1 - Introdução

A investigação científica em geral.

A investigação em Geografia.

O problema do método. O método indutivo e o método dedutivo.

O âmbito da análise. Abordagens extensivas e intensivas: objectivos, tipos de explicação, métodos, limitações, dados necessários e escala adequada.

2 – As Etapas do Trabalho Científico

A questão a tratar.

Seleção do tema e delimitação do âmbito da investigação. Formulação da questão de partida: clareza, exequibilidade e pertinência.

A preparação da investigação

As leituras: seleccionar, ler resumir e comparar

As entrevistas exploratórias. Identificar os peritos e os especialistas e as testemunhas.

Entrevistar: ouvir e decodificar os conteúdos.

Definir o problema

Conciliar leituras, entrevistas e coordenações.

Estabelecer um quadro teórico.

Explicitar a problema.

A construção do esquema básico de trabalho

Identificar e precisar os conceitos. Construir as hipóteses e os modelos.

A observação, a recolha da informação.

Delimitar o campo de observação. Determinar e testar os métodos de observação.

Recolher as informações

.A análise da informação

Arrumar os dados e prepará-los para análise.

Efectuar os procedimentos analíticos.

Comparar os resultados observados com os esperados.

A conclusão

3 – A Natureza da Informação Geográfica

A informação geográfica.

Matriz geográfica, Cubo de Dados de Cattell, Tipologias de informação e de tratamentos analíticos: matrizes de atributos, de proximidade, de dominância e de medição conjunta.

2 - ALGUMA MATEMÁTICA PARA GEÓGRAFOS

1 – Matemática básica

Potências, logaritmos, funções e matrizes.

2 – Índices

Quocientes e Percentagens, Taxas seccionais e de variação. Índices simples e compostos.

3 – Medidas de concentração

Curva de Lorenz. Índices de concentração: coeficiente de especialização, índices de Gini e de Schutz, ponto de igual partilha e coeficiente de Gini.

3 - RECOLHA DA INFORMAÇÃO

1 – Generalidades

2 – Fontes de informação

Fontes Documentais. Estatísticas Nacionais e Internacionais: das Nações Unidas, do Banco Mundial / PNUD, da União Europeia e de outros organismos.

Outras Fontes Documentais. Dados não publicados e informação em arquivos administrativos e de gestão. Integração de inquéritos estatísticos com fontes administrativas. Insuficiência das fontes documentais.

Observação. Exterior e participante. A análise documental.

Inquéritos: Entrevistas e questionários. Tipos, vantagens e limitações. Construção e realização de entrevistas e questionários.

Os grupos de discussão. Os métodos Delphi. As histórias de vida.

3 – Considerações finais

Etapas de uma operação de recolha e procedimentos posteriores.

4 – Amostragem

Introdução. Conceitos básicos. Amostra, valor aproximado, limites do erro ou de confiança, risco.

Estimativas.

Dimensionamento de amostras para populações infinitas e finitas.

Seleção dos elementos da amostra. Métodos probabilísticos de amostragem: aleatória, sistemática, estratificada, por agrupamentos, multi-etapas e multifases. Métodos não-probabilísticos: intencional, bola de neve, por quotas e itinerária.

4 - TABELAS DE CONTINGÊNCIA

1 – Introdução

2 – Transformação para tabelas de contingência.

Transformação por ponderação, quadros lógicos e tabelas de Burt: construção e análise.

3 - Análise de Tabelas a duas variáveis

Tabelas de contingência ($I \times c$) a n estados. Medidas de dependência, associação, e predição: qui quadrado, G quadrado, coeficiente ϕ , V de Cramer e de contingência; coeficiente gama, tau de Kendall, tau de Stuart; coeficientes simétricos e assimétricos D de Somer, de Incerteza, tau de Kruskal e λ .

Tabelas de contingência ($I \times c$) a 2 estados. Qui quadrado simples e corrigido, , coeficiente ϕ , Q de Yule, Y , alfa e logaritmo de alfa.

Apresentação, cálculo e interpretação.

4- Análise de Tabelas de Burt

5 - ANÁLISE DE VARIÂNCIA

1– Introdução.

Objectivos e condições de utilização.

2– Tipos de Análise de Variância.

Análise de Variância a um Factor e Multifactorial.

3– Análise de Covariância.

6 - REGRESSÃO LINEAR MÚLTIPLA

1– Revisão da Regressão Linear Simples.

Conceitos, ajustamento e erro, tendência e resíduos.

2– O Modelo de Regressão Linear Múltipla.

Coeficientes de Regressão: β s, betas e termo independente.

Coeficientes de Correlação e de Determinação Múltipla.

Significado da Regressão. Teste F , teste t , erro-padrão das estimativas e erro-padrão dos coeficientes de regressão.

Tipos de Regressão Múltipla. Simultânea, passo a passo, e outros.

Extensões do modelo. Variáveis mudas, modelos Arima e superfícies de tendência.

7 - ANÁLISE FACTORIAL

1– Introdução.

Objectivos e condições de utilização.

2– Tipos de Análise Factorial.

Análise em Componentes Principais, Análise Factorial propriamente dita e Análise Factorial de Correspondências Múltiplas. Diferenças de modelo matemático, de utilização e de procedimentos.

Preparação da matriz de associação. Extracção dos factores. Rotação dos eixos factoriais. Cálculo do valor dos factores.

3– Marchas Analíticas.

Sequências de uma análise em componentes principais e de uma análise factorial.

8 - CLASSIFICAÇÕES NUMÉRICAS**1– Apresentação do Problema.**

Escolha do tipo de classificação, da medida de associação e da estratégia de agrupamento. Dendogramas.

2– Medidas de Semelhança e de Distância a utilizar.

Medidas de Semelhança e de Distância a Utilizar para dados nominais binários dicotómicos: Coeficiente de Jaccard (1908), Coeficiente de Russel & Rao (1940) e Coeficiente de Dice (1915).

Medidas de Semelhança e de Distância a Utilizar para dados nominais binários alternativos: Coeficiente de Concordância Simples, Coeficiente de Sneath & Sokal, Coeficiente de Youle e Coeficiente de Pearson para dados nominais.

Medidas de Semelhança e de Distância a Utilizar para dados de intervalo e razão: Métricas de Minkowski: distância de Manhattan, euclidiana e de Chebychev; Coeficiente de Correlação de Pearson, Cosseno e distância de Qui-quadrado.

Medidas de Semelhança e de Distância a Utilizar para todos os tipos de dados: Transformações de Parker e de Gower, Funções de valor.

3– Estratégias de Agrupamento.

Para classificações não-hierárquicas (partições): Nuvens dinâmicas, K-médias e K-centróides.

Para classificações hierárquicas ascendentes: métodos do vizinho mais próximo, do vizinho mais distante, da média do grupo, da média simples,

do centróide, da mediana, de Ward e das distâncias internas. A fórmula geral de agregação.

4– Avaliação do Ajustamento.

Árvores de ligação, matrizes de valores e correlações cofenéticas.

9 – ANÁLISE DISCRIMINANTE

1– Introdução.

Objectivos e condições de utilização.

2– O Decurso da Análise.

As funções discriminantes: cálculo e interpretação.

As coordenadas e os centróides dos grupos.

10 – REDES E GRAFOS

1– A Linguagem Topológica: Redes, Grafos e Matrizes.

Redes, Grafos e Matrizes: arcos e nós. Grafos orientados e subgrafos. Matriz de conectividade.

Conceitos: Distância topológica, circuito, região, rede nula, rede em árvore, número associado, nó central e diâmetro da rede.

2– Medidas de Conectividade.

Número ciclomático. Índices beta, alfa e gama.

3– Medidas de Acessibilidade e Centralidade.

Índice de Shimbel, Índice de dispersão.

4– Matrizes Derivadas.

Matrizes C_i , T_i , TG_i e D_i .

5– Matrizes Ponderadas.

Limitações da análise de grafos. Formas de as ultrapassar.

Tratamento de matrizes de carga, simples e cumulativas, e de matrizes de interacção.

11 – GEOESTATÍSTICA

1– Introdução.

Geoestatística, Cartografia e Sistemas de Informação Geográfica.

2– Medidas Espaciais de Centralidade e Dispersão.

Centro de Gravidade, Distância-média, Distância-padrão, Raio Dinâmico e Elipse de Dispersão: ângulo de inclinação, dimensão dos eixos e excentricidade.

3– Potenciais.

Interacção à distância. Cálculo e significado do conceito de potencial. Determinação do efeito da distância.

4– Análise do Vizinho mais Próximo.

O Índice R: determinação do valor e teste do significado.

12 – MODELOS INTEGRADOS EM GEOGRAFIA

1– Introdução.

Os modelos em Geografia: tipos, princípios de construção e utilidade da modelização, na descrição, previsão e simulação de situações e de processos.

Relações entre a modelização e a análise de dados
Relações e funções. O uso de matrizes na descrição de sistemas e processos geográficos complexos.

2– Apresentação de modelos simples.

Descrição das relações fundamentais e da forma como se formalizaram dois modelos de utilização clássica em Geografia regional: um demográfico (*cohort-survival*) e outro económico (*input/output*).

3– Descrição e avaliação de dois modelos regionais integrados

O modelo REX: objectivos, princípios orientadores, relações fundamentais, organização e construção. Capacidades e limitações. Exemplos de aplicação. Crítica dos resultados.

O modelo ESPON114-WP5: objectivos, princípios orientadores, relações fundamentais, organização e construção. Capacidades e limitações. Exemplos de aplicação. Crítica dos resultados.

I.4 MÉTODOS DE ENSINO

Contrariando talvez uma opinião generalizada, consideramos que a leccionação das disciplinas metodológicas nos cursos de Geografia tem uma componente de maior exigência pedagógica que as disciplinas que se debruçam sobre aspectos mais concretos do conhecimento geográfico. Nomeadamente para as disciplinas de técnicas avançadas de análise de dados, sobre as quais os alunos dispõem de pouca informação, o docente tem de motivar os alunos para uma matéria cuja importância, mesmo quando entusiasticamente assumida⁵, é de natureza virtual, pois os alunos não estão ainda em condições de apreciar as vantagens que acarreta para a investigação.

É portanto um tipo de ensino que, para decorrer com sucesso, se torna muito exigente para todos os intervenientes no processo didáctico. Para professor, a quem compete a coordenação das actividades, a apresentação dos temas, a demonstração das metodologias, a indicação das fontes e a orientação do trabalho e a avaliação dos alunos. E para estes, a quem se pede que aprendam, num esforço organizado, responsável e intenso, em que também se deve fomentar a independência. Com efeito, se um dos objectivos principais do meu ensino é que os alunos aprendam a aprender, continuando a sua formação nos diversos assuntos por que se interessam ou de que possam vir a necessitar, muito para além do tempo em que são alunos, então é fundamental que se vão habituando a um trabalho independente, embora ainda devidamente apoiado em tudo o que for necessário.

A especificidade da matéria obriga à alternância de aulas teóricas de exposição e exemplificação temática, com aulas práticas de aplicação nas quais os alunos resolvem, de forma acompanhada, os problemas de cada técnica que lhes é sugerida pelo docente. Mas uma parte importante do trabalho do aluno deve decorrer fora da aula, executando exercícios, inicialmente totalmente resolvidos e depois em que são fornecidas as soluções, de modo a que possa, assim, ir controlando a sua própria

⁵ - O que raramente acontece.

aprendizagem.

As **aulas teóricas** serão lições magistrais, em que o professor expõe, discute e exemplifica as técnicas de análise de dados, e delimita as condições de aplicação. Os alunos participarão, por via de pedidos de esclarecimento e através de perguntas, de modo a que se obtenha o esclarecimento das dúvidas, mas o seu papel será bastante passivo, pelo quase completo desconhecimento que têm das temáticas abordadas. Será ao discutir as possibilidades e potencialidades das aplicações que se esperará a maior participação.

Todos os capítulos do programa serão inicialmente apresentados desta forma. Estas aulas permitem a transmissão do conhecimento das técnicas e da sua utilização, a visão crítica e a possibilidade de transmitir experiências e dificuldades já vividas pelo docente.

A preparação das aulas deve ser cuidadosa, pela novidade de que a maioria dos temas se reveste para os alunos. Por isso, deve evitar-se a prolixidade e a exemplificação excessiva, que podem vir a tornar confusa a exposição e mais difícil a compreensão.

Deve ter-se especial atenção à sequência de apresentação dos temas, dos problemas e dos exemplos. A regra é ir sempre do mais fácil para o mais difícil e sem grandes saltos na dificuldade, pelo que é muito desejável que a lição inclua mecanismos de controle do processo de aquisição de conhecimentos dos alunos, a fim de se poder recuperar eventuais incompreensões.

As **aulas práticas** são o espaço por excelência da participação dos alunos. Serão iniciadas por uma pequena apresentação do trabalho a realizar, que será de seguida efectuado de forma autónoma, sempre com auxílio de computador, por grupos de dois alunos.

As dificuldades terão de ser experimentadas pelos alunos e a ajuda, imprescindível, do docente nunca as deve antecipar. Para além deste processo de ensaio e erro em ambiente de aula, os alunos controlarão a sua aprendizagem através da execução individual dos exemplos e da resolução de problemas com soluções.

Como se disse, grande parte do trabalho decorre fora da aula, como é o caso da maioria dos exercícios de aplicação e da execução do trabalho prático de grupo. Mas com o objectivo de ajudar os alunos na organização e na calendarização do seu trabalho, parte de algumas aulas será dedicada à orientação do trabalho a fazer e à verificação do trabalho já efectuado, de modo a permitir a aprendizagem em tempo útil de cada uma das técnicas apresentadas.

O trabalho de grupo, para além de ser um valioso instrumento de avaliação, é também um processo pedagógico de consolidação de conhecimentos e uma prática indispensável para quem irá trabalhar em planeamento integrado em grandes ou pequenas equipas. Constará da aplicação das técnicas de análise de dados mais difundida a um caso concreto de investigação geográfica, seguindo todas as etapas, desde a escolha do tema, à recolha da informação de base e à execução dos cálculos, até à interpretação e à redacção e apresentação dos resultados finais. A escolha do tema tem-se revelado sempre o obstáculo maior para os alunos, pela dificuldade que têm tido em fazer corresponder os conceitos teóricos do tema que querem tratar com a capacidade em identificar e seleccionar as variáveis necessárias e suficientes para a sua compreensão.

I.5 AVALIAÇÃO

Pretende-se que os alunos, para além de serem capazes de saber usar as técnicas fundamentais e de interpretar os seus resultados, consigam utilizá-las com eficácia na investigação de temas geográficos. Em suma, pretende-se que saibam escolher as técnicas mais adequadas a cada investigação em concreto, e sejam capazes de as aplicar e de interpretar os resultados.

Por outro lado a avaliação constitui, em nosso entender, um elemento essencial no processo de aprendizagem, pois é durante os procedimentos de avaliação e de correcção que os alunos estão mais atentos e motivados e portanto mais capazes de reconhecer e emendar os erros que fizeram.

Importa por isso que a correcção dos elementos de avaliação intercalar seja sempre pública, atempada e exaustiva, de modo a que o aluno possa orientar o seu trabalho e avaliar o esforço que tem a fazer, de acordo com um processo que também tem algo de autoavaliação. A correcção dos elementos finais de avaliação também deve ser pública e exaustiva, embora o seu papel didáctico seja agora menos importante, pois a maioria dos alunos, satisfeitos com as classificações obtidas, já estão menos motivados e interessados. Não obstante pode ainda desempenhar um papel importante nos alunos que reprovaram, nos que estão entre o satisfatório e o não satisfatório e naqueles que desejam fazer melhoria de nota no próximo ano,

A avaliação destina-se a verificar até que ponto os objectivos da disciplina são alcançados. Como os objectivos respeitam a técnicas de análise de dados, o fundamental é avaliar se os alunos sabem utilizar os diversos métodos, interpretando os resultados, e se sabem aplicá-los aos diferentes temas da Geografia. Deste modo, a avaliação baseia-se em dois eixos fundamentais, o de saber utilizar os métodos e o de ser capaz de os aplicar a casos concretos da investigação geográfica.

A avaliação tem ainda de cumprir o que se encontra determinado no

Regulamento Geral de Avaliação da FLUL⁶. Nos termos deste regulamento, a avaliação deve ser apresentada aos alunos na primeira aula. E da apresentação devem constar os elementos de avaliação e a respectiva ponderação.

Para os alunos normais, que devem comparecer a pelo menos 75% das aulas, os elementos de avaliação que propomos são os seguintes:

Um teste presencial individual, a realizar no final do período de aulas, com a duração de duas horas, que abrange todos conteúdos da disciplina. Nele se procura que o aluno resolva problemas simples e interprete resultados de análises que lhe são apresentados. Uma parte do teste incidirá sobre os trabalhos de grupo que efectuaram.

Um trabalho de grupo, que é desenvolvido ao longo do período lectivo e é apresentado no final, sob a forma de um relatório escrito, e ou apresentado oralmente no final do semestre. O grupo de trabalho será constituído no máximo por três alunos (desejável 2) e terá uma nota conjunta para a parte escrita e individual para a discussão ou apresentação oral. Com este trabalho pretende-se que os alunos mostrem a sua capacidade de trabalhar em grupo, organizando-se para as diferentes tarefas e colaborando conjuntamente na definição do modo como estruturam o trabalho na execução dos cálculos e na interpretação dos resultados.

Participação nas aulas: discussão de exemplos, execução de exercícios, assiduidade e interesse nos temas tratados ao longo do semestre.

A ponderação de cada elemento de avaliação na definição da classificação final, é actualmente a seguinte:

- Teste individual	55%
- Trabalho de grupo	45%
- Participação	5%

⁶ O Regulamento Geral de Avaliação da FLUL foi estabelecido em 16 de janeiro de 2003 por despacho conjunto dos órgãos de gestão da Faculdade. O seu Artº. 7, nº 2, foi alterado pela deliberação do Conselhup Pedagógico de 23 de Abril de 2004.

Os outros alunos (voluntários, trabalhadores-estudantes, dirigentes associativos, atletas de alta competição, etc.) são avaliados de forma diferente, nos termos do Regulamento Geral de Avaliação. Não lhes é pedida assiduidade, mas terá de lhes ser exigido o mesmo que aos outros, embora por formas e prazos diferentes, mais adequados ao facto de não lhes ser possível frequentar regularmente as aulas.

Nesta disciplina, estes alunos geralmente fazem o mesmo teste e apresentam o trabalho nos mesmos prazos, salvo combinação em contrário com o docente, devido a incompatibilidades de calendário, devidamente justificadas. No mais, segue-se o Regulamento Geral de Avaliação e a legislação específica que lhes permite fazer exames em épocas especiais, sempre que o requeiram.

I.6 MATERIAIS DE APOIO

Dada a especificidade da disciplina, os instrumentos e materiais de apoio revestem-se de um interesse particular.

É necessário garantir que os alunos tenham acesso a meios mecânicos de cálculo, dentro e fora dos períodos de aulas.

Para as aulas práticas é indispensável a utilização das salas de informática da Faculdade, cujo equipamento é satisfatório para as necessidades da disciplina.

Para as aulas teóricas é muito conveniente a utilização permanente de uma máquina de calcular com funções estatísticas, que permita ir fazendo as pequenas operações de acompanhamento dos exemplos e dos problemas que são apresentados.

Fora da aula é indispensável ter acesso a um computador equipado com programas de tratamento estatístico avançado. Alguns dos alunos dispõem de um computador com estas condições, mas isso não acontece com a maioria deles. Torna-se assim necessário garantir a utilização das salas de informática fora do período das aulas. A solução que tem sido encontrada, e que se revela satisfatória, tem sido a de marcar a sala e disponibilizá-la para os trabalhos dos alunos, de forma apoiada, nos períodos que estão previstos para o docente fazer o atendimento aos alunos.

Mas para além do acesso às máquinas é necessário garantir a disponibilidade dos programas necessários. Grande parte da matéria da disciplina é coberta pelos programas comerciais de estatística avançada, como o SPSS⁷ ou o STATISTICA. São programas muito bons, mas bastante caros, pelo que a aquisição e actualização das licenças de utilização tem de estar prevista no orçamento da Faculdade e do Departamento, se for caso disso.

Há uma parte da matéria – os grafos e redes e a Geoestatística – para a qual não dispomos de programas comerciais. O problema tem sido

⁷ - SPSS – Statistical Package for Social Sciences.

resolvido com o recurso a exemplos de pequena dimensão e a programas próprios desenvolvidos no âmbito do Centro de Estudos Geográficos.

Ainda em relação aos materiais de apoio há a referir as tabelas estatísticas, indispensáveis para a validação dos resultados obtidos, mas que, por se encontrarem geralmente nos livros de apoio às metodologias, serão tratadas na bibliografia.

I.7 BIBLIOGRAFIA

A bibliografia que se apresenta é bastante extensa, mas essa foi uma opção que decidimos assumir.

Na realidade, o fundamental da disciplina decorre no trabalho que se faz em computador, utilizando programas comerciais de análise estatística avançada. Uma parte do apoio bibliográfico virá dos “*help*”⁸ e dos “*tutorial*”⁹ desses programas, dos textos que apresentamos sobre o desenvolvimento dos conteúdos dos capítulos deste programa (em Anexo).

A bibliografia que indicamos desempenha nesta disciplina um papel de apoio, supletivo, de esclarecimento, de exemplificação das técnicas e de motivação para novas experiências no campo da aplicação de técnicas quantitativas em Geografia.

Está organizada em três conjuntos,

Num **primeiro conjunto** agruparam-se as obras básicas e fundamentais, cuja leitura se aconselha, e que são o apoio seguro às dúvidas dos alunos. Estão marcadas com um asterisco (*) nas listas bibliográficas. Foram seleccionadas tendo em conta o rigor das apresentações e a clareza da exposição. Muitas delas foram escolhidas por serem obras clássicas em que estão apresentadas de forma definitiva as ideias originais, outras são exposições brilhantes e claras, elaboradas posteriormente. Procurou-se apresentar novas perspectivas, sempre que tal se justifica, tendo em conta a perturbação que muita novidade poderia acarretar para os alunos, dado o nível ainda incipiente dos seus conhecimentos. Até por uma questão de economia nas aquisições, evitou-se a indicação de algumas obras modernas que apenas repetem o que já foi dito inúmeras vezes, sem acrescentar qualquer outra mais valia de conhecimento.

⁸ - *Help* – programa de apoio existente nos programas comerciais que esclarecem o utilizador sobre o funcionamento destes. Muito desiguais, a sua qualidade varia entre o muito bom e o péssimo. O do *Statistica* é bastante bom.

⁹ - *Tutorial* – programa que acompanha por vezes os programas comerciais, constituídos por um conjunto de lições em que se ensina o modo de os utilizar e se esclarece, em maior ou menor grau, as técnicas de análise que nele existem.

Procurou-se que as obras indicadas existam na biblioteca de Geografia, ou então que possam ser disponibilizadas aos alunos pelo professor da disciplina.

Num **segundo grupo** integraram-se obras com aspectos complementares, que abordam com maior profundidade ou desenvolvimento parte dos conteúdos programáticos da disciplina, e/ou em que se apresentam exemplos de tratamento e aplicação a casos geográficos. Não é indispensável agora a sua leitura pelos alunos, embora em muitos casos seja conveniente. A razão fundamental para a sua inclusão decorre do interesse que os alunos possam vir a ter em desenvolver de forma autónoma maior competência nas análises de dados e o nosso entendimento de que esta bibliografia poderá vir a constituir um bom começo para tal.

O **terceiro conjunto** já foi referido. É constituído pelos textos de apoio do anexo e pela respectiva bibliografia, e encontra-se mais vocacionada para o apoio directo e concreto às situações didácticas que serão provocadas no ambiente da aula.

Como regra geral, o objectivo foi que:

- a bibliografia fosse perfeitamente adequada ao nível do conhecimento dos alunos, embora em alguns casos tal desiderato não tivesse sido possível de conseguir;
- a selecção dos exemplos contemplasse a diversidade no tempo, no espaço e nas diversas escalas de análise;
- que se evitasse indicar manuais técnicos com complexas formulações matemáticas, privilegiando os manuais de exemplificação em que o objectivo fundamental não é o docente mostrar quanto sabe, mas sim permitir que o aluno aprenda facilmente e em grande parte por si mesmo.

Consideramos ainda que os textos que apresentamos de forma experimental no anexo, possam vir a constituir, depois de mais amplamente discutidos, uma base para a elaboração de um manual de suporte para a leccionação desta disciplina e para a aprendizagem dos alunos interessados em desenvolver estas matérias.

Antes da apresentação da bibliografia que escolhemos, uma palavra ainda. É preciso estar consciente de que há uma enorme produção bibliográfica sobre estas matérias, e que as obras que se seleccionaram são apenas um pequeno segmento das possibilidades que se poderiam apresentar. Isto significa que é possível que os alunos tenham acesso a obras que não sendo referidas nestas bibliografias, podem constituir óptimos suportes para a aprendizagem. Deve portanto o docente dar todo o apoio aos alunos e este desenvolverem a máxima interacção com o docente no aspecto do aconselhamento bibliográfico, de modo a otimizar a utilização dos recursos que tenham disponíveis.

Bibliografia Geral

As referências bibliográficas indicadas com um asterisco () são as que consideramos mais importantes para a compreensão dos temas tratados.*

- BOSQUE SENDRA, Joaquín; MORENO JIMÉNEZ, Antonio (1994) – *Prácticas de Análisis Exploratório y Multivariante de Datos*, Oikos-Rau, Barcelona, 214 p.
- BRYMAN, Alan; CRAMER, Duncan (1993) – *Análise de Dados em Ciências Sociais. Introdução às técnicas utilizando o SPSS*, Celta Editores, Oeiras, 346 p.
- CICERI, M.F.; MARCHANT, B.; RIMBERT, S. (1977) – *Introduction a l'Analyse de l'Espace*, Masson, Paris, 173 p.
- CLIFFORD, Nicholas J.; VALENTINE, Gill (2003) – *Key Methods in Geography*, Sage Publications, London, 572 p.
- COLE, John P.; KING, Cuchlaine A. M. (1970) – *Quantitative Geography*, John Wiley & Sons, London, 692 p.
- D'HAINAUT, Louis (1992) – *Conceitos e Métodos da Estatística*. Volume 2, Fundação Calouste Gulbenkian, Lisboa, 366 p.
- DAVIS, John C. (1986) – *Statistics and Data Analysis in Geology*, 2nd edition, New York, John Wiley & Sons, 646 p.
- DAVIS, John C. (2002) – *Statistics and Data Analysis in Geology*, 3th edition, New York, John Wiley & Sons, 638 p.
- (*) EBDON, David (1996) – *Statistics in Geography*. 2nd edition, Blackwell, Oxford (UK), 232 p.
- FOTHERINGHAM, A. Stewart (2000) – *Quantitative Geography. Perspectives on Spatial Data Analysis*, Sage Publications, London, 270 p.

- HARMAN, Harry (1967) – *Modern Factor Analysis*, 2nd edition, The University of Chicago Press, Chicago, 474 p.
- HOGG, Robert V.; TANIS, Elliot A. (1997) – *Probability and Statistical Inference*, Prentice Hall, Upper Saddle River (NJ), 722 p.
- HOLLOWAY, Sarah L., et alia (2003) – *Key Concepts in Geography*, Sage Publications, London, 342 p.
- ISARD, Walter et alia (1998) – *Methods of Interregional and Regional Analysis*, Ashgate, Aldershot, 490 p.
- LAW, Averril M., KELTON, W. David (1991) – *Simulation Modelling & Analysis*, 2nd edition, McGraw-Hill International editions, New York, 759 p.
- LIGHTHILL, James (1978) – *Newer Uses of Mathematics*, Penguin Books, Harmondsworth, 431 p.
- LIMA Marinus P. (2000) – *Inquérito Sociológico. Problemas de Metodologia*, Editorial Presença, Lisboa, 95 p.
- MATHER, Paul (1976) – *Computational Methods of Multivariate Analysis in Physical Geography*, John Wiley & Sons, London, 532 p.
- MORRISON, Donald F. (1976) – *Multivariate Statistical Methods*, McGraw-Hill Kogakusha, Ltd., Tóquio, 415 p.
- NEWTON, P. W.; TAYLOR, M.A.P.; SHARPE, R. (Eds) (1988) – *Desktop Planning. Microcomputer applications for infrastructure and services planning and management*, Hargreen Publishing Co., Melbourne, 398 p.
- PELOSI, Marilyn K.; SANDIFER, Theresa M. (2002) – *Doing Statistics for Business with Excell*, 2nd edition, 858 p. + anexos.
- (*) PESTANA, Maria Helena; GAGEIRO, João Nunes (1998) – *Análise de Dados para Ciências Sociais. A complementaridade do SPSS*, Lisboa, Edições Sílabo, 478 p.
- PINDYCK, R. S.; RUBINFELD, Daniel L. (1991) – *Econometric Models & Economic Forecasts*, 3th edition, MaGraw-Hill Inc., New York, 596 p.
- (*) REIS, Elizabeth (1997) – *Estatística Multivariada Aplicada*, Lisboa, Edições Sílabo, 343 p.
- REIS, Elizabeth (1998) – *Estatística Descritiva*, Lisboa, Edições Sílabo, 245 p.
- ROBSON, Colin (1974) – *Experiment, Design and Statistics in Psychology*, Penguin Books, Harmondsworth, 174 p.
- THIESSEN, Heiner (1996) – *Measuring the real world. A textbook of Applied Statistical Methods*, John Wiley & Sons, Chicester, 268 p.
- VIGNERON, Emmanuel (1997) – *Géographie et Statistique*, P.U.F., Paris, 127 p.
- VOLLE, Michel (1978) – *Analyse des Donnees*, Economica, Paris, 267 p.
- WRIGLEY, N.; BENNETT, R.J. (Eds) (1981) – *Quantitative Geography*, Routledge & Kegan Paul, London, 419 p.
- YEATES, Maurice (1974) – *An Introduction to Quantitative Analysis in Human Geography*, MacGraw-Hil, New York, 300 p.

Bibliografia por Capítulos

1. Conceitos Básicos e de Enquadramento

- BRYMAN, Alan; CRAMER, Duncan (1993) – *Análise de Dados em Ciências Sociais. Introdução às técnicas utilizando o SPSS*, Celta Editores, Oeiras, p. 1-20; 233-264.
- FOTHERINGHAM, A. Stewart (2000) – *Quantitative Geography. Perspectives on Spatial Data Analysis*, Sage Publications, London, p. 1-14.
- HARVEY, David (1973) – *Explanation in Geography*, Edward Arnold, London, 521 p.
- HOLLOWAY, Sarah L., et alia (2003) – *Key Concepts in Geography*, Sage Publications, London, 342 p.

2. Alguma Matemática para Geógrafos

- CHIANG, Alpha (1982) – *“Matemática para Economistas”*, MacGraw-Hill do Brasil, São Paulo, 684 p.
- DAVIS, John C. (1986) – *Statistics and Data Analysis in Geology*, 2nd edition, New York, John Wiley & Sons, p. 107-140.
- EVANS, I.S.; JONES, k. (1981) *“Ratios and closed number Systems”*, in WRIGLEY, N.; BENNETT, R.J. (Eds) – *Quantitative Geography*, Routledge & Kegan Paul, London, p. 11-122.
- MATHER, Paul (1976) – *Computational Methods of Multivariate Analysis in Physical Geography*, John Wiley & Sons, London, p. 9-36.
- MORRISON, Donald F. (1976) – *Multivariate Statistical Methods*, McGraw-Hill Kogakusha, Ltd., Tóquio, p. 1-78.
- SMITH, David M. (1977) – *“Patterns in Human Geography”*, Penguin Books, Harmondsworth, 373 p.
- THIESSEN, Heiner (1996) – *Measuring the real world. A textbook of Applied Statistical Methods*, John Wiley & Sons, Chicester, p.10-186
- TURNER, J. C. (1970) – *Matemática moderna aplicada*, Alianza Editorial, Madrid, p. 17-106.
- WILSON, A. G.; KIRKBY. M.J. (1975) – *Mathematics for Geographers and Planners*, Claredon Press, Oxford, 325 p.

3. Recolha da Informação

- ABREU, Diogo de (1998) – *Recolha e Tratamento de Informação. Manual de Apoio ao 3º Curso de Técnicos Superiores de Planeamento na Administração Regional*. IGE/Partex, Luanda, p. 6-37.
- ALMEIDA, J. F. de, PINTO, J. M. (1976) – *A Investigação nas Ciências Sociais*, Lisboa, Editorial Presença, p. 105-113.
- BRYMAN, Alan; CRAMER, Duncan (1993) – *Análise de Dados em Ciências Sociais. Introdução às técnicas utilizando o SPSS*, Celta Editores, Oeiras, p. 121-140.

- CALOT, Gérard (1970) – *Curso de Estadística Descriptiva*, Paraninfo, Madrid, p. 94-99. É tradução castelhana do *Cours de Statistique Descriptive*, Dunod, Paris, 1965, p. 479-507.
- GARCIA BALLESTEROS, Aurora (Coord.) (1998) – *Métodos y Técnicas Cualitativas en Geografía Social*, Oikos-Tau, , Barcelona , 239 p.
- CAPLOW, Theodore (1972) – *La Investigación Sociológica*, Editorial Laia, Barcelona , p. 260-266
- CLIFFORD, Nicholas J.; VALENTINE, Gill (2003) – *Key Methods in Geography*, Sage Publications, London, 572 p.
- DUVERGER, Maurice (1964) – *Méthodes des Sciences Sociales*, 1ª ed., P.U.F., Paris, p. 251-266.
- (*) ESTEBANEZ ALVAREZ, J.; BRADSHAW, R.. P. (1978) – *Técnicas de Quantificación en Geografía*, Editorial Tebar Flores, Madrid, p. 277-336.
- LONGHURST, Robyn. (2003) – “*Semi-structured Interviews and Focus Groups*”, in Clifford & Valentine – *Key Methods in Geography*, Sage Publications, London, p. 117-132.
- MASON, Jennifer (2000) – *Qualitative Research*, Sage Publication, London, 223 p.
- MOSER, C. A.; KALTON, G. (1975) – *Survey Methods in Social Investigation*, Heinemann Educational Books, London, 549 p.
- REIS, Elizabeth *et alia* (2001) – *Estatística Aplicada*, Vol. 2, Edições Sílabo, Lisboa, 322 p.
- (*) VICENTE, Paula *et alia* (1996) – *Sondagens*, Edições Sílabo, Lisboa, 238 p.
- (*) VILARES, M. J. (1993) – *Sistemas de Informação Estatística*, ISEGI, Lisboa, p. 277-336.
- WRIGLEY, N.; BENNETT, R.J. (Eds) (1981) – *Quantitative Geography*, Routledge & Kegan Paul, London, p. 17-69.
- YIN, Robert K. (1964) – *Case Study Research. Design and Methods*. 2nd edition. SAGE Publication, London, 171 p.

4. Tabelas de Contingência

- BRYMAN, Alan; CRAMER, Duncan (1993) – *Análise de Dados em Ciências Sociais. Introdução às técnicas utilizando o SPSS*, Celta Editores, Oeiras, p. 265-278.
- D’HAINAUT, Louis (1992) – *Conceitos e Métodos da Estatística. Volume 2*, Fundação Calouste Gulbenkian, Lisboa, p. 70-119.
- MELO, F. Galvão de (1982) – *Modelos e Métodos Estatísticos em Psicologia*, Escola Nacional de Saúde Pública, Lisboa, p. 38-53.
- PELOSI, Marylin K.; SANDIFER, Theresa M. (2002) – *Doing Statistics for Business with Excell*, 2nd edition, p. 760-811.
- (*) PESTANA, Maria Helena; GAGEIRO, João Nunes (1998) – *Análise de Dados para Ciências Sociais. A complementaridade do SPSS*, Lisboa, Edições Sílabo, p. 91- 146.

WRIGLEY, N. (1981) "*Categorical Data Analysis*", in WRIGLEY, N.; BENNETT, R.J. (Eds) – *Quantitative Geography*, Routledge & Kegan Paul, London, p. 111-122.

5. Análise da Variância

BRYMAN, Alan; CRAMER, Duncan (1993) – *Análise de Dados em Ciências Sociais. Introdução às técnicas utilizando o SPSS*, Celta Editores, Oeiras, p. 233-264.

D'HAINAUT, Louis (1992) – *Conceitos e Métodos da Estatística. Volume 2*, Fundação Calouste Gulbenkian, Lisboa, p. 150-282.

DAVIS, John C. (1986) – *Statistics and Data Analysis in Geology*, 2nd edition, New York, John Wiley & Sons, p. 74-79.

(*) ESTEBANEZ ALVAREZ, J.; BRADSHAW, R.. P. (1978) – *Técnicas de Quantificación en Geografía*, Editorial Tebar Flores, Madrid, p. 381-386.

IVERSEN, G. R.; NORPOTH, H. (1976) – *Analysis of Variance*, Sage Publications, London, 95 p.

GMURMAN, V. E. (1984) – *Problemas em Probabilidades e Estatística*, Editora Mir, Moscovo, p. 236-245.

(*) KIM, Jae-On; KOHOUT, F. J. (1975) - "*Analysis of Variance and Covariance*" in NIE, Norman H. et alia - *SPSS -Statistical Package for Social Sciences*, 2th edition, McGraw-Hill Book Company, New York, p. 398-432.

MORRISON, Donald F. (1976) – *Multivariate Statistical Methods*, McGraw-Hill Kogakusha, Ltd., Tóquio, p. 170-229.

PELOSI, Marilyn K.; SANDIFER, Theresa M. (2002) – *Doing Statistics for Business with Excell*, 2nd edition, p. 695-759.

(*) PESTANA, M. H.; GAGEIRO, J. N. (1998) - *Análise de Dados para Ciências Sociais. A Complementaridade do SPSS*, Edições Sílabo, Lisboa, p. 177-243.

STEYN, A.G.W. et alia (1996) – *Modern Statistics in Practice*, J. L. van Schaik, Pretoria ,p. 508-546.

YEATES, Maurice (1974) – *An Introduction to Quantitative Analysis in Human Geography*, MacGraw-Hil, New York, p. 131-139.

6. Regressão Múltipla

BRYMAN, Alan; CRAMER, Duncan (1993) – *Análise de Dados em Ciências Sociais. Introdução às técnicas utilizando o SPSS*, Celta Editores, Oeiras, p. 285-298.

(*) CURWIN, Jon; SLATER, Roger (1996) – *Quantitative Methods for Business Decisions*, London, International Thomson Business Press, pp. 378 420, especialmente 397 a 420.

D'HAINAUT, Louis (1992) – *Conceitos e Métodos da Estatística. Volume 2*, Fundação Calouste Gulbenkian, Lisboa, p. 120-149.

DAVIS, John C. (1986) – *Statistics and Data Analysis in Geology*, 2nd edition, New York, John Wiley & Sons, p. 468-478.

- FERGUSON, Rob (1979) – *Linear Regression in Geography*, London, Institute of British Geographers, 44 p. (CATMOG, 15)
- LEWIS-BECK, Michael (1980) – *Applied Regression. An Introduction*, London, Sage Publications, 77 p. (SAGE, 22)
- (*) KIM, Jae-On; KOHOUT, F. J. (1975) – “Multiple Regression Analysis” in NIE, Norman H. et alia - *SPSS -Statistical Package for Social Sciences*, 2th edition, McGraw-Hill Book Company, New York, p. 320-367.
- MATHER, Paul (1976) – *Computational Methods of Multivariate Analysis in Physical Geography*, John Wiley & Sons, London, p. 39-198.
- PARIKH, Ashok; BAILEY, David (1990) – *Techniques of Economic Analysis with applications*, Harvester Wheatsheaf, New York, 357 p.
- PELOSI, Marylin K.; SANDIFER, Theresa M. (2002) – *Doing Statistics for Business with Excell*, 2nd edition, p. 512-630.
- (*) PESTANA, Maria Helena; GAGEIRO, João Nunes (1998) – *Análise de Dados para Ciências Sociais. A complementaridade do SPSS*, Lisboa, Edições Sílabo, p. 361-428.
- SILK, J. (1981) “The General Linear Model”, in WRIGLEY, N.; BENNETT, R.J. (Eds) – *Quantitative Geography*, Routledge & Kegan Paul, London, p. 75-85.
- YEATES, Maurice (1974) – *An Introduction to Quantitative Analysis in Human Geography*, MacGraw-Hil, New York, p. 98-122.

7. Análise Fatorial

- BENZÉCRI, J.-P. et collaborateurs (1973) – *L'Analyse des Données. 2 - L'Analyse des Correspondences*, Dunod, Paris, 619 p.
- BRYMAN, Alan; CRAMER, Duncan (1993) – *Análise de Dados em Ciências Sociais. Introdução às técnicas utilizando o SPSS*, Celta Editores, Oeiras, p. 307-323.
- (*) CICERI, M.F.; MARCHANT, B.; RIMBERT, S. (1977) – *Introduction a l'Analyse de l'Espace*, Masson, Paris, p. 13-42.
- DAVIS, John C. (1986) – *Statistics and Data Analysis in Geology*, 2nd edition, New York, John Wiley & Sons, p. 515-573.
- HARMAN, Harry (1967) – *Modern Factor Analysis*, 2nd edition, The University of Chicago Press, Chicago, 474 p.
- (*) KIM, Jae-On (1975) – “Factor Analysis” in NIE, Norman H. et alia - *SPSS - Statistical Package for Social Sciences*, 2th edition, McGraw-Hill Book Company, New York, p. 468-514.
- MORENO JIMÉNEZ, Antonio (1994) – “Análisis Factorial” in BOSQUE SENDRA, Joaquín; MORENO JIMÉNEZ, Antonio – *Prácticas de Análisis Exploratório y Multivariante de Datos*, Oikos-Rau, Barcelona, p. 87-131.
- MATHER, Paul (1976) – *Computational Methods of Multivariate Analysis in Physical Geography*, John Wiley & Sons, London, p. 215-306.
- MATHER, P.M. (1981) “Factor Analysis”, in WRIGLEY, N.; BENNETT, R.J. (Eds) – *Quantitative Geography*, Routledge & Kegan Paul, London, p. 144-150.

- MELO, F. Galvão de (1982) – *Modelos e Métodos Estatísticos em Psicologia*, Escola Nacional de Saúde Pública, Lisboa, p. 54-112.
- MORRISON, Donald F. (1976) – *Multivariate Statistical Methods*, McGraw-Hill Kogakusha, Ltd., Tóquio, p. 226-345.
- (*) PESTANA, Maria Helena; GAGEIRO, João Nunes (1998) – *Análise de Dados para Ciências Sociais. A complementaridade do SPSS*, Lisboa, Edições Sílabo, p. 321-344.
- (*) REIS, Elizabeth (1997) – *Estatística Multivariada Aplicada*, Lisboa, Edições Sílabo, p. 251-284.
- TORRENS-IBERN, J. (1972) – *Modèles et Méthodes de l'Analyse Factorielle*, Dunod, Paris, 202 p.
- VOLLE, Michel (1978) – *Analyse des Donnees*, Economica, Paris, p. 61-194 p.
- YEATES, Maurice (1974) – *An Introduction to Quantitative Analysis in Human Geography*, MacGraw-Hil, New York, p. 207-238.

8. Classificação Numérica

- BENZÉCRI, J.-P. et collaborateurs (1973) – *L'Analyse des Donnees. 1 – La Taxinomie*, Dunod, Paris.
- (*) BROCHADO, Ana Margarida (2002) – “A análise de clusters: técnica de classificação na análise espacial” in COSTA, José da Silva (Coord.), *Compêndio de Economia Regional*, Coimbra, A.P.D.R., p. 743 - 766.
- DAVIS, John C. (1986) – *Statistics and Data Analysis in Geology*, 2nd edition, New York, John Wiley & Sons, p. 502-514.
- MATHER, Paul (1976) – *Computational Methods of Multivariate Analysis in Physical Geography*, John Wiley & Sons, London, p. 309-329, 342-351.
- MORENO JIMÉNEZ, Antonio (1994) – “Classificación Multivariante” in BOSQUE SENDRA, Joaquín; MORENO JIMÉNEZ, Antonio – *Prácticas de Análisis Exploratório y Multivariante de Datos*, Oikos-Rau, Barcelona, 133-172 p.
- (*) PESTANA, Maria Helena; GAGEIRO, João Nunes (1998) – *Análise de Dados para Ciências Sociais. A complementaridade do SPSS*, Lisboa, Edições Sílabo, p. 345 – 354.
- (*) REIS, Elizabeth (1997) – *Estatística Multivariada Aplicada*, Lisboa, Edições Sílabo, p. 285 – 336.
- VOLLE, Michel (1978) – *Analyse des Donnees*, Economica, Paris, p. 195-265.

9. Análise Discriminante

- DAVIS, John C. (1986) – *Statistics and Data Analysis in Geology*, 2nd edition, New York, John Wiley & Sons, p. 602-606.
- (*) KLECKA, William R. (1975) – “Discriminant Analysis” in NIE, Norman H. et alia - *SPSS -Statistical Package for Social Sciences*, 2th edition, McGraw-Hill Book Company, New York, p. 434-448.
- MATHER, Paul (1976) – *Computational Methods of Multivariate Analysis in Physical Geography*, John Wiley & Sons, London, p. 420- 459.

- (*) REIS, Elizabeth (1997) – *Estatística Multivariada Aplicada*, Lisboa, Edições Sílabo, p. 199-249.

10. Redes e Grafos

- (*) CICERI, M.F.; MARCHANT, B.; RIMBERT, S. (1977) – *Introduction a l'Analyse de l'Espace*, Masson, Paris, p. 45-53.
- CLIFF, A.D.; HAGGETT, P. (1981) "Graph Theory", in WRIGLEY, N.; BENNETT, R.J. (Eds) – *Quantitative Geography*, Routledge & Kegan Paul, London, p. 225-234.
- HAGGETT, P.; CHORLEY, R. J. (1969) – *Network Analysis in Geography*, Edward Arnold, London. 348 p.
- POTTS, R.B. (1978) "Networks" in LIDTHILL – *Newer Uses of Mathematics*, Penguin Books, Harmondsworth, 431 p.
- SEGUÍ PONS, Joana M.; PETRUS BEY, Joana M. (1991) – *Geografía de Redes y Sistemas de Transporte*, Editorial Sintesis, Madrid, 241 p.
- SMITH, David M. (1977) – *Patterns in Human Geography*, Penguin Books, Harmondsworth, p. 279-285.
- TAAFFE, E. J.; GAUTHIER, H. L. (1973) – *Geography of Transportation*, Prentice-Hall, Englewood Cliffs, p. 249-283.

11. Geoestatísticas

- ABREU, Diogo de (1978) – *Estudo Estatístico de Algumas Características do Povoamento em Portugal*. Faculdade de Letras de Lisboa, Lisboa, 121 p.
- BRONSTEIN, I.; SEMENDIAEV, K. (1979) – *Manual de Matemática para Engenheiros e Estudantes*, 2ª edição, Editorial Mir, Moscovo, p.271.
- DAVIS, John C. (1986) – *Statistics and Data Analysis in Geology*, 2nd edition, New York, John Wiley & Sons, p. 295-311.
- (*) EBDON, David (1996) – *Statistics in Geography*. 2th edition, Blackwell, Oxford (UK), p. 128-169.
- (*) ESTEBANEZ ALVAREZ, J.; BRADSHAW, R.. P. (1978) – *Técnicas de Quantificación en Geografía*, Editorial Tebar Flores, Madrid, p. 391-465.
- THOMAS, R. W. (1981) "Point Pattern Analysis", in WRIGLEY, N.; BENNETT, R.J. (Eds) – *Quantitative Geography*, Routledge & Kegan Paul, London, p. 164-176.

12. Modelos Integrados em Geografia

- ABRAMI, Valerio Terra (1998) – *Le Previsioni Demografiche*, Il Mulino, Bologna, 186 p.
- ABREU, D. et alia (1986) - "Construção de um Modelo para Avaliação do Impacte Regional da Integração de Portugal na CEE", *IV Colóquio Ibérico de Geografia - Actas*. Coimbra, Instituto de Estudos Geográficos, p. 59-76

- ABREU, D.; JENSEN-BUTLER, Chris (1987) - *The Multiplicative Effects of Regional Investment in Portugal*, Aarhus, Geografisk Institut, 25 p. (Notat 66).
- (*) ABREU, D. (1988) - *Desenvolvimento Regional no Oeste. Problemas e Métodos*. Universidade de Lisboa, Faculdade de Letras, Lisboa, 526 p. Dissertação de Doutoramento em Geografia Humana.
- (*) ABREU, D. et alia (1988) - *Rex: A Regional Economic-Demographic Model for Portugal*, Aarhus, Geografisk Institut, 20 p. (Notat 70).
- ABREU, D. ; COSTA, N. (2002) - "Portugal: Perspectivas sobre um país em mudança. Reflexos territoriais de diferentes cenários de desenvolvimento" in C. Cavaco (Coord.) - *"Repensar Portugal na Europa. Perspectivas de um país periférico"*, Lisboa, C.E.G., p. 332-344 (EPRU 55).
- (*) ABREU, D.; GASPAR, J.; COSTA, N. (2005) - *ESPON Project 1.1.4 Annex B: Ageing, Labour shortage, and 'Replacement Migration*, C.E.G., Lisboa, CEG, 168 p.
- (*) ARMSTRONG, H.; TAYLOR, J. (1993) - *Regional economics & policy*, 2nd edition, Harvester Wheathsheaf, New York, 397 p.
- (*) ARMSTRONG, H.; TAYLOR, J. (2000) - *Regional economics & policy*, 3th edition, Blackwell Publishers, Malden (USA), 437 p.
- BANDEIRA, M. L. (2004) - *Demografia. Objecto, teorias e métodos*, Escolar Editora, Lisboa, 340 p.
- BONNEUIL, Noel (1997) - *Introduction à la modélisation démographique*, Armand Colin, Paris, 128 p.
- BOURGEOIS-PICHAT, Jean (1970) - *Main Trends in Demography*, George Allen & Unwin, London, 79 p.
- CASTRO, E. et alia (2002) - *Avaliação do impacto socio-económico de instrumentos legais para o controlo das emissões de gases com efeito de estufa, Relatório Final* Departamento de Ambiente e Ordenamento da Universidade de Aveiro, Aveiro.
- CURWIN, Jon; SLATER, Roger. (1996) - *Quantitative Methods for Business Decisions*, 4th edition, International Thoms Business Press, London, 667 p.
- DEELSA/ELSA (2002) - *Can Migration Help to Alleviate Labour Shortages?*. Organization for Economic Co-operation and Development.
- ESPON 2006 Programme Action 1.1.3 (2003) - *"2nd Interim Report", Particular effects of enlargement of the EU and beyond on the polycentric spatial tissue with special attention on discontinuities and barriers*, Espon, Luxembourg.
- ESPON 2006 Programme Action 1.1.4 (2003) - *"2nd Interim Report", The Spatial Effects of Demographic Trends and Migration*, Espon, Luxembourg.
- ESPON 2006 Programme Action 1.1.2 (2003) - *"3rd Interim Report", Urban-Rural Relations in Europe*, Espon, Luxembourg.
- ESPON 2006 Programme Action 3.1 (2003) - *"3rd Interim Report", Integrated Tools for European Spatial Development*. Espon, Luxembourg.
- European Commission (1998) - *Analysis and Forecasting of International Migrations by Major Groups (Part II)*. EC, Luxembourg.

- European Commission (1999) - *Sixth Periodic Report on the social and economic situation and development of the regions of the European Union*. EC, Luxembourg.
- European Commission (2002) - *National reports on the demographic situation in 12 Central European Countries, Cyprus and Malta in 2000*. EC, Luxembourg.
- European Commission (2004) - *A New Partnership for Cohesion - convergence competitiveness cooperation. Third Report on Economic and Social Cohesion*. EC, Luxembourg.
- European Commission (S/d) - *The European Labour Market in the light of demographic change*. Employment & Social Affairs.
- FERRÃO, J. et alia (1986) - *Construction of Regional and Interregional Input-Output Tables for Portuguese Regions. A technical report*, Lisbon, 59 p.
- (*) GASPAR, J. et alia (1989) - *PORTUGAL. Os Próximos 20 Anos. Vol. VI - Ocupação e Organização do Espaço. Uma Prospectiva*. Fundação Calouste Gulbenkian, Lisboa, 344 p.
- GASPAR, J. et alia (1997) - *"Plano Nacional da Água - População, Economia e Território: Cenários de desenvolvimento"*, relatório-parecer de natureza reservada, entregue no INAG - Instituto Nacional da Água, 152 p.
- (*) GASPAR, J. (Coord.) (2002) - *"Estudo de Impacte Sócio-Económico do Porto de Lisboa"*, APL - Administração do Porto de Lisboa, S.A., Lisboa, 158 p. e anexos.
- GASPAR, J. et alia (2003) - *"O Impacte Socio-Económico do Porto de Lisboa"*, in Sousa, João Figueira (coord.) - *"Portos, Transportes Marítimos e Território"*, Instituto da Dinâmica do Espaço, FCSH, Lisboa, ISBN 972-99124-0-8.
- GASPAR, J.; ABREU, D.; COSTA, N. et al. (2005 ?) - *Ageing. Labour Shortage and Replacement Migration*, Gaspar, Jorge.; Abreu, Diogo, Costa, Nuno et al. ESPON, Luxemburgo. ISBN: 91-975257-0-7b (no prelo).
- (*) ISARD, Walter et alia (1998) - *Methods of Interregional and Regional Analysis*, Ashgate, Aldershot, 490 p.
- (*) JOHANSSON Mats; RAUHUT, Daniel (eds.) (2005 ?) - *ESPON project 1.1.4 The Spatial Effects of Demographic Trends and Migration*, ESPON, Luxemburgo. ISBN: 91-974582-9-5 (no prelo).
- LAW, Averril M., KELTON, W. David (1991) - *Simulation Modelling & Analysis*, 2nd edition, McGraw-Hill International editions, New York, 759 p.
- MADSEN, Bjarne et alia (1996) - *Modelling the Economy and the Environment*, Springer Verlag, Berlin, 365 p.
- MARTINS, J. et alia (2002) - *"Avaliação do impacto económico de instrumentos legais de controlo das emissões de gases com efeito de estufa: análise Input - Output"*, Porto, Actas do II Encontro Nacional do Colégio de Engenharia do Ambiente - 29 e 30 de Novembro de 2002.
- MARTINS, J. et alia (2003) - *"Construção de uma Matriz Input-Output Rectangular e sua Aplicação à Análise de Questões Ambientais"*, Recife, V Encontro de Economistas de Língua Portuguesa, 5 - 7 de Novembro de 2003.

- MARTINS, J. *et alia* (2003) - "*Avaliação do Impacto Económico de Instrumentos Legais de Controle das Emissões de Gases com Efeito de Estufa: Aplicação da Análise Input-Output*", Lisboa, ISEG, Workshop sobre Novos Desenvolvimentos e Aplicações da Análise Input-Output, 9 de Maio de 2003.
- MARTINS, J. *et alia* (2005) - *OIKOMATRIX II - Avaliação do impacto socio-económico de instrumentos legais para o controlo das emissões de gases com efeito de estufa*, Departamento de Ambiente e Ordenamento da Universidade de Aveiro, Aveiro, 115 p.
- MUSU, Ignazio; SINISCALCO, Domenico (1988) - *Ambiente e contabilità nazionale*, Il Mulino, Bologna, 485 p.
- NAZARETH, J. Manuel (1996) - *Introdução à Demografia Teoria e Prática*, Editorial Presença, Lisboa, 194 p.
- NEWTON, P. W.; TAYLOR, M.A.P.; SHARPE, R. (Eds) (1988) - *Desktop Planning. Microcomputer applications for infrastructure and services planning and management*, Hargreen Publishing Co., Melbourne, p. 108-394.
- (*) REIF, Benjamin (1978) - *Modelos en la Planificacion de Ciudades y Regiones*, Instituto de Estudios de Administracion Local, Madrid, 422 p.
- SMITH, David M.(1977) - *Patterns in Human Geography*, Penguin Books, Harmondsworth, p. 2786-303.
- UN (1998) - *Rapport succinct sur le suivi des questions de population à l'échelle mondiale (1999): accroissement, structure et repartition de la population*. UN, New York.
- (*) UN (2000) - *Replacement Migration: Is it A Solution to Declining and Ageing Populations?* United Nations, Department of Economic and Social Affairs, United Nations Secretariat, New York.
- (*) UN (2001) - *World Population Ageing 1950-2050*. United Nations, Department of Economic and Social Affairs, United Nations Secretariat.
- UN (2001) - *World Population Prospects. The 2000 Revision*. United Nations Population Division.
- UN (2003) - *World Population Prospects. The 2002 Revision*. United Nations Population Division.
- (*) VINUESA, Julio (Ed.) (1997) - *World Population Ageing 1950-2050*. United Nations, Department of Economic and Social Affairs, United Nations Secretariat.

Os alunos são estimulados a utilizar a InterNet, onde existem mais de meio milhão de *sítes* relacionados com as técnicas avançadas de análise de dados, e em que abundam os manuais explicativos e os programas *freeware*, de tratamento da informação, que se podem carregar livremente.

Mas como a avaliação da qualidade destes produtos e das dificuldades da sua aplicação não deve ser feita em absoluto, mas sim em relação com as capacidades e competências próprias dos alunos, e com as especificidades dos problemas que pretendem resolver, pareceu preferível não indicar *sítes*, mas antes ajudar os alunos na busca dos elementos que achem necessários, de acordo com os seus interesses, com a qualidade dos produtos apresentados e com o nível de conhecimentos que tenham.

I.8 PLANIFICAÇÃO DAS AULAS

Para um total de 26 aulas seguidas por 4 sessões de avaliação.

Capítulos do Programa	Sessão n.º	Sumário da sessão
Apresentação 1 - Conceitos básicos e de enquadramento	1	Apresentação dos objectivos, do programa, da bibliografia básica e da avaliação. Preenchimento das fichas dos alunos. A investigação científica em geral e em Geografia. A escolha da abordagem metodológica: extensiva vs. intensiva. As etapas do trabalho de investigação: preparação, definição do problema, construção do esquema de análise, do tipo de observação e das análises a efectuar. A informação geográfica: modelos de dados; tipos de matrizes de dados.
2 – Matemática essencial para a disciplina.	2	Revisão dos conceitos que vão ser mais utilizados: potências e funções. Logaritmos e matrizes. Operações com matrizes. Medidas de concentração: curvas de Lorenz e índices de concentração.
	3 (+)	Exercícios de construção e interpretação de índices, quocientes, taxas. Construção de indicadores simples e compostos. Cálculo de concentrações com a curva de Lorenz e os respectivos índices.
3 – Recolha da informação	4	Discussão dos aspectos fundamentais da recolha de informação. Vantagens e inconvenientes absolutos e relativos das diversas formas de recolha. As amostragens: erro e risco aceitáveis, dimensão das amostras e selecção dos seus elementos.
4 – Tabelas de Contingência	5	A aplicação dos métodos de análise de tabelas a dados qualitativos e quantitativos. Transformações no tipo dos dados disponíveis: quadros lógicos e tabelas de Burt. Apresentação de medidas de dependência, associação e predição para tabelas de contingência, e discussão sobre as suas condições de utilização. Referência à possibilidade de usar AFCM (Análise factorial de correspondências múltiplas). Análise de tabelas de Burt.
5 – Análise de Variância	6	Objectivos e condições para a aplicação das análises de variância e covariância. Análise de um só factor e multifactorial. Análise de covariância.
<i>Exercícios</i>	7 (+)	Apresentação de exemplos de cálculo e execução de exercícios simulados sobre dimensionamento de amostras, tabelas de contingência, análise de variância Lançamento do trabalho prático de grupo.

6 – Regressão múltipla.	8	Revisão da regressão linear simples. Objectivos e condições para a aplicação das técnicas de regressão linear múltipla. O modelo de regressão múltipla: calibrar e interpretar os coeficientes de correlação e determinação múltiplas, os coeficientes de regressão (bês e betas); saber em que circunstâncias é conveniente determinar e interpretar os valores do teste F e do erro-padrão da estimativa, e os testes t e os erros-padrão dos coeficientes. Interpretar os resíduos. Os diversos tipos de regressão múltipla: simultânea e passo a passo. Utilização de variáveis mudas.
	9	
	10 (+)	Apresentação de exemplos de cálculo e execução de exercícios simulados sobre os diferentes aspectos que podem ter interesse nas análises de regressão múltipla.
7 – Análise factorial	11	Objectivos e condições para a aplicação das técnicas de análise factorial em geral. Distinção entre ACP (análise em componentes principais) e AF (análise factorial propriamente dita). Tipos de análises factoriais: métodos de extracção dos factores, rotações dos eixos e cálculo dos pesos dos factores nas unidades de análise; A identificação dos factores. Executar e interpretar uma sequência de resultados de análise factorial
	12	
	13 (+)	Apresentação de exemplos de cálculo e execução de exercícios simulados sobre os diferentes aspectos que podem ter interesse nas análises factoriais.
<i>Exercícios</i>	14	Acompanhamento dos trabalhos práticos de grupo.
8 – Classificação numérica	15	Objectivos e condições para a aplicação das técnicas de classificação multidimensional. Escolha da medida de distância, de proximidade ou de associação adequada à situação. As estratégias de agrupamento das unidades de análise. Avaliação dos resultados. Interpretação da classificação resultante.
	16	
	17 (+)	Apresentação de exemplos de cálculo e execução de exercícios simulados sobre os diferentes aspectos que podem aparecer quando se utilizam métodos de classificação numérica multidimensional.
9 – Análise discriminante	18	Objectivos e condições para a aplicação da análise discriminante. Determinação e significado das funções discriminantes. Determinação das coordenadas dos elementos e dos centróides dos grupos. Avaliação da probabilidade de pertencer aos grupos. Utilização da análise discriminante como técnica classificatória.

9 – Análise discriminante (continuação)	19 (+)	Apresentação de exemplos de cálculo e execução de exercícios simulados sobre os diferentes aspectos que podem aparecer quando se utilizam métodos de análise discriminante.
<i>Exercícios</i>	20	Acompanhamento dos trabalhos práticos de grupo.
10 – Redes e grafos	21	A topologia: redes, grafos e matrizes. Conceito e interpretação da ligação ou conectividade. Cálculo de medidas de conectividade. Conceito e interpretação da centralidade e da acessibilidade. Cálculo de medidas de acessibilidade e de centralidade. Significado de algumas matrizes derivadas da matriz de conectividade e de matrizes ponderadas.
	22 (+)	Apresentação de exemplos de cálculo e execução de exercícios sobre análise de redes e de grafos.
11 – Geoestatísticas	23	A importância da introdução da localização na análise de dados. Medidas espaciais de centralidade e de dispersão O conceito de potencial e o seu cálculo. Análise de proximidade e de vizinhança.
	24	Apresentação de exemplos de cálculo e execução de exercícios de Geoestatísticas.
12 – Modelos integrados em Geografia	25	O conceito de modelização, de previsão e de simulação e as suas relações com as análises de dados. Apresentação de alguns modelos simples de análise geográfica, ilustrando o modo como se fazem, as relações funcionais ou matriciais em que se baseia, os problemas e da sua adequação ao mundo real, Apresentação de alguns exemplos simples de modelos demográficos (<i>cohort-survival</i>) e económicos (<i>input/output</i>).
	26	Apresentação de modelos integrados demográfico-regionais: o modelo REX e o modelo Espon114-WP5.
<i>Avaliação de conhecimentos</i>	A1	Apresentação e discussão dos trabalhos práticos de grupo efectuados pelos alunos.
	A2	Apresentação e discussão dos trabalhos práticos de grupo efectuados pelos alunos.
	A3	Teste para avaliação de conhecimentos – 1ª chamada.
	A4	Teste para avaliação de conhecimentos – 2ª chamada.

(+) – As sessões assinaladas deverão decorrer na sala de informática, para que os alunos possam fazer todos os exercícios e seguir os exemplos propostos utilizando computadores.

Segunda Parte

DESENVOLVIMENTO DO

PROGRAMA

1. Conceitos Básicos e de Enquadramento

1. OBJECTIVOS A ATINGIR

1. Saber organizar um trabalho de investigação em Geografia.
2. Saber seleccionar o tema e delimitar o âmbito da investigação.
3. Ter a capacidade de construir o quadro teórico e explicitar o problema a tratar, de definir os conceitos e de estabelecer os modelos.
4. Conhecer os tipos de abordagem metodológicos, e saber em que circunstâncias se deve optar por cada uma delas.
5. Identificar os diversos tipos de informação geográfica, e conhecer os tratamentos analíticos que lhes são mais apropriados.

2. CONTEÚDOS A DESENVOLVER

A investigação científica em geral e em Geografia.

A escolha da abordagem metodológica: extensiva vs. intensiva.

As etapas do trabalho de investigação: preparação, definição do problema, construção do esquema de análise, do tipo de observação e das

análises a efectuar.

A informação geográfica: modelos de dados; tipos de matrizes de dados e tipologias de tratamentos analítico.

O desenvolvimento detalhado dos temas deste capítulo é feito no Anexo 1, que se apresenta mais à frente.

3. MEIOS E ESTRATÉGIAS A UTILIZAR

Apresentação das questões, pelo docente, em aula teórica.

Discussão alargada com os alunos, baseada em exemplos concretos de trabalhos de investigação geográfica.

4. BIBLIOGRAFIA

- BRYMAN, Alan; CRAMER, Duncan (1993) – *Análise de Dados em Ciências Sociais. Introdução às técnicas utilizando o SPSS*, Celta Editores, Oeiras, p. 1-20; 233-264.
- FOTHERINGHAM, A. Stewart (2000) – *Quantitative Geography. Perspectives on Spatial Data Analysis*, Sage Publications, London, p. 1-14.
- HARVEY, David (1973) – *Explanation in Geography*, Edward Arnold, London, 521 p.
- HOLLOWAY, Sarah L., et alia (2003) – *Key Concepts in Geography*, Sage Publications, London, 342 p.

2. ALGUMA MATEMÁTICA PARA GEÓGRAFOS

1. OBJECTIVOS A ATINGIR

1. Saber os fundamentos da matemática necessários para esta disciplina.
2. Conhecer os conceitos e saber efectuar operações com potências, logaritmos, funções e matrizes.
3. Saber calcular índices, quocientes e taxas.
4. Ser capaz de interpretar os diversos valores que são calculados para os índices, quocientes e taxas.
5. Conhecer as principais medidas de concentração de uma variável, as curvas de Lorenz e os Índices de Gini.
6. Saber calcular e interpretar as medidas de concentração de uma variável.

2. CONTEÚDOS A DESENVOLVER

Revisões de matemática básica: Potências, logaritmos, funções e

matrizes. Índices, Quocientes e Percentagens: Taxas seccionais e de variação. Índices simples e compostos.

Medidas de concentração. Curva de Lorenz. Índices de concentração: coeficiente de especialização, índices de Gini e de Schutz, ponto de igual partilha e coeficiente de Gini.

O desenvolvimento dos conteúdos deste capítulo é apresentado com maior detalhe mais adiante, no Anexo 2.

3. MEIOS E ESTRATÉGIAS A UTILIZAR

Apresentação, em aula de tipo expositivo, dos conceitos matemáticos e das operações com potências, logaritmos, funções e matrizes.

Exercícios simples para fazer manualmente em ambiente aula. Os mesmos exercícios e outros mais extensos e complicados para ser resolvidos por computador.

4. BIBLIOGRAFIA

- CHIANG, Alpha (1982) – *“Matemática para Economistas”*, MacGraw-Hill do Brasil, São Paulo, 684 p.
- DAVIS, John C. (1986) – *Statistics and Data Analysis in Geology*, 2nd edition, New York, John Wiley & Sons, p. 107-140.
- EVANS, I.S.; JONES, k. (1981) *“Ratios and closed number Systems”*, in WRIGLEY, N.; BENNETT, R.J. (Eds) – *Quantitative Geography*, Routledge & Kegan Paul, London, p. 11-122.
- MATHER, Paul (1976) – *Computational Methods of Multivariate Analysis in Physical Geography*, John Wiley & Sons, London, p. 9-36.
- MORRISON, Donald F. (1976) – *Multivariate Statistical Methods*, McGraw-Hill Kogakusha, Ltd., Tóquio, p. 1-78.
- SMITH, David M. (1977) – *“Patterns in Human Geography”*, Penguin Books, Harmondsworth, 373 p.
- THIESSEN, Heiner (1996) – *Measuring the real world. A textbook of Applied Statistical Methods*, John Wiley & Sons, Chicester, p.10-186
- TURNER, J. C. (1970) – *Matemática moderna aplicada*, Alianza Editorial, Madrid, p. 17-106.
- WILSON, A. G.; KIRKBY. M.J. (1975) – *Mahematics for Geographers and Planners*, Claredon Press, Oxford, 325 p.

3. Recolha da Informação

1. OBJECTIVOS A ATINGIR

1. Conhecer os diversos métodos de recolha de informação.
2. Identificar os métodos de recolha de informação mais adequados aos diferentes tipos de investigação geográfica.
3. Saber fazer a recolha da informação e ter a capacidade de avaliar criticamente o método utilizado.
4. Saber transformar as variáveis da forma mais adequada à utilização das diversas técnicas de análise de dados.
5. Saber dimensionar uma amostra e conhecer os limites de erro correspondentes.
6. Conhecer os diversos métodos de seleccionar os elementos de uma amostra, e saber qual aplicar em cada caso concreto

2. CONTEÚDOS A DESENVOLVER

Aspectos fundamentais da recolha de informação. Diferentes métodos e técnicas de recolha e cuidados a ter. Os problemas mais comuns e as maneiras de os evitar.

As fontes documentais, a observação (exterior e participante), a experimentação, a análise documental, os inquéritos (entrevistas e questionários), os grupos de discussão, os métodos Delphi e as histórias de vida.

A avaliação das diversas técnicas de recolha e a sua adequação ao problema a tratar e à abordagem teórica escolhida.

A transformação de variáveis e a construção de indicadores significativos. Os tipos de tratamento analítico possíveis e a maneira de escolher os mais adequados.

A amostragem: o erro e o risco, a determinação das dimensões das amostras e os métodos de escolha dos elementos que as constituem.

O desenvolvimento detalhado dos conteúdos deste capítulo é apresentado no Anexo 3.

3. MEIOS E ESTRATÉGIAS A UTILIZAR

Apresentação da base teórica, de casos de estudo e de exemplos.

Determinação da dimensões de amostras para situações simuladas. Discussão sobre os critérios a utilizar na escolha dos métodos de selecção dos elementos de amostras.

O docente deverá apresentar casos exemplares de sucesso e de insucesso de amostras, evidenciando as respectivas razões.

4. BIBLIOGRAFIA

ABREU, Diogo de (1998) – *Recolha e Tratamento de Informação. Manual de Apoio ao 3º Curso de Técnicos Superiores de Planeamento na Administração Regional*. IGE/Partex, Luanda, p. 6-37.

ALMEIDA, J. F. de, PINTO, J. M. (1976) – *A Investigação nas Ciências Sociais*, Lisboa, Editorial Presença, p. 105-113.

BRYMAN, Alan; CRAMER, Duncan (1993) – *Análise de Dados em Ciências Sociais. Introdução às técnicas utilizando o SPSS*, Celta Editores, Oeiras, p. 121-140.

CALOT, Gérard (1970) – *Curso de Estadística Descriptiva*, Paraninfo, Madrid, p. 94-99. É tradução castelhana do *Cours de Statistique Descriptive*, Dunod, Paris, 1965, p. 479-507.

- GARCIA BALLESTEROS, Aurora (Coord.) (1998) - – *Métodos y Técnicas Cualitativas en Geografía Social*, Oikos-Tau, , Barcelona , 239 p.
- CAPLOW, Theodore (1972) – *La Investigación Sociológica*, Editorial Laia, Barcelona , p. 260-266
- CLIFFORD, Nicholas J.; VALENTINE, Gill (2003) – *Key Methods in Geography*, Sage Publications, London, 572 p.
- DUVERGER, Maurice (1964) – *Méthodes des Sciences Sociales*, 1ª ed., P.U.F., Paris, p. 251-266.
- (*) ESTEBANEZ ALVAREZ, J.; BRADSHAW, R.. P. (1978) – *Técnicas de Quantificación en Geografía*, Editorial Tebar Flores, Madrid, p. 277-336.
- LONGHURST, Robyn. (2003) – “*Semi-structured Interviews and Focus Groups*”, in Clifford & Valentine – *Key Methods in Geography*, Sage Publications, London, p. 117-132.
- MASON, Jennifer (2000) – *Qualitative Research*, Sage Publication, London, 223 p.
- MOSER, C. A.; KALTON, G. (1975) – *Survey Methods in Social Investigation*, Heinemann Educational Books, London, 549 p.
- REIS, Elizabeth *et alia* (2001) – *Estatística Aplicada*, Vol. 2, Edições Sílabo, Lisboa, 322 p.
- (*) VICENTE, Paula *et alia* (1996) – *Sondagens*, Edições Sílabo, Lisboa, 238 p.
- (*) VILARES, M. J. (1993) – *Sistemas de Informação Estatística*, ISEGI, Lisboa, p. 277-336.
- WRIGLEY, N.; BENNETT, R.J. (Eds) (1981) – *Quantitative Geography*, Routledge & Kegan Paul, London, p. 17-69.
- YIN, Robert K. (1964) – *Case Study Research. Design and Methods*. 2nd edition. SAGE Publication, London, 171 p.

4. Tabelas de Contingência

1. OBJECTIVOS A ATINGIR

1. Reconhecer quais os problemas geográficos que podem ser resolvido com o auxílio de tabelas de contingência.
2. Ter a capacidade de construir tabelas de contingência com qualquer tipo de variáveis.
3. Saber fazer quadros ponderados, quadros lógicos e tabelas de Burt.
4. Executar, nos quadros de contingência, testes sobre a independência entre linhas e entre colunas, calcular medidas de associação e de correlação, de predição e de incerteza.
5. Reconhecer quais as medidas que se devem aplicar em cada circunstância.
6. Saber interpretar o significado dos valores calculados.
7. Saber analisar tabelas de Burt usando a técnica do qui-quadrado e interpretar os resultados.
8. Avaliar criticamente os métodos utilizados.

2. CONTEÚDOS A DESENVOLVER

A análise de tabelas de contingência enquanto forma de tratamento muito adequado para os dados expressos em escalas de tipo nominal e ordinal. A transformação da informação em tabelas de contingência: transformação por ponderação, quadros lógicos e tabelas de Burt.

Análise de tabelas a duas variáveis. Medidas de associação, predição e de incerteza. Cálculo e interpretação. Análise de tabelas a mais de duas variáveis (tabelas de Burt). Medida de qui-quadrado. Cálculo e interpretação.

O desenvolvimento detalhado dos conteúdos deste capítulo apresenta-se mais adiante, no Anexo 4.

3. MEIOS E ESTRATÉGIAS A UTILIZAR

Apresentação em aula do modo de construir e de analisar tabelas de contingência. Exposição e exemplificação do modo de calcular as diversas medidas de associação, predição e incerteza e interpretação do seu significado.

Exercícios de aplicação do cálculo e da respectiva interpretação.

4. BIBLIOGRAFIA

- BRYMAN, Alan; CRAMER, Duncan (1993) – *Análise de Dados em Ciências Sociais. Introdução às técnicas utilizando o SPSS*, Celta Editores, Oeiras, p. 265-278.
- D'HAINAUT, Louis (1992) – *Conceitos e Métodos da Estatística. Volume 2*, Fundação Calouste Gulbenkian, Lisboa, p. 70-119.
- MELO, F. Galvão de (1982) – *Modelos e Métodos Estatísticos em Psicologia*, Escola Nacional de Saúde Pública, Lisboa, p. 38-53.
- PELOSI, Marylin K.; SANDIFER, Theresa M. (2002) – *Doing Statistics for Business with Excell*, 2nd edition, p. 760-811.
- (*) PESTANA, Maria Helena; GAGEIRO, João Nunes (1998) – *Análise de Dados para Ciências Sociais. A complementaridade do SPSS*, Lisboa, Edições Sílabo, p. 91- 146.
- WRIGLEY, N. (1981) "Categorical Data Analysis", in WRIGLEY, N.; BENNETT, R.J. (Eds) – *Quantitative Geography*, Routledge & Kegan Paul, London, p. 111-122.

5. Análise de Variância

1. OBJECTIVOS A ATINGIR

1. Identificar os problemas geográficos que podem ser resolvidos através das técnicas de análise de variância e de covariância.
2. Saber quais os tipos de dados que melhor se adaptam à análise com recurso a técnicas de variância e covariância.
3. Saber aplicar a casos concretos de investigação geográfica análises de variância a um a vários factores e análises de covariância.
4. Avaliar o significado das relações encontradas.
5. Saber interpretar os resultados.
6. Avaliar criticamente as análises efectuadas.

2. CONTEÚDOS A DESENVOLVER

A pesquisa de relações entre variáveis qualitativas independentes e variáveis quantitativas. Relações entre duas e mais variáveis de vários tipos.

Análises de variância com uma e mais variáveis independentes e análise de covariância. Variância dos grupos e entre os grupos. Capacidade de explicação do modelo. O teste F e o grau de significação das relações.

O desenvolvimento detalhado dos conteúdos é apresentado no Anexo 5, adiante.

3. MEIOS E ESTRATÉGIAS A UTILIZAR

Exposição pelo docente dos fundamentos teóricos das análises de variância e covariância e dos procedimentos informáticos necessários para os efectuar.

Execução de exemplos de aplicação em Geografia.

Avaliação crítica e comparativa das vantagens em usar as análises de variância e covariância.

4. BIBLIOGRAFIA

- BRYMAN, Alan; CRAMER, Duncan (1993) – *Análise de Dados em Ciências Sociais. Introdução às técnicas utilizando o SPSS*, Celta Editores, Oeiras, p. 233-264.
- D'HAINAUT, Louis (1992) – *Conceitos e Métodos da Estatística. Volume 2*, Fundação Calouste Gulbenkian, Lisboa, p. 150-282.
- DAVIS, John C. (1986) – *Statistics and Data Analysis in Geology*, 2nd edition, New York, John Wiley & Sons, p. 74-79.
- (*) ESTEBANEZ ALVAREZ, J.; BRADSHAW, R. P. (1978) – *Técnicas de Quantificación en Geografía*, Editorial Tebar Flores, Madrid, p. 381-386.
- IVERSEN, G. R.; NORPOTH, H. (1976) – *Analysis of Variance*, Sage Publications, London, 95 p..
- GMURMAN, V. E. (1984) – *Problemas em Probabilidades e Estatística*, Editora Mir, Moscovo, p. 236-245.
- (*) KIM, Jae-On; KOHOUT, F. J. (1975) - "Analysis of Variance and Covariance" in NIE, Norman H. et alia - *SPSS -Statistical Package for Social Sciences*, 2th edition, McGraw-Hill Book Company, New York, p. 398-432.
- MORRISON, Donald F. (1976) – *Multivariate Statistical Methods*, McGraw-Hill Kogakusha, Ltd., Tóquio, p. 170-229.
- PELOSI, Marilyn K.; SANDIFER, Theresa M. (2002) – *Doing Statistics for Business with Excell*, 2nd edition, p. 695-759.
- (*) PESTANA, M. H.; GAGEIRO, J. N. (1998) - *Análise de Dados para Ciências Sociais. A Complementaridade do SPSS*, Edições Sílabo, Lisboa, p. 177-243.

- STEYN, A.G.W. *et alia* (1996) – *Modern Statistics in Practice*, J. L. van Schaik, Pretoria ,p. 508-546.
- YEATES, Maurice (1974) – *An Introduction to Quantitative Analysis in Human Geography*, MacGraw-Hil, New York, p. 131-139.

6. Regressão Linear Múltipla

1. OBJECTIVOS A ATINGIR

1. Conhecer quais os problemas geográficos que podem ser resolvidos através das técnicas de regressão múltipla.
2. Saber escolher entre regressão simples e regressão múltipla.
3. Reconhecer os tipos de dados que podem ser analisados com técnicas de regressão múltipla.
4. Conhecer os limites de uma regressão múltipla feita a partir de uma amostra.
5. Saber distinguir entre uma regressão múltipla descritiva e uma regressão múltipla feita para previsão.
6. Saber interpretar os resultados de uma regressão múltipla descritiva.
7. Saber interpretar os resultados e conhecer os limites de uma regressão múltipla para previsão.
8. Saber calibrar um modelo de regressão múltipla.
9. Ser capaz de fazer regressões múltiplas “passo a passo”, escolhendo as necessárias opções nos programas comerciais mais comuns.

10. Saber escolher o número de variáveis independentes a incorporar na relação de regressão múltipla, face à sua respectiva capacidade explicativa.
11. Saber calcular e interpretar os resíduos da regressão.
12. Ter uma noção de auto-correlação temporal e espacial.

2. CONTEÚDOS A DESENVOLVER

A regressão linear simples enquanto medida da proporcionalidade das variações de fenómenos geográficos. O coeficiente de correlação linear de Pearson (Bravais) e o ajustamento à associação entre variáveis.

A regressão linear múltipla, como formalização de modelos causais. Significado dos coeficientes de regressão (bês e betas). A medida do ajustamento: o coeficiente de correlação múltipla, simples e ajustado e o coeficiente de determinação.

O significado da regressão. O teste F e o erro-padrão das estimativas; os testes t e os limites de confiança dos coeficientes de regressão.

Os diversos tipos de regressão: simultânea e passo-a-passo, para a frente e para trás (*forward* e *backward*).

As extensões do modelo. As variáveis mudas: utilização e significado.

Interpretação e avaliação dos resultados.

O desenvolvimento detalhado dos conteúdos está no Anexo 6, adiante.

3. MEIOS E ESTRATÉGIAS A UTILIZAR

Exposição pelo docente dos fundamentos teóricos dos métodos, dos procedimentos informáticos a desenvolver e das possibilidades de interpretação dos diversos passos.

Apresentação de exemplos de aplicação a problemas e processos diversificados de natureza geográfica.

Execução pelos alunos de exemplos detalhados e de exercícios em regime de auto-avaliação.

Discussão sobre o benefício resultante da utilização das técnicas de regressão múltipla.

4. BIBLIOGRAFIA

- BRYMAN, Alan; CRAMER, Duncan (1993) – *Análise de Dados em Ciências Sociais. Introdução às técnicas utilizando o SPSS*, Celta Editores, Oeiras, p. 285-298.
- (*) CURWIN, Jon; SLATER, Roger (1996) – *Quantitative Methods for Business Decisions*, London, International Thomson Business Press, pp. 378 420, especialmente 397 a 420.
- D’HAINAUT, Louis (1992) – *Conceitos e Métodos da Estatística. Volume 2*, Fundação Calouste Gulbenkian, Lisboa, p. 120-149.
- DAVIS, John C. (1986) – *Statistics and Data Analysis in Geology*, 2nd edition, New York, John Wiley & Sons, p. 468-478.
- FERGUSON, Rob (1979) – *Linear Regression in Geography*, London, Institute of British Geographers, 44 p. (CATMOG, 15)
- LEWIS-BECK, Michael (1980) – *Applied Regression. An Introduction*, London, Sage Publications, 77 p. (SAGE, 22)
- (*) KIM, Jae-On; KOHOUT, F. J. (1975) - “Multiple Regression Analysis” in NIE, Norman H. et alia - *SPSS -Statistical Package for Social Sciences*, 2th edition, McGraw-Hill Book Company, New York, p. 320-367.
- MATHER, Paul (1976) – *Computational Methods of Multivariate Analysis in Physical Geography*, John Wiley & Sons, London, p. 39-198.
- PARIKH, Ashok; BAILEY, David (1990) – *Techniques of Economic Analysis with applications*, Harvester Wheatsheaf, New York, 357 p.
- PELOSI, Marylin K.; SANDIFER, Theresa M. (2002) – *Doing Statistics for Business with Excell*, 2nd edition, p. 512-630.
- (*) PESTANA, Maria Helena; GAGEIRO, João Nunes (1998) – *Análise de Dados para Ciências Sociais. A complementaridade do SPSS*, Lisboa, Edições Sílabo, p. 361-428.
- SILK, J. (1981) “The General Linear Model”, in WRIGLEY, N.; BENNETT, R.J. (Eds) – *Quantitative Geography*, Routledge & Kegan Paul, London, p. 75-85.
- YEATES, Maurice (1974) – *An Introduction to Quantitative Analysis in Human Geography*, MacGraw-Hil, New York, p. 98-122.

7. Análise Factorial

1. OBJECTIVOS A ATINGIR

1. Ter a capacidade de reconhecer um problema geográfico que possa ser resolvido através das técnicas de análise factorial.
2. Saber em que condições escolher entre uma análise em componentes principais e uma análise factorial propriamente dita.
3. Conhecer as circunstâncias em que se deve aplicar uma análise factorial de correspondências múltiplas.
4. Reconhecer os tipos de dados que podem ser analisados com as técnicas de análise em componentes principais, de análise factorial e de análise factorial de correspondências múltiplas.
5. Saber escolher o número de factores a seleccionar para as análises e como fazer essa selecção.
6. Conhecer o significado dos valores próprios.
7. Saber justificar a escolha do método de extracção de factores.
8. Justificar a opção da rotação de factores pelo método Varimax.
9. Ter a capacidade de interpretar os coeficientes de correlação entre as variáveis e factores (*loadings*).
10. Saber identificar os factores seleccionados pela análise.
11. Calcular os valores dos factores nas unidades de análise (*scores*).

12. Fazer análises factoriais usando programas comerciais comuns.
13. Saber interpretar os resultados de uma análise em componentes principais.
14. Saber interpretar os resultados de uma análise factorial.
15. Avaliar criticamente os resultados.

2. CONTEÚDOS A DESENVOLVER

As análises factoriais como método de simplificação da estrutura de variáveis: identificação de grupos de variáveis associadas, redução do seu número e eliminação da redundância.

Os diversos tipos: análise em componentes principais (ACP), análises factoriais propriamente ditas (AFs) e análise factorial de correspondências múltiplas (AFC). Distinção entre Análise em componentes principais e Análises Factoriais: no modelo matemático, na utilização e nos procedimentos analíticos.

Métodos de extracção dos factores, tipos de rotações e cálculos dos valores dos factores nas unidades de análise.

Interpretação do significado dos valores próprios e da correlação entre factores e variáveis. Interpretação e avaliação dos resultados.

O desenvolvimento detalhado dos conteúdos está no Anexo 7, adiante.

3. MEIOS E ESTRATÉGIAS A UTILIZAR

Exposição pelo docente dos fundamentos teóricos das metodologias de análise factorial, dos procedimentos informáticos e do modo como devem ser interpretados os diversos passos. Apresentação de exemplos diversificados de aplicação a temas de Geografia.

Execução pelos alunos de exemplos de demonstração variados e de exercícios resolvidos. Interpretação dos resultados de várias análises factoriais apresentados pelo docente.

Execução de análises factoriais e interpretação dos seus resultados no âmbito dos trabalhos de grupo.

Avaliação das vantagens de utilização das técnicas de análise factorial.

4. BIBLIOGRAFIA

- BENZÉCRI, J.-P. et collaborateurs (1973) – *L'Analyse des Données. 2 - L'Analyse des Correspondences*, Dunod, Paris, 619 p.
- BRYMAN, Alan; CRAMER, Duncan (1993) – *Análise de Dados em Ciências Sociais. Introdução às técnicas utilizando o SPSS*, Celta Editores, Oeiras, p. 307-323.
- (*) CICERI, M.F.; MARCHANT, B.; RIMBERT, S. (1977) – *Introduction a l'Analyse de l'Espace*, Masson, Paris, p. 13-42.
- DAVIS, John C. (1986) – *Statistics and Data Analysis in Geology*, 2nd edition, New York, John Wiley & Sons, p. 515-573.
- HARMAN, Harry (1967) – *Modern Factor Analysys*, 2nd edition, The University of Chicago Press, Chicago, 474 p.
- (*) KIM, Jae-On (1975) – “Factor Analysis” in NIE, Norman H. *et alia* - *SPSS - Statistical Package for Social Sciences*, 2th edition, McGraw-Hill Book Company, New York, p. 468-514.
- MORENO JIMÉNEZ, Antonio (1994) – “Análisis Factorial” in BOSQUE SENDRA, Joaquín; MORENO JIMÉNEZ, Antonio – *Prácticas de Análisis Exploratório y Multivariante de Datos*, Oikos-Rau, Barcelona, p. 87-131.
- MATHER, Paul (1976) – *Computational Methods of Multivariate Analysis in Physical Geography*, John Wiley & Sons, London, p. 215-306.
- MATHER, P.M. (1981) “Factor Analysis”, in WRIGLEY, N.; BENNETT, R.J. (Eds) – *Quantitative Geography*, Routledge & Kegan Paul, London, p. 144-150.
- MELO, F. Galvão de (1982) – *Modelos e Métodos Estatísticos em Psicologia*, Escola Nacional de Saúde Pública, Lisboa, p. 54-112.
- MORRISON, Donald F. (1976) – *Multivariate Statistical Methods*, McGraw-Hill Kogakusha, Ltd., Tóquio, p. 226-345.
- (*) PESTANA, Maria Helena; GAGEIRO, João Nunes (1998) – *Análise de Dados para Ciências Sociais. A complementaridade do SPSS*, Lisboa, Edições Sílabo, p. 321-344.
- (*) REIS, Elizabeth (1997) – *Estatística Multivariada Aplicada*, Lisboa, Edições Sílabo, p. 251-284.
- TORRENS-IBERN, J. (1972) – *Modèles et Méthodes de l'Analyse Factorielle*, Dunod, Paris, 202 p.
- VOLLE, Michel (1978) – *Analyse des Donnees*, Economica, Paris, p. 61-194 p.
- YEATES, Maurice (1974) – *An Introduction to Quantitative Analysis in Human Geography*, MacGraw-Hil, New York, p. 207-238.

8. Classificação Numérica

1. OBJECTIVOS A ATINGIR

1. Identificar os problemas geográficos que podem ser resolvido utilizando técnicas de taxionomia numérica.
2. Conhecer o tipo de dados necessários para aplicar as diferentes técnicas de classificação numérica.
3. Ter a capacidade de fazer transformações de dados que permitam a comparação de informação medida em qualquer tipo de escala de medida.
4. Saber quando é necessário padronizar a informação de base.
5. Conhecer quando se deve aplicar uma partição ou uma classificação hierárquica.
6. Saber optar entre classificações hierárquicas ascendentes e descendentes.
7. Conhecer as medidas de proximidade ou de distância mais importantes que se podem utilizar, e ser capaz de justificar a escolha de uma delas.

8. Conhecer as estratégias de agrupamento e ter a capacidade de justificar a razão por que se deve optar por uma delas, em cada caso concreto.
9. Ser capaz de fazer classificações numéricas usando os programas comerciais mais comuns, escolhendo os procedimentos mais adequados.
10. Saber fazer dendogramas e interpretá-los.
11. Calcular as distâncias cofenéticas e ter a capacidade de avaliar a qualidade da solução escolhida.

2. CONTEÚDOS A DESENVOLVER

Classificar elementos a partir dos valores de muitas variáveis. O significado da ponderação das variáveis.

Partições e classificações hierárquicas ascendentes e descendentes. Como escolher ?

As medidas de distância ou de proximidade entre os elementos, para dados nominais, binários dicotômicos ou alternativos, de intervalo ou de razão. Coeficientes de Jaccard, Russel, Dice e concordância simples, Sneath e Sokal, Youle, e Pearson. Distâncias de Minkowski, euclidiana, de Manhattan e de Chebychev. Coeficiente de Pearson, cosseno e qui-quadrado,

As estratégias de agrupamento para partições e para classificações hierárquicas. Nuvens dinâmicas, K-médias K-centróides. Métodos do vizinho mais próximo, mais distante, da média do grupo, da média simples, do centróide, da mediana, de Ward e das distâncias internas.

Dendogramas. A avaliação do ajustamento. Distâncias cofenéticas e respectivo coeficiente de correlação cofenético.

O desenvolvimento detalhado dos conteúdos deste capítulo é apresentado mais à frente, no Anexo 8.

3. MEIOS E ESTRATÉGIAS A UTILIZAR

Apresentação em aula dos diversos métodos de efectuar classificações, a uma e a várias variáveis, das medidas de distância e de proximidades mais utilizadas em Geografia e das estratégias de agrupamento usuais.

Explicação dos procedimentos informáticos, justificação das opções a tomar e interpretação dos resultados.

Apresentação de exemplos diversificados de aplicação a problemas de classificação em Geografia.

Apresentação pelo docente e execução pelos alunos de exemplos de demonstração variados e de exercícios resolvidos.

Interpretação dos resultados de várias taxionomias apresentados pelo docente, execução de classificações e dendogramas e interpretação dos seus resultados pelos alunos no âmbito dos seus trabalhos de grupo.

4. BIBLIOGRAFIA

- BENZÉCRI, J.-P. et collaborateurs (1973) – *L'Analyse des Donnees. 1 – La Taxinomie*, Dunod, Paris.
- (*) BROCHADO, Ana Margarida (2002) – “A análise de clusters: técnica de classificação na análise espacial” in COSTA, José da Silva (Coord.), *Compêndio de Economia Regional*, Coimbra, A.P.D.R., p. 743 - 766.
- DAVIS, John C. (1986) – *Statistics and Data Analysis in Geology*, 2nd edition, New York, John Wiley & Sons, p. 502-514.
- MATHER, Paul (1976) – *Computational Methods of Multivariate Analysis in Physical Geography*, John Wiley & Sons, London, p. 309-329, 342-351.
- MORENO JIMÉNEZ, Antonio (1994) – “Classificación Multivariante” in BOSQUE SENDRA, Joaquín; MORENO JIMÉNEZ, Antonio – *Prácticas de Análisis Exploratório y Multivariante de Datos*, Oikos-Rau, Barcelona, 133-172 p.
- (*) PESTANA, Maria Helena; GAGEIRO, João Nunes (1998) – *Análise de Dados para Ciências Sociais. A complementaridade do SPSS*, Lisboa, Edições Sílabo, p. 345 – 354.
- (*) REIS, Elizabeth (1997) – *Estatística Multivariada Aplicada*, Lisboa, Edições Sílabo, p. 285 – 336.
- VOLLE, Michel (1978) – *Analyse des Donnees*, Economica, Paris, p. 195-265 p.

9. Análise Discriminante

1. OBJECTIVOS A ATINGIR

1. Identificar os problemas geográficos que podem ser resolvidos através das técnicas de análise discriminante.
2. Reconhecer os tipos de informação que podem ser analisados com o recurso a esta técnica.
3. Conhecer as vantagens de utilização desta técnica em Geografia.
4. Ser capaz de fazer análises discriminantes “passo a passo”, escolhendo as necessárias opções nos programas comerciais mais comuns
5. Saber interpretar os resultados de uma análise discriminante.
6. Saber escolher o número de variáveis independentes a incorporar nas funções discriminantes, face à sua respectiva capacidade explicativa.
7. Saber calcular e interpretar os valores da função discriminante em cada unidade de análise.
8. Ser capaz de avaliar a utilização de análises discriminantes.

2. CONTEÚDOS A DESENVOLVER

O conceito de pertença a um grupo e a função discriminante das

variáveis. A probabilidade de pertencer ao grupo e o facto de poder estar mal classificado.

As funções discriminantes. Cálculo, significado e interpretação.

As coordenadas dos casos e os centróides dos grupos ou classes. Interpretação. Probabilidade de pertencer a uma classe.

Análise discriminante “passo-a-passo”.

A apresentação do desenvolvimento detalhado dos conteúdos deste capítulo está feita no Anexo 9.

3. MEIOS E ESTRATÉGIAS A UTILIZAR

Apresentação da base teórica da análise discriminante, dos conceitos e dos procedimentos analíticos e informáticos pelo professor em aula de tipo expositivo.

Apresentação de exemplos e execução de exercícios em aula e fora dela. Análise de casos exemplares da bibliografia geográfica de regionalização.

4. BIBLIOGRAFIA

- DAVIS, John C. (1986) – *Statistics and Data Analysis in Geology*, 2nd edition, New York, John Wiley & Sons, p. 602-606.
- (*) KLECKA, William R. (1975) - “*Discriminant Analysis*” in NIE, Norman H. *et alia* - *SPSS -Statistical Package for Social Sciences*, 2th edition, McGraw-Hill Book Company, New York, p. 434-448.
- MATHER, Paul (1976) – *Computational Methods of Multivariate Analysis in Physical Geography*, John Wiley & Sons, London, p. 420- 459.
- (*) REIS, Elizabeth (1997) – *Estatística Multivariada Aplicada*, Lisboa, Edições Sílabo, p. 199-249.

10. Redes e Grafos

1. OBJECTIVOS A ATINGIR

1. Reconhecer as relações teóricas e analíticas entre as ligações e os fluxos entre os territórios e os conceitos de rede, grafo e matriz de conectividade.
2. Saber estabelecer o grafo e a matriz de conectividade para uma qualquer rede de ligação ou de fluxo.
3. Calcular e interpretar medidas de conectividade, acessibilidade e centralidade de redes e interpretar os seus respectivos valores.
4. Saber utilizar programas de análise de redes topológicas.
5. reconhecer as insuficiências da análise topológica e conhecer os meios de as ultrapassar.
6. Saber efectuar tratamentos simples de análise de matrizes de carga, aditivas e multiplicativas,
7. Interpretar e avaliar os resultados obtidos pelas diversas análises.

2. CONTEÚDOS A DESENVOLVER

As redes e a linguagem topológica. Grafos e matrizes. Arcos e nós. Grafos orientados e subgrafos.

Matrizes de conectividade. Conceitos de distância topológica, circuito, região, rede nula, rede em árvore, número associado, nó central e diâmetro da rede.

Medidas de conectividade em grafos: número ciclomático e índices beta, alfa e gama. Medidas de acessibilidade e de centralidade: índices de Shimbel e de dispersão.

Matrizes derivadas da de conectividade: Matrizes C_i , T_i , TG_i e D_i . Cálculo e interpretação.

Limitações da análise de grafos. Formas de as ultrapassar. As matrizes ponderadas, tratamento de carga, simples e cumulativas, aditivas e multiplicativas.

O desenvolvimento detalhado destes conteúdos é efectuado no Anexo 10.

3. MEIOS E ESTRATÉGIAS A UTILIZAR

Apresentação da base teórica, dos conceitos e dos procedimentos analíticos e informáticos de análise de redes, pelo professor em aula de tipo expositivo.

Disponibilização pelo docente de programas de análise de redes, inexistentes na secção de informática do Departamento.

Apresentação de exemplos e execução de exercícios em aula.

Análise e discussão de casos exemplares da investigação geográfica de redes.

4. BIBLIOGRAFIA

- (*) CICERI, M.F.; MARCHANT, B.; RIMBERT, S. (1977) – *Introduction a l'Analyse de l'Espace*, Masson, Paris, p. 45-53.
- CLIFF, A.D.; HAGGETT, P. (1981) "Graph Theory", in WRIGLEY, N.; BENNETT, R.J. (Eds) – *Quantitative Geography*, Routledge & Kegan Paul, London, p. 225-234.
- HAGGETT, P.; CHORLEY, R. J. (1969) – *Network Analysis in Geography*, Edward Arnold, London. 348 p.
- POTTS, R.B. (1978) "Networks" in LIGHTHILL – *Newer Uses of Mathematics*, Penguin Books, Harmondsworth, 431 p.
- SEGUÍ PONS, Joana M.; PETRUS BEY, Joana M. (1991) – *Geografía de Redes y Sistemas de Transporte*, Editorial Sintesis, Madrid, 241 p.
- SMITH, David M. (1977) – *Patterns in Human Geography*, Penguin Books, Harmondsworth, p. 279-285.
- TAAFFE, E. J.; GAUTHIER, H. L. (1973) – *Geography of Transportation*, Prentice-Hall, Englewood Cliffs, p. 249-283.

11. Geoestatísticas

1. OBJECTIVOS A ATINGIR

1. Reconhecer o interesse da introdução explícita da dimensão espacial nos procedimentos estatísticos de análise de dados.
2. Conhecer as medidas espaciais de centralidade e de dispersão, saber calculá-las e interpretar os valores determinados.
3. Saber calcular potenciais de variáveis.
4. Reconhecer o interesse dos potenciais em todos os tipos de análise geográfica e especialmente na análise de evoluções temporais de territórios em que existiram alteração de limites.
5. Identificar a importância do conceito de vizinhança em Geografia
6. Saber caracterizar entre distribuições regulares, aleatórias e concentradas de localizações territoriais.
7. Saber efectuar análise de vizinhança e avaliar o seu significado.

2. CONTEÚDOS A DESENVOLVER

A importância da dimensão espacial na análise estatística de variáveis geográficas. A localização enquanto factor fundamental na explicação da

variação. Relações entre Geoestatística, Cartografia e Sistemas de Informação Geográfica.

Medidas espaciais de centralidade e de dispersão. Centro de gravidade, distância-média, distância-padrão, e raio dinâmico. Elipse de dispersão: inclinação, dimensão dos eixos e excentricidade e área.

O conceito de potencial de uma variável e a sua utilidade na investigação geográfica, como expressão da interacção em Geografia Humana.

Análises de vizinhança. O conceito de vizinho mais próximo. As distribuições pontuais de localizações: concentradas, aleatórias e regulares. O índice R e a seu significado.

O desenvolvimento detalhado dos conteúdos programáticos deste capítulo é efectuado no Anexo 11.

3. MEIOS E ESTRATÉGIAS A UTILIZAR

Apresentação dos conceitos e dos métodos de cálculo elementares da Geoestatística pelo professor em aula. Apresentação de exemplos execução de cálculos manuais e utilização de programas no cálculo de parâmetros de geoestatística.

Execução de exercícios de aplicação, na aula e em sistema de auto-avaliação.

Análise e discussão de casos e da sua relação com a análise em SIGs.

4. BIBLIOGRAFIA

- ABREU, Diogo de (1978) – *Estudo Estatístico de Algumas Características do Povoamento em Portugal*. Faculdade de Letras de Lisboa, Lisboa, 121 p.
- BRONSTEIN, I.; SEMENDIAEV, K. (1979) – *Manual de Matemática para Engenheiros e Estudantes*, 2ª edição, Editorial Mir, Moscovo, p.271.
- DAVIS, John C. (1986) – *Statistics and Data Analysis in Geology*, 2nd edition, New York, John Wiley & Sons, p. 295-311.

- (*) EBDON, David (1996) – *Statistics in Geography*. 2th edition, Blackwell, Oxford (UK), p. 128-169.
- (*) ESTEBANEZ ALVAREZ, J.; BRADSHAW, R.. P. (1978) – *Técnicas de Quantificación en Geografía*, Editorial Tebar Flores, Madrid, p. 391-465.
- THOMAS, R. W. (1981) “*Point Pattern Analysis*”, in WRIGLEY, N.; BENNETT, R.J. (Eds) – *Quantitative Geography*, Routledge & Kegan Paul, London, p. 164-176.

12. Modelos Integrados em Geografia

1. OBJECTIVOS A ATINGIR

1. Motivar os alunos para as potencialidades da utilização de modelos em Geografia e em Planeamento.
2. Ilustrar como os conhecimentos adquiridos anteriormente na disciplina podem ser úteis na construção de modelos.
3. Conhecer os princípios fundamentais de construção de modelos.
4. Saber as bases teóricas gerais dos modelos demográficos “*cohort-survival*” e de entradas/saídas (*input/output*).
5. Conhecer os princípios teóricos e o esquema básico do funcionamento do modelo integrado REX.
6. Avaliar a sua importância para a simulação de processos de desenvolvimento regional.
7. Conhecer os princípios teóricos e o esquema básico do funcionamento do modelo demográfico Espon114-WP5.
8. Avaliar a sua importância para prever alterações na estrutura demográfica regional.

9. Aprofundar o conhecimento dos processos demográficos regionais, nomeadamente o envelhecimento, o despovoamento e as migrações.

10. Compreender as relações entre a demografia e a economia no âmbito regional.

11. Reflectir sobre o papel dos modelos na melhor compreensão dos processos geográficos.

12. Acentuar o papel pedagógico de que se reveste a utilização de modelos na investigação e em planeamento, por permitir efectuar simulações em que se exageram as condicionantes e as restrições, como forma didáctica de acentuar também as suas consequências.

2. CONTEÚDOS A DESENVOLVER

Os modelos em Geografia: tipos e princípios de construção. Utilidade da modelização na descrição, previsão e simulação de situações e de processos.

Relação entre a modelização e a análise de dados Relações e funções. O uso de matrizes na descrição de sistemas e processos geográficos complexos

Apresentação de modelos simples. Descrição das relações fundamentais e da forma como se formalizaram dois modelos de utilização clássica em Geografia regional, um demográfico (*cohort-survival*) e outro económico (*input/output*).

O modelo integrado REX: objectivos, princípios orientadores, relações fundamentais, organização e construção. Capacidades e limitações. Exemplos de aplicação. Crítica dos resultados.

O modelo demográfico ESPON114-WP5: objectivos, princípios orientadores, relações fundamentais, organização e construção. Capacidades e limitações. Exemplos de aplicação. Crítica dos resultados.

3. MEIOS E ESTRATÉGIAS A UTILIZAR

Exposição pelo docente dos princípios orientadores da utilização de modelos e da sua importância em Geografia, e apresentação, como exemplo, dos modelos REX e Espo114-WP5.

Discussão, liderada pelo docente, sobre a avaliação dos modelos apresentados.

4. BIBLIOGRAFIA

- ABRAMI, Valerio Terra (1998) – *Le Previsioni Demografiche*, Il Mulino, Bologna, 186 p.
- ABREU, D. *et alia* (1986) - "Construção de um Modelo para Avaliação do Impacte Regional da Integração de Portugal na CEE", *IV Colóquio Ibérico de Geografia - Actas*. Coimbra, Instituto de Estudos Geográficos, p. 59-76
- ABREU, D.; JENSEN-BUTLER, Chris (1987) - *The Multiplicative Effects of Regional Investment in Portugal*, Aarhus, Geografisk Institut, 25 p. (Notat 66).
- (*) ABREU, D. (1988) - *Desenvolvimento Regional no Oeste. Problemas e Métodos*. Universidade de Lisboa, Faculdade de Letras, Lisboa, 526 p. Dissertação de Doutoramento em Geografia Humana.
- (*) ABREU, D. *et alia* (1988) - *Rex: A Regional Economic-Demographic Model for Portugal*, Aarhus, Geografisk Institut, 20 p. (Notat 70).
- ABREU, D. ; COSTA, N. (2002) - "Portugal: Perspectivas sobre um país em mudança. Reflexos territoriais de diferentes cenários de desenvolvimento" in C. Cavaco (Coord.) – *"Repensar Portugal na Europa. Perspectivas de um país periférico"*, Lisboa, C.E.G., p. 332-344 (EPRU 55).
- (*) ABREU, D.; GASPAR, J.; COSTA, N. (2005) - *ESPON Project 1.1.4 Annex B: Ageing, Labour shortage, and 'Replacement Migration'*, C.E.G., Lisboa, CEG, 168 p.
- (*) ARMSTRONG, H.; TAYLOR, J. (1993) – *Regional economics & policy*, 2nd edition, Harvester Wheatsheaf, New York, 397 p.
- (*) ARMSTRONG, H.; TAYLOR, J. (2000) – *Regional economics & policy*, 3th edition, Blackwell Publishers, Malden (USA), 437 p.
- BANDEIRA, M. L. (2004) – *Demografia. Objecto, teorias e métodos*, Escolar Editora, Lisboa, 340 p.
- BONNEUIL, Noel (1997) – *Introduction à la modélisation démographique*, Armand Colin, Paris, 128 p.
- BOURGEOIS-PICHAT, Jean (1970) – *Main Trends in Demography*, George Allen & Unwin, London, 79 p.
- CASTRO, E. *et alia* (2002) - *Avaliação do impacto socio-económico de instrumentos legais para o controlo das emissões de gases com*

- efeito de estufa, Relatório Final* Departamento de Ambiente e Ordenamento da Universidade de Aveiro, Aveiro.
- CURWIN, Jon; SLATER, Roger. (1996) – *Quantitative Methods for Business Decisions*, 4th edition, International Thoms Business Press, London, 667 p.
- DEELSA/ELSA (2002) – *Can Migration Help to Alleviate Labour Shortages?*. Organization for Economic Co-operation and Development.
- ESPON 2006 Programme Action 1.1.3 (2003) – “2nd Interim Report”, *Particular effects of enlargement of the EU and beyond on the polycentric spatial tissue with special attention on discontinuities and barriers*, Espon, Luxembourg.
- ESPON 2006 Programme Action 1.1.4 (2003) – “2nd Interim Report”, *The Spatial Effects of Demographic Trends and Migration*, Espon, Luxembourg.
- ESPON 2006 Programme Action 1.1.2 (2003) – “3rd Interim Report”, *Urban-Rural Relations in Europe*, Espon, Luxembourg.
- ESPON 2006 Programme Action 3.1 (2003) – “3rd Interim Report”, *Integrated Tools for European Spatial Development*. Espon, Luxembourg.
- European Commission (1998) – *Analysis and Forecasting of International Migrations by Major Groups (Part II)*. EC, Luxembourg.
- European Commission (1999) – *Sixth Periodic Report on the social and economic situation and development of the regions of the European Union*. EC, Luxembourg.
- European Commission (2002) – *National reports on the demographic situation in 12 Central European Countries, Cyprus and Malta in 2000*. EC, Luxembourg.
- European Commission (2004) – *A New Partnership for Cohesion - convergence competitiveness cooperation. Third Report on Economic and Social Cohesion*. EC, Luxembourg.
- European Commission (S/d) – *The European Labour Market in the light of demographic change*. Employment & Social Affairs.
- FERRÃO, J. et alia (1986) – *Construction of Regional and Interregional Input-Output Tables for Portuguese Regions. A technical report*, Lisbon, 59 p.
- (*) GASPAR, J. et alia (1989) – *PORTUGAL. Os Próximos 20 Anos. Vol. VI - Ocupação e Organização do Espaço. Uma Prospectiva*. Fundação Calouste Gulbenkian, Lisboa, 344 p.
- GASPAR, J. et alia (1997) – “*Plano Nacional da Água - População, Economia e Território: Cenários de desenvolvimento*”, relatório-parecer de natureza reservada, entregue no INAG - Instituto Nacional da Água, 152 p.
- (*) GASPAR, J. (Coord.) (2002) – “*Estudo de Impacte Sócio-Económico do Porto de Lisboa*”, APL - Administração do Porto de Lisboa, S.A., Lisboa, 158 p. e anexos.
- GASPAR, J. et alia (2003) – “*O Impacte Socio-Económico do Porto de Lisboa*”, in Sousa, João Figueira (coord.) – “*Portos, Transportes Marítimos e Território*”, Instituto da Dinâmica do Espaço, FCSH, Lisboa, ISBN 972-99124-0-8.

- GASPAR, J.; ABREU, D.; COSTA, N. et al. (2005 ?) - *Ageing. Labour Shortage and Replacement Migration*, Gaspar, Jorge.; Abreu, Diogo, Costa, Nuno et al. ESPON, Luxemburgo. ISBN: 91-975257-0-7b (no prelo).
- (*) ISARD, Walter et alia (1998) - *Methods of Interregional and Regional Analysis*, Ashgate, Aldershot, 490 p.
- (*) JOHANSSON Mats; RAUHUT, Daniel (eds.) (2005 ?) - *ESPO project 1.1.4 The Spatial Effects of Demographic Trends and Migration*, ESPON, Luxemburgo. ISBN: 91-974582-9-5 (no prelo).
- LAW, Averril M., KELTON, W. David (1991) - *Simulation Modelling & Analysis*, 2nd edition, McGraw-Hill International editions, New York, 759 p.
- MADSEN, Bjarne et alia (1996) - *Modelling the Economy and the Environment*, Springer Verlag, Berlin, 365 p.
- MARTINS, J. et alia (2002) - "Avaliação do impacto económico de instrumentos legais de controlo das emissões de gases com efeito de estufa: análise Input - Output", Porto, Actas do II Encontro Nacional do Colégio de Engenharia do Ambiente - 29 e 30 de Novembro de 2002.
- MARTINS, J. et alia (2003) - "Construção de uma Matriz Input-Output Rectangular e sua Aplicação à Análise de Questões Ambientais", Recife, V Encontro de Economistas de Língua Portuguesa, 5 - 7 de Novembro de 2003.
- MARTINS, J. et alia (2003) - "Avaliação do Impacto Económico de Instrumentos Legais de Controle das Emissões de Gases com Efeito de Estufa: Aplicação da Análise Input-Output", Lisboa, ISEG, Workshop sobre Novos Desenvolvimentos e Aplicações da Análise Input-Output, 9 de Maio de 2003.
- MARTINS, J. et alia (2005) - *OIKOMATRIX II - Avaliação do impacto socio-económico de instrumentos legais para o controlo das emissões de gases com efeito de estufa*, Departamento de Ambiente e Ordenamento da Universidade de Aveiro, Aveiro, 115 p.
- MUSU, Ignazio; SINISCALCO, Domenico (1988) - *Ambiente e contabilità nazionale*, Il Mulino, Bologna, 485 p.
- NAZARETH, J. Manuel (1996) - *Introdução à Demografia Teoria e Prática*, Editorial Presença, Lisboa, 194 p.
- NEWTON, P. W.; TAYLOR, M.A.P.; SHARPE, R. (Eds) (1988) - *Desktop Planning. Microcomputer applications for infrastructure and services planning and management*, Hargreen Publishing Co., Melbourne, p. 108-394.
- (*) REIF, Benjamin (1978) - *Modelos en la Planificacion de Ciudades y Regiones*, Instituto de Estudios de Administracion Local, Madrid, 422 p.
- SMITH, David M. (1977) - *Patterns in Human Geography*, Penguin Books, Harmondsworth, p. 2786-303.
- UN (1998) - *Rapport succinct sur le suivi des questions de population à l'échelle mondiale (1999): accroissement, structure et repartition de la population*. UN, New York.
- (*) UN (2000) - *Replacement Migration: Is it A Solution to Declining and Ageing Populations?* United Nations, Department of Economic and Social Affairs, United Nations Secretariat, New York.

- (*) UN (2001) - *World Population Ageing 1950-2050*. United Nations, Department of Economic and Social Affairs, United Nations Secretariat.
- UN (2001) - *World Population Prospects. The 2000 Revision*. United Nations Population Division.
- UN (2003) - *World Population Prospects. The 2002 Revision*. United Nations Population Division.
- (*) VINUESA, Julio (Ed.) (1997) - *World Population Ageing 1950-2050*. United Nations, Department of Economic and Social Affairs, United Nations Secretariat.

ANEXOS

CONTEÚDOS A DESENVOLVER

(Capítulos 1 a 11)

UNIVERSIDADE DE LISBOA



**FACULDADE DE LETRAS
DEPARTAMENTO DE GEOGRAFIA**

ANEXOS
ao programa de
ANÁLISE DE DADOS II

Capítulos 1 a 11
Conteúdos a desenvolver

DIOGO JOSÉ BROCHADO DE ABREU

ÍNDICE

JUSTIFICAÇÃO	A - 11

ANEXO 1 – CONCEITOS BÁSICOS E DE ENQUADRAMENTO	A - 13

..1. – Introdução	A - 15
1.1. – A investigação científica em geral	A - 15
1.2. – A investigação em Geografia	A - 16
1.2.1 – <u>O problema do método</u>	A - 16
1.2.2 – <u>O âmbito da análise</u>	A - 18
2. – As Etapas do Trabalho Científico	A - 19
2.1. – A questão de partida	A - 19
2.1.1 – <u>Seleccção do tema</u>	A - 19
2.1.2 – <u>Delimitação do âmbito da investigação</u>	A - 19
2.1.3 – <u>Formulação da questão de partida</u>	A - 19
2.1.3.1 – Clareza	A - 19
2.1.3.2 – Exequibilidade	A - 20
2.1.3.3 – Pertinência	A - 20
2.2. – Preparação da investigação	A - 20
2.2.1 – <u>As leituras</u>	A - 20
2.2.1.1. – Seleccionar	A - 20
2.2.1.2 – Ler, resumir e comparar	A - 21
2.2.2 – <u>As entrevistas exploratórias</u>	A - 21

2.2.2.1. – Identificar peritos, especialistas e testemunhas	A - 21
2.2.2.2. – Entrevistar: ouvir e decodificar os conteúdos	A - 21
2.3. – Definir o problema	A - 22
2.3.1 – <u>Conciliar leituras, entrevistas e coordenações</u>	A - 22
2.3.2 – <u>Estabelecer um quadro teórico</u>	A - 22
2.3.3 – <u>Explicitar a problemática</u>	A - 22
2.4. – A Construção do esquema	A - 22
2.4.1 – <u>Identificar e precisar os conceitos</u>	A - 22
2.4.2 – <u>Construir as hipóteses e os modelos</u>	A - 22
2.5. – A observação	A - 22
2.5.1 – <u>Delimitar o campo de observação</u>	A - 23
2.5.2 – <u>Determinar e testar os métodos de observação</u>	A - 23
2.5.3 – <u>Recolher as informações</u>	A - 23
2.6. – A análise das informações	A - 24
2.6.1 – <u>Arrumar os dados e prepará-los para análise</u>	A - 24
2.6.2 – <u>Efectuar os procedimentos analíticos</u>	A - 24
2.6.3 – <u>Comparar os resultados observados com os esperados</u>	A - 24
2.7. – A conclusão	A - 25
3. – A Natureza da Informação Geográfica	A - 25
3.1. – A informação geográfica	A - 25
3.1.1 – <u>A matriz geográfica</u>	A - 26
3.1.2 – <u>O cubo de dados de Cattell</u>	A - 27
3.1.3 – <u>Breve tipologia de matrizes e tratamentos</u>	A - 28

ANEXO 2 – ALGUMA MATEMÁTICA PARA GEÓGRAFOS ..	A - 31
1. – Matemática Básica	A - 33
1.1. – <u>Potências</u>	A - 33
1.2. – <u>Logaritmos</u>	A - 34
1.3. – <u>Funções</u>	A - 34
1.4. – <u>Matrizes</u>	A - 35
2. – Índices	A - 35
2.1. – <u>Quocientes</u>	A - 36
2.2. – <u>Taxas seccionais</u>	A - 37
2.3. – <u>Taxas de variação</u>	A - 37
2.4. – <u>Exercícios</u>	A - 38
2.5. – <u>Bibliografia</u>	A - 38
3. – Medidas de Concentração	A - 39
3.1. – <u>Curva de Lonrenz</u>	A - 39
3.2. – <u>Índices de Concentração</u>	A - 46
3.2.1 – <u>Coeficiente de Especialização</u>	A - 46
3.2.2 – <u>Índice de Gini</u>	A - 47
3.2.3 – <u>Índice de Schutz</u>	A - 49
3.2.4 – <u>Ponto de Igual Partilha</u>	A - 50
3.2.5 – <u>Coeficiente de Concentração de Gini</u>	A - 51
3.3. – <u>Considerações</u>	A - 53
3.4. – <u>Exemplos</u>	A - 54
3.5. – <u>Exercícios</u>	A - 57
3.6. – <u>Bibliografia</u>	A - 59

ANEXO 3 – RECOLHA DA INFORMAÇÃO	A - 61
1. – Generalidades	A - 63
2. – Fontes de Informação	A - 63
2.1. – Fontes Documentais	A - 64
2.1.1 – <u>Estatísticas Nacionais</u>	A - 65
2.1.2 – <u>Estatísticas Internacionais</u>	A - 66
2.1.2.1 – Das Nações Unidas	A - 66
2.1.2.2 – Do Banco Mundial / PNUD	A - 67
2.1.2.3 – Da União Europeia	A - 67
2.1.2.4 – De outros organismos	A - 67
2.1.3 – <u>Outras Fontes Documentais</u>	A - 68
2.1.3.1 – Dados não publicados	A - 68
2.1.3.2 – Informação em arquivos administrativos e de gestão	A - 69
2.1.3.3 – Integração de inquéritos estatísticos com fontes administrativas	A - 70
2.1.3.4 – Insuficiência das fontes documentais	A - 71
2.2. – Observação	A - 72
2.2.1 – <u>A Observação Exterior</u>	A - 72
2.2.2 – <u>A Observação Participante</u>	A - 73
2.3. – Experimentação	A - 74
2.4. – Análise Documental	A - 74
2.5. – Inquéritos	A - 75
2.5.1 – <u>Entrevistas</u>	A - 75
2.5.1.1 – Tipos de Entrevista	A - 75
2.5.1.2 – Vantagens e Limitações das Entrevistas.....	A - 77
2.5.1.3 – O Esquema da Entrevista	A - 77
2.5.1.4 – A Realização da Entrevista	A - 77
2.5.2 – <u>Questionários</u>	A - 78
2.5.2.1 – Tipos de Questionários	A - 79
2.5.2.2 – Vantagens e Limitações dos Questionários	A - 81
2.5.2.3 – O Desenho dos Questionários	A - 82
2.5.2.4 – A Realização dos Questionários	A - 82
2.6. – Os Grupos de Discussão	A - 84
2.7. – Os métodos Delphi	A - 85
2.8. – As histórias de vida	A - 86
3. – Considerações Finais	A - 86
3.1. – Etapas de uma Operação de Recolha	A - 87
3.2. – Procedimentos Posteriores	A - 88
4. – Amostragem	A - 90
4.1. – Introdução	A - 90
4.2. – Conceitos básicos	A - 90
4.2.1 – <u>Amostra</u>	A - 91
4.2.2 – <u>Valor aproximado</u>	A - 91
4.2.3 – <u>Limites do erro</u>	A - 91
4.2.4 – <u>Risco</u>	A - 91
4.2.5 – <u>Notações</u>	A - 92
4.3 – Estimativas	A - 92
4.4. – Tamanho das Amostras	A - 93
4.4.1 – <u>Populações Infinitas</u>	A - 93
4.4.2 – <u>Populações Finitas</u>	A - 94
4.4.3 – <u>Exemplos de cálculo da dimensão de amostras</u>	A - 94
4.5. – Selecção dos Elementos de Amostras	A - 98
4.5.1 – <u>Métodos de Amostragem</u>	A - 98

4.5.1.1 – Métodos probabilísticos	A - 98
4.5.1.2 – Métodos não probabilísticos	A - 101
5. – Tabelas	A - 103
5.1. – Tabelas da Dimensão de Amostras	A - 103
5.2. – Tabela de Números Aleatórios	A - 106
6. – Bibliografia	A - 107

ANEXO 4 – TABELAS DE CONTINGÊNCIA	A - 109
1. – Introdução	A - 111
2. – Transformação para Tabelas de Contingência	A - 113
2.1. – Modificação por ponderação	A - 114
2.2. – Quadros Lógicos	A - 116
2.3. – Tabelas de Burt	A - 118
2.3.1 – <u>Construção de uma tabela de Burt</u>	A - 120
2.3.2 – <u>Generalidades sobre análise de tabelas</u>	A - 121
3. – Análise de Tabelas a duas variáveis	A - 121
3.1. – Tabelas de contingência (I x c) a n estados	A - 122
3.1.1 – <u>Medidas de dependência</u>	A - 123
3.1.2 – <u>Medidas de associação</u>	A - 126
3.1.3 – <u>Medidas de associação e predição</u>	A - 129
3.1.4 – <u>Medidas de predição</u>	A - 131
3.2. – Tabelas de contingência (I x c) a 2 estados	A - 135
3.2.1 – <u>Qui quadrado</u>	A - 136
3.2.2 – <u>Qui quadrado corrigido por Yates</u>	A - 136
3.2.3 – <u>Coeficiente Fi</u>	A - 136
3.2.4 – <u>Coeficiente Q de Yule</u>	A - 137
3.2.5 – <u>Coeficiente Y</u>	A - 137
3.2.6 – <u>Razão dos produtos cruzados (α)</u>	A - 137
3.2.7 – <u>Logaritmo de α</u>	A - 138
3.3. – Exemplos	A - 138
3.3.1 – <u>Medidas de dependência para 2 variáveis a n estados</u>	A - 138
3.3.2 – <u>Medidas de associação para 2 variáveis a n estados</u>	A - 142
3.3.3 – <u>Medidas de associação e predição para 2 variáveis a n estados</u> ..	A - 144
3.3.4 – <u>Medidas de predição para 2 variáveis a n estados</u>	A - 145
3.3.5 – <u>Tabelas de contingência de 2 variáveis a 2 estados</u>	A - 148
4. – Análise de Tabelas de Burt	A - 150
4.1. – Análise de χ^2	A - 150
4.1.1 – <u>Exemplo</u>	A - 150
5. – Tabela de valores críticos de Qui quadrado	A - 154
6. – Bibliografia	A - 155

ANEXO 5 – ANÁLISE DE VARIÂNCIA	A - 157
1. – Introdução	A - 159
2. – Tipos de Análise de Variância	A - 160
2.1. – Análise de Variância a um Factor	A - 160
2.2. – Análise de Variância Multifactorial	A - 164
3 – Análise de Covariância	A - 167
4 – Considerações Gerais	A - 169
5 – Bibliografia	A - 170

ANEXO 6 – REGRESSÃO LINEAR MÚLTIPLA	A - 171
1. - Revisão da Regressão Linear Simples	A - 174
2. - O modelo de Regressão Linear Múltipla	A - 174
2.1. - Coeficientes de Regressão	A - 174
2.2. - Coeficiente de Correlação Múltipla	A - 175
2.3. - Significado da Regressão	A - 176
2.3.1 - <u>Teste F</u>	A - 176
2.3.2 - <u>Teste t</u>	A - 177
2.4. - Erro-padrão da estimativa	A - 179
2.5. - Erro-padrão dos coeficientes de regressão	A - 179
2.6. - Tipos de regressão múltipla	A - 180
2.6.1 - <u>Simultânea</u>	A - 180
2.6.2 - <u>Passo-a-passo</u>	A - 180
2.6.3 - <u>Outros</u>	A - 181
2.7. - Extensões do modelo	A - 181
3 - Problemas Resolvidos	A - 182
3.1. - Enunciados	A - 182
3.2. - Soluções	A - 182
4. - Exemplos	A - 183
4.1. - Tipos de cultivo e rendimento em explorações agrícolas ...	A - 183
4.2. - Criminalidade nos Estados Unidos, 1978	A - 187
4.3. - Evotranspiração em Quelimane (Moçambique)	A - 190
5. - Bibliografia	A - 191

ANEXO 7 – ANÁLISE FACTORIAL	A - 193
1. - Introdução	A - 195
2. -Tipos de Análise Factorial	A - 196
2.1. - Preparação da Matriz de Correlação	A - 197
2.2. - A Extracção dos Factores	A - 198
2.3. - A Rotação dos Factores	A - 199
2.4. - O Cálculo do valor dos Factores	A - 200
3. - Marchas de Análise	A - 200
3.1. - Sequência de uma Análise em Componentes Principais	A - 201
3.2. - Sequência de uma Análise Factorial	A - 201
4. - Exemplo	A - 203
5. - Bibliografia	A - 206

ANEXO 8 – CLASSIFICAÇÃO NUMÉRICA	A - 209
1. - Apresentação do Problema	A - 211
2. - Medidas de Semelhança e de Distância a Utilizar	A - 214
2.1. - Para Dados Nominais	A - 214
2.1.1 - <u>Para Dados Binários Dicotómicos</u>	A - 215
2.1.1.1 - Coeficiente de Jaccard (1908)	A - 215
2.1.1.2 - Coeficiente de Russel & Rao (1940)	A - 215
2.1.1.3 - Coeficiente de Dice (1915)	A - 215
2.1.2 - <u>Para Dados Binários Alternativos</u>	A - 215
2.1.2.1 - Coeficiente de Concordância Simples	A - 215

2.1.2.2 – Coeficiente de Sneath & Sokal	A - 216
2.1.2.3 – Coeficiente de Youle	A - 216
2.1.2.4 – Coeficiente de Pearson para dados nominais	A - 216
2.2. – Para Dados de Intervalo e Razão	A - 216
2.2.1 – <u>Métricas de Minkowski</u>	A - 218
2.2.1.1 – Distância de Manhattan	A - 218
2.2.1.2 – Distância Euclidiana	A - 219
2.2.1.3 – Coeficiente de Chebychev	A - 219
2.2.2 – <u>Coeficiente de Correlação de Pearson</u>	A - 220
2.2.3 – <u>Cosseno</u>	A - 220
2.2.4 – <u>Distância de Qui-quadrado</u>	A - 221
2.2.5 – <u>Exemplos de cálculo</u>	A - 221
2.3. – Para Dados Nominais, Ordinais, de Intervalo e Razão	A - 224
3. Estratégias de Agrupamento	A - 225
3.1. Para Classificações não hierárquicas (partições)	A - 225
3.1.1 – <u>Nuvens dinâmicas</u>	A - 225
3.1.2 – <u>K-médias</u>	A - 226
3.1.3 – <u>K-centróides</u>	A - 226
3.2. – Para Classificações Hierárquicas Ascendentes	A - 226
3.2.1 – <u>Método do vizinho mais próximo</u>	A - 226
3.2.2 – <u>Método do vizinho mais distante</u>	A - 227
3.2.3 – <u>Método da média do grupo</u>	A - 227
3.2.4 – <u>Método da média simples</u>	A - 228
3.2.5 – <u>Método do centróide</u>	A - 228
3.2.6 – <u>Método da mediana</u>	A - 228
3.2.7 – <u>Método de Ward</u>	A - 229
3.2.8 – <u>Método das distâncias internas</u>	A - 229
3.2.9 – <u>A fórmula geral de agregação</u>	A - 229
3.2.10 – <u>Exemplificação da agregação</u>	A - 230
3.2.11 – <u>Exemplos</u>	A - 231
4. – Avaliação do Ajustamento	A - 237
5. Bibliografia	A - 238
 ANEXO 9 – ANÁLISE DISCRIMINANTE	 A - 239
1. – Introdução	A - 241
2. – O Decurso da Análise	A - 241
2.1. – As Funções Discriminantes	A - 242
2.2. – As Coordenadas e os Centróides dos Grupos	A - 242
3. – Exemplo	A - 243
4. – Bibliografia	A - 246
 ANEXO 10 – REDES E GRAFOS	 A - 247
1. – A Linguagem Topológica: Redes, Grafos e Matrizes	A - 249
1.1. – Grafos	A - 250
1.2. – Matriz de Conectividade	A - 250
1.3. – Conceitos	A - 251
2. – Medidas de Ligação e Conectividade	A - 251
3. – Medidas de Acessibilidade e Centralidade	A - 254
4. – Matrizes Derivadas	A - 256

5. – Matrizes Ponderadas	A - 260
6. – Bibliografia	A - 262
ANEXO 11 – GEOESTATÍSTICAS	A - 263
1. – Introdução	A - 265
2. – Medidas Espaciais de Centralidade e Dispersão	A - 265
2.1. – Centro de Gravidade	A - 265
2.2. – Distância-média	A - 268
2.3. – Distância-padrão	A - 269
2.4. – Raio Dinâmico	A - 270
2.5. – Elipse de Dispersão	A - 272
2.5.1 – <u>Ângulo de Inclinação</u>	A - 273
2.5.2 – <u>Dimensão dos Eixos</u>	A - 273
2.5.3 – <u>Excentricidade</u>	A - 274
2.5.4 – <u>Área da Elipse</u>	A - 274
3. – Potenciais	A - 276
4. – Análise do Vizinho mais Próximo	A - 280
4.1. – O Índice R	A - 280
4.2. – Teste de Significado	A - 282
4.3. – Considerações Finais	A - 282
4.4. – Exemplo de Cálculo	A - 283
5. – Leitura do “Output”	A - 284
6. – Exercícios	A - 286
7. – Bibliografia	A - 288

JUSTIFICAÇÃO

É conveniente que o ensino inicial dos Métodos Quantitativos em Geografia tenha como base manuais básicos, em que os temas matemáticos a tratar sejam apresentados de forma adequada aos objectivos de ensino e à formação com que os alunos chegam ao curso de Geografia. Para isso é fundamental que os diversos temas sejam expostos com rigor e simplicidade e que se apresentem exemplificações dos problemas que as técnicas ajudam a resolver e das interpretações geográficas que permitem.

Não se pode esquecer que o que se ensina é sempre Geografia e que os métodos matemáticos, simples ou complexos, são apenas instrumentos que devem tornar mais evidente e rigorosa a compreensão dos fenómenos geográficos. Se assim não for, se os métodos quantitativos forem apenas uma dificuldade mais, então será inútil ministrar o seu ensino. Nos cursos superiores actuais, cuja duração se torna cada vez menor, a eficácia terá de ser a regra, não se podendo admitir que as matérias que se leccionam não tenham utilidade para os alunos.

Por outro lado, no decorrer do processo didáctico deve ser dada ao aluno a capacidade de controlar os progressos que vai fazendo na sua aprendizagem, através de problemas resolvidos que exemplificam os procedimentos e de problemas em que é dada a solução. A análise pormenorizada dos exemplos e o controle que permite o confronto com a solução permite avaliar a correcção dos

processos, compreender em que condições podem ser utilizados e sugerir como devem ser interpretados os resultados obtidos. A execução de exercícios com soluções dá-lhes a certeza de que estão a executar correctamente os procedimentos matemáticos e informáticos necessários.

Neste volume de anexos exemplifica-se o modo como podem ser elaborados esses manuais. Os anexos que aqui se apresentam têm a ver com o programa da disciplina de **Análise de Dados II**, que é a disciplina cujo programa e métodos de ensino se apresentam a estas provas, mas consideramos que o mesmo tipo de textos se deve utilizar no ensino da disciplina de **Análise de Dados I**, que respeita ao ensino dos métodos quantitativos mais elementares.

Na nossa opinião os textos de apoio que se seguem poderão vir a configurar um manual para o segundo nível do ensino de métodos quantitativos em Geografia.

Nele se apresentam textos referentes aos Capítulos 1 a 11 do programa de **Análise de Dados II**, que são os capítulos mais técnicos, em que se trata das técnicas e métodos que os alunos têm de saber formular, calcular e interpretar.

O Capítulo 12 ficou de fora destes textos. É o capítulo com que se encerra o programa, em que são apresentados exemplos de aplicação das técnicas e métodos, integrados em modelos complexos, que foram efectivamente utilizados em estudos de planeamento territorial. Não se pretende que os alunos consigam formular ou calcular estes modelos. Pretende-se apenas que compreendam o modo como foram feitos e avaliem criticamente as condições e o interesse da sua utilização. Nessas condições não tem sentido a sua apresentação nestes anexos de apoio ao programa. A estratégia de apresentação e discussão será outra, baseada em documentos originais e em súmulas exemplificativas da estrutura, organização, funcionamento e utilização dos modelos, apresentadas pelo docente.

Anexo 1

Conceitos Básicos e de Enquadramento

1. - INTRODUÇÃO

1.1. A investigação científica em geral

Nesta ocasião, em que se inicia o estudo da disciplina de **Análise de Dados II**, passada que foi já a fase inicial da aprendizagem de técnicas quantitativas básicas e com os alunos mais integrados nos princípios e nas práticas da investigação geográficas, é altura de colocar a questão da utilização dos métodos quantitativos de Análise de Dados em Geografia.

O actual pensamento científico¹, que pode ser apreciado em Popper², considera que o mundo real é organizado, que é possível conhecer aspectos de sua organização e que uma das mais eficientes maneiras de o fazer é através do pensamento científico, que deve respeitar três exigências:

- o estabelecimento de conjecturas, que são ideias simples, inovadoras e poderosas que estabelecem relações entre os fenómenos;
- a reprodutibilidade das conjecturas, isto é, a possibilidade de se repetir as experiências e ou as observações que levaram à conjectura, com os mesmos resultados, ou com diferenças explicáveis pelas condições do meio;
- e a utilidade da conjectura, no sentido de que as suas predições se realizam, e resiste a experiências e observações que procuram a sua falsidade.

A Análise de Dados desempenha um papel fundamental em todo o processo de produção de conhecimento científico. É com a Análise de Dados que se encontram os padrões e as regularidades nos acontecimentos e as repetições e recorrências que permitem e suscitam o aparecimento das conjecturas – em análise de tipo exploratório – e é

96

¹ - Que é apenas uma forma de aceder ao conhecimento, pese embora a grande valorização social de que desfruta, em grande parte devida à sua utilidade e capacidade preditiva.

² - POPPER, Karl R. (1959) – **The Logic of Scientific Discovery**, London, Hutchinson (tradução do alemão **Logik der Firschung** (1934), Springer-Verlag, Viena)

também com a Análise de Dados que se constroem as teorias e os modelos que permitem as predições que confirmam a sua respectiva validade, não já baseada em ideias de senso comum, mas na verificação rigorosa das suas ocorrências.

1.2. A investigação em Geografia

Em Geografia, como noutros ramos do saber, a abordagem científica é dominante, mesmo tendo em conta a impossibilidade de repetir muitos dos factos, acontecimentos e processos que são objecto de observação e interpretação em Geografia.

A investigação científica baseia-se no estabelecimento de conjecturas (as hipóteses de trabalho), que reflectem a existência de recorrências ou de repetições nos padrões e nos processos, que levam à intuição de relações entre os fenómenos analisados.

1.2.1. O problema do método

Uma das questões que se colocam ao investigador é pois a da escolha do método de análise que vai utilizar na investigação. É costume distinguir classicamente entre dois tipos distintos de análise científica das questões, o método indutivo e o método dedutivo.

Uma boa apresentação destes métodos foi feita por David Harvey e pode ser apreciada sinteticamente no esquema da Figura 1.1.

De uma forma simplificada, pode afirmar-se que o método indutivo, partindo da observação das realidades concretas, formula, por indução, as hipóteses, enquanto o método dedutivo estabelece as hipóteses a partir de leis e teorias gerais pré-existentes.

De facto, a realidade é um pouco mais complexa, e a abordagem ao problema do conhecimento científico, mesmo em Geografia, mais integradora. Na realidade, a observação, passo primeiro e fundamental nas abordagens indutivas, é sempre orientada pelo conhecimento pré-existente, e nesse sentido, é sempre consequência de formulações teóricas do observador, mais ou menos explícitas. E no método dedutivo, é o confronto com realidades diversas das que foram previstas a partir das teorias existentes, que leva a novas formulações (de tipo indutivo) que procuram corrigir, adaptar ou substituir as formulações teóricas anteriores.

Tem vindo assim a ganhar força uma perspectiva empírico-dedutiva integrada, em que se aliam as formulações indutivas com as dedutivas, sendo o predomínio de uma ou de outra mais fruto dos fins e das condições objectivas em que decorre a investigação, do que de posições teórico-filosóficas apriorísticas dos investigadores.

O que importa sublinhar aqui é a importância da Análise de Dados nas duas abordagens metodológicas, tanto na ordenação e classificação da informação bruta (desordenada) e na constatação de regularidades

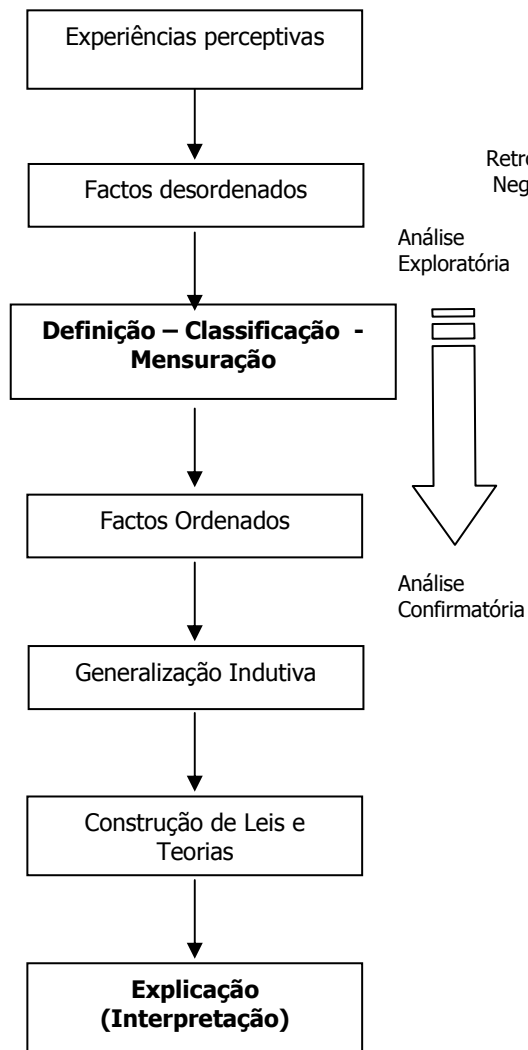
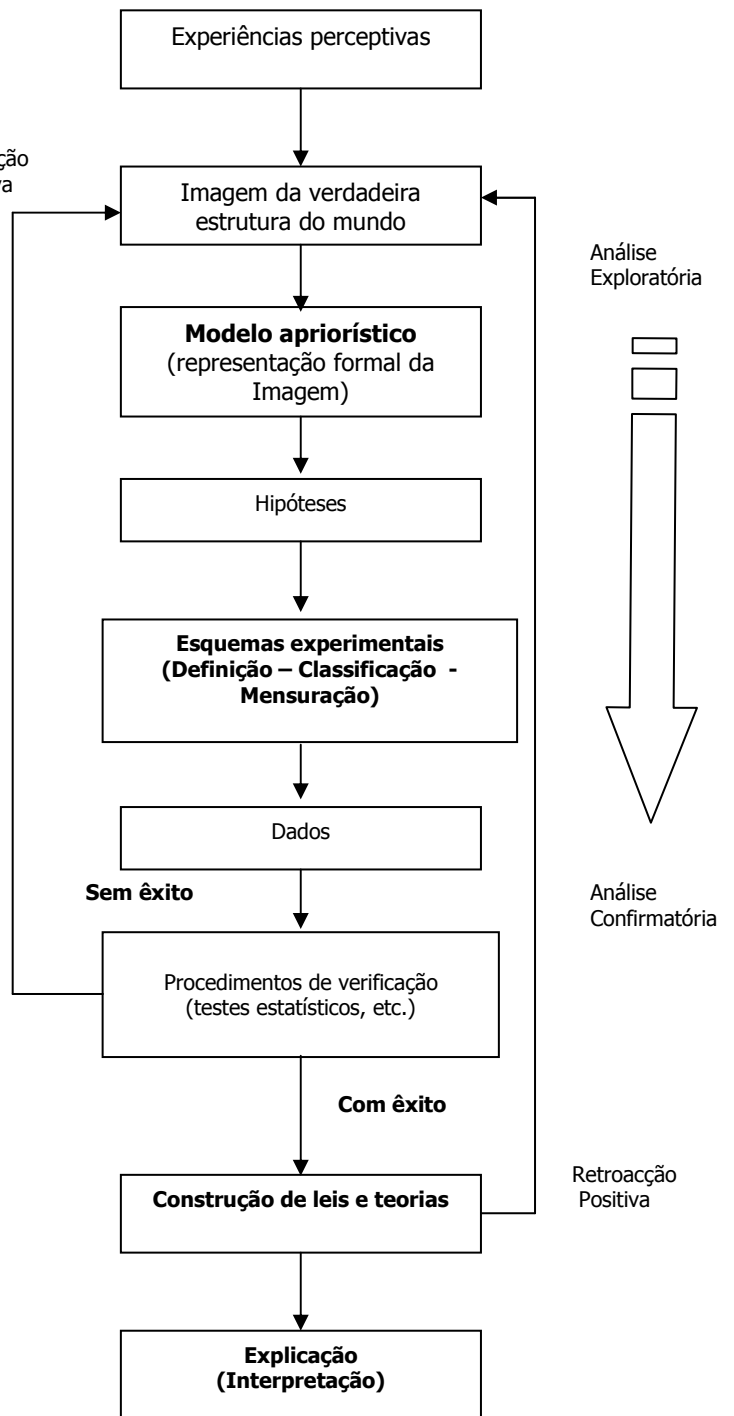
a) método indutivo**b) método dedutivo**

Figura 1.1 – Métodos de Investigação Científica (adaptado)

Fonte: Gerardi E. H. de Oliveira; Silva, B.-C. Nentwig (1981:6), baseado em Harvey, D. (1969).

da abordagem indutiva, como na definição dos esquemas experimentais, na recolha de dados e nos procedimentos de verificação próprios dos métodos dedutivos.

1.2.2. O âmbito da análise

Outra questão que se coloca ao investigador é a da escolha do âmbito de investigação.

	Abordagem Extensiva	Abordagem Intensiva
Objectivo	Descrição das regularidades dos dados. Procura de padrões e leis.	Descrição detalhada de situações específicas.
Âmbito	Grandes conjuntos, de modo a assegurar a "representatividade"	Um ou poucos "case-studies"
Questões de partida	Que representatividade tem a característica, atributo ou padrão da população ?	Quem? O quê? Por quê? No ou nos poucos casos estudados.
Métodos típicos de investigação.	Questionários. Grandes amostras. Análise estatística.	Observação e Participação. Casos de estudo. Análise qualitativa
Limitações	A explicação é uma generalização, muitas vezes difícil de relacionar com os casos concretos ou individuais. A generalização é possível apenas para a população em causa.	Não são "representativas". As relações descobertas não são passíveis de generalização.
Tipos de Explicação	Há uma estrutura latente que provoca as regularidades observadas. As generalizações são o produto de observações repetidas ou de grandes amostras.	Há estruturas latentes que provocam os acontecimentos observados. As causas surgem como consequências de exames e interpretações aprofundados dos casos individuais.
Tipos de dados necessários	Grandes conjuntos de dados já publicados ou gerados a partir de outras fontes. Dados resultantes de inquéritos a grandes populações	Muita informação sobre poucos casos.
Âmbito geográfico	Internacional, Nacional, ou regional: conjuntos de países, regiões ou sub-regiões administrativas. Grandes conjuntos de instituições similares, tais como empresas, ONGs, etc.	Local ou quando muito sub-regional. Indivíduos, ou pequenos organismos, instituições, povoações e espaços.

Baseado em Clifford & Valentine, 2003.

Quadro 1.1 - Abordagens extensivas *versus* abordagens intensivas

2. – As Etapas do Trabalho Científico

A disciplina de Análise de Dados II assume-se como a última disciplina obrigatória do curso de geografia, em que se fornecem os conhecimentos teóricos e práticos que permitem um tratamento suficientemente competente do ponto de vista técnico e interpretativo da informação necessária à generalidade dos trabalhos geográficos.

Revela-se assim de especial importância um sólido conhecimento da metodologia geral de execução de trabalhos científicos, cujas principais etapas se relembram neste momento do programa e se exigem em todo o decorrer da disciplina.

2.1.- A Questão de partida

2.1.1 – Selecção do Tema

A escolha do tema a tratar deve ser feita criteriosamente de acordo com as competências e capacidades do investigador ou com competências que este possa mobilizar facilmente, com os seus próprios interesses e com os das instituições a que pertence.

2.1.2.- Delimitação do âmbito da investigação

Logo de início deve ser tomado em consideração o âmbito temático, espacial e temporal da investigação e deve ser verificado com rigor se é possível formular a questão de partida para o tema e para os âmbitos escolhidos.

2.1.3.- Formulação da questão de partida

É nesse contexto que se tem de formular a questão de partida, a hipótese sobre a qual se irá estruturar a investigação. Esta hipótese é uma pergunta, uma interrogação, a que a investigação vai procurar dar resposta. Por vezes esta hipótese pode também surgir sob a forma de tese, ou seja, como uma afirmação, que se vai procurar provar.

É sobre a pergunta de partida que se articula toda a investigação. Por isso importa avaliar aprofundadamente as suas qualidades de clareza, de exequibilidade e de pertinência.

2.1.3.1.- Clareza

O investigador tem de ter sempre presente que a questão de partida deve ser formulada com a maior clareza possível. É indispensável que corresponda, de forma adequada, a uma perfeita compreensão do tema e do âmbito e dos diversos conceitos teóricos e práticos envolvidos na investigação.

2.1.3.2.- Exequibilidade

Outro aspecto fundamental que deve ser analisado logo nas fases iniciais da investigação é a avaliação crítica das dificuldades que podem vir a acontecer no decorrer da investigação.

Tem de se conhecer, desde o início, as insuficiências teóricas, os problemas de dados e as dificuldades de contacto com os actores e agentes envolvidos e os modos e possibilidades de as ultrapassar.

2.1.3.3.- Pertinência

O interesse pessoal, académico, social e, eventualmente, até o económico, da investigação pode ser um factor fundamental na definição exacta da pergunta a que a investigação vai procurar dar resposta.

2.2.- Preparação da Investigação

2.2.1 – As leituras

A leitura de obras relevantes é a base de todo o trabalho científico. Com efeito, é através das leituras que é possível conhecer o que os diversos especialistas têm vindo a desenvolver na área de investigação em causa, que diferentes perspectivas e escolas de pensamento existem e quais são as teorias explicativas e os modelos fundamentais que sobre o tema têm sido propostos.

A leitura obriga ainda ao confronto crítico das opiniões próprias com as teorias dominantes e à contextualização do trabalho que se vai desenvolver e permite conhecer e avaliar as metodologias que foram seguidas em situações similares.

2.2.1.1.- Seleccionar

A selecção dos textos é uma tarefa de grande importância em qualquer investigação. É a capacidade de encontrar a informação bibliográfica fundamental mais adequada que distingue os investigadores competentes.

É necessário ter técnica e experiência na organização da pesquisa bibliográfica. Conhecer as obras, os autores e as editoras fundamentais, assim como quem são aqueles que dão voz a tendências minoritárias, é uma competência absolutamente necessária do investigador.

Tradicionalmente as leituras a fazer tinham por base documentos em suporte papel – livros e revistas – e era necessário um conhecimento aprofundado das melhores bibliotecas e da organização dos respectivos fundos. Actualmente, a mesma capacidade e o mesmo nível de conhecimento também são exigidos relativamente à documentação disponibilizada na InterNet, cujo acesso é muito mais fácil, mas cuja validação é simultaneamente muito mais difícil.

É necessário que os alunos conheçam os manuais escolares da disciplina, obras em que a generalidade dos temas são apresentados de

forma simples e didáctica, mas também as obras de referência em que foram apresentadas pela primeira vez as ideias mais importantes que constituem hoje o corpo teórico e metodológico dos diferentes temas. E também, para cada tema, saber em que revistas e em que editoras são apresentadas as novidades e os temas polémicos, campos de debate hoje de muitas das fecundas ideias do futuro.

2.2.1.2.- Ler, resumir e comparar

Mas se é necessário ler com método a bibliografia básica e fundamental, tal facto por si só não é suficiente.

É necessário resumir os textos, de forma alargada e compreensiva, isto é, é preciso que o aluno saiba extrair da leitura a informação relevante, devidamente organizada e hierarquizada. E é também necessário que saiba comparar os textos entre si, que tome consciência das contradições, e as saiba, eventualmente, resolver.

2.2.2 – As entrevistas exploratórias

A leitura de obras relevantes, que é a base de todo o trabalho científico como se disse, não esgota toda a preparação que se deve fazer num trabalho de investigação. Deve procurar-se outros apoios teóricos e tem de se confrontar as descrições, as teorias e os modelos dos textos lidos, com a realidade concreta da situação em que se vai desenrolar a pesquisa.

2.2.2.1.- Identificar peritos, especialistas e testemunhas

A melhor maneira de o fazer é falando, discutindo com os peritos e especialistas do tema em análise.

Ao nível académico deve procurar-se os diversos docentes, coordenadores, investigadores e orientadores que habitualmente tratam a matéria em questão e em ambiente real, os indivíduos que conhecem melhor a realidade local, os actores, agentes e personalidades melhor informados sobre os processos em análise, que podem complementar aspectos teóricos relevantes no caso concreto que se vai estudar.

Não devem ser esquecidas as testemunhas, isto é, os indivíduos implicados nos processos que se quer analisar, que mesmo tendo poucos conhecimentos teóricos, são profundos conhecedores da realidade local.

2.2.2.2.- Entrevistar: ouvir e descodificar os conteúdos

Todas estas entrevistas devem ser encaradas como entrevistas semi-estruturadas, isto é, devem ser preparadas, fazendo uma lista de perguntas e de temas a discutir, mas não se deve interromper o fluxo de informação nova que o entrevistado venha a fornecer, até por que é esse o objectivo da entrevista.

Neste tipo de entrevistas deve-se ser ouvinte e procurar identificar os temas e conteúdos principais da conversa que se vai fazendo com o entrevistador.

2.3.- Definir o Problema

2.3.1 – Conciliar leituras, entrevistas e coordenações

Efectuadas as leituras e as entrevistas exploratórias o aluno, sempre com o apoio do coordenador, tem de resumir e conciliar a informação que recebeu. É uma fase do trabalho difícil, que exige opções, que se inicia pelo estabelecimento de um quadro teórico de referência e termina na elaboração de um pre-índice detalhado do trabalho a realizar.

2.3.2.- Estabelecer um quadro teórico

O estabelecimento do quadro teórico de referência consiste em organizar e definir os conceitos fundamentais que vão ser utilizados no trabalho e escolher as perspectivas mais relevantes com que irão ser tratados.

2.3.3.- Explicitar a problemática

Tem a ver com a definição clara dos objectivos, do quadro teórico de referência, das perspectivas e dos níveis de análise e do nível da investigação (descrição, variação simultânea ou mecanismos de causalidade).

2.4.- A Construção do Esquema

A construção do esquema básico do trabalho tem de tomar em conta o nível da investigação e o tipo de abordagem, mais ou menos extensivo.

2.4.1 – Identificar e precisar os conceitos

Efectuadas as leituras e as entrevistas exploratórias o aluno, sempre com o apoio do orientador, tem de explicitar todos os conceitos teóricos com que vai observar a realidade, precisando especialmente os indicadores e as variáveis que vai utilizar nas diversas dimensões da investigação.

2.4.2.- Construir as hipóteses e os modelos

A partir dos conceitos, estabelecem-se, para todos os conceitos relevantes, as relações fundamentais que se consideram apropriadas para a compreensão do tema em estudo, formulam-se as hipóteses estruturantes e constroem-se os modelos que descrevem os processos analisados e as relações entre as hipóteses.

2.5.- A Observação

2.5.1 – Delimitar o campo de observação

Definido o quadro teórico enquadratório e o esquema de trabalho que deve ser seguido, a investigação em Geografia procura geralmente o confronto com a realidade.

O trabalho de campo obriga à definição da área, habitualmente restrita, em que se vão observar os fenómenos. Pode ser uma área geográfica contígua, uma região administrativamente estabelecida ou um território qualquer definido *ad hoc*, ou um conjunto de territórios não contíguos organizados em rede, como a rede urbana, ou ainda um conjunto de entidades económicas, políticas e sociais (empresas, associações, etc.), associadas à sua respectiva localização no território.

Em vez de uma área de observação ou de uma rede, pode optar-se pela análise de casos de estudo, que podem ser locais, organizações ou instituições que se observam intensiva e profundamente.

2.5.2.- Determinar e testar os métodos de observação

A observação tanto pode ser extensiva como intensiva. A opção por um destes métodos tem de ser justificada, pelo objecto do estudo, pelos seus objectivos, pela capacidade do investigador e pelos meios de que dispõe.

Importa também escolher o tipo de unidades de análise em que se vai fazer a observação e proceder à recolha de informação (a escala de análise). Podem ser utilizadas unidades territoriais de menor dimensão, instituições, organizações, entidades ou indivíduos, isoladamente ou associados conjuntamente. O que importa é que se tenham identificado e escolhido os actores e as entidades significativas.

Nos processos de recolha de informação em grandes conjuntos, ainda se coloca a questão de se observarem todos, alguns ou poucos. E nestes dois últimos casos, que processos de selecção se deverão escolher: amostragem ou selecção dirigida. As amostragens obrigam ainda a definir quantos e quais serão os observados.

Sempre que seja possível, é muito conveniente testar os métodos de recolha, fazendo pré-inquéritos, entrevistas piloto para avaliação dos esquemas de entrevistas estruturados que se prepararam, e entrevistas com potenciais membros de grupos de discussão e de entidades a ser alvo de case-studies ou de observação participante, a fim de determinar a validade da observação futura que se pretende desenvolver.

2.5.3.- Recolher as informações

Na recolha deve ter-se em conta as variáveis sobre as quais se quer ter informação, os tipos de escalas de medida e/ou, as nomenclaturas que se vai utilizar.

Os procedimentos de recolha devem ser os mais correctos e adequados à informação que se pretende, à forma da observação que se utiliza e aos objectivos finais do estudo (veja-se o anexo sobre a recolha).

2.6.- A Análise das Informações

2.6.1 – Arrumar os dados e prepará-los para análise

Os dados recolhidos devem ser guardados em matrizes, folhas de cálculo ou bases de dados, de uma forma que permita o fácil acesso e tratamento.

Assume especial importância e deve ser dada grande atenção à codificação e classificação das variáveis de tipo nominal, ao eventual desdobramento de alguns ou todos dos seus estados em variáveis dicotómicas ou à sua combinação em tabelas de Burt. Para as variáveis de tipo ordinal e de intervalo e razão, é necessário avaliar a conveniência em classificá-las e definir para que limiares devem ser padronizadas ou transformadas por qualquer função de valor.

Todos estes possíveis procedimentos das variáveis iniciais, devem ter sempre em conta o tipo de tratamentos que lhes vão ser aplicados. E é com o mesmo objectivo que se coloca a questão de construir índices resumo e indicadores por via da combinação de variáveis. Os factores a ter em conta nestas transformações são o aumento da legibilidade e a correspondência com aspectos estruturantes da realidade.

2.6.2 – Efectuar os procedimentos analíticos

A informação recolhida pode ser sujeita a dois tipos de procedimentos analíticos: a modelização e a análise de dados, exploratória ou de confirmação.

Nas situações em que a realidade a estudar não é suficientemente conhecida nos seus aspectos teóricos ou nas condições contextuais específicas, utiliza-se a análise exploratória de dados. Por outro lado, quando a realidade é mais conhecida pode usar-se a modelização ou a análise confirmatória de dados, após verificar a sua validade.

Os métodos de análise de dados mais utilizados em Geografia são apresentados nos capítulos seguintes. E em relação às técnicas de modelização, é feita uma introdução exemplificativa no último capítulo deste programa.

2.6.3 – Comparar os resultados observados com os esperados

A análise dos dados recolhidos produz resultados que se devem comparar com os resultados que se esperavam. Do confronto entre os dois pode vir a corroboração das hipóteses e teses iniciais, ou muito pelo contrário, pode surgir a necessidade de proceder à sua revisão.

Esta fase do trabalho de investigação é crucial e constitui uma das mais ricas fontes de nova informação, que é necessário validar, esclarecer, completar e integrar com e nas teorias existentes.

2.7.- A Conclusão

Na conclusão resume-se o tema a tratar e os objectivos, as hipóteses e teses de que se partiu, recapitula-se os procedimentos e apresentam-se os resultados obtidos, tendo especialmente em atenção os novos conhecimentos – o que a investigação trouxe de novo – e as suas consequências – o modo ou modos como esse novo conhecimento pode ser integrado e utilizado no contexto espacial da investigação efectuada e eventualmente a maneira como poderá ser transposto para outras situações.

3. – A Natureza da Informação Geográfica

Todo o trabalho de investigação em Geografia se baseia na *informação*. É com a informação que se fazem os mapas e os gráficos, é com informação que se identificam os padrões e as regularidades, e se fazem as análises comparativas, e é com informação que se estabelecem as relações e as inferências entre os tipos de actividade e de modo de vida das populações e os diversos elementos das paisagens.

A informação umas vezes é recolhida e publicada por institutos especializados (informação secundária) e outras vezes é recolhida pelos próprios investigadores (informação primária).

Em Geografia, como no conjunto das Ciências Sociais e Naturais, são muitas as maneiras de obter a informação necessária. Porém é preciso não esquecer que existe um balanço crítico entre a adequação da informação, que queremos sempre a melhor possível, e o seu respectivo custo, tanto material como em trabalho, que pode atingir valores muito elevados. É preciso ter a capacidade de optar pelos balanços adequação/custo favoráveis, mas que permitam uma boa compreensão do assunto em análise.

3.1.- A Informação Geográfica

No contexto das ciências, tanto físico-naturais como sociais, a perspectiva geográfica introduz uma valorização da dimensão espacial e, conjuntamente com a História, da dimensão temporal. Pode assim dizer-se que um facto ou acontecimento geográfico tem sempre associado uma localização e um tempo. Embora em muitas análises estatísticas esse facto não seja completamente integrado, os estudos de Geoestatísticas e o recente desenvolvimento dos SIGs, incorporam a localização de forma explícita da análise, e permitem trabalhar noções de pertença, vizinhança, contiguidade, entre outras.

3.1.1 – A Matriz Geográfica

Estebanez e Bradshaw (1978:6), seguindo Berry e Haggett, referem que um facto ou acontecimento geográfico é o resultado da observação de uma característica ou variável, numa unidade territorial de análise, ou de uma forma mais abrangente, a geografia pode ser encarada como uma enorme matriz em que as linhas são todos os lugares ou localizações possíveis e as colunas todas as características desses lugares ou localizações.

Teoricamente esta matriz geográfica (Fig. 1.2) descreve todo o conhecimento geográfico de todas as regiões do mundo.

	Características/Variáveis				
Unidades de análise	Coluna j				
Linha i					
			a_{ij}		

Figura 1.2 – A Matriz Geográfica de Berry

A observação da figura 1.2 permite compreender as diversas possibilidades da análise geográfica:

- A análise de uma célula (**a_{ij}**) é a descrição do valor concreto da variável **j** na unidade de análise **i**.
- A análise da linha **i** dá-nos as características de uma única região ou lugar **i**. É a perspectiva de monografia regional da Geografia clássica.
- a análise por colunas evidencia a variação espacial de uma única característica. É a perspectiva própria da Geografia sectorial.

O estudo da matriz pode ser alargado a conjuntos de linhas ou de colunas, constituindo os estudos multidimensionais.

3.1.2 – O Cubo de Dados de Cattell

Cattell foi um investigador que propôs um cubo de dados, em que o valor geral é o resultante do cruzamento das colunas (variáveis) e das linhas (unidades de análise) da matriz frontal do cubo, de algum modo similar à matriz geográfica com a dimensão tempo, que é representada pela profundidade do cubo.

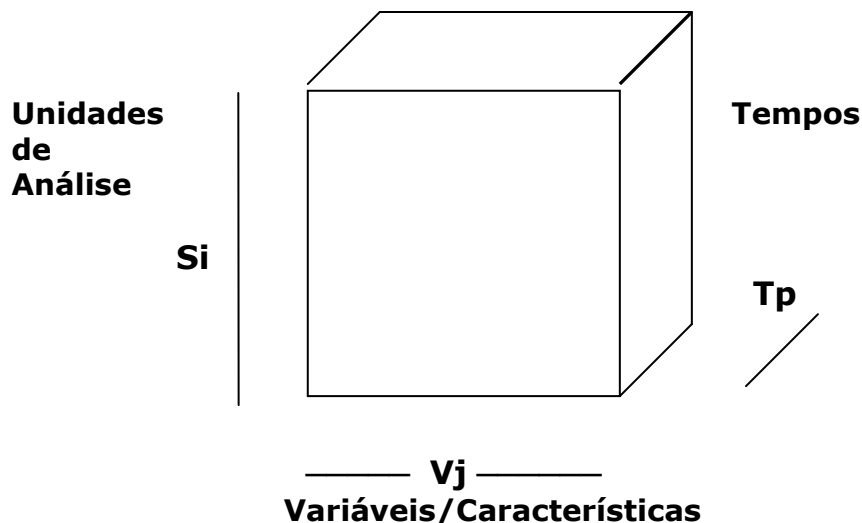


Figura 1.3 – O Cubo de Dados de Cattell

O termo geral tem pois a forma **a_{ijp}** que representa o valor que a variável **j** assume na unidade de análise **i** no tempo **p** .

Cattell propôs ainda que a análise deste cubo de dados fosse feita segundo seis (6) matrizes, correspondentes a seis modos, os modos R, Q, P, O, S e T, que representam todas as combinações possíveis de estabelecer entre variáveis, unidades de análise e tempo.

O modo R e o Q representam matrizes de atributos, em que se encontram os valores que as variáveis assumem nas unidades de análise para um dado tempo. No caso dos estudos geográficos são muitas vezes unidades territoriais. O modo R é similar à matriz geográfica de Berry e é a forma como habitualmente se representam as matrizes de atributos em ciências sociais. O modo Q é habitualmente usado pelos investigadores e pelas metodologias da biologia e das ciências da vida.

Os modos P e O são a representação simultânea de várias séries temporais, uma para cada variável. O interessante é ver se a evolução é simultânea ou não, ou se existe uma separação temporal constante entre a evolução dos valores das diversas séries.

Os modos S e T representam a evolução que uma única variável apresenta no tempo em distintas unidades de análise. Estes quatro últimos modos são típicos das análises multi-espectrais.

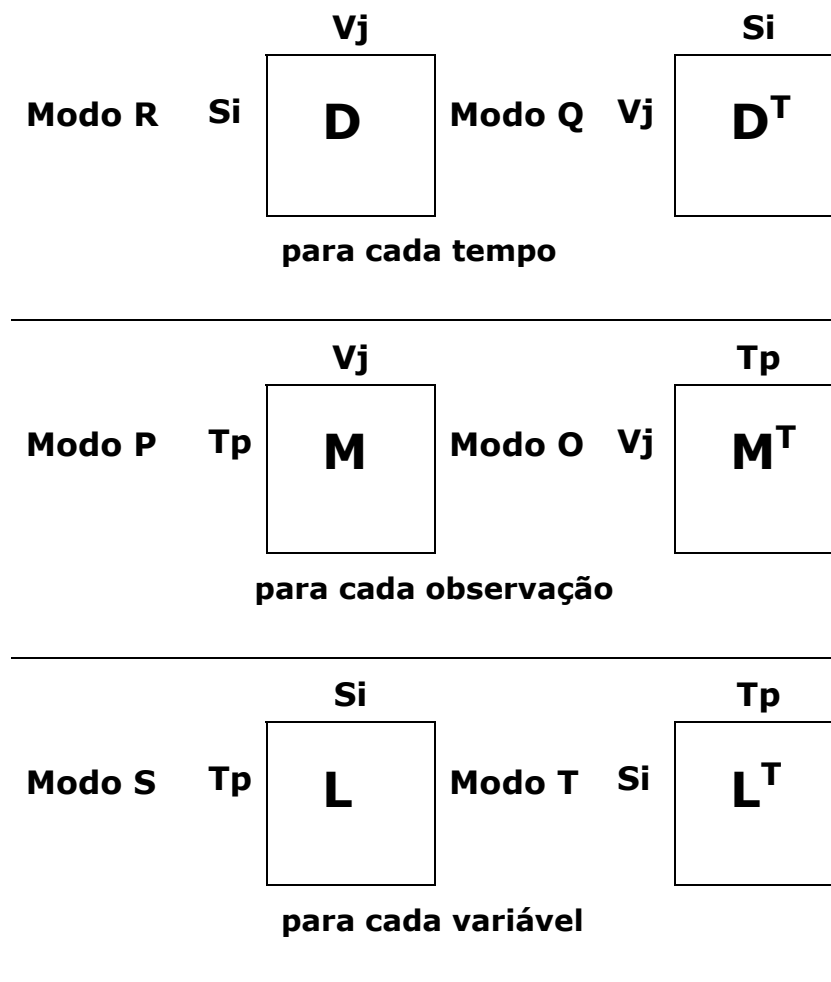


Figura 1.4 – Os modos de Cattell

3.1.3 – Breve tipologia de matrizes e tratamentos

De um modo geral pode dizer-se que as matrizes de atributos são as matrizes básicas em que se recolhe informação de tipo geográfico. Existem dois outros tipos de arrumação da informação que podem ter interesse nos trabalhos de Geografia: as matrizes de distância, muito difundidas, e as de dominância. Importa ainda referir um caso especial, que são as matrizes de medição conjunta, que parecem uma matriz de atributos, mas na realidade são representações de quadros de frequência

ou tabelas de contingência em que se encontram as frequências conjuntas de diversos estados de duas variáveis. No Quadro 1.2 são apresentadas estas matrizes, o seu respectivo significado, o modelo matemático a que correspondem e alguns exemplos de tratamentos que lhes podem ser aplicados.

NOME	TIPO	SIGNIFICADO	MODELO	TRATAMENTOS
Matriz de Atributos	n unidades de análise x m variáveis	O valor da variável j na unidade de análise i	As n unidades de análise são pontos num espaço com m dimensões	- Análise factorial - Taxionomias - Regressões - Análise discriminante
Matriz de Proximidade	n unidades de análise x n unidades de análise	Valor da afinidade, distância ou proximidade entre a unidade de análise i e unidade de análise j	A afinidade está relacionada monotonicamente com as distâncias entre os n pontos num espaço multidimensional subjacente.	- Análise em coordenadas principais de Gower - Escalonamento multidimensional (MDS) métrico e não métrico.
Matriz de Dominância	n unidades de análise x n unidades de análise	Valor do grau em que a unidade de análise i é escolhida, preferida ou domina a unidade de análise j	Cada uma das n unidades de análise é representada por um ponto numa escala pluri ou monodimensional de tal modo que, quanto maior é a dominância, maior será a diferença em pontuação nessas escalas.	- Análise Unidimensional de dominância (Thurstone) - Análise Multidimensional de dominância (Slater)
Matrizes de Medição Conjunta (quadros de frequência)	n estados da variável 1 x m estados da variável 2	Número de vezes (ou de unidades de análise) em que ocorre, ao mesmo tempo, o estado i da variável 1 e o estado j da variável 2	Cada valor é uma combinação simples dos valores em duas escalas unidimensionais, correspondentes a cada uma das variáveis.	- Análise de tabelas de contingência - Análise factorial de correspondências múltiplas AVP ou AF em matrizes de variância/covariância

Quadro 1.2 – Tipologia de matrizes de dados geográficos (D. Abreu)

Anexo 2

Alguma Matemática para Geógrafos

Toda a investigação científica utiliza a análise de dados para descrever e compreender as realidades que estuda. O processo de compreensão, baseado na exploração e confirmação das conjecturas, baseia-se, para além da busca de padrões na distribuição das variáveis e da existência de correlações entre elas, na classificação e na comparação dos valores das variáveis.

Em Geografia, quase toda a investigação exige recolha rigorosa de informação. Como o âmbito da investigação geográfica é muito diversificado, a informação recolhida é de todos os tipos, como por exemplo a medição de temperaturas, da pluviosidade e da velocidade do vento, dos tipos de solo e de sedimentos, dos quantitativos populacionais e de rendimento, do número e tipo de empresas, de sondagens e resultados eleitorais.

1. – Matemática Básica

Neste capítulo devem rever-se os conceitos fundamentais da matemática que estão insuficientemente apreendidos e ensinar alguns temas novos.

Em seguida, apresentam-se alguns conceitos que importa rever.

1.1. – Potências

Ao produto de um número **b** multiplicado por si mesmo **n** vezes dá-se o nome de **n**-ésima potência do número **b**.

As operações com potências gozam das seguintes propriedades:

1. Multiplicação de potências da mesma base: $a^n \times a^m = a^{n+m}$
2. Multiplicação de potências com o mesmo expoente: $a^n \times b^n = (a \times b)^n$
3. Divisão de potências da mesma base: $a^n : a^m = a^{n-m}$
4. Divisão de potências com o mesmo expoente: $a^n : b^n = (a : b)^n$
5. Potência de uma potência: $(a^n)^p = a^{n \times p}$
6. Potência de expoente nulo: $a^0 = 1$ ($a \neq 0$)
7. Potência de expoente inteiro negativo: $a^{-n} = \frac{1}{a^n}$

1.2. – Logaritmos

O logaritmo de um número positivo expressa-se como uma potência de uma base fixa, isto é, o expoente y tal que o valor de x seja b^y , segundo:

Logaritmo de um valor x ($x > 0$) na base b ($b > 0, b \neq 1$) é o valor y

$$Y = \log_b x, \text{ tal que } x = b^y$$

São duas as bases utilizadas: para os logaritmos neperianos (**ln**) é o número de Neper (**e**), um número real com o valor aproximado de 2,71828... e para os decimais (**log**) o número 10, mais adequado a cálculos no sistema decimal.

1. $\log 1 = 0$	$\log_{10} 1 = 0$
2. $\log 10 = 1$	$\log_{10} 10 = 1$
3. $\log 100 = 2$	$\log_{10} 100 = 2$

....

Os logaritmos têm a propriedade de baixar a dificuldade das operações algébricas, para o que basta passar a execução das operações dos números para os seus logaritmos.

- Para $x * y$ basta saber que $\log (x * y) = \log x + \log y$
- Para $x : y$ basta saber que $\log (x/y) = \log x - \log y$
- Para b^y basta saber que $\log (b^y) = y * \log b$

- Para $\sqrt[y]{b}$ basta saber que $\log (\sqrt[y]{b}) = \frac{\log b}{y}$

- Para $\frac{1}{x}$ basta saber que $\log (\frac{1}{x}) = - \log x$

1.3. – Funções

Uma função é uma relação entre duas variáveis **x** e **y** tal que a cada valor de X corresponde apenas um valor de Y, embora a cada valor de Y possam corresponder vários Xs.

Em Geografia, são especialmente interessantes as funções polinomiais (lineares, quadráticas, cúbicas, ...), exponenciais e logarítmicas.

As funções polinomiais lineares são linhas rectas, ou seja, relações proporcionais, as quadráticas têm uma “bossa” ou uma “depressão”, as cúbicas uma bossa e uma depressão, as de grau quatro ou uma bossa e

duas depressões ou duas bossas e uma depressão e assim sucessivamente.

1.4. – Matrizes

Matrizes são conjuntos de números dispostos em linhas e em colunas. Já se referiram vários tipos de matrizes geográficas, de dados, etc.. As matrizes têm o seguinte aspecto:

$$A = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{vmatrix}$$

Os índices indicam a posição dos elementos nas linhas e colunas da matriz. O primeiro respeita à linha e o segundo à coluna.

Os alunos devem conhecer o que são:

- um vector (linha e coluna);
- uma matriz quadrada
- uma matriz diagonal
- uma matriz escalar
- uma matriz identidade
- uma matriz nula

e saber fazer as seguintes operações:

- transpor matrizes
- somar matrizes
- subtrair matrizes
- multiplicar uma matriz por um escalar
- multiplicar matrizes
- inverter matrizes

2. – Índices

Segundo Thiessen (2000:13), a informação de que se dispõe em Geografia é constituída por grandezas ou dados primários, a informação que se recolhe especificamente de cada vez que dela se necessita e as grandezas ou dados secundários, a informação que já foi recolhida e eventualmente publicada e para aceder à qual apenas é necessário consultar as respectivas fontes.

Por outro lado importa também distinguir entre variáveis de estado e de fluxo. As primeiras caracterizam uma situação e respeitam a um

ponto específico no tempo, enquanto as segundas se referem ao que aconteceu à variável de estado no decorrer de um dado período de tempo.

Na análise comparativa, os quocientes, percentagens e índices permitem, juntamente com os parâmetros estatísticos mais usuais, expressar, resumir e classificar os dados.

Por exemplo, pode-se enunciar a número de idosos de uma dada região, mas é mais expressivo dizer qual a sua importância na população total. Ou, em vez de dizer que a população cresceu tantos milhares de pessoas, enunciar a sua respectiva taxa de variação.

Os índices facilitam a classificação, pois permitem organizar de forma mais sintética os dados, facilitam a comparação entre diversas regiões e distintos momentos, e resumem de forma simples informação de base que era originalmente mais complexa.

2.1.- Quocientes

Quocientes são grandezas ou variáveis secundárias que resultam da divisão de dois valores ou da divisão de duas séries de valores de duas variáveis, uma pela outra.

O objectivo da determinação do quociente é obter um valor para o tipo de relação que as variáveis estabelecem entre si. Por exemplo, dividindo a população de um território (nº de habitantes) pela respectiva área (em Km²), obtém-se o valor da densidade de população, medido em habitantes por Km². Do mesmo modo, quando se divide o PIB (Produto Interno Bruto) de uma região num dado ano, pela sua população, no mesmo ano, determina-se o valor do PIB/capita desse ano.

Como se vê, é fundamental considerar as unidades de medida das variáveis primárias, para determinar correctamente a unidade de medida em que o quociente se expressa.

Por exemplo, a velocidade, que pode ser expressa em quilómetros por hora, ou a densidade, em hab por Km².

Quando o valor do quociente resulta muito pequeno, para não utilizar muitas casas decimais, multiplica-se por 100 ou 1000 ou até mesmo por 10000. Nesse caso trata-se de valores que se denominam por **taxas**, como são, em demografia, as taxas de natalidade e mortalidade (x1000) ou os índices de envelhecimento e de dependência (X 100).

Segundo Estabanez & Bradshaw (1978:96), a construção de um índice na forma de quociente ou de taxa:

- deverá ser simples e de fácil percepção;
- deverá variar, tanto quanto possível, entre dois valores fixos. Os melhores serão entre 0 e 1 ou 0 e 100;
- a variação do valor do índice deve ser proporcional à variação das características que ele pretende medir;

- deverá ter elementos de referência, que permitam analisar os resultados obtidos. Por exemplo, o valor 0 poderá indicar a ausência total, enquanto 100 indicará a presença completa;
- deverá ser aplicável a uma gama variada e ampla de situações.

O quociente também pode ser calculado entre as partes de uma variável e o seu total. Neste caso, o quociente representa o valor relativo de cada parte, não tem unidades (pois é a divisão de duas grandezas medidas nas mesmas unidades) e geralmente é expresso em **percentagem**.

2.2- Taxas seccionais

Para Chaves *et al* (2000:4), algumas taxas podem ter como base um determinado período, que é considerado como referência. Diz-se então que os dados são seccionais quando os valores se referem a um certo período de tempo (um ano ou outro período específico) e são dados temporais quando implicam uma medição ao longo do tempo.

A construção de taxas fez-se pela selecção de um elemento de referência, a partir do qual se exprimem, em termos relativos, as outras informações (Chaves *et al.*, 2000:48). A escolha do elemento de referência deve ter em conta a finalidade da investigação, reflectir a natureza dos dados e facilitar a comparação.

2.3- Taxas de variação

Quando o conceito de taxa é alargado à evolução de uma variável ao longo do tempo, trata-se de uma taxa de variação.

Quando se dispõe de uma sucessão de valores anuais deve distinguir-se entre taxas anuais de variação (1) e as taxas de variação média anual (2) e de variação efectiva anual (3) para o conjunto do período.

$$1) \quad V_{x_{t+n}, t} = \frac{X_{t+n} - X_t}{X_t}$$

$$2) \quad V_{m_{t+n}, t} = \frac{V_{t+n, t}}{n}$$

$$3) \quad r_{t+n, t} = \left(\frac{x_{t+n}}{x_t} \right)^{1/n} - 1$$

2.4- Exercícios

1).- Considere os seguintes valores para doze países da OCDE em 1990:

Países	População (milhões)	PIB ppp (milhões \$US)
Austrália	16,9	271 262
Bélgica	9,8	160 534
Canadá	26,5	509 648
França	56,1	976 421
Alemanha	77,6	1 413 329
Itália	57,1	907 319
Japão	123,5	2 175 576
Holanda	15,0	235 425
Suécia	8,4	142 918
Suiça	6,6	137 768
Reino Unido	57,2	903 989
Estados Unidos	249,2	5 345 091

Fonte: PNUD, (1992,1993)

- Calcule os respectivos PIB/cap. Interprete os resultados.

2) Dada a seguinte evolução do PIB em Portugal de 1995-2000

Ano	1995	1996	1997	1998	1999	2000
PIB _{pm} (milhares de contos) Preços 1977	1.453.563	1.494.262	1.546.562	1.624.171	1.672.896	1.723.083

Fonte: MATEUS, A. (2001) – Economia Portuguesa. Verbo Editora.

- Calcule as taxas de variação anuais, a taxa de variação média anual e a taxa de variação efectiva anual para o período;

- Calcule os índices anuais do PIB, assumindo o ano de 1995 como base 100.

2.5 - Bibliografia

CHAVES, C. *et alia* (2000) – **Instrumentos Estatísticos de Apoio à Economia**, MacGraw-Hill.

(*) ESTEBANEZ ALVAREZ, J.; BRADSHAW, R.. P. (1978) – **Técnicas de Quantificación en Geografía**, Madrid, Editorial Tebar Flores.

GERARDI, Lúcia H. O.; SILVA, Barbara-Christine N. (1981) – **Quantificação em Geografia**, São Paulo, Difel, p. 117-124.

3. – Medidas de Concentração

Os índices de concentração são medidas resumo muito utilizadas em Geografia, assim como noutras ciências físicas e sociais.

A noção de concentração deve-se a Corrado Gini, estatístico italiano, que a desenvolveu a propósito da distribuição dos salários e dos rendimentos. Respeitava ao facto de haver, ou não, muitos indivíduos com valores semelhantes de rendimento ou de salário e poucos indivíduos com valores de rendimento ou salário muito diferentes.

A concentração é um conceito aparentemente simples, que se opõe às distribuições igualitárias ou uniformes. Medir a concentração numa distribuição é pois conhecer o afastamento que se verifica entre essa distribuição e duas situações padrão: uma em que todas as unidades de análise têm a mesma parcela do total da variável (i.e. são iguais) e outra em que o total da variável se encontra atribuído a uma única unidade de análise.

Na realidade a concentração é geralmente avaliada por comparação entre a distribuição efectiva que uma variável tem e a distribuição igualitária. É medida através da distância ou diferença que existe entre os correspondentes valores de cada.

Os índices de concentração são por vezes referidos por outras designações. Desigualdade, dissemelhança e segregação são apenas algumas delas.

Deve salientar-se que uma das formas clássicas de mostrar a concentração se deve a *Lorenz*, que em 1905 propôs uma curva para a representar, curva que é denominada desde então por Curva de Lorenz.

3.1.- Curva de Lorenz

A curva de Lorenz é uma forma de representação gráfica que ilustra as desigualdades na distribuição de variáveis sociais, económicas e ambientais. A desigualdade é muitas vezes espacial, quando as unidades de análise são unidades territoriais - administrativas, estatísticas ou outras - mas também pode ser social, quando as unidades de análise são grupos ou classes sociais, ou de outra natureza, consoante as variáveis em análise.

Nos espaços territoriais e sociais são muitas as variáveis que apresentam grandes desigualdades na sua distribuição espacial, como sejam os rendimentos, as produções, os salários, as densidades populacionais, os povoamentos culturais ou arbóreos, as dimensões das empresas, das explorações ou das propriedades agrícolas, dos prédios urbanos, das famílias, etc..

A curva de Lorenz, tanto permite comparar a diversidade regional num dado momento, analisando a curva respectiva, como avaliar a sua evolução no tempo, comparando várias curvas. E o mesmo sucede com os índices de concentração que adiante se estudarão.

A curva de Lorenz representa os valores da distribuição de uma variável. Esta variável é geralmente uma variável secundária, resultante da razão entre outras duas variáveis iniciais. Por exemplo, estuda-se a distribuição do rendimento/capita de um conjunto de regiões, variável que é a razão entre o rendimento e a população de cada região. A densidade populacional é a razão entre a população e a área; e o mesmo acontece com a dimensão média da propriedade, razão da área de uma dada classe de propriedades com o número de propriedades que nela existe.

É a comparação entre as proporções das distribuições das duas variáveis iniciais que estão na origem da variável resultante, que permite construir a curva de Lorenz e calcular os índices de concentração.

Como exemplo, apresenta-se um caso de Portugal, a distribuição do PIB por habitante nos dezoito distritos do continente, em 1970. É uma variável secundária, que resulta da razão entre o PIB de cada distrito (a variável numerador) e a sua respectiva população (a variável denominador), como se pode ver no Quadro 2.3.1

Para fazer a curva de Lorenz é preciso ordenar por ordem crescente a variável em estudo (t_i , - o PIB/cap) e depois calcular as percentagens simples e acumuladas das duas variáveis primárias correspondentes (r_i - o PIB regional e s_i - a população).

Nº.	Nome	PIB Mil contos	População	PIB/cap Contos	Ordem
i		r_i	s_i	t_i	k
1	Aveiro	11 661,9	535 815	21,76	16
2	Beja	2 971,8	199 335	14,91	9
3	Braga	8 549,9	600 660	14,23	7
4	Bragança	2 437,8	175 185	13,92	6
5	Castelo Branco	3 252,2	247 855	13,12	5
6	Coimbra	7 643,0	391 710	19,51	14
7	Évora	2 745,7	172 380	15,93	11
8	Faro	3 431,0	263 145	13,04	4
9	Guarda	2 094,7	205 235	10,21	3
10	Leiria	7 004,3	370 155	18,92	13
11	Lisboa	51 545,4	1 545 860	33,34	17
12	Portalegre	2 234,7	140 640	15,89	10
13	Porto	25 701,0	1 287 625	19,96	15
14	Santarém	7 382,4	416 630	17,72	12
15	Setúbal	15 380,2	460 865	33,37	18
16	V. Castelo	2 092,7	249 075	8,40	1
17	Vila Real	2 400,9	259 395	9,26	2
18	Viseu	5 943,4	400 330	14,85	8
Σ	TOTAL	164 473,0	7 921 895	22,46	-----

Quadro 2.3.1 – Produto Interno Bruto por habitante nos distritos de Portugal continental, em 1970.

Dum modo mais formal, é preciso calcular

r_i variável inicial numerador: que neste exemplo é o PIB;

s_i variável inicial denominador: a População;

t_i variável resultante: o PIB/cap;

p_k percentagem de r_i

q_k percentagem de s_i

e depois de ordenar o quadro de valores por ordem crescente de t_i , calcula-se os valores de y_k e x_k , percentagens cumulativas de p_k e q_k , conforme se mostra no Quadro 2.3.2.

Ordem	Nº.	PIB	População	% PIB	% Pop	% cum PIB	% cum Pop
k	i	r_i	s_i	p_k	q_k	y_k	x_k
1	16	2 092,7	249 075	1,272	3,144	1,272	3,144
2	17	2 400,9	259 395	1,460	3,274	2,732	6,418
3	9	2 094,7	205 235	1,274	2,591	4,006	9,009
4	8	3 431,0	263 145	2,086	3,322	6,092	12,331
5	5	3 252,2	247 855	1,977	3,129	8,089	15,460
6	4	2 437,8	175 185	1,482	2,211	9,551	17,671
7	3	8 549,9	600 660	5,198	7,582	14,749	25,253
8	18	5 943,4	400 330	3,614	5,053	18,363	30,306
9	2	2 971,8	199 335	1,807	2,516	20,170	32,822
10	12	2 234,7	140 640	1,359	1,775	21,529	34,597
11	7	2 745,7	172 380	1,669	2,176	23,198	36,773
12	14	7 382,4	416 630	4,489	5,259	27,687	42,032
13	10	7 004,3	370 155	4,259	4,673	31,946	46,705
14	6	7 643,0	391 710	4,647	4,945	36,593	51,650
15	13	25 701,0	1 287 625	15,626	16,254	52,219	67,904
16	1	11 661,9	535 815	7,090	6,764	59,309	74,668
17	11	51 545,4	1 545 860	31,340	19,514	90,649	94,182
18	15	15 380,2	460 865	9,351	5,818	100,000	100,000
Σ		164 473,0	7 921 895	100,000	100,000	-----	-----

Quadro 2.3.2 – Quadro auxiliar para o cálculo da curva de Lorenz e alguns índices de concentração. PIB/cap nos distritos do continente, 1970.

A curva de Lorenz é a curva que representa os valores de y_k em função de x_k , conforme se pode observar no gráfico da Figura 2.3.1, em que os sucessivos valores de y_k e de x_k são as coordenadas dos diversos pontos da curva.

Como facilmente se vê na fig. 2.3.2, quanto mais uniforme (ou igualitária) a distribuição, mais a curva está junto à diagonal e quanto mais concentrada for, mais se aproxima do eixo dos x 's e do limite direito do quadrado em que se insere.

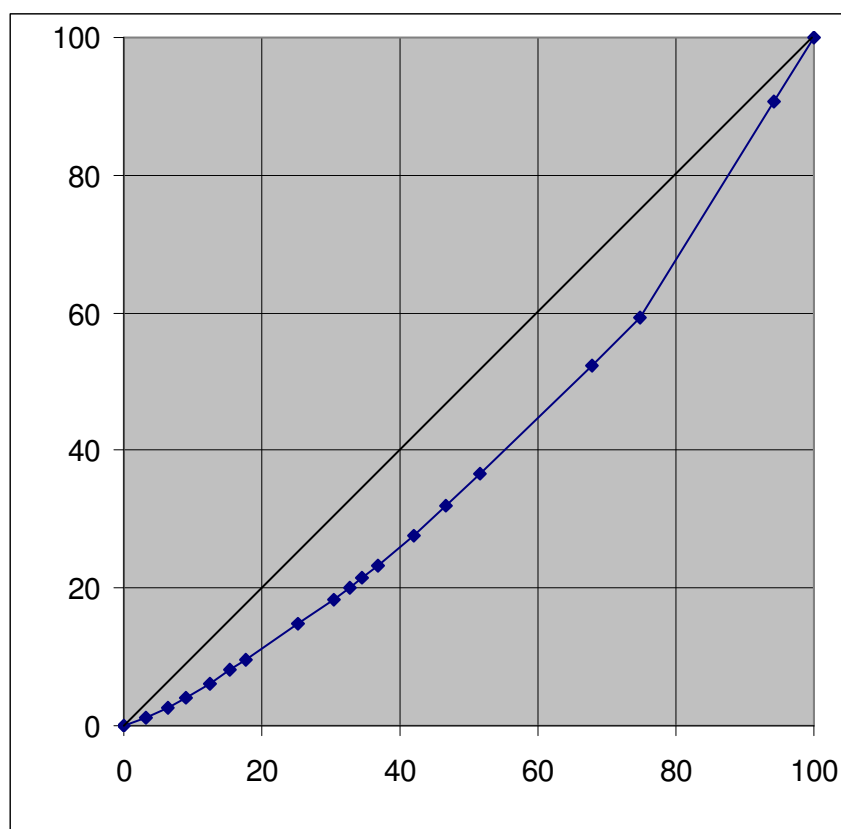


Figura 2.3.1 – Curva de Lorenz da distribuição distrital do PIB/capita, 1970.

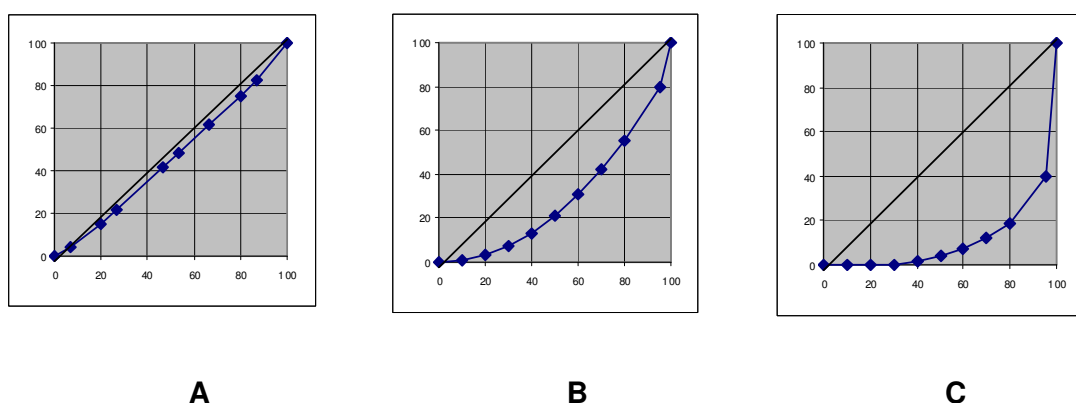


Figura 2.3.2 – Curvas de Lorenz para diferentes graus de concentração.

A distribuição da esquerda (A) tem uma pequena concentração (ou uma grande igualdade), a do meio (B) uma concentração média (ou uma desigualdade média) e a da direita (C) uma grande concentração ou uma grande desigualdade dos valores.

Embora seja habitual construir a curva de Lorenz para uma variável secundária, resultante da razão de outras duas, como até agora se mostrou, também é possível construir curvas de Lorenz e calcular os índices de concentração para uma variável isolada, ou independente, que não seja resultante da razão de outras duas. Pode fazer-se a curva e calcular índices para uma variável isolada – por exemplo, a distribuição dos activos por sectores de actividade industrial – desde que se considere que a cada grupo ou classe existente deve corresponder uma parcela igual. A curva e os índices assim calculados mostram então se existe ou não concentração em alguns sectores, comparando a distribuição existente com a que se teria se os grupos tivessem todos iguais quantitativos.

Seguindo outro exemplo - o número de activos na indústria transformadora de Barcelona em 1973 - e de um modo semelhante ao procedimento anterior, calcula-se a percentagem sobre o total de cada unidade de análise (dados do Quadro 2.3.3).

Nº	NOME	Nº de Activos	"Par- cela"	"Variável resultante"	Nº Ordem
i		r _i	s _i	t _i	k
1	Alimentar	53 155	1	53 155	4
2	Têxteis	181 654	1	181 654	8
3	Couro e Calçado	67 330	1	67 330	6
4	Madeira e Cortiça	30 916	1	30 916	2
5	Papel e Gráficas	54 206	1	54 206	5
6	Química	89 030	1	89 030	7
7	Cerâmica e Vidro	33 709	1	33 709	3
8	Minerais metálicos	10 923	1	10 923	1
9	Máquinas/Prod. Metálicos	266 803	1	266 803	9
	TOTAL	787 726	9	787 226	-----

Quadro 2.3.3 – Activos na Indústria transformadora em Barcelona, 1973.

Após ordenar os valores da variável resultante (que é igual à inicial r_i) calcula-se a percentagem simples p_k e cumulativa y_k. O valor de x_k é também fácil de calcular, pois como é a percentagem acumulada de q_k, o seu valor é igual a $k \times (100 / n)$. O valor de n neste exemplo é 9 (Quadro 2.3.4).

Nº Ordem	% de Activos (ordenada)	% de Activos (acumulada)	% de grupos	% de grupos (acumulada)
k	p_k	y_k	q_k	x_k
1	1,3866	1,3866	11,1111	11,1111
2	3,9247	5,3113	11,1111	22,2222
3	4,2793	9,5906	11,1111	33,3333
4	6,7479	16,3385	11,1111	44,4444
5	6,8813	23,2198	11,1111	55,5556
6	8,5474	31,7672	11,1111	66,6667
7	11,3022	42,0694	11,1111	77,7778
8	23,0606	66,1300	11,1111	88,8889
9	33,8700	100,0000	11,1111	100,0000

Quadro 2.3.4 – Quadro auxiliar para o cálculo da curva de Lorenz e alguns índices de concentração. Activos na Indústria Transformadora de Barcelona em 1973.

A representação no gráfico cartesiano dos pontos correspondentes aos valores de y_n e x_n , e a sua união por segmentos de recta, permite obter a curva de Lorenz da distribuição.

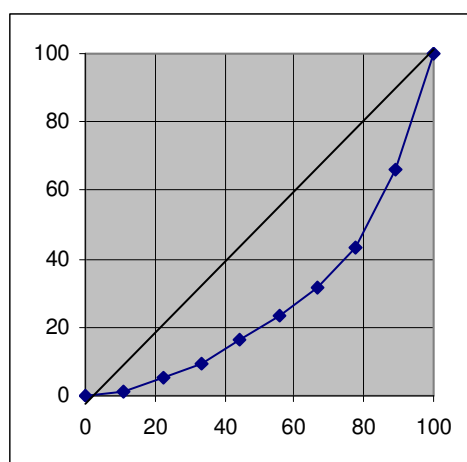


Figura 2.3.3 – Curva de concentração dos Activos na Indústria Transformadora de Barcelona em 1973.

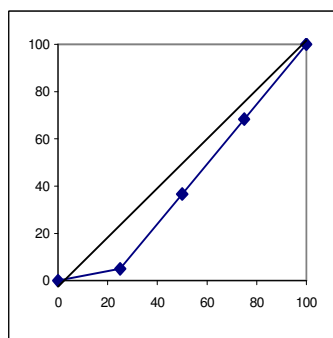
Relembrando os dois exemplos tratados, pode afirmar-se que a curva de Lorenz compara a distribuição duma variável qualquer com a distribuição teórica que ocorre quando:

- a) todos os valores são iguais (no caso duma variável primária isolada);

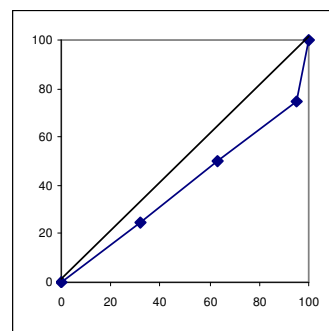
- b) os incrementos duma das variáveis iniciais correspondem sempre a incrementos proporcionais da outra variável inicial (no caso duma variável secundária resultante da razão entre duas primárias).

O menor ou maior afastamento da curva face à diagonal principal do quadrado é que quantifica o grau de concentração (ou desigualdade) da distribuição em estudo.

As curvas de Lorenz são muitas vezes difíceis de comparar, especialmente quando em umas se verificam concentrações nos valores altos da variável (Figura 2.3.4.A), e noutras concentrações nos valores menores (Figura 2.3.4.B), caso em que não é fácil descortinar visualmente as diferenças no grau de concentração ou de desigualdade da distribuição.



A



B

Figura 2.3.4 – Curvas de Lorenz diferentes com graus de concentração iguais.

Como se disse, quanto maior o afastamento da curva da diagonal, ou seja, quanto maior a área do gráfico que se situa entre a curva de Lorenz e a diagonal, maior é a concentração da variável.

Para medir com precisão o valor das distâncias da curva à diagonal ou da área compreendida entre a diagonal e a curva, foram propostos vários índices e coeficientes por diversos autores:

- o coeficiente de especialização;
- o índice de GINI;
- o índice de Schutz;
- o ponto de igual partilha;
- e o coeficiente de concentração de GINI.

3.2.- Índices de Concentração

Todos os índices de concentração são valores que medem o grau de concentração, desigualdade, dissimilaridade ou segregação da distribuição.

Antes de os apresentar, recorda-se a notação que será utilizada, tendo como exemplo a variável densidade populacional, que é resultante da razão entre a população (variável numerador) e a área (variável denominador):

NOTAÇÃO	SIGNIFICADO	EXEMPLO
t_i	Variável secundária resultante	Densidade populacional
r_i	Variável inicial (numerador)	População
s_i	Variável inicial (denominador)	Área
p_k	% da variável numerador	% da população
q_k	% da variável denominador	% da área
y_k	% cumulativa de p_k	% cumulativa da população
x_k	% cumulativa de q_k	% cumulativa da área

Quadro 2.3.5 – Notação utilizada para o cálculo da curva de Lorenz e dos índices de concentração.

Dentre os vários índices que têm sido apresentados, a seguir indicam-se os mais utilizados nas diversas ciências sociais.

3.2.1 – Coeficiente de Especialização

$$C = \frac{1}{2} \times \sum_{k=1}^n |q_k - p_k|$$

Este é um dos coeficientes de concentração que se deve a Gini. Os sociólogos chamam-lhe índice de dissimilaridade (Duncan & Duncan, 1995), e os economistas e geógrafos coeficiente de especialização.

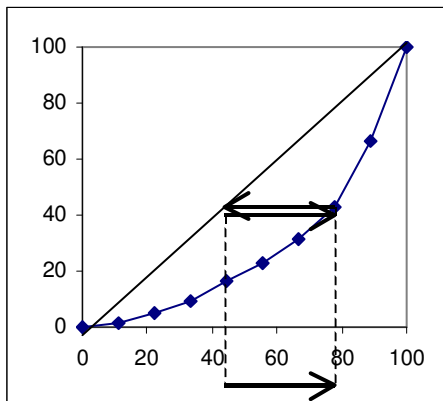


Figura 2.3.5 - Esquema da "barriga" da curva de Lorenz.

Mede, em percentagem, a "barriga" da curva de Lorenz, pois é metade da soma dos afastamentos para a esquerda e para a direita da curva, relativamente à diagonal.

Assim, correspondendo a metade do dobro do afastamento máximo entre a diagonal e a curva de Lorenz, dá uma aproximação à dimensão da área de concentração, compreendida entre a diagonal e a curva.

Exemplo de cálculo:

População por sectores industriais (Barcelona, 1973)

q_k	p_k	$(q_k - p_k)$	$ q_k - p_k $
11,1111	1,3866	9,7245	9,7245
11,1111	3,9247	7,1864	7,1864
11,1111	4,2793	6,8318	6,8318
11,1111	6,7479	4,3632	4,3632
11,1111	6,8813	4,2298	4,2298
11,1111	8,5474	2,5637	2,5637
11,1111	11,3022	- 0,1910	0,1910
11,1111	23,0606	- 11,9494	11,9494
11,1111	33,8700	- 22,7589	22,7589
100,0000	100,0000	0,0000	69,7988

Quadro 2.3.6 – Quadro auxiliar para o cálculo de C

$$C = \frac{1}{2} \times 69,7988 \quad C = 34,8994 \quad \dots$$

3.2.2 – Índice de Gini

$$G = \frac{\sum_{k=1}^{n-1} (x_k - y_k)}{\sum_{k=1}^{n-1} (x_k)} \times 100$$

Este índice “calcula” a área de concentração (entre a diagonal e a curva de Lorenz), através dum método de aproximação.

Compara a soma da parte das ordenadas que corresponde à área de concentração ($x_k - y_k$) com a soma total das correspondentes abcissas da diagonal (x_k).

Deve notar-se que a abcissa da diagonal é igual à ordenada da curva, pois para todos os pontos da diagonal, a abcissa é sempre igual à ordenada.

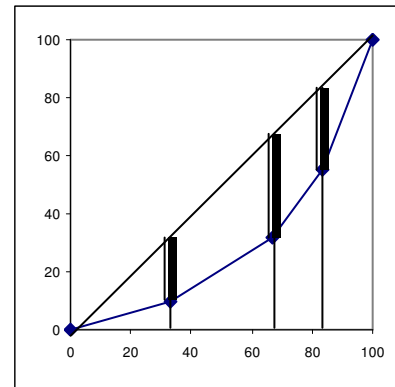


Figura 2.3.6 – Esquema da comparação entre as abcissas da diagonal e da curva de Lorenz.

Exemplo de cálculo:

População por sectores industriais (Barcelona, 1973)

q_k	p_k	x_k	y_k	$(x_k - y_k)$
11,1111	1,3866	11,1111	1,3866	9,7245
11,1111	3,9247	22,2222	5,3113	16,9109
11,1111	4,2793	33,3333	9,5906	23,7427
11,1111	6,7479	44,4444	16,3385	28,1059
11,1111	6,8813	55,5556	23,2198	32,3357
11,1111	8,5474	66,6667	31,7672	34,8994
11,1111	11,3022	77,7778	42,0694	34,7084
11,1111	23,0606	88,8889	66,1300	22,7589
11,1111	33,8700	~~100,0000~~	~~100,0000~~	~~0,0000~~
100,0000	100,0000	400,0000	-----	203,1863

Quadro 2.3.7 – Quadro auxiliar para o cálculo de G

$$G = \frac{203,1863}{400,000} \times 100 \quad G = 50,79$$

3.2.3 – Índice de Schutz

É uma medida de concentração, proposta por Schutz em 1951.

$$S = \sum_{k=1}^r \left[\left(1 - \frac{p_k}{q_k} \right) \times q_k \right]$$

onde r é o número de elementos ou valores em que p_k é menor que q_k , ou seja, em que a relação p_k / q_k é menor que 1.

Este índice, tal como o índice C , mede também o valor da “barriga” máxima da curva de Lorenz. Deve ser tomado em conta que $(1 - p_k / q_k) \times q_k = (q_k - p_k)$, e portanto que este índice apenas soma as diferenças entre as ordenadas e as abcissas da curva de Lorenz até ao ponto em que se verifica o maior afastamento da diagonal, ou seja, até à altura em que a curva de Lorenz, que até aqui se ia afastando da diagonal, se começa a aproximar dela de novo (o que sucede quando o seu declive passa a ser maior do que 1) e que por isso corresponde a metade da soma do módulo de todas as diferenças.

Exemplo de cálculo:

População por sectores industriais (Barcelona, 1973)

q_k	p_k	p_k / q_k	$(1 - p_k/q_k)$	$(1 - p_k/q_k) \times q_k$
11,1111	1,387	0,124798	0,875202	9.72446
11,1111	3,925	0,353224	0,646776	7.18640
11,1111	4,279	0,385135	0,614865	6.83183
11,1111	6,748	0,607311	0,392689	4.36321
11,1111	6,881	0,619319	0,380681	4.22978
11,1111	8,547	0,769265	0,230735	2.56372
11,1111	11.302	1,017194	-----	-----
11,1111	23,061	2,075450	-----	-----
11,1111	33,870	3,048302	-----	-----
100,000	100,000	-----	-----	34,89940

Quadro 2.3.8 – Quadro auxiliar para o cálculo de S . Notar que $r = 6$.

$$S = 34,89940$$

Nota: O índice de Schutz (S) é na realidade uma outra forma de calcular o coeficiente de concentração (C). O índice de Schutz é mais difícil de calcular manualmente, mas mais fácil de calcular por programas de cálculo automático.

3.2.4– Ponto de Igual Partilha

Calcula-se a partir da seguinte fórmula:

$$F = \sum_{k=1}^r q_k$$

onde r é o número de classes, grupos ou valores em que p_k é menor que q_k , (em que p_k / q_k é menor que 1).

É uma medida do grau de desigualdade da distribuição, que se pode definir como a ordenada do ponto em que a curva de Lorenz mais se afasta da diagonal (ou a ordenada da “barriga” da curva).

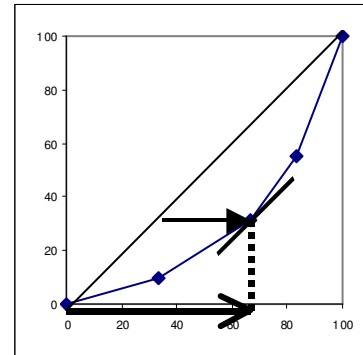


Figura 2.3.7 – Esquema da ordenada da “barriga” da curva de Lorenz.

Pode ser interpretado como a percentagem de indivíduos, membros, unidades, áreas ou regiões (conforme o caso), que integram os grupos que:

- estão abaixo do valor médio geral, no caso de uma variável secundária, resultante da razão de outras duas;
- ou são os mais desvantajosos, no caso de uma única variável. Consideram-se os mais desvantajosos os que têm menos do que a proporção que lhes competiria se todos os valores da distribuição fossem iguais.

Compreende-se assim a razão por que este índice é designado na bibliografia de origem inglesa por *fair-share point* (ponto de partilha equilibrada).

Exemplo de cálculo:

Activos por sectores industriais (Barcelona, 1973)

q_k	p_k	p_k / q_k	q_k
11,1111	1,3866	0,124798	11,1111
11,1111	3,9247	0,353224	11,1111
11,1111	4,2793	0,385135	11,1111
11,1111	6,7479	0,607311	11,1111
11,1111	6,8813	0,619319	11,1111
11,1111	8,5474	0,769265	11,1111
11,1111	11,3022	1,017194	-----
11,1111	23,0606	2,075450	-----
11,1111	33,8700	3,048302	-----
100,0000	100,0000	-----	66,6667

Quadro 2.3.9 – Quadro auxiliar para o cálculo de F

$$F = 66,67 \%$$

Interpretando o valor obtido, pode dizer-se que em Barcelona em 1973, dois terços (66,67 %) dos nove sectores de actividade têm menos activos que o que seria de esperar, se a distribuição dos activos por todos os sectores fosse uniforme.

3.2.5– Coeficiente de Concentração de Gini

Todos os índices e coeficientes que até agora se apresentaram fornecem imagens diferentes, e às vezes complementares, da concentração de valores existentes na distribuição em estudo.

São estimativas, valores aproximados, da verdadeira área de concentração, que corresponde à área entre a diagonal e a curva de Lorenz.

Mas é possível calcular exactamente o valor dessa área, ou melhor ainda, a sua importância relativamente à do triângulo definido pela diagonal e os dois lados do quadrado. Assim sendo, este valor variará entre 0 (zero), quando a distribuição é uniforme, com todos os valores iguais à média e a curva de Lorenz a coincidir com a diagonal, e 1 (ou 100%), quando um só dos elementos, grupos, classes ou áreas detém o total dos valores e os outros são todos iguais a zero.

O método de cálculo tem por base o cálculo de áreas de triângulos e rectângulos, como se pode ver na Fig. 2.3.8. A área do triângulo de referência, definido pela diagonal do quadrado (cuja dimensão é de 100 % por 100 %), pelo eixo dos x's e pelo limite direito (paralelo ao eixo dos y's) é:

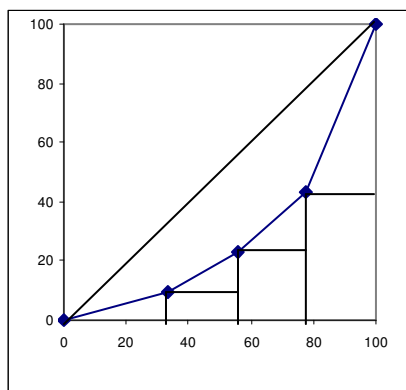
$$A = \frac{100 \times 100}{2} = 5000$$

A área de concentração é a que fica situada entre a diagonal e a curva de Lorenz. Calcula-se por diferença entre o triângulo da diagonal e a área que se situa abaixo da curva, que se calcula muito facilmente a partir dos triângulos e rectângulos respectivos.

Na Figura 2.3.8, exemplifica-se o cálculo da área de concentração, a partir dos diversos rectângulos e triângulos.

Conhecida a área de não concentração é fácil calcular o valor da área de concentração, que é a diferença entre a área do grande triângulo e aquela.

O Coeficiente da Área de Concentração de Gini, ou mais simplesmente Coeficiente de Gini (CG) compara a importância relativa da área de concentração em relação à área total.



$$\begin{aligned} & \frac{1}{2} \times q_1 \times p_1 \\ & \frac{1}{2} \times q_2 \times p_2 + q_2 \times y_1 \\ & \frac{1}{2} \times q_3 \times p_3 + q_3 \times y_2 \\ & \frac{1}{2} \times q_4 \times p_4 + q_4 \times y_3 \end{aligned}$$

Figura 2.3.8 – Esquema do cálculo de áreas de triângulos e rectângulos

Abaixo apresenta-se a fórmula que permite calcular o Coeficiente de Gini tendo sempre em conta que y_0 é igual a zero.

$$CG = \frac{5000 - \sum_{k=1}^n \left[q_k \times \left(\frac{1}{2} \times p_k + y_{k-1} \right) \right]}{5000} \times 100$$

Este coeficiente mede exactamente a área de concentração e é por isso o melhor dos coeficientes a usar. É o verdadeiro Coeficiente de Gini, do qual os outros são fórmulas aproximadas e simplificadas, que apenas se utilizaram quando os processos de cálculo eram morosos e pouco fidedignos, o que já não acontece.

Exemplos de cálculo: População por sectores industriais (Barcelona, 1973)

q_k	p_k	$1/2 p_k$	$y_{(k-1)}$	$1/2 p_k + y_{(k-1)}$	$q_k (1/2 p_k + y_{(k-1)})$
11,111	1,387	0,693	0,000	0,6933	7,7036
11,111	3,925	1,962	1,387	3,3490	37,2112
11,111	4,279	2,140	5,311	7,4510	82,7889
11,111	6,748	3,374	9,591	12,9646	144,0511
11,111	6,881	3,441	16,344	19,7792	219,7690
11,111	8,547	4,274	23,220	27,4936	305,4841
11,111	11,302	5,651	31,767	37,4183	415,7593
11,111	23,061	11,536	42,069	54,5997	606,6633
11,111	33,870	16,935	66,130	83,0650	922,9443
100,000	100,000	-----	-----	-----	2 742,3749

Quadro 2.3.10– Quadro auxiliar para o cálculo de CG

$$CG = \frac{5000 - 2742,3749}{5000} \times 100 = \frac{2257,6251}{5000} \times 100 = 45,1525...$$

CG = 45,15...

3.3.— Considerações

Quando se estudam concentrações territoriais de distribuições, ou quando se avalia o grau de especialização duma população dividida em vários tipos de estratos ou condições, tanto de natureza social, como económica, como outras, geralmente é suficiente para o investigador o conhecimento das curvas de Lorenz e do coeficiente de Gini (e eventualmente do coeficiente F).

A curva de Lorenz ilustra bem o grau de concentração da distribuição, e é especialmente indicada no caso de comparações entre regiões ou anos diferentes.

Tem um grande valor na apresentação de resultados pois permite uma fácil comparação entre as distribuições. Tem uma forte capacidade de captação visual, e por isso deve ser utilizada especialmente quando as distribuições a comparar são muito contrastantes.

Por outro lado, é uma forma de representação que não é de leitura fácil. Deve ser utilizada em relatórios destinados a especialistas ou para um público muito interessado e com formação técnica superior.

Não é aconselhável para textos de divulgação nem para obras técnicas de carácter geral.

O coeficiente de Gini mede o valor exacto do grau de concentração. A comparação dos valores e da sua variação em situações ou anos diferentes pode ser esclarecedora na identificação da força das tendências para a concentração ou para a igualdade das situações e processo analisados.

O coeficiente F permite uma leitura imediata. Mede a percentagem de valores da variável numerador que está abaixo da média da distribuição da variável razão. Não deve ser indicado em texto simplesmente como "o valor do coeficiente F é tanto", mas de uma forma mais descritiva, indicando expressamente qual a percentagem das entidades que têm menos do que a média.

As outras medidas, os coeficientes de especialização e os índices de Schutz e de Gini, são simples aproximações ao valor do coeficiente de Gini; foram muito utilizadas na época do cálculo manual, pois facilitavam as operações matemáticas a efectuar.

Ainda agora são aqui referidas apenas para facilitar a comparação de valores actuais com distribuições de valores da mesma natureza que foram estudados e apresentados dessa forma e com esses índices na bibliografia especializada, que é relativamente abundante em alguns ramos do conhecimento.

3.4.— Exemplos

Exemplo 1 - Moçambique - Densidade de População

Considere a distribuição por províncias da população residente em Moçambique em 1998, conforme o quadro anexo.

Províncias	População	Área (Km ²)	Densidade (hab/km ²)
Niassa	822000	129061	6.3691
Cabo Delgado	1374000	82625	16.6293
Nampula	3280000	81606	40.1931
Zambézia	3452000	105008	32.8737
Tete	1294000	100724	12.8470
Manica	1094000	61656	17.7436
Sofala	1479000	68018	25.7442
Inhambane	1156000	68615	16.8476
Gaza	1082000	75709	14.2916
Maputo	1883000	26358	71.4394
Total	16916000	799380	21.1614

Fonte: Moçambique em Números, 1999, Maputo, INE, pg.7 e 9.

- 1)- Sabendo que
- | | |
|-------------------------------|----------|
| Coeficiente de Gini (CG) = | 32,15571 |
| Coeficiente F = | 64,84901 |
| Índice de Gini (G) = | 32,39694 |
| Coef. de especialização (C) : | 24,52032 |
| Índice de Schutz (S) : | 25,52032 |

interprete os resultados.

- 2)- Faça a curva de Lorenz (pode utilizar os valores auxiliares do quadro seguinte, que foram calculados por um programa estatístico).

% cum y	% cum x
0,0000	0,0000
4,8593	16,1451
12,5089	29,7454
18,9052	38,2164
27,0277	48,5525
33,8614	57,1360
40,3287	64,8490
49,0719	73,3578
69,4786	86,4940
88,8685	96,7027
100,0000	100,0000

- 3)- Procure dados e faça os cálculos para os distritos de Portugal, para vários anos. Compare os resultados e interprete.

Resposta:

1) - O coeficiente de Gini é de 32,15, o que significa que a concentração da distribuição da população das províncias pelo território é moderadamente concentrada (afastando-se cerca de um terço da igualdade). O valor de F igual a 64,85 significa que existem províncias correspondentes a 64,85% do território com valores de densidade inferiores a média geral nacional, tendo as restantes províncias (ocupando 35,15% do território) valores superiores.

2) e 3) - A resposta conjunta a estas duas questões facilita a interpretação.

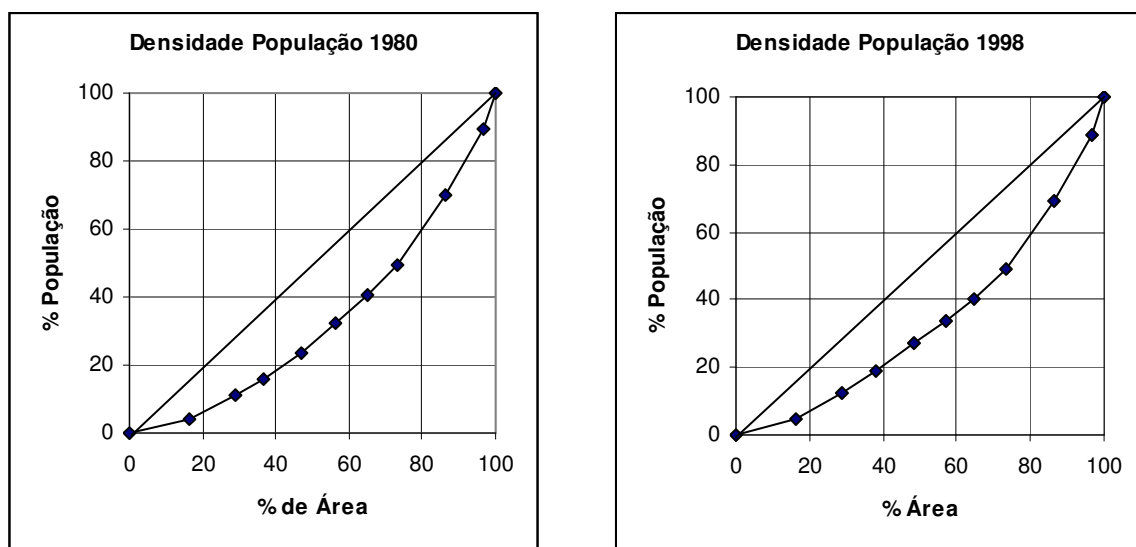


Figura 2.3.9 - Curvas de Lorenz para a distribuição da densidade populacional por províncias. Moçambique, 1980 e 1998.

Relativamente aos valores dos índices, tem-se:

	1980	1998
Coeficiente de Gini	32,16	32,95
Coeficiente F	64,77	64,85

A primeira conclusão é a de uma relativa estabilidade estrutural. O grau de concentração da distribuição da população por províncias é muito similar nos dois anos. As diferenças que se verificaram na respectiva variação não alteraram significativamente a estrutura geral da distribuição (veja-se que, embora as taxas de crescimento médias anuais da densidade populacional tenham variado entre 0,54 e 3,51 %, a correlação entre os valores de 1980 e os de 1998 é de $r = 0,989$).

Se o objectivo é estudar as alterações conjunturais da densidade de população, a análise deve incidir sobre os valores da variação e não sobre a comparação da população residente nas províncias.

Exemplo 2 - Precipitação em Maputo e Beira (Moçambique) e Buenos Aires (Argentina)

Considere a precipitação média mensal (normal de trinta anos) para as estações de Maputo e Beira em Moçambique e Buenos Aires na Argentina.

Meses	Maputo	Beira	Buenos Aires
Janeiro	129,9	275	75
Fevereiro	124,2	255	63
Março	96,9	263	115
Abril	64,1	113	75
Maio	27,6	62	70
Junho	26,9	40	68
Julho	12,5	32	56
Agosto	12,7	32	60
Setembro	38,1	23	75
Outubro	46,1	31	90
Novembro	86,1	127	70
Dezembro	102,6	240	98
Total	767,7	1493	915

Os resultados do programa para estes três casos são:

Índices	Maputo	Beira	Buenos Aires
Coeficiente de Gini (CG)	35,6226	42,8053	10,938
Coeficiente F	50,0000	58,3333	75,0000
Coeficiente G	39,4721	47,3300	12,2702
Coeficiente S e C	28,6505	36,0293	8,1147

A precipitação anual em Moçambique é muito mais concentrada que em Buenos Aires (coeficiente CG e curvas de Lorenz). Dos coeficientes F vê-se que em Maputo metade dos meses têm precipitações abaixo da média (e outra metade acima), enquanto na Beira são 58,33% (7 meses em doze) e em Buenos Aires 75%, isto é, nove meses no ano têm menos precipitação do que a média, embora a diferença seja pequena.

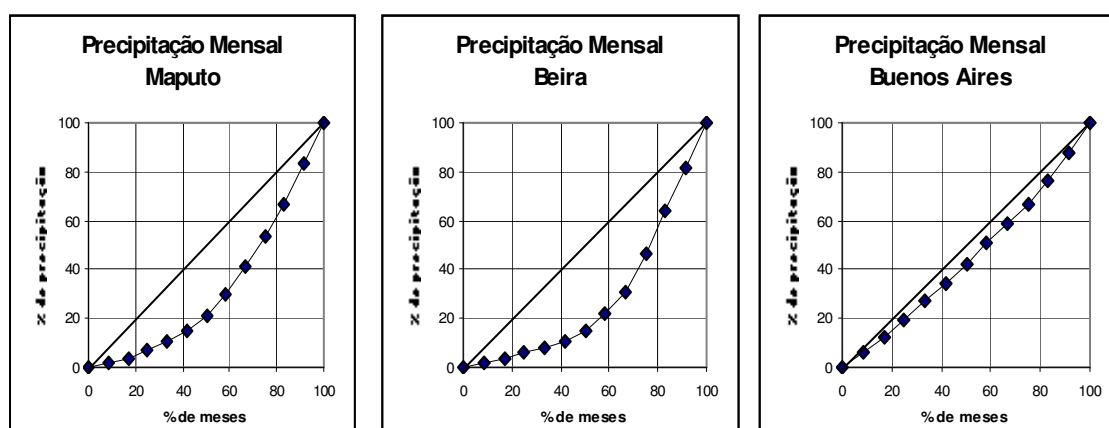


Figura 2.3.10 - Curvas de Lorenz das precipitações médias mensais em Maputo, Beira e Buenos Aires.

3.5.- Exercícios

1).- A população da província de Maputo tem a seguinte distribuição por sexos e por distritos/cidades.

Distritos/cidade	Homens	Mulheres	Total
Cidade da Matola	204083	220579	424662
Distrito de Boane	26697	30006	56703
Distrito de Magude	18160	24628	42788
Distrito de Manhiça	57800	72551	130351
Distrito de Marracuene	19721	21956	41677
Distrito de Matutuine	17312	17849	35161
Distrito de Moamba	20411	22985	43396
Distrito da Namaacha	15605	15836	31441
TOTAL	379789	426390	806179

Fonte: INE, II Recenseamento Geral da População. Resultados Definitivos. Província de Maputo.

Pede-se que estude a concentração da população feminina. Para isso:

- Quais são as unidades de análise ?
- Qual é a variável razão ?
- Qual é a variável inicial numerador ?
- Qual a variável inicial denominador ?
- Que técnicas deve usar para estudar a concentração ?
- Determine os coeficientes e desenhe a curva (Sugestão: use um programa adequado).
- Interprete os resultados

2).- Procure a mesma informação sobre a população feminina dos 18 distritos do Continente para dois anos censitários.

- Calcule os mesmos índices e desenhe a curva de concentração de Lorenz.
- Interprete os resultados, comparando-os uns com os outros e com o exercício anterior.

3).- A precipitação mensal em Maputo no ano de 1976 foi:

Janeiro	434,1	Maio	23,5	Setembro	16,2
Fevereiro	128,6	Junho	4,2	Outubro	37,6
Março	37,4	Julho	0,0	Novembro	61,5
Abril	75,1	Agosto	0,2	Dezembro	106,4

Fonte: Serviço Meteorológico de Moçambique - Anuário de Observações Ano de 1976, I Parte - Observações de superfície, Maputo.

Estude a concentração da distribuição (Sugestão: Desenhe a curva de Lorenz e calcule os coeficientes F e CG de Gini).

- a).- Qual é a variável resultante ou de razão ?
- b).- Qual a variável inicial numerador ?
- c).- Qual a variável inicial denominador ?
- d).- Interprete os resultados obtidos.

4).- Compare a pluviosidade mensal de 1976 em Maputo com a de Mabote, abaixo indicadas

Janeiro	434,1	Maio	23,5	Setembro	16,2
Fevereiro	128,6	Junho	4,2	Outubro	37,6
Março	37,4	Julho	0,0	Novembro	61,5
Abril	75,1	Agosto	0,2	Dezembro	106,4

Fonte: Serviço Meteorológico de Moçambique - Anuário de Observações Ano de 1976, I Parte - Observações de superfície, Maputo.

Faça os cálculos e desenhe a curva, compare os resultados e interprete.

5).- Procure nas Estatísticas dos Transportes e Comunicações, o número de mortos em acidentes de viação por distrito (Continente), para alguns anos diferentes. Calcule o nº de mortes em acidente de viação por mil veículos registados.

Estude a concentração desta variável para cada ano.

- a).- Calcule os coeficientes F e CG e desenhe a curva de Lorenz.
- b).- Interprete os resultados.

6).- Veja a estrutura dos trabalhadores por ramo de actividade nos anos de 1980 e 1991.

- a).- Calcule os coeficientes de concentração CG e F e desenhe a respectiva curva de Lorenz. Estude a concentração desta variável por província.
- b).- Compare os dois anos e interprete os resultados.

RESPOSTAS:

- 1).-a) A cidade da Matola e os sete distritos administrativos.
- 1).-b) A proporção de mulheres na população de cada distrito/cidade.
- 1).-c) O número de mulheres de cada distrito/cidade.
- 1).-d) A população total de cada distrito/cidade.
- 1).-e) A curva de Lorenz e os coeficientes CG e F.
- 1).-f) A proporção média de mulheres na região é de 0,5289.

Os valores para o curva de Lorenz são:

0,0000	0,0000	71,8183	73,1406
3,7140	0,9000	77,2089	78,5235
7,9001	8,2614	94,2241	94,6925
59,6318	60,9373	100,0000	100,0000
CG = 1,11701		F = 66,107	

A concentração é mínima, como seria de esperar dada a natureza da variável; sessenta e seis vírgula um por cento da população vive em distritos/cidade em que a percentagem de mulheres na população é inferior à percentagem média na região.

3).-a) A precipitação mensal

3).-b) A precipitação.

3).-c) Cada mês (é uma constante sempre igual a um).

3).-d) CG = 62,61 F = 75,00

A precipitação ocorre muito concentrada em poucos meses no ano (CG= 62,6%). Em 75 % dos meses chove menos do que a média.

4).- A precipitação total em Mabote é menor que em Maputo (811,4 mm). Os meses de chuva estão um pouco mais concentrados, o F corresponde a 7 meses com precipitações abaixo da média de precipitações por mês, portanto a concentração verifica-se nos mais pluviosos.

5).-a) o número de mortes por acidente rodoviário por mil veículos registados na respectiva província.

5).-b) O número de mortes em acidentes rodoviários.

5).-c) O número de veículos em milhares (pois a variável razão é por mil veículos).

3.6.- Bibliografia

- CALOT, Gérard (1970) – **Curso de Estadística Descriptiva**, Madrid, Paraninfo, p. 94-99;
- COLE, J. P.; KING, C. A .M. (1968) – **Quantitative Geography**, London, John Wiley & Sons, p. 243-245.
- (*) ESTEBANEZ ALVAREZ, J.; BRADSHAW, R.. P. (1978) – **Técnicas de Quantificación en Geografía**, Madrid, Editorial Tebar Flores, p. 77-80 e 108-117.
- GERARDI, Lúcia H. O.; SILVA, Barbara-Christine N. (1981) – **Quantificação em Geografia**, São Paulo, Difel, p. 117-124.

Anexo 3

Recolha da Informação

1. – GENERALIDADES

2. - FONTES DE INFORMAÇÃO

A informação a recolher tem a ver fundamentalmente com a natureza da investigação. São muitas as fontes e formas de recolher informação. Consoante o tipo de trabalho a fazer, umas serão mais adequadas e outras menos.

Em qualquer país, a maior parte da informação estatística é produzida por um conjunto de organismos públicos - o Sistema Estatístico Nacional - que têm a competência legal de a produzir e divulgar em condições técnicas aceitáveis. A informação produzida pelo Sistema Estatístico Nacional deve ser, conforme Vilares (1993):

- pertinente para os utilizadores,
- disponibilizada atempadamente,
- precisa,
- isenta,
- comparável (no tempo, e com outros sistemas),
- divulgada continuamente e sem rupturas,
- e contribuir para a diminuição da incerteza.

Os geógrafos, como outros investigadores de ciências naturais e sociais, devem procurar utilizar todas as potencialidades do Sistema Estatístico Nacional para os seus trabalhos. Umas vezes isso não é possível e em muitas outras não é suficiente. Assim, os investigadores devem conhecer os diversos métodos de recolha de informação e ser treinados na sua utilização.

Nos processos de investigação científica, o tipo e origem da informação revestem geralmente as seguintes formas:

- a).- Fontes Documentais
- b).- Observação
- c).- Experimentação
- d).- Análise documental
- e).- Inquéritos

A recolha de informação em Geografia utiliza pouco a experimentação. Em Geografia Humana usam-se principalmente as fontes

documentais, a observação e a inquirição, enquanto em Geografia Física predomina a observação, embora a entrevista e o inquérito tenham cada vez mais utilização nos procedimentos de avaliação e mitigação de perigos e riscos.

Por outro lado, tanto em Geografia como noutras ciências sociais, tem-se vindo a recorrer cada vez mais à opinião de interessados e de especialistas, traduzida pela recolha de informação através de:

- f)- "Focus-group"
- g)- Métodos "Delphi"

Assim como à análise detalhada de situações típicas ou excepcionais, que decorreram ou decorrem tanto no tempo como no espaço, através de:

- h)- Os estudos de caso
- i)- As histórias de vida

2.1.- Fontes Documentais

As fontes documentais constituem um modo de recolha fundamental em Geografia que dada a facilidade de obtenção e o baixo custo de aquisição, deve ser sempre utilizada.

Com efeito, grande parte do trabalho de investigação em Geografia é feito a partir de dados publicados. Isto é tanto mais certo quanto mais se estudam as pequenas escalas, aos níveis regional, nacional e internacional.

Neste caso, geralmente os geógrafos utilizam informação publicada, que foi recolhida por outras entidades ou no âmbito de outros projectos de investigação. E assim sendo, pouco ou nada tiveram a ver nem com os critérios nem com os procedimentos de recolha.

Mas noutros casos, embora mais raramente, o próprio investigador, pela sua proximidade funcional a organismos e entidades produtores de estatísticas, ou pelo seu papel de coordenação em grandes projectos de investigação, é obrigado a garantir a adequação e a qualidade da informação que recolhe e publica e que todos podem vir a utilizar. Em muitos países, como em Portugal ou em Moçambique, é o caso dos censos e de grandes operações de inquéritos básicos.

Existem em todos os países instituições cuja função é recolher e disponibilizar informação estatística geral, como é o caso do D.N.E., o Departamento Nacional de Estatística em Moçambique, ou do I.N.E., o Instituto Nacional de Estatística em Portugal. Para além destas instituições gerais, existem outras organizações de análise e gestão especializada, que também publicam informação estatística sobre o seu âmbito de acção, como é o caso dos serviços meteorológicos, hidrológicos, de agricultura, pescas, etc..

Tem de se salientar a importância de toda esta informação estatística ser disponibilizada a toda a sociedade. Embora a informação seja originalmente solicitada e paga pelo Estado, e os reduzidos orçamentos das

instituições tentem a venda (cara) da informação, a realidade é que a sua difusão alargada e a baixo preço é um poderoso factor de desenvolvimento das sociedades, pelo melhor conhecimento das sociedades e territórios que permite e pela a mais eficaz utilização dos recursos.

A organização estatística dos diversos países segue dois modelos possíveis:

- Os centralizados, em que todos os organismos oficiais (ou quase) e privados, são obrigados a fornecer informação ao instituto central, que depois a publica, ou não. É o caso de Portugal.

- Os descentralizados, em que cada organismo ou instituição faz e publica as suas próprias estatísticas. Neste caso, muitos ministérios e entidades publicam estatísticas (Ex. o Ministério da Agricultura publica Estatísticas Agrícolas, o Instituto das Pescas, as Estatísticas das Pescas, a Universidade as suas estatísticas, etc.). Existe sempre um instituto central que publica, ou fornece, as estatísticas gerais.

2.1.1 - Estatísticas Nacionais

Qualquer que seja o modelo adoptado e a capacidade financeira dos institutos estatísticos e dos governos, geralmente dispõe-se de:

Censos da População, com as principais características demográficas da população de um país e das suas regiões mais importantes. Idealmente devem realizar-se de dez em dez anos (geralmente nos anos que terminados em 1). Em alguns países mais desenvolvidos os censos da população já são efectuados de cinco em cinco anos, enquanto noutros o controle do registo da população é tal que é possível conhecer as principais características da população *on line*, em qualquer momento que se deseje. Por outro lado, existem países sem censos e com muito pouca informação demográfica de confiança (como a China, por exemplo).

Os Censos ou Recenseamentos da População são inquéritos exaustivos a todos os indivíduos que residem e estão presentes num dado território numa certa ocasião - o momento censitário. O número de perguntas é geralmente reduzido, mas deve ser o adequado às capacidades da organização censitária e às necessidades do território que o organiza. A recolha, tratamento e difusão da informação são efectuados através de operações de grande envergadura, que envolvem muita gente e grande volume de meios (de impressão, transporte, armazenagem e tratamentos). São operações muito caras, mesmo quando parte significativa do trabalho é garantido por via administrativa³.

Recenseamentos da Actividade Produtiva, relativos às actividades agrícolas, pecuárias e florestais, à indústria e ao comércio e serviços. A periodicidade varia com as necessidades do país e a sua capacidade de realização, mas geralmente é mais ou menos decenal. Baseiam-se em inquéritos, geralmente exaustivos para as maiores unidades e exaustivos ou

³ - A título de exemplo refere-se que o pessoal envolvido no Censo de 1991 em Portugal Continental, ultrapassou 17.000 indivíduos, contratados por períodos que variaram de algumas semanas a vários anos.

por amostragens bastante significativas para as unidades menores, muito mais numerosas.

Informação Estatística geral sobre o país, geralmente de periodicidade anual, que tem por origem os procedimentos administrativos a que os cidadãos, empresas e outras pessoas colectivas estão obrigados pela lei. Cada organismo do estado é obrigado a produzir estatísticas resumo da sua actividade, ou a fornecer os dados aos institutos de estatísticas que os tratam tecnicamente e os publicam.

Entre esta informação salienta-se a informação demográfica, sobre o movimento da população: óbitos, nascimentos, casamentos, divórcios. É importante para a previsão da população e para os cenários de evolução nacionais e é relativamente fácil de recolher, pois correspondem a movimentos que são objecto de procedimentos administrativos. A sua qualidade depende do controle administrativo da população e da eficácia dos sistemas de administração pública.

Geralmente também se publica outra informação estatística respeitante às diversas actividades dos organismos de controle do Estado. Respeitam aos Transportes, à Pesca, à Construção e Obras Públicas, ao Rendimento e Bem-estar, a assuntos eleitorais, ao Consumo de Bens e Serviços, à Energia, ao Comércio Externo, à Cultura e Desporto, à Alimentação, à Saúde, Doença e Morte, à Justiça, à Administração Pública, à Produção Agrícola e Industrial, para além de informação climatológica e hidrológica quando os organismos têm capacidade para tal.

Antes de execução de um trabalho de qualquer tipo ou âmbito é necessário proceder ao inventário de todas as fontes documentais publicadas existentes, históricas e recentes, que de algum modo se relacionem com o trabalho a realizar.

2.1.2 - Estatísticas Internacionais

Para além das publicações estatísticas nacionais, que quase sempre incluem também alguma desagregação regional, existem as estatísticas internacionais. Estas publicações são muito úteis para comparar as situações dos diversos países e para situar o grau de desenvolvimento dos territórios que se analisam.

São publicações caras que devem ser procuradas nas delegações dos organismos internacionais que as publicam.

As mais importantes são:

2.1.2.1 - Das Nações Unidas

- O Anuário Estatístico (Statistical Yearbook), publicação anual com informação sistemática sobre os diversos países do mundo; para além da informação básica geral que consta todos os anos, o número de cada ano é dedicado a um tema de interesse e actualidade específico.

- As Estatísticas Demográficas (Demographic Yearbook), publicação anual com informação sobre o estado e os movimentos da população; é especialmente interessante para os países que não

dispõem de censos ou de institutos estatísticos, porque os técnicos de demografia da O.N.U. procuram fazer estimativas com os melhores métodos indirectos e mistos.

- Estatísticas temáticas, sem periodicidade de publicação, sobre assuntos diversos: Agricultura, Indústria, Transportes, Turismo, etc.

- Do Programa das Nações Unidas para o Desenvolvimento (P.N.U.D.) é publicado anualmente o importantíssimo Relatório do Desenvolvimento Humano, com estudos monográficos e informação estatística comparativa.

2.1.2.2 - Do Banco Mundial / PNUD

- Para além de estatísticas temáticas e de estudos analíticos sobre territórios e projectos, publica anualmente relatórios de grande qualidade. O mais interessante para os geógrafos é o relatório anual sobre o desenvolvimento no mundo, que actualmente é publicado pela PNUD, organismo especializado da ONU.

2.1.2.3 - Da União Europeia

O organismo estatístico da União Europeia é o EUROSTAT, localizado no Luxemburgo. Compete-lhe a harmonização dos sistemas estatísticos nacionais dos países da União Europeia, e a armazenagem e produção de séries estatísticas comparativas entre os países da Europa. Para além das estatísticas nacionais dispõe de muita informação regional. Também se dedica à cartografia normal e automática, tratando imagens de satélite e fotografia aéreas.

Publica e disponibiliza em base de dados e disquetes muita informação com grande diversidade e boa resolução espacial.

A sua importância é muito grande. Na União Europeia é a partir de critérios estatísticos do EuroStat sobre o rendimento regional que as regiões têm ou não direito a receber fundos estruturais para o seu desenvolvimento.

2.1.2.4 - De outros organismos

- Da O.C.D.E.

A Organização para a Cooperação e Desenvolvimento Económico dedica-se especialmente às economias dos países mais desenvolvidos do mundo. Publica regularmente estudos sobre cada país e tem algumas publicações e séries sobre problemas económicos em geral.

- Da F.A.O.

A Organização para a Alimentação e Agricultura das Nações Unidas desenvolve uma série de estudos e projectos sobre agricultura e alimentação. Para além do trabalho no terreno publica relatórios anuais, estatísticas e monografias, que podem ter muito interesse, especialmente para países em desenvolvimento.

- Da O.M.M.

Outro organismo das Nações Unidas com grande interesse para os estudos geográficos é a Organização Meteorológica Mundial. Para lá de estudos e análises publica também alguma informação estatística sobre o clima.

Muitas outras organizações internacionais, como a Organização dos Estados Africanos, a OPEP, a Cruz Vermelha, a Organização Mundial de saúde, etc., publicam estatísticas comparativas internacionais.

A importância das estatísticas internacionais deriva de servirem para enquadrar os valores que encontramos nos nossos trabalhos e análises e permitir avaliá-los, em termos da sua ocorrência no mundo e na região.

2.1.3 - Outras fontes documentais

Apresentaram-se anteriormente as publicações fundamentais em que é mais comum procurar e recolher informação.

Importa não esquecer outras fontes menos tradicionais de informação publicada, como são os mapas, as fotografias aéreas e as imagens de satélite, assim como as monografias e os estudos sectoriais e territoriais.

Mas é preciso não esquecer que só uma pequena parte da informação recolhida é publicada.

Os institutos de estatística recolhem muito mais informação do que aquela que tratam; e da que tratam apenas publicam uma pequeníssima parte. Em alguns casos vendem essa informação inédita, noutras casos ela fica inacessível.

O acesso à informação é um direito das sociedades, especialmente importante nos países em desenvolvimento.

Os processos de recolha são caros e o dinheiro que os viabiliza vem do orçamento nacional e por isso a informação não é propriedade dos institutos estatísticos especializados, mas de todos, e esses institutos têm o dever de disponibilizar a informação a todos os interessados. Os dois grandes argumentos são a democracia e a reciprocidade.

Como conclusão, pode afirmar-se que "Como todos dão informação, todos têm direito a conhecê-la", e que "Um país só se pode desenvolver se estiver bem informado sobre si próprio".

2.1.3.1 - Dados não publicados

Quanto aos organismos estatísticos, para além da completa exploração dos dados publicados, o investigador deve procurar ter acesso aos dados não publicados.

Todas as instituições de recolha, tratamento e divulgação da informação recolhem mais informação do que a que tratam e muito mais do que a que publicam.

Publicam apenas uma pequena parte da informação total que dispõem, a que mais interessa ao maior número de indivíduos e entidades, pois a publicação é cara e só é possível para grandes edições. Muita da informação que existe e interessa a estudantes e investigadores pode ter interesse para apenas para um reduzido número de interessados, pelo que a sua publicação seria um grave erro financeiro.

Mas deve ser disponibilizada para consulta, com os devidos cuidados, de que se salienta a garantia da confidencialidade da informação relativa a indivíduos e empresas isoladamente (o "segredo estatístico") e a garantia da segurança da informação (segurança física, tanto contra acções dolosas, como involuntárias).

Obviamente que a acessibilidade à informação deve ser total quando os trabalhos se desenrolam no interior dos organismos estatísticos, por conta do estado, servindo objectivos nacionais.

2.1.3.2 — Informação em arquivos administrativos e de gestão

Os arquivos de gestão corrente dos diversos organismos (e empresas) são outra fonte possível de informação. Nas sociedades actuais cada organismo de regulação administrativa, todas as entidades públicas e todas as empresas têm de manter arquivos dos documentos que manuseiam e processam. Cada instituição gera um conjunto de documentos que são o reflexo do seu trabalho; por exemplo, a Universidade gera processos de alunos, de professores, de funcionários e passa certidões dos actos mais importantes que nela se processam.

Os progressos que se têm verificado nos últimos anos nas formas e meios de tratamento dos actos administrativos permitiram também o aproveitamento crescente destes actos para fins estatísticos. A utilização, que os organismos públicos têm vindo a fazer, de equipamentos informáticos com capacidade de cálculo e de armazenagem cada vez maiores e a constituição de sistemas genéricos e padronizados de identificação (como os números de identidade ou os de pessoa colectiva e as diversas nomenclaturas estatístico-administrativas) abriram perspectivas completamente novas para o aproveitamento dos actos administrativos para fins estatísticos.

Os diversos países têm vindo a desenvolver a cooperação institucional entre os órgãos da administração pública e os organismos responsáveis pela elaboração das estatísticas nacionais, o que deriva do facto das fontes administrativas apresentarem, face aos inquéritos estatísticos, grandes vantagens. Têm custos muito inferiores (por vezes nulos), dispensam a consulta dos informadores (os indivíduos ou empresas), cobrem todo o universo a que respeitam e, quando os procedimentos de exploração informática o permitem, têm altos níveis de disponibilidade (disponíveis em relação a qualquer momento e para qualquer nível de desagregação espacial).

Mas o aproveitamento dos actos administrativos também apresenta dificuldades. As mais relevantes são:

- As definições e os conceitos utilizados nos actos administrativos são muitas vezes pouco adaptados para as finalidades estatísticas;

- A qualidade da informação nem sempre é a desejável. O problema principal deriva das formas de tratamento da informação e da falta de validação dos apuramentos;

- As alterações legislativas modificaram obviamente a informação que é recolhida. Para alguns analistas que privilegiam a análise de séries, isto pode ser um problema. Mas é preciso ter em conta que na maioria dos casos são as alterações que se verificaram de facto na sociedade, o melhor indicador para a temática que se estuda no sistema.

A título de exemplo pode referir-se o problema da educação básica. Na generalidade dos países a educação básica obrigatória tem vindo legalmente a aumentar. Em Portugal começou por ser de três anos (anos 40), depois quatro (anos 60), foi de seis na década de setenta e actualmente é de nove. Coloca-se a questão: o que será mais interessante, conhecer o volume de alunos matriculados na terceira classe, de 1940 até à actualidade, ou ter o volume de alunos inscritos no ensino obrigatório, seja qual for o número de anos que este tem ?

2.1.3.3 — Integração de inquéritos estatísticos com fontes administrativas

Segundo Vilares (*op. cit.*), o aproveitamento directo dos ficheiros administrativos tem sido efectuados segundo três perspectivas complementares, que traduzem uma crescente cooperação entre os organismos públicos e a entidade nacional estatística.

A primeira perspectiva consiste em atribuir à entidade nacional de estatística a gestão técnica dos ficheiros administrativos que têm interesse estatístico. É então a esta que compete definir tecnicamente os conceitos, as nomenclaturas, os número de identificação e outros elementos técnicos que permitem que o acto administrativo satisfaça ao mesmo tempo os objectivos administrativos e os estatísticos.

É um modelo comum nos países escandinavos. Na Dinamarca a lei dá à entidade estatística nacional a capacidade de conectar a informação dos ficheiros administrativos, com a ajuda dos números de identificação nacionais. Neste país o sistema baseia-se em três números nacionais: o número do indivíduo, o do imóvel e o da empresa. Sem cada um destes número é impossível desenvolver qualquer actividade no país. A legislação permite à entidade estatística nacional juntar e relacionar os diversos ficheiros, relacionando os indivíduos entre si, com o trabalho e a residência. Nestes países é possível efectuar automaticamente inquéritos exaustivos à população, residência e trabalho em qualquer altura.

A segunda forma de abordar o problema é a constituição de comissões mistas com a função de criar e manter os meios técnicos de identificação comuns aos inquéritos estatísticos e aos ficheiros administrativos. Com delegações técnicas nos ministérios para gerir os ficheiros administrativos, possibilitam a fácil utilização estatística dos ficheiros para a gestão e a compatibilização da informação com as necessidades do sistema nacional. Esta cooperação reveste duas formas. Uma é a colocação de perguntas de carácter estatístico nos formulários administrativos, outra é o uso dos actos administrativos para construir os ficheiros de base dos inquéritos: o registo geral da população, o ficheiro de entidades colectivas, os cadernos eleitorais, os registos de prédios e de

alojamentos. Deve ter-se cuidado em não abusar da primeira destas formas, transformando os formulários em grandes inquéritos que por vezes nem são tratados posteriormente.

É o modelo utilizado em França. Existem várias comissões de coordenação (sendo de realçar a Comissão Nacional de Nomenclaturas de Actividades e Produtos e o Centro de Registo e de Revisão dos Formulários Administrativos) e os quadros superiores do instituto estatístico nacional estão destacados em vários serviços onde assumem a responsabilidade de exploração estatística dos ficheiros administrativos (Vilares, op. cit.).

A terceira perspectiva consiste na integração dos ficheiros administrativos com os inquéritos estatísticos, tanto para validação da informação, como para estabelecimento e controle das operações de recolha, quer ainda para actualização. Esta integração está adoptada em quase todos os países em que existem sistemas estatísticos desenvolvidos e juntam a informação administrativa exaustiva, com inquéritos por amostragem mais desenvolvidos.

É o caso de Portugal e do Canadá, em se cruzam inquéritos ao emprego, desemprego e rendimentos, com as declarações de trabalho, fiscais e da segurança social, conforme os casos.

2.1.3.4 - Insuficiência das fontes documentais

É evidente a importância que os ficheiros administrativos podem ter no melhor conhecimento dos sistemas estudados e a necessidade em utilizar a informação que contêm. Verifica-se que os organismos estatísticos nacionais procuram utilizar essa informação, segundo diferentes formas e modelos. A sua utilização directa pelos investigadores é mais complicada. Por um lado, os serviços não estão preparados para fornecer informação a particulares, mesmo quando com o estatuto de investigadores. Por outro lado, muitas vezes não estão autorizados, tanto pela existência restritiva de segredo profissional ou de segredo estatístico, como pela existência de legítimas protecções legais e restrições de segurança.

O acesso dos particulares é muitas vezes difícil e, mesmo quando é possível, coloca problemas éticos ao investigador quanto à sua divulgação. A norma é que os dados devem ser apresentados sempre muito agregados, tornando impossível a identificação ou até mesmo a sugestão da atribuição individual da informação a pessoas ou entidades, excepto quando exista expressa autorização destas.

Toda a informação já recolhida em ficheiros de administração e de gestão é importante e, numa investigação, antes de se proceder à recolha directa importa verificar a sua existência e as possibilidades e condições de utilização.

Mas apesar do que se disse, as fontes documentais, quer as publicadas quer as não publicadas, são muitas vezes insuficientes. E são insuficientes porque a informação que contêm não é adequada:

- umas vezes porque apresentam pouca especificação temática. Foi recolhida a informação que interessa à investigação, mas foi publicada ou tratada apenas segundo grandes temas, que não foram convenientemente esmiuçados para as necessidades da investigação;

- outras vezes a escala que não é a conveniente. O tema foi perguntado, mas o tratamento ou apresentação só foi efectuado para as pequenas escalas, o conjunto nacional ou as grandes regiões. Se o objecto espacial da investigação for uma área reduzida, por exemplo uma cidade ou concelhos, e a informação documental existir apenas sobre as províncias, não pode ser utilizada.

- outras ainda por que é inexistente. A informação sobre o tema que se pretende analisar nunca foi recolhida.

Todas estas circunstâncias obrigam a que o investigador tenha de proceder por si próprio, ou segundo a sua coordenação, à recolha directa da informação.

Tem a enorme vantagem de poder recolher a informação que quiser, segundo a forma que lhe seja mais conveniente. Porém, deve ter-se sempre em conta o grande volume de trabalho que acarreta e os elevadíssimos custos financeiros a que obriga.

Dependendo da legislação nacional, a recolha directa da informação pode ainda necessitar de autorizações especiais.

Por tudo isto, é especialmente adaptado a estudos locais ou de âmbito territorial reduzido e, mesmo assim, deve ter-se o cuidado de preparar e avaliar muito bem o plano de recolha, a fim de evitar o grave problema de não ter capacidade técnica ou financeira para concluir a operação de recolha.

2.2.- Observação

Muitas vezes pouco lembrada, a observação é dos melhores métodos de recolha de informação, sendo mesmo o mais utilizado nos processos de investigação em Geografia Física.

Em Ciências Sociais assume duas formas: a observação exterior e a observação participante

2.2.1 - A observação exterior

Na observação exterior, ou externa, o investigador não interfere com os acontecimentos ou processos que observa. Tem uma atitude passiva. No caso de processos sociais não interfere e é "invisível" ao meio social que está a observar.

A observação exterior de processos da flora, da fauna e das sociedades humanas exige ainda mais cuidados para não influir nos processos, o que é devido à maior interacção com o meio envolvente que esses sistemas apresentam.

É aplicável em todos os ramos da Geografia, Humana e Física. É o método mais utilizado para recolha de informação sobre processos naturais. Geralmente inerte nos processos geomorfológicos, climatológicos e

hidrológicos, pode em alguns casos concretos necessitar de cuidados especiais para não influenciar ou perturbar os processos observados.

Consiste em observar os factos relevantes da natureza, de indivíduos, de comunidades ou de organizações, e em registá-los de forma adequada. O registo é efectuado construindo uma ou várias matrizes de dados, que são posteriormente analisadas.

É um processo ao mesmo tempo objectivo e subjectivo, por se basear em escalas de categorias padrão, que são atribuídas pelo investigador.

Exige formação técnica adequada.

O objectivo é obter informação, que permita reconhecer relações e regularidades.

O investigador pode ser comparado a uma testemunha – um observador que está próximo mas fora do processo, como no caso de um acidente rodoviário.

2.2.2 - A observação participante

A observação participante é típica das Ciências Sociais, como é o caso da Etnologia, da Antropologia, da Sociologia e da Geografia Humana, Económica, Social e Política.

O investigador vive no sistema que observa, geralmente durante bastante tempo. Consiste em estar, viver e trabalhar com indivíduos, ou comunidades a fim de as compreender. É um procedimento natural, em que o observador se torna progressivamente membro da comunidade, permanecendo o mais próximo possível dos fenómenos e, desse modo, vai conhecendo e interiorizando os processos que observa e vive.

É um método baseado na observação e na participação, em que se tiram notas e se capta informação. É subjectivo. O investigador é, ao mesmo tempo, um observador e um participante. Colige informação, mas influi no meio em que passou a viver e tem as suas referências alteradas por ele. Originalmente a ênfase estava mais posta na observação, hoje, a maioria dos investigadores valoriza a participação.

O objectivo é conseguir um comentário compreensivo e alargado e uma apreciação dos comportamentos.

O investigador pode ser comparado a um comentador – um observador próximo, que comenta, avalia e esclarece o que se passa num processo em que participou, como é o caso de um ex-jogador a comentar um desafio.

As vantagens do método são a possibilidade de observar muito para além do que é visível, passando a conhecer e eventualmente a compreender as motivações dos procedimentos. O perigo é a possibilidade de observador se "converter" completamente, perdendo a distanciação e a capacidade de reflexão sobre os acontecimentos e processos que está vendo ou vivendo.

Deve tomar-se em atenção que a técnica de recolha de informação da observação não consiste apenas em ver. A observação é o meio pelo qual se dão respostas concretas (que devem constar de um questionário previamente elaborado) e se alteram algumas das questões, se o decorrer da observação vier a mostrar que foram inicialmente mal formuladas. O resultado final dum processo de observação é a atribuição, a cada unidade de análise ou elemento observado, de respostas a todas as perguntas que se colocaram.

2.3.- Experimentação

É uma técnica de recolha de informação sobre fenómenos ou processos que se controlam.

Quase que não se utiliza em Geografia mas é muito utilizada noutras ciências sociais, como a Psicologia, a Sociologia e as Ciências da Educação.

É um processo em que as populações em estudo são divididas em vários grupos e algumas das variáveis são manipuladas em alguns deles. Neste caso toma duas formas: uma mais controlada, que decorre em ambiente próprio, e outra que se desenrola no terreno, na rua ou no ambiente real da situação analisada, com mais imprevisto.

As experiências no terreno são representações teatralizadas, em que se simula as situações que se pretendem estudar. São sempre provocatórias, pois pretendem analisar a reacção dos indivíduos aos estímulos. Por vezes correm mal, no sentido que a reacção dos elementos no terreno que ignoram estar a ser sujeitos à experiência podem exceder as possibilidades de controle.

Esta técnica é também muito utilizada em ciências da natureza, em laboratórios em que se reproduzem e simulam fenómenos e processos naturais, sendo objecto dum ramo específico do conhecimento, a modelística.

2.4.- Análise Documental

É um tipo de recolha de informação que analisa preferencialmente o texto dos documentos produzidos por pessoas, organizações e comunidades.

Estuda os textos produzidos pelos meios de comunicação social de massa, os jornais, revistas, televisão, mas também textos produzidos para tipos de documentos próprios, como declarações, discursos e comunicados.

De uma forma geral, qualquer tipo de texto pode ser objecto de uma análise documental.

É uma técnica por vezes muito interessante, embora seja pouco utilizada em Geografia. Por exemplo, a escola de Geografia de Lisboa fez

uma análise geográfica do conteúdo dos discursos dos candidatos nas primeiras eleições livres em Portugal (para a Assembleia Constituinte, em 25 de Abril de 1975).

A análise documental apresenta-se segundo os seguintes tipos:

a).- Análise qualitativa - em que se analisam as citações, o vocabulário, as figuras de estilo ou o tipo de construção das frases utilizadas.

b).- Análise quantitativa - que se divide em semântica quantitativa, na qual se contam o vocabulário, ou as estruturas verbais utilizadas, e em análise de conteúdo, baseada na ocorrência de palavras, conceitos e ideias-força presentes nos textos a analisar.

2.5.- Inquéritos

É o processo de recolha de informação em que o investigador se dirige directamente às populações, perguntando-lhes o que elas sabem e ele deseja conhecer. O inquérito pode ser dirigido à população em geral, ou a agentes determinados, que detêm posições particulares no sistema que se pretende analisar, como directores, presidentes, chefes, gestores, funcionários, professores, etc., agentes que se admite terem conhecimentos específicos sobre o funcionamento do sistema em que estão inseridos.

Os inquéritos podem ser entrevistas ou questionários.

2.5.1 - Entrevistas

As entrevistas são uma técnica de recolha de informação em que o investigador conversando com o entrevistado, procura obter deste a informação que lhe interessa.

Todas as entrevistas são interacções verbais, em que uma pessoa, o investigador, procura extrair informação de outras pessoas, fazendo-lhes perguntas.

Segundo Chisnall (1986), citado por Vicente *et al.* (1996,120) uma entrevista é "Uma conversa com um propósito".

É um processo que também é empregue fora do domínio da pesquisa científica, constituindo um dos processos essenciais do jornalismo escrito e falado.

2.5.1.1 - Tipos de Entrevistas

Segundo Duverger (1964, 252), as entrevistas podem ser documentais ou de opinião. São documentais quando se interrogam os indivíduos sobre o que eles são ou sobre o que fazem e de opinião quando se procura saber o que sabem. Estas últimas são importantes em regiões, sociedades ou circunstâncias em que a falta de registos escritos obriga a

utilizar as memórias das pessoas como fonte principal de informação. Pode assim classificar-se fenómenos naturais ou sociais, como cheias, tremores de terra, secas, ou acontecimentos como eleições, conflitos sociais, etc., de que não existem quaisquer registos ou medidas.

A mesma classificação aparece noutros autores, com as designações de entrevistas de opinião, quando se quer conhecer a posição do entrevistado sobre um estímulo ou acontecimento qualquer, ou entrevistas de facto, quando se interroga sobre acontecimentos.

As entrevistas podem ser estruturas, semi-estruturadas e livres.

A entrevista estruturada consiste numa série de perguntas fixas, predeterminadas e estandardizadas (constantes numa lista escrita ou decorada) que o entrevistador vai fazendo sempre na mesma ordem, numa forma muito aproximada do questionário. O objectivo é preencher uma folha de respostas, fisicamente existente ou apenas na memória do investigador. Quando a resposta directa não foi respondida ou o investigador considera que a resposta dada não é suficiente, a entrevista pode estender-se por outras perguntas, até que seja obtida a resposta desejada.

Nas entrevistas semi-estruturadas, o investigador prepara também uma lista de perguntas, com que se inicia a entrevista. Mas o investigador não se limita às perguntas que leva e permite que a conversa com o entrevistado evolua para outros aspectos que eles considerem relevantes. Feita geralmente a um só indivíduo, pode ser aplicada a grupos pequenos (2 a 4 pessoas no máximo).

O objectivo é obter informação tanto para as questões que se levava preparadas como para aquelas que se revelaram importantes no decurso da conversação.

Nas entrevistas informais ou livres a conversa decorre consoante a vontade dos intervenientes. Não existe lista de perguntas, mas apenas um tema para discussão ou apresentação. É típica da recolha de histórias ou da descrição de acontecimentos, tanto sociais como naturais (ex. descrição do que se sentiu perante uma catástrofe natural). O decorrer da entrevista, tanto no que respeita à ordem de apresentação, como à escolha e valorização relativa dos diversos sub-temas, é mais determinado pelo entrevistado do que pela eventual lista de interesses do entrevistador.

Em alguns casos, a entrevista pode ser enviada por escrito algum tempo antes a fim do entrevistado preparar as respostas.

Convém salientar que todos os tipos de entrevistas são procedimentos em que se procura recolher informação, residindo a diferença na forma de recolha, que está relacionada com o tipo de abordagem da investigação, mais extensiva ou mais intensiva.

2.5.1.2 - Vantagens e limitações das entrevistas

A grande vantagem da entrevista está na flexibilidade da interacção entre entrevistador e entrevistado. Podem ser tiradas dúvidas e dados esclarecimentos complementares, tanto sobre as perguntas como sobre as respostas. Por isso, as entrevistas são muito ricas de informação e muito precisas.

A principal limitação deriva da dificuldade em classificar ou tipificar as respostas obtidas. Tão ricas são que obrigam o investigador a simplificações, por vezes radicais, da informação obtida.

As respostas obtidas por entrevista de opinião são sempre subjectivas, mas as entrevistas de facto também o são. Isto pode ser uma fragilidade, se o objectivo era obter um "retrato fiel do acontecimento", mas é uma vantagem se o objectivo for conhecer o modo como as sociedades vivem os acontecimentos.

Uma das mais importantes vertentes dos estudos geográficos, mesmo em aspectos da natureza, é isso mesmo, não conhecer como as coisas são, mas como os diferentes indivíduos julgam ou acreditam que são.

Por exemplo, na década de setenta procurou-se estudar o incómodo que o trânsito de aviões no aeroporto de Lisboa provocava na população da cidade. O estudo iniciou-se no LNEC através de medições da intensidade dos ruídos (com decibelímetros) efectuados por físicos e engenheiros, mas cedo se verificou que a intensidade do ruído pouco tinha a ver com as queixas das populações, ou seja, pouco tinha a ver com a incomodidade que a população sentia. O estudo evoluiu para um inquérito à população efectuado por psicólogos que verificaram que o incómodo tinha a ver com a classe social (os ricos queixavam-se mais que os pobres) e que o maior incómodo se verificava nas áreas de fronteira, onde o ruído era sentido somente de vez em quando e não nas áreas onde era mais forte.

2.5.1.3 – O Esquema da Entrevista

A entrevista pode ser realizada directamente ou por via telefónica. E, embora possam ser gravadas, em geral não vale a pena. Só quando se esperam entrevistas muito ricas em informação, como é o caso das histórias de vida, é que vale a pena efectuar gravações.

O mais adequado é utilizar um formulário com as questões para que se procura resposta (as variáveis que se irão tratar posteriormente) que o entrevistador procura preencher, orientando a conversa que faz com o entrevistado nesse sentido.

O melhor preenchimento do formulário, a que tem de obedecer o esquema da entrevista, é o objectivo fundamental do entrevistador.

2.5.1.4 - A realização da entrevista

A realização da entrevista obriga a ter alguns cuidados nas diversas fases do trabalho.

É necessário encontrar os indivíduos a entrevistar, tal como consta do plano de amostragem ou do plano de selecção de actores que se seguiu,

desenvolvendo todos os esforços para obter respostas, assegurando que se cumpre esse plano.

Deve realizar-se as perguntas segundo as instruções, procurando respostas válidas e fiáveis, que devem ser imediata e correctamente registadas, o que é difícil quando a entrevista decorre com vivacidade e interesse.

Tem de se minimizar os erros. A fonte principal de erros tanto vem do entrevistado como do entrevistador. Entre muitas outras, o entrevistado pode ser incapaz de responder, por ignorância, falta de memória (Vicente, *op cit*, 124) ou incapacidade de verbalização adequada à descrição do que se pergunta. Pode ainda esconder aspectos que julgue serem prejudiciais ou pouco abonatórios para si, ou pelo contrário sobrevalorizar aspectos a que dê especial relevância. Em caso de ignorância há uma forte tendência para responder consoante o que pensa ser habitual.

Numa série de entrevistas efectuadas numa área rural junto a Lisboa sobre a que localidades se deslocavam quando precisavam de ir à farmácia, os indivíduos que nunca tinham ido a nenhuma farmácia respondiam com o nome de localidades que eles consideravam ter importância para ter farmácia, embora mais tarde se viesse a verificar que não a tinham.

Outras fontes de erros são a relutância de alguns indivíduos em responder a perguntas de natureza pessoal, a falta de tempo, a antipatia ou o desejo de agradar ao entrevistador.

Na medida do possível, deve procurar esclarecer-se o entrevistado, tendo em conta que nas entrevistas de opinião o fundamental é deixar o entrevistado expressar a sua opinião. Por isso não se lhe devem dar esclarecimentos, nem tentar "canalizá-lo para o tema certo da questão", pois o que interessa é mesmo a sua opinião e não a opinião corrigida pelo entrevistador.

2.5.2 - Questionários

O questionário consiste num conjunto de perguntas, geralmente escritas, que são colocadas ao inquirido e cujas respostas são registadas pelo próprio ou pelo inquiridor.

Os questionários presenciais são na realidade uma entrevista especial, com perguntas e comportamentos fixos. É apenas por facilidade de exposição que é costume distinguir o questionário como uma categoria independente de inquérito.

Difere da entrevista no sentido de que o papel do inquiridor é neutro. Não deve explicar as perguntas, não deve ajudar a responder, não deve esclarecer. A sua intervenção deve ser mínima e sempre igual para todos os inquiridos.

Os questionários consistem numa série de perguntas. Deve tomar-se em atenção que a ordem em que as perguntas são feitas é um factor que afecta significativamente as respostas, conforme demonstrou Duverger (1964, p. 199-233).

A preparação dum questionário é uma operação complexa e delicada.

2.5.2.1 - Tipos de questionários

Quanto à presença dos inquiridores o questionário pode ser:

questionário presencial, quando o inquiridor está presente. O contacto entre quem inquirir e quem é inquirido é pessoal. Geralmente é o inquiridor que faz as perguntas e regista as respostas. Obriga a deslocações e estadias que podem ser significativas se os elementos a contactar estão dispersos por uma grande região. É o melhor método de recolha, embora seja demorado, trabalhoso e caro.

questionário telefónico, é similar ao presencial, mas o inquirido é contactado pelo telefone. Tem as vantagens da facilidade e baixo custo, porque evita deslocações e permite efectuar um número muito elevado de contactos em pouco tempo. Os inconvenientes são a dependência total de formas de expressão oral, não sendo possível comunicar gráfica ou cartograficamente, a menos eficaz comunicação que se estabelece entre inquirido e inquiridor (muitos consideram uma forma pouco nobre e educada de recolher informação) e principalmente a dependência da existência de redes telefónicas. Este último facto faz com o método só possa ser utilizado com algum sucesso quando o uso de telefone é banal no universo inquirido (pode ser o caso de inquiridos a grandes empresas ou a categorias especiais da população, embaixadores, ministros, etc.). Para inquiridos à população em geral só devem ser usados em países ou regiões em que a posse e uso do telefone seja um facto banal.

- questionário não presencial, quando é o inquirido que preenche o questionário. Obriga a que o próprio saiba ler e escrever, ou que alguém das suas relações o faça por ele. O formulário do inquirido, depois de preenchido, é devolvido. Pode ser enviado e recolhido através de agentes ou pelo correio. Quando se utilizam agentes, estes podem ser funcionários administrativos ou indivíduos com posições importantes nas comunidades, como padres, professores, chefes. É muito comum e eficaz a distribuição da entrega e a recolha ser feita através dos alunos das escolas, mesmo quando as perguntas se dirigem às famílias.

- questionário pelo correio, que é um caso especial de questionário não presencial, em que as respostas são enviadas pelo correio. A entrega dos questionários tanto pode ser feita por correio, como pessoalmente, por agentes do investigador, de uma forma sistemática. Pela forma distante como o inquirido lhe chega às mãos, geralmente o inquirido não se sente muito motivado para responder. O inquiridor não está presente nem na altura do preenchimento do inquirido nem na ocasião do envio, o inquirido sente-se em geral muito pouco interessado, o que leva a uma taxa de respostas muito baixa.

Relativamente às perguntas do questionário, estas podem ser:

- perguntas fechadas, quando o indivíduo a quem ela se dirige apenas pode responder por sim ou por não, ou é colocado perante alternativas explícitas. Por exemplo a pergunta "É aluno da Universidade ?" tem apenas como resposta sim ou não.

Um caso particular de perguntas fechadas são as perguntas com alternativas, em que se coloca ao inquirido uma série de alternativas, dentre as quais ele faz a sua escolha. As alternativas podem ser exclusivas, quando apenas se pode escolher uma do conjunto (e o inquirido escolhe aquela a que mais se adapta), ou de escolha múltipla, em que o inquirido tem a liberdade de optar por mais do que uma das alternativas de resposta que lhe são apresentadas. Este último caso obriga a alguns cuidados no apuramento e tratamento dos dados.

perguntas abertas, quando a resposta é livre e o inquirido pode responder como entender. Por exemplo, à pergunta "O que acha do sistema de transportes públicos da cidade?" a resposta pode variar desde "É bom" ou "É mau", até um texto apreciativo de várias páginas.

perguntas semiabertas, quando existe uma série de respostas possíveis fechadas, mas se liberta o inquirido, quer deixando-lhe a alternativa de "Outras situações:", em que ele pode optar por categorias que não foram consideradas na pergunta fechada, quer permitindo-lhe justificar as opções que tomou ou foi obrigado a tomar.

As perguntas abertas e as fechadas têm vantagens e inconvenientes. As perguntas abertas são mais ricas e menos superficiais, permitem perceber melhor as atitudes e opiniões dos indivíduos e as suas causas. As perguntas fechadas, por seu lado, são muito mais fáceis de apurar, não colocam problemas de interpretação ao investigador. As perguntas fechadas são incómodas para os inquiridos porque os obrigam a um esforço de integração numa grelha que eles, por vezes, não consideram adaptada a si próprios, mas as perguntas abertas vão colocar o mesmo problema de classificação ao investigador. As perguntas abertas têm ainda uma vantagem adicional, que é a de permitirem utilizar no relatório pedaços do texto das respostas que os inquiridos escreveram, o que pode enriquecer bastante o trabalho de análise a que se procede.

As perguntas semi-abertas devem ser consideradas, especialmente quando não foram efectuados pré-inquéritos de levantamento exaustivo das situações. Podem ser causa de alguma distorção, pois alguns dos inquiridos procurarão cingir-se e adaptar-se às escolhas que o inquérito lhes coloca, enquanto outro tipo de indivíduos está sempre mais preocupado em integrar-se em situações muito diferenciadas dos outros.

Na prática, um bom questionário deve conter perguntas dos diversos tipos, pois as fechadas permitem caracterizar facilmente as situações tipo que se verificam no sistema e as abertas permitem completar e corrigir as outras e perceber melhor por que razão os inquiridos fazem aquelas escolhas nas perguntas de resposta fechada.

Duverger (1964, 203-206) refere ainda que as questões podem ser diferenciadas consoante a sua natureza em:

perguntas de acção ou de facto, quando se pretende saber alguma característica ou acontecimento do sistema em estudo, que o inquirido conhece e o inquiridor não. Pergunta-se às pessoas factos como a idade, a profissão, o seu rendimento, o nível de instrução,

onde compra bens, etc. As repostas são fáceis e precisas e geralmente a sinceridade das repostas é grande. Tudo factos ou acontecimentos que ele pratica ou conhece. Nas perguntas de acção, em que se interrogam os indivíduos se eles fizeram ou não determinada acção, como se votou, quando se inscreveu, se foi ou não a determinado local em tal dia, que opções tomou numa ou outra ocasião, as dificuldades em obter respostas fiáveis são geralmente maiores.

perguntas de opinião ou de intenção, quando o objectivo não é saber como o inquirido agiu (o que seria uma pergunta de facto), mas como ele procederia em determinadas circunstâncias se tivesse oportunidade para isso. A resposta é habitualmente menos precisa e por vezes é difícil ao inquirido posicionar-se face a situações hipotéticas que nunca experimentou na realidade.

perguntas-teste, que são uma variedade especial de perguntas, em que o importante não é a resposta aparente, mas o significado profundo ou oculto. Por exemplo, quando se quer saber o nível de rendimento dos indivíduos, mas a população é reticente em fornecer essa informação, pode interrogar-se sobre os hábitos de consumo: se tem rádio, televisão, telemóvel, carro, etc. Por outro lado convém colocar perguntas teste que permitam aferir do grau de sinceridade das respostas dos sujeitos, ou perguntando a mesma coisa de diferentes maneiras, ou fazendo duas ou mais perguntas cujas respostas se sabe terem elevadíssimas correlações entre si.

Para efeitos práticos convém ainda distinguir as perguntas de avaliação (Duverger, 1964: 202), em que se solicita ao inquirido que se posicione face a uma situação com um certo grau de intensidade, mais ou menos preciso. É uma pergunta em que a resposta obriga a uma quantificação, uma ponderação, segundo uma escala subjectiva. A escala pode ser ordinal, do tipo "Gosto muito, pouco, nada", ou de razão, onde o indivíduo dá uma pontuação à situação de referência, geralmente numa escala de 0 a 20, ou de 0 a 100 (percentagens).

2.5.2.2 - Vantagens e limitações dos questionários

A grande vantagem dos questionários reside na sistematização que permitem na fase de recolha da informação e na facilidade com que se efectuam os apuramentos.

Os questionários com perguntas fechadas são os melhores nestes dois aspectos. Ao forçarem a resposta dentre um conjunto explícito de alternativas, fazem com que o inquirido não se esqueça de nenhuma das alternativas. Por outro lado obrigam a resposta num universo de possibilidades que não é o deles, mas sim o de quem concebeu o inquérito.

As perguntas abertas exigem um trabalho de pesquisa ao inquirido, que tem de procurar inventariar todas as possibilidades e, como se disse atrás a propósito das entrevistas, são muito difíceis de tipificar e analisar posteriormente.

De uma forma geral, os questionários devem ser utilizados em inquéritos à população e nos casos em que os universos são relativamente homogêneos., enquanto as entrevistas se mostram mais adaptadas na recolha de informação referente a indivíduos que saiam da normalidade ou de alguma forma se isolem no conjunto em análise.

2.5.2.3 – O Desenho dos Questionários

Os questionários podem ser feitos directamente, por telefone ou por via postal.

Tem de se utilizar sempre um formulário escrito, com as questões para que se procura resposta (as variáveis que se irão tratar posteriormente), que o entrevistador ou o entrevistado preenche.

A apresentação das perguntas do questionário, quer seja feita pessoalmente, quer seja através de um pequeno texto introdutório, deve ser clara e sucinta e, no caso de haver entrevistador, ser feita sempre de modo semelhante. Os esclarecimentos complementares devem ser evitados, pois também podem enviesar as respostas, de acordo com as interpretações e perspectivas do entrevistador.

No caso das entrevistas pessoais, directas ou por telefone, é conveniente ter chaves de codificação abreviadas para as respostas, a fim de evitar quebras de tempo no decorrer do processo de recolha de informação.

2.5.2.4 - A Realização dos Questionários

O uso de questionários exige um razoável conhecimento das populações a inquirir e a sua elaboração, que tem de ser adequada ao objectivos e fins que se pretendem alcançar e às características da população a que se destina, é um factor fulcral para o sucesso de qualquer investigação.

Alguns factores simples devem ser tomados em conta:

- A compreensão - Tendo em conta que a formulação das perguntas influi sobre as respostas, o texto das perguntas deve ser simples e redigido usando termos que sejam familiares e compreensíveis para os inquiridos.

- A facilidade - As perguntas a colocar devem ser hierarquizadas e dispostas numa sequência lógica. Deve começar-se pelas questões mais abrangentes e passar depois para as mais específicas.

- A referenciação - Não se deve esquecer a identificação geográfica e/ou administrativa do local onde se fez o inquérito, assim como é indispensável anotar a data em que foi respondido.

- A clareza - Sempre que o inquiridor não é o próprio investigador que elaborou o inquérito, deve haver notas explicativas e instruções de resposta suficientemente detalhadas, a fim de estabelecer condições de relativa igualdade para as respostas obtidas por diferentes inquiridores, ou para auxiliar as respostas não presenciais.

- A identificação - Em muitos casos a identificação nominal precisa não é necessária na fase de apuramento. Pelo contrário, como o que interessa é identificar características ou opiniões de um conjunto de pessoas através de amostragens, não importa em absoluto o conhecimento pessoal da opinião de cada um dos respondentes. No mesmo sentido vai a garantia de confidencialidade das respostas: se não se sabe quem respondeu melhor se garante a confidencialidade.

Mas, na fase de recolha, por vezes é conveniente saber se determinado indivíduo, família ou alojamento já foi ou não inquirido. Outras vezes a ocorrência de respostas extraordinárias pode aconselhar a verificação ou repetição do inquérito. Nestes casos torna-se muito conveniente dispor da identificação do inquirido.

Nos inquéritos à população a identificação é um problema em aberto que deve ser resolvido conforme as circunstâncias locais o aconselhem.

No caso de recolha de informação de instituições (empresas, organismos oficiais) ou de agentes influentes no sistema a identificação precisa do respondente é indispensável, até como forma de salvaguarda do investigador e de garantia da fiabilidade da informação recolhida.

A adequação - O inquérito deve estar muito bem adequado ao que se quer saber. Devem evitar-se erros de principiante: querer saber tudo e fazer demasiadas perguntas, o que vai desmotivar quem responde. As consequências são ou o aumento de não-respostas ou as respostas rápidas e desinteressadas. Também é de evitar fazer perguntas insuficientes, quer seja em número, quer seja na forma como são expressas, pois não vão permitir obter a informação necessária.

A caracterização - Um inquérito faz-se para obter determinada informação e é sobre ela que se devem construir as perguntas a colocar. Mas, para além disso, o conhecimento prévio do sistema em análise, as teorias enquadratórias do tema a estudar e mesmo o senso comum, sugerem comportamentos diferenciados na população em estudo, consoante as suas características estruturais básicas, como podem ser o sexo, o nível de rendimento, a idade, a instrução, a dimensão das empresas, a localização, a filiação em associações cívicas, religiosas ou políticas, etc.. Torna-se então conveniente conhecer essas diferenças ou clivagens da população de modo a ser possível estabelecer e verificar hipóteses de causalidade com as diferentes opiniões, atitudes ou comportamentos.

Muitos outros aspectos podem ser considerados na formulação das questões, por influírem nas respostas: a tendência para responder "sim", o medo de algumas palavras e a submissão aos estereótipos, a influência positiva ou negativa de afirmações recentes de personalidades conhecidas, as simpatias e antipatias, etc. etc.

Quanto à execução dos trabalhos de recolha de informação propriamente dita, deve tomar-se em consideração o momento de execução do inquérito. Sendo especialmente importante nas grandes operações censitárias, é também aconselhável procurar que a execução dos trabalhos

de campo para recolha de informação decorram numa época conveniente, relativamente homogénea e bem determinada. Esta definição é um valor acrescentado à operação de recolha e aumenta a fiabilidade na informação.

Deve também ter-se em atenção o pessoal auxiliar da recolha de informação. Na maioria dos trabalhos de investigação académica não existe pessoal auxiliar de recolha de informação, a qual é recolhida directamente pelo investigador responsável pelo trabalho. Porém em projectos de maior dimensão e em muitas tarefas de recolha da vida real, é necessário utilizar pessoal auxiliar para recolher informação. Deve ter-se especial atenção à sua selecção e formação, assim como à forma de pagamento (à peça ou à hora e, neste caso, definição de qual o volume de trabalho médio a realizar) e muito especialmente com a verificação do trabalho feito e com o controle de qualidade da informação recolhida.

É muito conveniente que o pessoal auxiliar de recolha pertença ao mesmo grupo social e económico da população a inquirir. Por um lado porque pode ser uma ajuda económica não desprezível para a população em causa. Por outro porque é um factor importante para que a comunidade saiba mais sobre o processo de inquérito de que vai ser objecto, o que atenua as desconfianças ao mesmo tempo que permite um melhor controle da população sobre todo o processo, aumentando a democraticidade e a participação da população, que pode chegar a fazer sugestões sobre melhorias nos procedimentos. É claro que se assume que dos trabalhos de investigação, mesmo os mais académicos, deverão advir melhorias para a situação geral das populações envolvidas.

O trabalho de campo deve sempre ser acompanhado pelo investigador responsável pelo trabalho, que deste modo detectará os problemas e procurará a sua mais eficiente resolução.

2.6. – Os Grupos de Discussão

Os Grupos de Discussão ou de foco, o “*focus group*” dos autores de língua inglesa, constituem um método de recolha de informação que tem grandes semelhanças com as entrevistas semi-estruturadas. Consistem numa discussão, em grupo, de um tema sobre o qual se necessita de recolher informação, sobre o qual se pretende saber mais.

Falar com diversos indivíduos ao mesmo tempo é uma óptima maneira de recolher informação.

Um grupo de *focus* ou de foco é um grupo de pessoas, geralmente mais de 5 e menos de 10, que se encontram para falar de um determinado assunto que lhe é posto pelo investigador. A denominação de Grupo de discussão ou de foco, resulta do facto de se basear numa discussão focalizada sobre um tema restrito. O tema é apresentado inicialmente pelo investigador, que depois solicita a participação dos membros do grupo e modera as intervenções. O investigador deve assumir um papel não dirigista, mantendo a discussão dentro do tema, mas não encaminha a discussão, permitindo que o assunto seja abordado e explorado com o maior número de perspectivas diferentes possível.

Os grupos tanto podem ser homogéneos como ser constituídos à partida por indivíduos que perfilham opiniões e posições diferentes. Há vantagem em serem constituídos por indivíduos que se conhecem, por vezes até mesmo por amigos, por facilitar a discussão, com à vontade, das diversas perspectivas. As reuniões do grupo de foco demoram geralmente de uma a duas horas. (Longhurst,2003:120)

A sua principal característica é a mais valia que decorre da interacção que se estabelece entre os elementos do grupo. É essa interacção que é o elemento fundamental desta técnica de investigação e que a diferencia das entrevistas, em que a interacção é entre entrevistador e entrevistado. Os participantes influenciam-se uns aos outros pelas respostas que dão às afirmações e ideias que se colocam durante a discussão, estimulados pelos comentários ou questões do moderador (o investigador ou outra pessoa).

A informação fundamental produzida por esta técnica é transcrita das discussões do grupo (notas ou gravação), acrescida das anotações e reflexões do moderador e de outros observadores, caso existam.

Esta técnica de recolha de informação, que é geralmente muito viva, provoca um grande envolvimento nos participantes, que pode chegar até à mobilização para a intervenção social activa, nos temas ou processos discutidos. Uma grande vantagem é a possibilidade de gerar informação nova, informação de que, em muitos casos, nem o próprio investigador suspeitava.

A informação obtida, embora também possa ser de natureza quantitativa, é geralmente de natureza qualitativa.

Os geógrafos têm utilizado este método para obter informação variada. Em 1996, Zeigler, Brunn, and Johnson (1996) usaram-no para perceber as respostas das pessoas aos procedimentos de emergência durante um grande furacão; Burgess, em 1996, utilizou grupos de discussão e observação participante, para obter informação acerca dos factores que poderiam ser eventualmente inibitórios da visita a áreas húmidas e Armstrong e Lucinda Fonseca estão actualmente a utilizá-los para compreender os processos que conduzem ao sucesso das acções de empresarialidade nos imigrantes na área metropolitana de Lisboa.

2.7. – Os métodos Delphi

Os métodos de Delphi, ou de Delfos⁴, são procedimentos de recolha de informação típicos para a projecção e avaliação da evolução futura de processos em que os métodos mais tradicionais de previsão podem ser problemáticos, devido à dificuldade em integrar modificações bruscas do sistema, ou à instabilidade que tem vindo a apresentar no passado recente.

Consistem na consulta a uma série alargada de especialistas, que se pronunciam de forma sistematizada sobre um tema que lhes é proposto pelo coordenador da investigação.

⁴ - A denominação deriva do oráculo que na antiga Grécia existia em Delfos e fazia previsões sobre o futuro.

O processo decorre em várias fases. Numa primeira etapa é pedida a opinião individual de cada componente do grupo, geralmente sob a forma de um questionário escrito. Embora a resposta possa ser identificada, há por vezes vantagem em ser anónima. De seguida, as respostas são comunicadas a todos os intervenientes, quer por consulta directa de cópias dos questionários, quer através de súmulas preparadas pelo coordenador, quer dos dois modos em simultâneo. Com esta informação, inicia-se um ciclo de reuniões de discussão (preferencialmente presenciais, mas que podem ser à distância – por correspondência, áudio e video-conferência ou InterNet - em que discute, justifica, questiona e aprofunda as diversas perspectivas que surgiram nas respostas da fase inicial. Assim que se considera que a discussão ilustrou e justificou suficientemente estas perspectivas, é novamente apresentado a todos os intervenientes um questionário semelhante ao inicial, em que é dada aos membros do painel, a oportunidade de nele integrar os consensos e os diferendos de opinião que a discussão revelou.

2.8. – As histórias de vida

As histórias de vida ou de acontecimentos são na actualidade uma forma muito interessante de conhecer e preservar factos e processos sociais e territoriais que habitualmente não deixam registo histórico. É assim possível obter e guardar informação sobre processos que afectaram as sociedades, por vezes de forma muito intensa, mas de que pouco se conhece, devido à iliteracia dos participantes ou à própria natureza dos factos, muitas vezes de natureza privada, clandestina ou até ilegal.

Este método consiste na recolha detalhada de descrições e de testemunhos, em primeira ou segunda mão, de factos ou acontecimentos insuficientemente documentados. Pode efectuar-se através de notas recolhidas pelo investigador ou por gravação integral, áudio ou vídeo, do conteúdo da história que é contada. A recolha pode ser feita em ambiente de grupo (pequeno), mas geralmente é feita individualmente.

É especialmente interessante para conhecer e compreender aspectos da Geografia Humana tais como a imigração, lutas e participações sociais e conflitos territoriais, a diversas escalas.

3. – Considerações finais

A recolha da informação consiste na série de operações que se inicia com a procura e selecção das fontes adequadas e termina com a validação da informação recolhida. Isto tanto se verifica quando se utilizam fontes documentais, observação, experimentação, questionários e entrevistas individualmente, ou qualquer combinação destas operações.

3.1.- Etapas de uma Operação de Recolha

Inicia-se pelo estudo do problema e pela identificação prévia da informação que se considera necessária obter. O nível da investigação (mais descritivo ou compreensivo) vai permitir definir melhor a informação de que se necessita.

Depois, começa o trabalho de pesquisa da informação documental existente. De acordo com ela e com as necessidades, seleccionam-se as áreas de estudo e as unidades de análise que vão ser utilizadas em cada subconjunto do trabalho de investigação: população em geral, agentes qualificados, empresas, organizações, entidades e territórios.

O plano da observação consiste na definição das características das diversas recolhas a que se vai proceder. Tanto nas recolhas documentais como nas directas, tem de se definir as unidades de análise de cada área observada, os *case-studies*, assim como as variáveis e/ou indicadores e as respectivas escalas e unidades de medida. Importa ainda definir para cada subconjunto se se vai utilizar inquéritos exaustivos ou por amostragem e proceder à elaboração dos questionários-base que permitirão a construção de variáveis e indicadores. Ainda tem de se definir o método de amostragem, se for caso disso e seleccionar os elementos a inquirir ou observar.

Com o plano de recolha de informação deve fazer-se um plano prévio de apuramentos, enumerando os tipos de suportes e estruturas de armazenagem de dados e os quadros de frequência e as matrizes de dados a construir. E também os tratamentos analíticos, lógicos, matemáticos, gráficos e cartográficos, com que se espera vir a explorar a informação recolhida. É óbvio que este plano será muito alterado com o decorrer do trabalho de análise, influenciado pelas diversas descobertas que se forem fazendo.

A seguir efectuam-se as operações de recolha de informação, eventualmente diferentes para cada subconjunto, tendo em conta a necessidade de recorrer ou não a pessoal auxiliar.

Toda a informação recolhida, seja qual for o nível ou forma da operação de recolha, assume a forma geral de matriz de dados (ver o capítulo correspondente), geralmente de atributos. A sua armazenagem pode, portanto, ser feita em bases de dados, folhas de cálculo e quadros ou matrizes de dados em papel.

Nesta fase importa codificar e validar a informação. O controle de qualidade obriga a verificar o correcto preenchimento das folhas de recolha, as não respostas e as respostas "estranhas", se o trabalho de recolha não foi efectuado pelo investigador.

A codificação, já preparada anteriormente, ocorre nesta fase e pode tornar-se difícil no caso das entrevistas ou das perguntas abertas dos questionários. É muito importante a utilização de grelhas de classificação, bem adaptadas ao caso em estudo e com algum grau de generalidade,

podendo aproveitar-se as nomenclaturas estatísticas nacionais ou internacionais⁵, se tal for conveniente.

As operações de classificação levam à construção de matrizes em que os dados se apresentam sob a forma de estados, enquanto a conservação da informação nas escalas em que foram medidos conduz a matrizes de valores (Fig. 3.1).

3.2.- Procedimentos Posteriores

Os procedimentos posteriores são operações que fazem já parte dos tratamentos e vão incidir especialmente (Figura 3.1):

- Sobre a transformação de variáveis, em que se muda a natureza das escalas de medida, se agrupa ou classifica-se ou se procede à construção de indicadores mais ou menos complexos (Fig. 3.1);

- Sobre uma variável de cada vez (uma coluna das matrizes), em que se faz mapas e gráficos ou se calcula os parâmetros da distribuição estatística de cada variável;

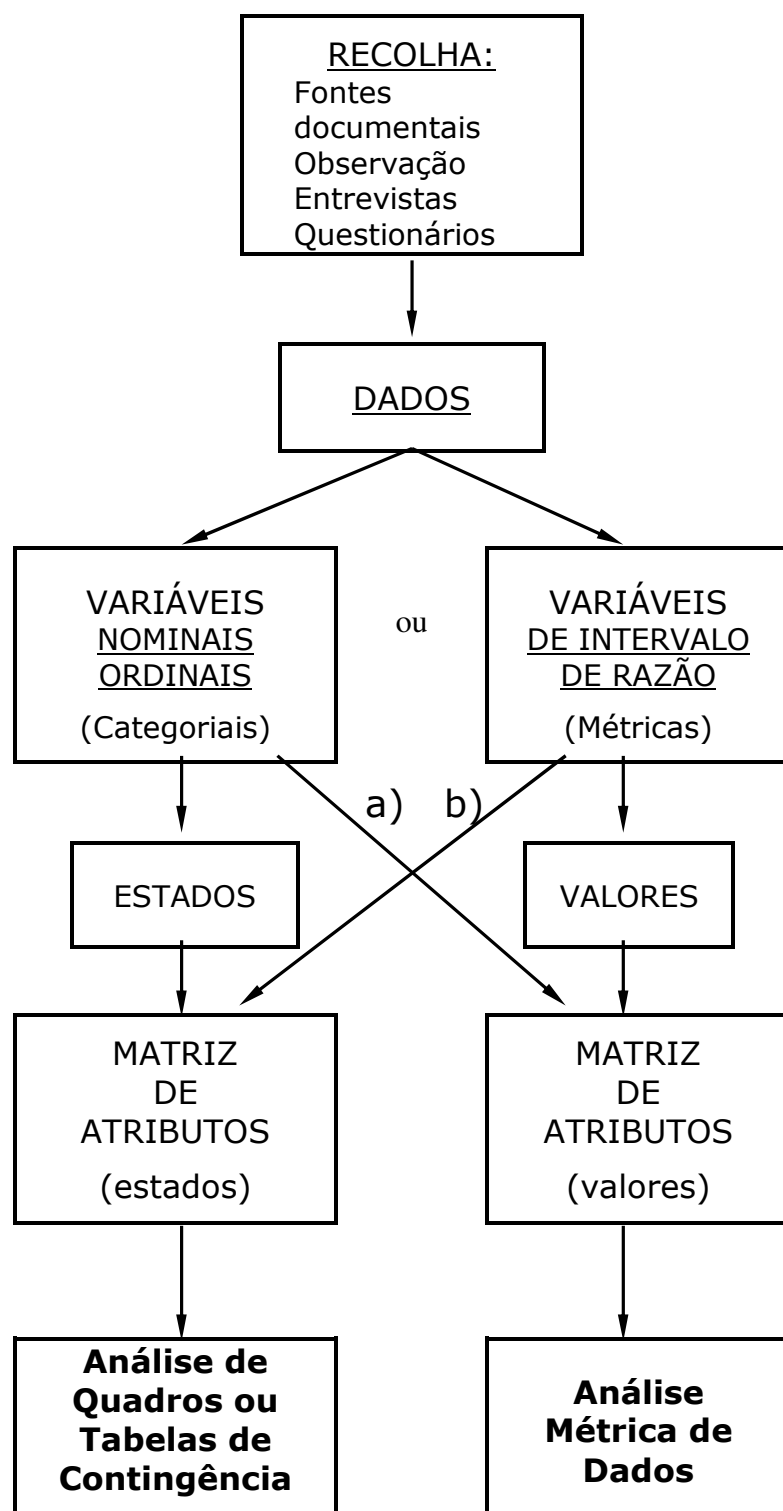
- Sobre duas variáveis ao mesmo tempo (duas colunas das matrizes), calcula-se a associação ou as relações entre as variáveis, estudam-se as tendências e os resíduos, faz-se gráficos e mapas com a combinação das duas variáveis;

- Sobre muitas variáveis simultaneamente (colunas das matrizes), usam-se métodos de análise multivariada, como a regressão múltipla, análise factorial, Q-análise ou outros;

- Sobre as unidades de análise (as linhas das matrizes), analisa-se as proximidades e as diferenças, estabelece-se grupos e classes a partir das variáveis, ou procura-se quais as variáveis que contribuem para a separação em grupos;

- Sobre quadros de frequência de estados das variáveis ou das unidades de análise calcula-se medidas de proximidade, associação ou predição.

⁵ - As nomenclaturas são classificações gerais sobre diversos aspectos estatísticos, algumas adoptadas internacionalmente por um grande número de países, outras estabelecidas pelos estados nacionais, por proposta das respectivas entidades estatísticas. Constan de conceitos e definições unívocas e bem definidos que permitem uma fácil comparação com os dados nacionais e internacionais.



a) – os estados são assumidos como valores
b) – classificação (os valores passam a estados)

Fig. 3.1 – Operações de Recolha da Informação

4. - Amostragem

4.1. – Introdução

As populações infinitas, as que têm um número infinito ou muito grande de elementos, não podem ser analisadas na sua totalidade. É preciso estudar estas populações através de subconjuntos representativos, as amostras. Mesmo em populações finitas é muitas vezes inútil inquirir todos os seus elementos.

Quando se desejam conhecer as características de uma população recolhe-se informação disponível ou fazem-se observações ou inquéritos.

Se o objectivo é conhecer o valor exacto dessas características tem de se recolher informação sobre todos os elementos da população.

Mas se o universo é infinito⁶ ou se o seu número de elementos é muito grande ou ainda se nos satisfazemos com um valor aproximado para as características, não se torna necessário inquirir todos os elementos da população, mas apenas uma parte, que em alguns casos pode ser muito pequena.

Importa porém clarificar o que se entende por valor aproximado de uma característica. É um valor que não é o exacto, diferindo deste por uma certa quantidade - o erro.

$$\text{Valor exacto} = \text{Valor aproximado} \pm \text{Erro}$$

O objectivo dos métodos estatísticos neste capítulo da amostragem vai ser, então, o de conhecer um valor aproximado e os limites do erro que lhe está associado.

Na prática verifica-se que para erros razoavelmente pequenos, a dimensão da amostra, o seu número de unidades de análise pode ser muito pequeno relativamente à população.

Assim se manifesta o grande interesse dos estudos sobre amostragem, pois recolher informação é muito trabalhoso e caro e em muitos casos, para os objectivos dos nossos estudos, são perfeitamente aceitáveis valores aproximados, desde que os limites do erro sejam pequenos. Se recolher informação a demasiadas unidades de análise leva a diminuir os erros, a verdade é que acarreta gastos muito maiores e um apreciável aumento dos trabalhos de armazenagem e apuramento, sendo o balanço custos/benefícios raramente positivo para as recolhas exaustivas.

4.2. – Conceitos básicos

Vamos definir alguns dos conceitos básicos que a seguir se vão utilizar, na definição dos procedimentos de amostragem.

⁶ - População infinita em sentido estatístico tanto é a população matematicamente estatística como aquela que tem um número de elementos ou unidades de análise muito grande.

4.2.1 – Amostra

A amostra de uma população é um subconjunto de elementos dessa população que tem as mesmas características que a população original. Nesse sentido, é costume dizer-se que uma amostra é um subconjunto representativo duma população.

Se um subconjunto não é representativo da população, se não tem as mesmas características que esta, não é uma amostra. Como geralmente não se conhecem os valores exactos da população, não se podem comparar (*a posteriori*) com os valores determinados no subconjunto, e por isso ignora-se se é ou não uma amostra. Apenas a utilização de procedimentos standardizados poderá garantir (sem verificação) que o subconjunto de elementos que temos é uma amostra. Estes procedimentos respeitam à determinação do número de elementos da amostra e aos processos de selecção desses elementos dentro da população.

Só com a utilização destes procedimentos se garante que se está perante uma amostra, isto é, perante um subconjunto representativo da população em análise.

4.2.2 – Valor aproximado

O valor aproximado de uma variável para uma dada população é o valor que resulta do cálculo dessa variável numa amostra da população. Como se disse, difere do valor exacto, que é geralmente desconhecido, por uma certa quantidade, o erro. Desconhecendo-se o valor exacto é óbvio que não se conhece o erro. O que se pode é colocar limites para os erro, através da utilização de procedimentos estatísticos.

4.2.3 – Limites do erro

Ao definir a dimensão de uma amostra, está-se também a definir os limites do erro, a diferença aceitável entre o valor estimado de uma variável determinado a partir da amostra e o valor exacto que a variável tem na população.

O erro é medido nas mesmas unidades de medida que a variável.

Os limites do erro são valores acima e abaixo do valor aproximado que determinámos, que definem uma área na qual se crê que o valor exacto se situa, embora exista sempre uma possibilidade, reduzida e calculável, de que na realidade não seja assim, isto é, de que o valor exacto esteja fora dos limites do erro.

4.2.4 – Risco

O risco é precisamente a probabilidade que existe de que o valor exacto não se encontre no intervalo definido pelos limites do erro.

É medido em desvios-padrões (z-scores), na tabela da lei normal, geralmente em termos de significação de 0,95% e 0,99%.

Ao risco de 0,95% corresponde o facto geral médio de em 100 medições se encontrarem 95 em que o valor exacto se encontra dentro dos

limites do erro e 5 em que a diferença é maior. Em Estatística denomina-se este nível de certeza nos resultados, risco de errar, como significativo.

O risco de 0,99% garante que, em cada conjunto médio de 100 medições, em 99 a diferença entre os valores exactos e os da amostra é inferior ao limite do erro e apenas numa se excede esse valor. A este nível de certeza (ou de risco) dá-se o nome de muito significativo.

4.2.5 – Notações

Nas considerações que se seguem importa pois ter em consideração:

N - o tamanho da população

n - o tamanho da amostra

e - o erro (nas mesmas unidades de medida da variável)

Vc - o valor calculado (na amostra)

Ve - o valor exacto (da amostra)

z - o risco, ou seja, a probabilidade de que o valor exacto (**Ve**) e o valor calculado (**Vc**) se situem a uma distância superior ao erro (**e**).

4.3 – Estimativas

Como se viu, as características calculadas numa amostra não são exactamente iguais às características da respectiva população. Por exemplo, a média mostral não é geralmente igual à média da população de que foi retirada a amostra. Contudo, os erros, ou seja, as diferenças entre estas duas médias, numa grande série de amostra, segue a lei de distribuição normal.

Assim sendo, considera-se a noção de erro-padrão da estimativa, valor que indica a probabilidade da média verdadeira da população se encontrar num intervalo dado em torno da média mostral.

Quando as populações são infinitas (quer dizer, em sentido estatístico, quando são mesmo infinitas ou quando são muito grandes) o erro-padrão (**σ_m**) calcula-se do seguinte modo:

$$\sigma_m = \frac{\sigma}{\sqrt{n}} \quad \text{com } \sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

Quando as populações são pequenas, calcula-se do seguinte modo:

$$\sigma_m = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \quad \text{com } \sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

As diferenças nas fórmulas derivam do facto de nas grandes populações (as infinitas) a "saída" de um caso não alterar as probabilidades de os outros "saírem" posteriormente (significativamente, no caso das grandes e nada, no caso das realmente infinitas), enquanto nas pequenas populações essa possibilidade de alteração existe e assume valores importantes.

4.4. – Tamanho das Amostras

É a partir das relações anteriores que é possível determinar as dimensões das amostras, relativamente ao risco que se está disposto a correr.

4.4.1 – Populações Infinitas

Nas considerações que se seguem, importa ter em conta que nas populações infinitas as observações mostrais são independentes e seguem a lei normal, pelo que, se for:

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

em que σ_m erro-padrão da média
 σ desvio-padrão
 n tamanho da amostra

sendo o erro igual a

$$e = z \times \frac{\sigma}{\sqrt{n}}$$

onde **e** erro admissível em unidades da variável (para + e para -)
z o risco em z-scores (a probabilidade de não ocorrer).

deduz-se então que:

$$z^2 \times \sigma^2 = e^2 \times n$$

e torna-se possível determinar a dimensão da amostra, resolvendo a equação em ordem a **n**:

$$n = \frac{z^2 \times \sigma^2}{e^2}$$

4.4.2 – Populações Finitas

Neste caso, em que as populações são finitas, as fórmulas que permitem determinar o erro (**e**) são diferentes. Seguindo o mesmo procedimento e sabendo que

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

e que:

$$e = z \times \left(\frac{\sigma}{\sqrt{n}} \times \frac{\sqrt{N - n}}{\sqrt{N - 1}} \right)$$

através da combinação das duas expressões, obtém-se a expressão:

$$e^2 = z^2 \times \sigma^2 \times \frac{N - n}{n \times (N - 1)}$$

que, resolvida em ordem a **n**, permite determinar a dimensão da amostra, que é dada pela equação:

$$n = \frac{N \times z^2 \times \sigma^2}{(N - 1) \times e^2 + z^2 \times \sigma^2}$$

4.4.3 – Exemplos de cálculo da dimensão de amostras

De seguida exemplifica-se, através do cálculo de casos concretos, como se determina a dimensão de algumas amostras.

Exemplo 1

Tem-se uma população infinita que apresenta um desvio-padrão igual a 8,26. Se for aceite um erro de 2 e um nível de confiança de 95%, qual deverá ser a dimensão da amostra?

Como o valor de Z equivalente a 95% é 1,960,

$$n = Z^2 \times \frac{s^2}{e^2} \qquad n = 1,96^2 \times \frac{8,26^2}{2^2} = 65,5258 \approx 66$$

Resposta: a dimensão da amostra deverá ser de 66 indivíduos.

Exemplo 2

Mas se, para o mesmo exemplo, a população em vez de infinita for constituída por 200 elementos, qual será então a dimensão da amostra? O desvio-padrão é 8,26, o erro 2 e o nível de confiança de 95%, com um valor de Z equivalente de 1,960:

$$n = \frac{Z^2 \times s^2 \times N}{e^2 \times (N-1) + s^2 \times Z^2} =$$
$$n = \frac{1,96^2 \times 8,26^2 \times 200}{2^2 \times 199 + 1,96^2 \times 8,26^2} = 49,54208 \approx 50$$

Resposta: a amostra será de 50 indivíduos

Exemplo 3

É preciso determinar a dimensão de uma amostra para cada um dos casos abaixo indicados (A e B), de tal modo que a probabilidade de ter erros máximos de 0,5 seja igual a 0,001 (um por mil).

As características específicas de cada caso são:

Caso A).- População infinita; desvio-padrão mostral de 2,1;

Caso B).- População 18 distritos; desvio-padrão mostral de 2,1.

Qual a dimensão que devem ter as amostras?

Cálculos e resultado do caso A

Para $p = 0,001$ e $z = 3$ (correspondente a 0.001)

$$n = \frac{3^2 \times 2,1^2}{0,5^2} = \frac{39,69}{0,25} \approx 160$$

Resultado do caso B

Para um risco de $p = 0,001$ e $z = 3$ (correspondente a 0.001)

$$n = \frac{18 \times 3^2 \times 2,1^2}{17 \times 0,5^2 + 3^2 \times 2,1^2} = \frac{714,42}{43,94} = 16,26 \approx 16$$

Note-se as diferenças na dimensão das amostras, quer entre si, quer relativamente à população a que respeitam. Para o mesmo desvio-padrão e idênticos níveis de erro e risco, numa população infinita basta uma amostra de 160 elementos, enquanto que para uma população de dezoito distritos é necessário conhecer o valor de dezassete.

Exemplo 4

Para facilitar a determinação do tamanho de amostras, vai exemplificar-se de seguida como se constroem tabelas que relacionam as amostras para diversas populações, com diferentes erros (em desvios-padrões), para três tipos de riscos:

- 1.- 5 vezes em 100 (z = 1,9596)
- 2.- 1 vez em 100 (z = 2,5749)
- 3.- 1 vez em 1000 (z = 3,2901)

Nota: Os valores de z são retirados da tabela da distribuição normal.

1.- Para um risco de 95% (1 vez em 20) o valor de Z é igual a 1,9596.

Para construir a parte da tabela relativa a um erro de 0.50 desvios-padrões tem de se assumir $e = 0,05 \sigma$

$$n = \frac{N \times 3.84003216 \times \sigma^2}{(N - 1) \times 0,25 \sigma^2 + 3.84003216 \times \sigma^2} = \frac{N \times 3.84003216}{(N - 1) \times 0.25 + 3.84003216}$$

Sendo **N** a população, a dimensão da amostra será **n**

N=10	6,30543...	n = 6
N=18	8,5439...	n = 9
N=20	8,94067...	n = 9
N=50	11.93295...	n = 12
N=50000000	15,360124..	n = 15
N=infinito	15,36028...	n = 15

2.- Para um erro de 0.20 desvios-padrões

$$n = \frac{N \times 3.84003216}{(N - 1) \times 0.04 + 3.84003216}$$

e por consequência...

N=10	9,14286...	n = 9
N=18	15,29025...	n = 15
etc.		

3.- Para um risco de 99% (1 em 100) e um erro de 0.50 σ a relação entre o tamanho da população e o da amostra é de

$$n = \frac{N \times 6.63011001 \times \sigma^2}{(N - 1) \times 0,25 \sigma^2 + 6.63011001 \times \sigma^2} = \frac{N \times 6.63011001}{(N - 1) \times 0.25 + 6.63011001}$$

N=10	7,466...	n = 7
N=18	10,9698..	n = 11
N=208	11,652...	n = 12
etc.		

As tabelas completas, calculadas desta forma, que se apresentam no capítulo 5 do presente anexo, consideram limites de erro que variam desde 0,5 (metade) até 0,01 (um centésimo) de desvio-padrão. Para ter uma ideia do que se significam estes valores, considere-se uma distribuição de valores em percentagem.

O caso mais desfavorável, aquele em que o desvio-padrão é maior, ocorre quando metade dos elementos é 0 e a outra metade 100. O desvio-padrão desta distribuição limite é de 50 e os limites de erro a que respeitam são respectivamente ± 25 e $\pm 0,5$ pontos percentuais.

Porém se for assumido que as diversas distribuições de percentagens que se apresentam na vida real tendem a seguir uma distribuição de tipo normal, o valor mais provável para o seu respectivo desvio-padrão variará entre 25 e 15 pontos percentuais.

Para esses valores um limite de erro de $\pm 2,5$ pontos percentuais corresponde geralmente a probabilidades que variam entre 0,20 e 0,10 e limites de erro de ± 1 ponto percentual correspondem a probabilidades de cerca de 0,05.

Pode ainda acrescentar-se que, para níveis significativos (risco de erro de 5 em 100), a relação entre limites de erro e dimensão da amostra será aproximadamente de:

- para limites de erro de $\pm 2,5$ e população infinita - amostra $\approx \pm$ de 96 a 384

- e para limites de erro de ± 1 e população infinita- amostra $\approx \pm$ de 1540

Considerando-se níveis muito significativos (com riscos de erro de 1 em 100), a relação será então:

- para limites de erro de $\pm 2,5$ e população infinita amostra $\approx \pm$ de 271 a 1083

- e para limites de erro de ± 1 , e população infinita amostra $\approx \pm$ de 4300

Estes valores podem considerar-se geralmente adequados e as tabelas poderão sê-lo, especialmente quando não se tem informação sobre a constituição e as características da população.

4.5. – Selecção dos Elementos de Amostras

Como se viu, o inquérito por amostragem permite, com um menor investimento, obter resultados com significado estatístico. Os ganhos relativos decorrentes da inquirição exaustiva do universo não compensam, na maior parte dos casos, os investimentos acrescidos que exige a sua execução. Para além dos custos, o tempo de preparação e de execução de um projecto de recolha exaustiva da informação não é compatível com a necessidade de ter informação recente e fiável.

A utilização de amostragens permite ter elevados graus de certeza nos resultados, mesmo quando se recolhe informação em muito menos unidades de análise.

Depois de definida a dimensão da amostra importa identificar os elementos a inquirir.

Só a utilização de métodos de amostragem pode garantir que a selecção das unidades de análise a inquirir dê origem a uma amostra, isto é, a um subconjunto representativo da população.

É melhor ter uma amostra pequena, mas bem seleccionada, do que uma amostra grande em que foram utilizados métodos de selecção de indivíduos deficientes.

Como exemplo desta afirmação, Moser & Kalton (1975, 79) referem a sondagem feita pelo *Literary Digest* para as eleições presidenciais americanas de 1936. Tendo sido inquirida pelo correio uma enorme amostra de 10.000.000 pessoas, erraram completamente nas previsões. As razões principais residiram no facto da sondagem ter sido feita a partir das listas telefónicas (quando na altura o telefone era ainda um bem que estava longe de ter uma difusão universal), tendo ficado excluídos os estratos mais pobres da população e na baixa taxa de retorno de respostas, de apenas 20%, o que provocou o predomínio dos indivíduos com maior consciência cívica.

4.5.1 – Métodos de Amostragem

Nos manuais vêm referidas diversas maneiras de fazer amostragens, mas apenas se deve dar o nome de amostragem aos métodos que garantem que todo e qualquer membro do conjunto de unidades de análise a inquirir tem a mesma probabilidade de ser seleccionado.

Na bibliografia de origem anglo-americana estes métodos são chamados de métodos probabilísticos. Na realidade existem apenas dois: o aleatório e o sistemático, sendo os outros, ou métodos que não são verdadeiros métodos de amostragem, ou combinações mais ou menos complexas destes dois.

4.5.1.1 – Métodos probabilísticos

Como se viu, são aqueles em que o método de selecção garante a qualquer elemento da população a mesma probabilidade de ser seleccionado para a amostra.

- Amostragem aleatória

Uma amostra aleatória é aquela em que os elementos são seleccionados segundo uma lei de acaso.

Obriga a que sejam conhecidos todos os elementos que constituem a população. No caso mais fácil, parte-se duma listagem numerada de todos os elementos da população-alvo.

Os elementos podem ser seleccionados segundo dois procedimentos: o da lotaria e o dos números aleatórios.

No método da lotaria cada elemento da população é representado por um pequena ficha que é colocada numa urna. As fichas são bem misturadas na urna e depois extraem-se tantas fichas quantos os elementos que devem constituir a amostra. O processo pode apresentar variantes, mas todas se baseiam nos mesmos processos das lotarias. Podem sortear-se sucessivamente dígitos de 0 a 9, até se obterem os números relativos aos elementos seleccionados. Por exemplo, se a população tem 37456 elementos, tira-se à sorte um primeiro número entre 0 e 9 para as unidades, depois faz-se o mesmo para as dezenas, as centenas e os milhares e por último extrai-se um número entre 0 e 3 que indicará as dezenas de milhar. Se o número resultante ultrapassar 37456 não é válido e repete-se o processo.

No método dos números aleatórios, o processo de tirar à sorte os dígitos por meio de uma urna ou tómbola é substituído pela utilização de uma tabela de números aleatórios (como a do Capítulo 5.2) ou são extraídas por um sistema qualquer informatizado automático.

Estas tabelas são apresentadas em folhas com grupos de números de 0 a 9 dispostos em linhas e colunas e são construídas de tal modo que, em qualquer parte da folha, a probabilidade de "sair" um determinado número ou combinação de números é a mesma.

O conjunto de elementos donde se vai extrair a amostra apresenta-se segundo a forma de uma série ordenada com n elementos. Primeiro verifica-se quantos dígitos tem o número n . Esse número de dígitos é que vai definir o tamanho dos grupos a extrair da tabela. A extracção inicia-se em qualquer ponto da tabela, mas depois de iniciado, usa-se exhaustivamente toda a tabela, até chegar ao mesmo ponto. Se ainda falta extrair elementos, continua-se a extracção repetindo o procedimento: escolha de um novo ponto de começo e de um novo processo. O processo de extracção consiste em extrair sequencialmente (da esquerda para a direita, de cima para baixo, da direita para a esquerda ou de baixo para cima) grupos que têm sempre o mesmo número de dígitos. Se porventura o número extraído for superior ao n , o número de elemento da população a amostrar, não é válido e continua-se o processo.

- Amostragem sistemática

As amostras sistemáticas são aquelas em que os elementos foram seleccionados da população sempre do mesmo modo, segundo uma lei ou regra, que dê a cada um dos elementos igual probabilidade de ser seleccionado.

Partindo de uma população que está organizada numa série ordenada, calcula-se primeiro o intervalo de amostragem, isto é, calcula-se a dimensão do grupo de elementos da população em que tem de se extrair um elemento da amostra,

$$a = \frac{N}{n}$$

em que **a** é a dimensão de cada grupo em que se vai extrair um elemento, **N** a população e **n** a amostra.

De seguida escolhe-se um elemento do primeiro grupo, de qualquer modo, até mesmo por um qualquer processo aleatório e depois em todos os outros, sempre na mesma posição na sequência de cada grupo, por exemplo o terceiro ou o primeiro ou o último. Do mesmo modo, a partir do elemento seleccionado no primeiro grupo, se podem obter todos os outros, simplesmente adicionando-lhe sucessivos intervalos de amostragem.

- Amostragem estratificada

Nos dois métodos anteriores a população é amostrada como um todo. No método estratificado, a população é inicialmente dividida em conjuntos que são o resultado da divisão da população segundo grupos que nela efectivamente existem, como é o caso do sexo, da idade ou de qualquer outra característica que se possa considerar relevante.

Após a divisão em estratos, dentro de cada um deles faz-se uma amostragem aleatória ou sistemática.

A estratificação, quando correspondente a características que existem com aquela dimensão na população inicial e que têm relevância para a análise que se efectua, pode conduzir a resultados melhores, mais precisos. Mas quando não se tem a certeza da real proporção dos estratos, é melhor não a executar, por que se pode estar a introduzir um enviesamento involuntário.

- Amostragem por agrupamentos

Neste tipo de amostragem, em vez de se considerar a unidade de análise elementar que constitui a população e sobre a qual incide a análise numa primeira fase consideram-se agrupamentos ou classes de unidades de análise. Por exemplo, para conhecer algumas características dos alunos de uma escola do ensino secundário do País, começa-se por trabalhar com as turmas, que são consideradas exaustivamente. É nelas que se faz a selecção primária e só depois, dentro de cada uma e de forma sistemática ou aleatória, se chega aos alunos.

O processo decorre do seguinte modo: a partir da população, seleccionam-se os grupos, que irão servir como unidades de amostragem, sistemática ou aleatoriamente. Depois inquiram-se todos os elementos de cada um dos grupos que foi seleccionado.

- Amostragem multi-etapas

É a extensão do método dos agrupamentos, em que a selecção por grupos se repete em sucessão.

Suponhamos que queremos conhecer as características dos alunos do ensino secundário em Portugal. Fazia-se uma primeira amostragem por agrupamentos, considerando a escola. Mas depois e por que as escolas são ainda muito grandes, fazia-se uma amostragem por agrupamentos, de turmas e só então é que se inquiriam todos os elementos das turmas. Teríamos assim uma amostragem multi-etapas em duas etapas, na primeira seleccionando escolas e na segunda escolhendo turmas. Se em vez de inquirir todos os alunos das turmas seleccionadas, se fizesse, dentro de cada uma delas, uma outra amostragem, teríamos então três etapas.

- Amostragem multifases

Neste método, ao contrário do anterior, os grupos são sempre os mesmos (Chisnall, 1986, citado por Vicente *et alia*, 1996: 60-61); acontece é que uns são sujeitos a um inquérito mais aprofundado do que outros. O procedimento é o seguinte: a partir da série ordenada dos elementos da população selecciona-se uma amostra, que é objecto de recolha de informação que, depois de analisada, permite seleccionar, intencionalmente, uma segunda amostra, cujos elementos vão ser objecto de uma recolha de informação mais detalhada e com maior profundidade.

4.5.1.2 – Métodos não probabilísticos

Enquanto os métodos probabilísticos garantem que o subconjunto de elementos da população que foi seleccionado é estatisticamente representativo desta, isto é, tende a apresentar as mesmas características em função da respectiva dimensão, estes métodos não probabilísticos, pelo contrário, não garantem tal facto.

- Amostragem intencional

Neste método, os elementos são escolhidos pelo investigador, quer por que este julga que apresentam as características médias da população, quer por os achar representativos da sua totalidade. A amostra pode não ser representativa da população, mas o investigador julga que sim.

Baseia-se na identificação do grupo ou grupos que se crê serem relevantes para o estudo em causa e na sua posterior selecção para análise. É geralmente muito económica, pois a introdução do conhecimento prévio do investigador, pode permitir identificar os grandes grupos que vão ser influentes no tema em estudo e, portanto, recolhe-se informação apenas destes.

Como exemplo, pode referir-se a utilização de amostragens em freguesias “tipo” para efeitos de previsão eleitoral.

É um método que dá excelentes resultados, enquanto se mantiverem as razões subjacentes a essa tipicidade.

- Amostragem bola de neve

É uma variante do método intencional, em que a selecção dos indivíduos a inquirir é feita por indicação de anterior inquiridos.

É típica de ambientes sociais muito fechados sobre si próprios ou em que o investigador e a sua equipa não têm um conhecimento suficiente da população. Neste método corre-se o risco de os primeiros inquiridos, os que dão origem à bola de neve, enviesarem a amostra, canalizando a sequência de entrevistas para a sua rede de relacionamento.

- Amostragem por conveniência

É aquela em que os elementos da amostra são escolhidos por que se disponibilizam, de uma forma mais ou menos activa, a ser inquiridos. Como exemplo, pode referir-se elementos da população que respondem a apelos para inquérito efectuados por jornais e revistas, estações de rádio e de televisão. Como a participação é voluntária, as respostas obtidas não são representativas da população e os seus resultados devem ser analisados com cuidado, embora possam ser úteis para identificar ideias gerais e problemas nas fases preliminares dos estudos. Como exemplo de utilização deste método, pode referir-se a recolha de informação que veio a dar origem ao célebre Relatório Hite, desencadeado por um pedido de informação em revistas feministas.

- Amostragem por quotas

A amostragem por quotas tem algumas semelhanças com a amostra estratificada, embora seja menos segura quanto aos seus resultados do que esta.

De acordo com os estratos em que se encontra dividida a população, organiza-se a amostra, de tal modo que as diversas categorias de cada variável – por exemplo, grupos etários, sexos, tipos de habitações, de residência, etc. - tenham o mesmo peso na amostra que têm na população a inquirir – a quota.

Existe a possibilidade de fazer amostragens por quotas independentes ou relacionadas. Nas primeiras, a selecção dos elementos da população faz-se de acordo com uma só característica: o sexo, a idade ou a profissão, por exemplo. Nas amostragens por quotas relacionadas, os elementos seleccionados correspondem a um conjunto que obedece a várias características simultaneamente. O investigador procura elementos da população e cada um que encontra é entrevistado e é “abatido” nas diversas quotas, ao mesmo tempo. Por exemplo, um jovem lisboeta trabalhador na construção civil, conta para as quotas do sexo masculino, dos jovens e dos operários.

- Amostragem itinerária

É um tipo de amostragem bastante utilizado em estudos locais, que consiste no entrevistador partir de um local dado e seguir um itinerário a partir de regras simples, seleccionando pelo caminho os elementos da amostra.

O investigador escolhe primeiro o ou os locais de partida dos entrevistadores, a quem dá instruções para a escolha do itinerário e para a selecção dos elementos a inquirir. Por exemplo, parte da praça principal para Sul, escolhe casas alternadas, sendo a andar em sequência para cima até ser possível e depois para baixo, sempre do mesmo lado ou alternadamente.

5. – Tabelas

5.1. – Tabelas da Dimensão de Amostras (95 e 99%)

RELAÇÃO DA DIMENSÃO DA POPULAÇÃO COM A AMOSTRA (Tabela a 95%)

Popula- ção	Erro aceite (em desvio-padrões)							
	0.5	0.2	0.1	0.05	0.04	0.03	0.02	0.01
5	4	5	5	5	5	5	5	5
10	6	9	10	10	10	10	10	10
15	8	13	14	15	15	15	15	15
18	9	15	17	18	18	18	18	18
20	9	17	19	20	20	20	20	20
25	10	20	24	25	25	25	25	25
30	10	23	28	29	30	30	30	30
40	11	28	36	39	39	40	40	40
50	12	33	44	48	49	49	50	50
60	12	37	52	58	59	59	60	60
70	13	41	59	67	68	69	70	70
80	13	44	66	76	77	79	79	80
90	13	47	73	85	87	88	89	90
100	13	49	80	94	96	98	99	100
150	14	59	108	137	141	145	148	149
200	14	65	132	177	185	191	196	199
250	15	70	152	215	227	236	244	248
275	15	71	160	233	247	258	267	273
500	15	81	217	377	414	448	475	494
1000	15	88	278	606	706	810	906	975
2000	15	92	322	869	1091	1362	1655	1901
4030	15	94	351	1112	1504	2073	2839	3647
5000	15	94	357	1175	1622	2302	3288	4424
10000	15	95	370	1332	1936	2991	4898	7934
20000	15	96	377	1427	2143	3517	6487	13151
50000	15	96	381	1490	2290	3931	8054	21720
100000	15	96	383	1513	2344	4092	8759	27746
200000	15	96	383	1524	2372	4178	9160	32215
500000	15	96	384	1531	2389	4231	9419	35662
1000000	15	96	384	1534	2394	4249	9509	36980
2000000	15	96	384	1535	2397	4258	9554	37677
5000000	15	96	384	1536	2399	4263	9582	38108
10000000	15	96	384	1536	2399	4265	9591	38253
20000000	15	96	384	1536	2400	4266	9595	38327
50000000	15	96	384	1536	2400	4266	9598	38371
Infinito	15	96	384	1536	2400	4267	9600	38400

RELAÇÃO DA DIMENSÃO DA POPULAÇÃO COM A AMOSTRA (Tabela a 99%)

Popula- ção	Erro aceite (em desvio-padrões)							
	0.5	0.2	0.1	0.05	0.04	0.03	0.02	0.01
5	4	5	5	5	5	5	5	5
10	7	9	10	10	10	10	10	10
15	10	14	15	15	15	15	15	15
18	11	16	18	18	18	18	18	18
20	12	18	19	20	20	20	20	20
25	13	22	24	25	25	25	25	25
30	14	26	29	30	30	30	30	30
40	16	32	38	39	40	40	40	40
50	18	39	47	49	49	50	50	50
60	19	44	55	59	59	60	60	60
70	19	49	63	68	69	69	70	70
80	20	54	71	78	79	79	80	80
90	21	59	79	87	88	89	90	90
100	21	63	87	96	98	99	99	100
150	23	79	122	142	145	147	149	150
200	24	91	154	186	191	195	198	199
250	24	100	182	229	236	242	246	249
275	24	104	195	249	258	265	271	274
500	25	125	285	421	446	468	485	496
1000	26	142	399	726	806	881	943	985
2000	26	153	498	1140	1349	1573	1785	1941
4030	26	159	569	1600	2043	2605	3242	3799
5000	26	160	585	1733	2266	2979	3841	4649
10000	26	163	622	2096	2930	4242	6237	8690
20000	26	164	642	2342	3433	5384	9064	15365
50000	27	165	654	2519	3827	6421	12449	28504
100000	27	165	659	2584	3979	6861	14219	39868
200000	27	166	661	2617	4060	7105	15307	49794
500000	27	166	662	2638	4110	7260	16043	58539
1000000	27	166	663	2645	4127	7313	16305	62179
2000000	27	166	663	2649	4135	7340	16439	64174
5000000	27	166	663	2651	4140	7356	16521	65433
10000000	27	166	663	2651	4142	7361	16548	65864
20000000	27	166	663	2652	4143	7364	16562	66082
50000000	27	166	663	2652	4143	7366	16570	66213
Infinito	27	166	663	2652	4144	7367	16575	66301

RELAÇÃO DA DIMENSÃO DA POPULAÇÃO COM A AMOSTRA (Tabela a 99,9%)

Popula- ção	Erro aceite (em desvio-padrões)							
	0.5	0.2	0.1	0.05	0.04	0.03	0.02	0.01
5	5	5	5	5	5	5	5	5
10	8	10	10	10	10	10	10	10
15	11	14	15	15	15	15	15	15
18	13	17	18	18	18	18	18	18
20	14	19	20	20	20	20	20	20
25	16	23	24	25	25	25	25	25
30	18	27	29	30	30	30	30	30
40	21	35	39	40	40	40	40	40
50	23	42	48	49	50	50	50	50
60	25	49	57	59	59	60	60	60
70	27	56	66	69	69	70	70	70
80	28	62	75	79	79	79	80	80
90	29	68	83	88	89	89	90	90
100	30	73	92	98	99	99	100	100
150	34	97	132	145	147	148	149	150
200	36	115	169	191	194	197	199	200
250	37	130	203	236	241	245	248	249
275	38	137	219	259	264	269	272	274
500	40	176	342	448	466	480	491	498
1000	42	213	520	813	871	923	964	991
2000	42	238	703	1368	1544	1715	1862	1964
4030	43	254	853	2088	2526	3019	3508	3885
5000	43	257	890	2321	2875	3532	4220	4779
10000	43	264	977	3022	4036	5460	7302	9154
20000	43	267	1027	3559	5056	7511	11501	16881
50000	43	269	1060	3985	5959	9695	17559	34202
100000	43	270	1071	4150	6337	10736	21298	51980
200000	43	270	1077	4238	6544	11345	23837	70234
500000	43	270	1080	4293	6675	11745	25672	88983
1000000	43	271	1081	4311	6720	11885	26349	97675
2000000	43	271	1082	4321	6743	11956	26701	102690
5000000	43	271	1082	4326	6756	11999	26916	105954
10000000	43	271	1082	4328	6761	12013	26989	107088
20000000	43	271	1082	4329	6763	12020	27025	107665
50000000	43	271	1082	4330	6765	12025	27047	108014
Infinito	43	271	1082	4330	6765	12028	27062	108248

5.2. – Tabela de Números Aleatórios

26169	36649	07723	75848	48330	13207	54974	30717	43280	73909	48752	14447
33486	81536	17992	22643	74587	37447	85845	92848	99579	55842	55822	69241
56976	88393	51771	90337	35182	81702	62801	09418	12556	05880	13559	96241
27732	47101	24901	68843	34428	94511	88226	82765	02189	68550	80862	09221
03476	51882	64621	05398	61251	76754	16945	22659	02578	35097	62846	59115
06026	16043	80067	78932	21557	78174	59231	71469	60652	05268	95937	73756
73923	06871	04981	49307	19255	80243	68828	22003	91680	91582	73681	34145
88837	14054	46633	90722	64086	24450	18088	94456	70934	34222	86211	13272
79233	67708	47718	22395	41392	63726	05516	80799	24659	76728	68564	30910
94765	04616	28860	98267	14857	00983	83662	85912	34238	47018	75434	72511
26117	42759	09642	26062	17680	41268	06312	69062	76558	00610	84066	84481
78406	22814	57139	08353	23602	99345	10025	34734	46184	05497	04184	95641
32620	96527	96925	31067	95921	45459	58039	60276	66357	91649	05127	43825
78789	48048	48574	15092	16691	46801	51070	66349	26563	60329	19629	50051
58159	41117	31018	07131	62855	81440	19899	44521	48854	24576	94750	15559
75846	68658	27788	77947	68418	52628	13767	17138	92887	50454	12208	08244
09790	30045	58024	19528	41140	57779	00514	38244	03988	67932	82843	92769
93963	56074	75510	75024	74486	59041	67124	09067	05133	10660	05495	01033
73666	66477	07480	40397	45263	02162	61771	24585	17642	20310	58493	72798
26029	48474	25424	91915	84312	11526	12889	97821	99513	20989	01663	35532
41827	81680	59046	16597	11484	99984	25438	10865	40218	42966	57125	86876
07029	98357	82520	99713	95021	80659	13970	87741	33287	44343	14308	99018
78802	08321	90111	06357	22570	55392	82899	02770	55352	34860	79314	33582
76169	39432	83966	28580	92049	46997	14063	02312	80298	24119	05961	63946
44352	60045	12507	29909	80888	93051	78690	09767	14071	27278	56282	75661
01087	18294	62514	69138	57154	85784	21097	32606	24794	04827	36060	18608
74617	71377	02750	64474	57153	34295	39969	89066	86493	16383	15480	52079
07773	49773	04579	65763	18195	62413	36659	43059	75881	14847	73989	49814
83944	29188	86542	83380	75494	47877	03352	97085	18364	06701	59201	91325
18561	74550	97998	47856	76172	82729	02359	24387	25858	73078	20615	41371
06300	20377	99431	52221	70002	27635	94362	63230	29886	44573	47384	47252
82341	65436	46364	99478	97854	31267	61587	02844	37181	09798	95083	74173
45103	90157	93259	86148	04670	96400	65140	13015	55179	24828	43997	49633
95348	46460	76299	88228	12409	97278	35660	00163	74172	37841	68938	66486
15508	85489	19805	90912	59209	61812	89629	05861	34309	03596	91865	77344
41896	63581	89630	37419	72033	79748	96311	08351	81310	16193	91254	53471
90458	58416	64617	40730	19809	14669	74646	52712	83096	98100	45150	07184
91542	97197	27367	74230	09093	00451	66221	55341	09291	83854	96511	40143
23113	08453	13666	47539	48102	98514	15953	44669	00598	70473	09051	25055
07327	91607	03434	44801	10146	35289	55241	75863	37460	04756	88394	93512
10342	11322	84860	33342	30498	89573	08766	52669	27252	00858	84093	43047
55852	97050	36507	51249	49631	70201	76402	23996	46043	76398	51185	64633
57215	47407	93154	39338	15351	93459	57862	88073	09961	01390	35326	55132
54940	67775	06648	07338	25638	49669	92227	35705	38263	62655	75669	21469
45636	17401	63672	00883	74159	98455	21130	04525	10853	13358	25433	70238
96741	96651	28116	75864	86055	17176	02545	10278	60193	06696	72770	75748
28762	96269	58742	05290	46910	59179	08621	61324	85817	49913	63517	43580
89563	12677	32944	54127	39748	79912	02017	31132	08566	43833	92450	74100
25804	41498	31874	10224	06248	95711	27430	31429	84528	61351	40658	82088
12708	59022	84713	05530	22881	96469	09552	29158	47717	86648	70655	36449
91944	86465	23876	03148	97765	78247	64875	04573	54264	76210	49569	79391
19803	70953	19974	45679	44779	07044	73436	03826	70328	72199	89932	98787
50776	02858	75709	67007	60660	59558	38202	59689	98016	76209	89735	00681
34330	21326	98219	00313	24823	12810	85106	23442	64356	34786	55080	36865
18205	50090	01964	21268	18459	35299	01674	55755	98936	90781	62352	20389
48331	70644	29509	09796	30929	69999	82619	83948	89537	24004	83415	45345
43573	61376	71954	03410	41719	16131	86380	17129	44756	76935	31694	78124
02764	16068	33873	51326	28524	58542	86456	23483	23688	50290	72732	21287
39007	89265	36299	39310	00513	48126	13532	62916	94190	14045	18329	98100
30086	56168	12063	29374	50891	94820	39649	78326	11778	70129	16994	62751

6.- BIBLIOGRAFIA

- ABREU, Diogo de (1998) – ***Recolha e Tratamento de Informação. Manual de Apoio ao 3º Curso de Técnicos Superiores de Planeamento na Administração Regional.*** IGE/Partex, Luanda, p. 6-37.
- ALMEIDA, J. F. de, PINTO, J. M. (1976) – ***A Investigação nas Ciências Sociais***, Lisboa, Editorial Presença, p. 105-113.
- CALOT, Gérard (1970) – ***Curso de Estadística Descriptiva***, Paraninfo, Madrid, p. 94-99. É tradução castelhana do ***Cours de Statistique Descriptive***, Dunod, Paris, 1965
- CAPLOW, Theodore (1972) – ***La Investigación Sociológica***, Editorial Laia, Barcelona , p. 260-266
- CLIFFORD, Nicholas J.; VALENTINE, Gill (2003) – ***Key Methods in Geography***, Sage Publications, London, 572 p.
- DUVERGER, Maurice (1964) – ***Méthodes des Sciences Sociales***, 1ª ed., P.U.F., Paris, p. 251-266.
- (*) ESTEBANEZ ALVAREZ, J.; BRADSHAW, R.. P. (1978) – ***Técnicas de Quantificación en Geografía***, Editorial Tebar Flores, Madrid, p. 277-336.
- LONGHURST, Robyn. (2003) – “***Semi-structured Interviews and Focus Groups***”, in Clifford & Valentine – ***Key Methods in Geography***, Sage Publications, London, p. 117-132.
- MOSER, C. A.; KALTON, G. (1975) – ***Survey Methods in Social Investigation***, Heinemann Educational Books, London.
- (*) VICENTE, Paula et alia (1996) – ***Sondagens***, Edições Sílabo, Lisboa, 238 p..
- (*) VILARES, M. J. (1993) – ***Sistemas de Informação Estatística***, ISEGI, Lisboa, p. 277-336.
- YIN, Robert K. (1964) – ***Case Study Research. Design and Methods. 2nd edition.*** SAGE Publication, London, 171 p

Anexo 4

Tabelas de Contingência

1. – Introdução

Em análise de dados é muito frequente o aparecimento de quadros ou tabelas de contingência, tanto por que a informação disponível para análise já se encontra sob esta forma, como porque é uma maneira muito fácil de efectuar a sua descrição.

De facto, é sob esta forma, que muita da informação estatística é apresentada ou disponibilizada pelos grandes organismos produtores de informação, e os próprios investigadores muitas vezes guardam e apresentam os resultados dos seus inquéritos deste mesmo modo.

As Tabelas de Contingência podem construir-se com todos os tipos de informação, seja qual for o tipo de escala de medida utilizado.

Um quadro ou tabela de contingência é uma matriz [$n \times m$] que contem os frequências (o número de ocorrências) de associação entre as diferentes modalidades (ou estados) de duas variáveis qualitativas, ou que para o efeito se consideram como tal (classificadas). É pois aquilo a que vulgarmente se chama uma tabela cruzada ou de dados cruzados.

Um exemplo pode ser a tabela de contingência das situações na profissão, segundo a Classificação Nacional das Profissões e das freguesias de Lisboa. Em cada célula tem-se o número de indivíduos da categoria profissional j na freguesia i .

Os quadros de contingência são assim quadros de frequência entre duas variáveis, cada uma delas dividida em vários estados ou em várias classes.

Considere-se duas variáveis, a variável C , com três estados, modalidades ou categorias e a variável L , com quatro estados, modalidades ou categorias, como consta do Quadro 1.

Define-se O_{ij} como sendo o número de unidades de análise, sujeitos, casos ou de indivíduos em que ocorre simultaneamente o estado i da variável L e o estado j da variável C .

		V_C		
		C_1	C_2	C_3
V_L	L_1			
	L_2			
	L_3		O_{ij}	
	L_4			

Quadro 4.1 - Tabela de contingência

As tabelas ou quadros de contingência podem ser quadrados, quando o número de estados da variável linha é igual ao número de estados da variável coluna, ou rectangulares quando são diferentes. O caso particular de maior simplicidade é o dos quadrados de dois estados ($l=c=2$).

Dum ponto de vista formal, os *quadros de frequência* da análise unidimensional de dados, podem ser considerados como tabelas ou quadros de contingência a uma dimensão, em que o número de linhas é igual ao de classes ($l=nc$) e o número de estados da variável coluna é 1.

Embora mais difíceis de analisar, podem ser construídas tabelas de contingência de n variáveis (ou n dimensões). O esquema é o mesmo. Para três, ou mais¹, variáveis, por exemplo, a variável **C**, com três estados, modalidades, categorias ou classes, a variável **L**, com quatro estados, modalidades, categorias ou classes e a variável **E**, com dois estados, modalidades, categorias ou classes, constituindo um quadro de linhas e colunas (conforme o Quadro 4.2), em que a variável **L** é a variável linha, **C** a variável coluna e **E** a variável estrato.

		V_C					
		C₁		C₂		C₃	
		E₁	E₂	E₁	E₂	E₁	E₂
V_L	L₁						
	L₂						
	L₃				O_{ijk}		
	L₄						

Quadro 4.2 - Tabela de contingência de três variáveis

O número de células é o produto do número de estados das diferentes variáveis, que neste caso é três vezes dois vezes quatro, ou seja vinte e quatro.

O_{ijk} é o número de unidades de análise, sujeitos, casos ou indivíduos em que ocorrem simultaneamente o estado **i** da variável linha, o estado **j** da variável coluna e o estado **k** da variável estrato.

Podem considerar-se mais de três variáveis, a que correspondem mais variáveis estrato. Muitos dos quadros estatísticos publicados pelos organismos oficiais são deste tipo, com muitas variáveis estratos.

Quando as variáveis estrato são muitas, a legibilidade dos quadros diminui bastante.

Nas *tabelas de contingência* (que também se denominam *quadros de medição conjunta*, ou *quadros de frequência n-dimensionais*) em cada uma das células da matriz existe um valor de frequência. Mas estes

¹ - Tabelas de contingência com grande número de variáveis são bastante raras.

valores de frequência não são os únicos elementos a tomar em conta. Devem considerar-se também os *totais marginais*, que são a soma dos elementos de uma linha ou de uma coluna, e o *total geral*, que é a soma de todos os elementos da tabela.

Dado o tabela de contingência que consta do Quadro 4.3, em que está relacionado um conjunto de população quanto à raça (com os estados de branca e não branca) e quanto ao rendimento mensal, (mais ou menos de 80 dólares americanos), os totais marginais são, respectivamente, de 570 para branca e 1772 para não branca e 1689 para os que ganham menos de 80 \$US e 653 para os que ganham mais.

O total geral é de 2342, o número total de unidades de análise, sujeitos, casos, ou indivíduos.

	Branca	N. branca	Σ
<80	9	1680	1689
>80	561	92	653
Σ	570	1772	2342

Quadro 4.3 - Exemplo de tabela de contingência

O total geral é igual à soma de todos os elementos, ou à soma dos totais marginais das linhas, ou das colunas, segundo esta notação

$$\sum_j O_{ij} = \text{Total de Colunas}$$

$$\sum_i O_{ij} = \text{Total de Linhas}$$

$$\sum_{..} O_{ij} = \text{Total Geral}$$

2. – Transformação para Tabelas de Contingência

É necessário ter em conta que uma grande parte dos quadros estatísticos não são quadros de contingência. Mas sempre que a soma das colunas e das linhas tenha sentido, podem ser assumidos como quadros de contingência, e como tal analisados.

Algumas vezes também é possível, transformar os quadros de dados (que são quase sempre de atributos) o que vai permitir analisá-los tratá-los como tabelas de contingência.

Estas transformações têm alguns riscos, e implicações na interpretação dos resultados.

De seguida apresentam-se exemplos desses tipos de transformação.

2.1. – Modificação por ponderação

O processo consiste em transformar as variáveis (colunas) de modo a que a soma das linhas tenha sentido, o que permite de algum modo considerá-la como "uma frequência" ou "peso".

Tendo um índice 100 para a produção em Janeiro de diversos produtos, e um valor que compara a produção em cada sector de cada mês com a que se verificou em Janeiro (Quadro 4.4)

É uma matriz de atributos, em que se descreve a percentagem de variação em cada mês, relativamente à produção de Janeiro.

Meses	Agricultura	Ind. Text	Ind. Met
Janeiro	100	100	100
Fevereiro	98	110	130
Março	97	107	143
Abril	90	105	179
.....			

Quadro 4.4 - Produção por sectores de actividade

Nesta matriz a soma dos valores das colunas não tem qualquer sentido.

Mas ... se multiplicarmos cada um dos valores mensais, pelo valor base do índice (o valor real da produção de Janeiro), que poderá ser

Índice de Janeiro	Agricult	Ind. Text.	Ind. Metalúrgica
100	30	25	400

medidos em milhões de escudos, obtemos então uma matriz de produção, como consta no Quadro 4.5, que se segue

Meses	Agricultura	Ind. Textil	Ind. Metalúrgica
Janeiro	30,0	25,0	400,0
Fevereiro	29,4	27,5	520,0
Março	29,1	26,8	572,0
Abril	27,0	26,2	716,0
.....			

Quadro 4.5 - Produção por sectores de actividade, em milhões de escudos

em que a soma das linhas tem sentido, pois é o valor da produção dos três sectores, e a soma das colunas também o tem, devido a ser o valor da produção de cada ramo no período de tempo considerado.

Este tipo de quadro pode tratar-se como se fosse uma tabela de contingência (***I x c***) de ***I*** linhas por ***c*** colunas, o que podendo não ser melhor forma de tratamento analítico, é porém uma das formas possíveis, e que não acarreta erros ou mal interpretações significativas.

Veja-se outro exemplo: Uma série de alunos teve um conjunto de notas em três testes diferentes. (Quadro 4.6, abaixo)

Alunos	Teste A	Teste B	Teste C
1	12	14	13
2	17	15	16
3	13	15	12
4	8	9	12
5	14	13	13
6	11	12	10
7	12	14	14
8	10	11	13
9	9	12	11
10	7	9	10
11	12	11	13
12	13	15	14
13	11	14	12
14	10	13	11
15	12	14	11

Quadro 4.6 - Notas dos alunos de uma turma

A soma das linhas só terá sentido se os testes forem equivalentes, o que não é o caso do exemplo. Mas se atribuirmos a cada teste um peso, por exemplo, considerando que cada teste tem a seguinte ponderação

<u>Ponderação</u>	<u>Teste 1</u>	<u>Teste 2</u>	<u>Teste 3</u>
inicial	2	1	1,5
percentagem	44,4	22,2	33,3

obtemos um outro quadro, com o seguinte aspecto:

i	Teste A	Teste B	Teste C	Contributo de A	Contributo de B	Contributo de C	Soma dos contributos
1	12	14	13	5.328	3.108	4.329	12.765
2	17	15	16	7.548	3.330	5.328	16.206
3	13	15	12	5.772	3.330	3.996	13.098
4	8	9	12	3.552	1.998	3.996	9.546
5	14	13	13	6.216	2.886	4.329	13.431
6	11	12	10	4.884	2.664	3.330	10.878
7	12	14	14	5.328	3.108	4.662	13.098
8	10	11	13	4.440	2.442	4.329	11.211
9	9	12	11	3.996	2.664	3.663	10.323
10	7	9	10	3.108	1.998	3.330	8.436
11	12	11	13	5.328	2.442	4.329	12.099
12	13	15	14	5.772	3.330	4.662	13.764
13	11	14	12	4.884	3.108	3.996	11.988
14	10	13	11	4.440	2.886	3.663	10.989
15	12	14	11	5.328	3.108	3.663	12.099
Tot	171	191	185	75,924	42,402	61,605	179,931

Quadro 4.7 - Quadro auxiliar de notas dos alunos de uma turma

Pode considerar-se que, se as pontuações ou notas do primeiro sujeito em cada teste foram 12, 14 e 13, e tomando em conta o peso específico de cada um, a "aptidão global" deste primeiro sujeito é de 12.765, com uma "contributo específica" de cada um dos testes de respectivamente 5,328, 3,108 e 4,329. O mesmo para os sujeitos dois e três, que têm uma aptidão global de 16,206 e 13,098, respectivamente baseada nas aptidões específicas parcelares de 7,584, 3,330 e 5,328 para o segundo e de 5,772, 3,330 e 3,996 para o terceiro.

Por outro lado pode dizer-se que a potência total do primeiro teste para a classificação destes quinze alunos foi de 75,924, de 42,402 para o segundo e de 61,605 para o terceiro.

Ambas as somas, de linhas e de colunas, têm sentido, e por isso também se podem tratar como se de um quadro de contingência se tratasse.

2.2. – Quadros Lógicos

Os quadros lógicos são tabelas de unidades de análise² em que se refere a presença ou ausência de cada estado elementar (ou classe) que a variável pode assumir.

São quadros em que só ocorrem valores de **0** e de **1**. O primeiro representa a ausência do estado ou desse valor de classe na unidade de análise, e o segundo a sua presença.

Os dados originais podem ser nominais (por exemplo, de uma variável que descreve o sexo, passa-se a duas variáveis, uma "ser ou não homem" e outra que é "ser ou não ser mulher"), ordinais (caso em que para cada valor da variável inicial se passa a ter uma nova variável que é "ter ou não ter esse valor"), ou de intervalo/razão, em que obviamente se torna necessário proceder à discretização/classificação da variável, e depois transformar as diversas classes em variáveis do tipo "pertence à classe tal".

Por exemplo,

para as variáveis nominais cada estado da variável constitui uma variável-coluna de presença/ausência;

para as variáveis ordinais se são poucos os estados, cada um constitui uma variável-coluna de presença/ausência;
se são muitos, agrupam-se (em classes) e depois com cada uma das classes forma-se uma variável-coluna de presença/ausência;

para variáveis de intervalo e de razão classificam-se sempre, e com cada classe constitui-se uma variável-coluna de presença/ausência.

² - indivíduos, casos ou sujeitos

Geralmente, o quadro lógico resultante terá em cada linha tantos 1, quantas eram as variáveis iniciais, pois em geral, a unidade de análise, caso ou sujeito apenas assume um estado, ou valor, em cada variável.

Exemplificando com duas variáveis, *Idade* e *Sexo*, para um conjunto de seis sujeitos ou unidades de análise (Quadro 4.8)

Sujeito	Idade	Sexo
1	35	M
2	54	M
3	18	F
4	45	M
5	32	F
6	21	M

Quadro 4.8 - Quadro de valores

Aceitando o seguinte código de transformação de variáveis do Quadro 4.8

Sexo	M	F
Idade	J	< de 25 anos
	M	26 a 49 anos
	V	> de 50 anos

obtem-se um quadro lógico de cinco estados (três relativos à idade e dois ao sexo) que são respectivamente

- 1 - "Ser ou ser do sexo masculino"
- 2 - "Ser ou ser do sexo feminino"
- 3 - "Ser ou ser de idade jovem (menos de 25 anos)"
- 4 - "Ser ou ser de idade média (entre 26 e 49 anos)"
- 5 - "Ser ou ser de idade velha (mais de 50 anos)"

U. Análise i	Sexo		Idade			Soma (a)
	M	F	J	M	V	
1	1	0	0	1	0	2
2	1	0	0	0	1	2
3	0	1	1	0	0	2
4	1	0	0	1	0	2
5	0	1	0	1	0	2
6	1	0	1	0	0	2
Soma	4(b)	2	2	3	1	12(c)

Quadro 4.9 - O mesmo quadro de valores, como quadro lógico

- (a) - é o valor do total de cada linha, que no caso presente igual a 2. E é também o número de variáveis iniciais, se o sujeito apenas tem um estado em cada variável;
- (b) - é o número de vezes que cada estado ocorre. A soma das ocorrências em todos os estados de uma variável inicial é igual ao número de sujeitos ou de unidades de análise.
- (c) - É o número de sujeitos vezes o número de variáveis, se cada unidade de análise apenas assumir um estado em cada variável.

Este tipo de tabela também pode ser tratada como um quadro de contingência, que, no caso, é de 6 unidades de análise por 5 estados de 2 variáveis.

2.3 – Tabelas de Burt

Um outro tipo de quadros muito importante na análise de tabelas de contingência são as *tabelas* ou *quadros de Burt*.

Quando se fazem teste, inquéritos ou se recolhe informação de qualquer outra forma, a informação obtida é geralmente sobre muitas variáveis ou itens, e parte dela, muitas vezes sob a forma de estados. Mesmo quando se trata de variáveis contínuas é muitas vezes mais interessante e simples considerá-las em classes ou estados³.

Tem-se assim uma população traduzida por **N** unidades de análise, casos ou indivíduos, em que foram medidas **M** variáveis, ou caracteres, através de **K_i** estados, modalidades ou classes diferentes para cada um dos caracteres **i**.

É importante ter informação prévia sobre quais podem revelar-se as relações mais importantes e mais interessantes, especialmente quando o número de variáveis e de estados é muito grande, e se forma um emaranhado de relações.

Como exemplo duma estrutura organizada em unidades de análise, variáveis e estados, veja-se o Quadro 4.10.

POPULAÇÃO	Variável 1 - Idade Variável 2 - Sexo Variável 3 - Rendimento	3 grupos de idade 2 grupos sexo: M, F 4 grupos rendimento
N unidades de análise	M variáveis	K _i K1=3 K2=2 K3=4

Quadro 4.10 - Tipos de elementos numa população classificada

Chamam-se *Quadros ou Tabelas de Burt*, aos grandes quadros de frequência, quadrados e simétricos, que cruzam entre si todos os estados de todas as variáveis.

São quadros de frequência que apresentam o número de unidades de análise em que ocorrem simultaneamente as combinações de todos os

³ - Não se estranhe que muitas vezes a riqueza de informação (muito diferenciada) de variáveis contínuas possa ser uma máscara duma realidade subjacente, possa ser um inconveniente e não uma vantagem. Por exemplo, numa estrutura hierarquizada, organizada em poucos patamares, ou níveis, as diferenças entre elementos que se encontram num mesmo patamar podem esconder o facto, muito mais significativo, de que pertencem à mesma classe ou grau.

estados de todas as variáveis, duas a duas. Nesse sentido, contêm em si toda a estrutura de relações do universo que se está a estudar.

As tabelas de Burt apresentam o aspecto do Quadro 11.

		V1			V2		V3			
		e ₁	e ₂	e ₃	e ₁	e ₂	e ₁	e ₂	e ₃	e ₄
V1	e ₁ e ₂ e ₃									
V2	e ₁ e ₂									
V3	e ₁ e ₂ e ₃ e ₄									

Quadro 4.11 - Esquema tipo de uma tabela de Burt

Ao longa da diagonal situam-se os *quadros diagonais* em que estão as tabelas de frequência de cada uma das variáveis isoladamente, e que só têm valores na diagonal, no caso mais frequente dos valores da variável serem mutuamente exclusivos, isto é, na situação em que cada unidade de análise apenas pode assumir um único valor, como é o caso, por exemplo, da idade de um indivíduo.

Mas se cada unidade de análise pode assumir, ao mesmo tempo, dois valores numa mesma variável, então existirão também valores fora da diagonal, como pode ser o caso na variável "*Em que países gostaria de passar férias ?*". Então, se algumas pessoas apenas respondem Espanha, ou França, outras haverá que poderão responder Espanha e França, e então surgirão valores fora da diagonal

Veja-se como seriam então neste caso, as respectivas tabelas

	França	Espanha	Portugal
França	32	0	0
Espanha	0	42	0
Portugal	0	0	138

Quadro 4.12 - Exemplo 1

	França	Espanha	Portugal
França	12	18	2
Espanha	18	32	15
Portugal	2	15	133

Quadro 4.13- Exemplo 2

No exemplo 1 (quadro 4.12) de um total de 212 indivíduos inquiridos sobre o país em que gostariam de passar férias, numa pergunta que apenas admite a escolha de um país dos três que foram apresentados à escolha, houve 32 que declararam escolher a França, 42 a Espanha e 138 Portugal.

No exemplo 2 (quadro 4.13) do mesmo total de 212 indivíduos, mas agora numa pergunta em que poderiam escolher um ou dois países, 12 escolheram apenas a França, 18 a Espanha e a França, e 2 França e

Portugal; foram 32 os que preferiram só a Espanha, e 133 os que optaram apenas Portugal, enquanto 15 escolheram Portugal e Espanha em conjunto.

Voltando ao Quadro 4.11, fora da diagonal situam-se todas as tabelas de contingência simples bidimensionais, possíveis de existir, como resultado das combinações das variáveis duas a duas. No total deve ter-se em conta que os valores se encontram todos duplicados, pois o quadro é simétrico e tanto contabiliza a ocorrência de a com b como a de b com a.

2.3.1 – Construção de uma tabela de Burt

Exemplificando o modo como se constrói uma tabela de Burt, considere-se o seguinte quadro de dados, de 4 unidades de análise, alunos, e 3 variáveis, idade, sexo e nota no exame (Quadro 4.14),

i	Idade	Sexo	Exame final
1	35	M	90
2	54	M	110
3	18	F	98
4	45	F	137

Quadro 4.14 - Dados exemplificativos

respectivamente com 3, 2 e 3 estados cada, a saber, jovens (**Jv**) com menos de 20 anos, médios (**Md**) com idades entre 20 e 50 anos e velhos (**Ve**) com mais de cinquenta, do sexo masculino (**Ma**) e do feminino (**Fe**), e Reprovados (**Rep**) com menos de 100 no teste, sofrível (**Sf**) com nota entre 100 e 135 e bom (**Bm**) com mais de 135.

		Idade			Sexo		Teste		
		Jv	Md	Ve	Ma	Fe	Rep	Sf	Bm
Idade	Jv	1	0	0	0	1	1	0	0
	Md	0	2	0	1	1	1	0	1
	Ve	0	0	1	1	0	0	1	0
Sexo	Ma	0	1	1	2	0	1	1	0
	Fe	1	1	0	0	2	1	0	1
Teste	Rep	1	1	0	1	1	2	0	0
	Sf	0	0	1	1	0	0	1	0
	Bm	0	1	0	0	1	0	0	1

Quadro 4.15 - Tabela de Burt para os dados exemplificativos

Sendo as variáveis mutuamente exclusivas, a soma de cada linha ou de cada coluna é igual ao número de variáveis (**M**), a soma de cada célula é igual ao número de unidades de análise (**N**), e a soma total do quadro é igual ao produto do número de unidades de análise pelo quadrado do número de variáveis (**N x N x M**).

2.3.2 – Generalidades sobre análise de tabelas

Apresenta-se de seguida uma breve resenha de alguns dos tratamentos mais utilizados na análise de dados, quando estes se encontram sob a forma de tabelas de contingência.

As tabelas de duas variáveis ($I \times c$) podem ser de três tipos. Segundo o grau de generalidade, as mais gerais são as rectangulares quando $I \neq c$, depois as quadradas em que $I = c$ e as menos gerais são as quadradas a dois estados, quando $I = c = 2$, que é um caso espacial das quadradas.

Para analisar as relações existentes nestas tabelas de duas variáveis, existem uma série de métodos alguns que serão apresentados, outros que os alunos devem procurar na bibliografia especializada, se as quiserem utilizar.

As tabelas de mais de duas variáveis, podem ser justapostas (segundo linhas, segundo colunas ou segundo ambas), caso em que as variáveis são divididas em estratos que se repetem para cada estado da linha ou coluna - o que nem sempre é fácil de interpretar - ou são quadros de Burt.

Os quadros de Burt podem ser analisados pelo método do χ^2 com o fim de identificar as relações mais interessantes, ou por outros métodos também adaptados a este tipo de informação., de que se salientam:

- 1.- Análise Factorial de Correspondências;
- 2.- Análise de Variância;
- 3.- Análise de Regressão Múltipla;
- 4.- Modelo Log-linear;
- 5.- Dinâmica de Poliedros (Q-Analysis).

3. – Análise de Tabelas a duas variáveis

As tabelas de duas variáveis são tabelas do tipo ($I \times c$) , em que I é o número de estados da variável-linha e c o número de estados da variável-coluna.

As metodologias para analisar as relações existentes nestas tabelas de duas variáveis, permitem efectuar:

- Testes de independência entre linhas e colunas, ou testes de igualdade de proporções;
- Medidas de associação e de correlação;
- Medidas de predição e incerteza.

Os especialistas têm apresentado muitas medidas diferentes. Um problema que se coloca então é o de escolher qual é a mais adequada.

Não existem uma resposta única, devendo ser analisadas casos a caso. Como regras gerais, convém não esquecer que:

as medidas a utilizar deverão ser escolhidas de acordo com o "nível de medida" das variáveis (isto é, de acordo com o tipo das escalas de medida com que se recolheram). Testes aplicáveis a escalas de "nível fraco" podem ser usados a níveis mais fortes, mas a inversa quase nunca resulta;

as medidas estatísticas não probabilísticas só podem ser usadas em comparações entre si, não se podem comparar umas com as outras;

e se deve ter sempre em conta se os resultados são, ou não, influenciados pela dimensão da população.

3.1 – Tabelas de Contingência ($I \times c$) a n estados

São tabelas (ou quadros) do tipo ($I \times c$), em se relacionam I estados de uma variável (linha) com c estados de outra variável (coluna). O número de estados (ou classes) das duas variáveis pode ser igual ou diferente. Quando só existem dois estados é um caso particular que se analisará à parte.

Relembra-se a notação matemática

		V_c							
		C_1	C_2	..	..	C_j	..		
V_I	I_1							L_1	
	I_2								
	I_3								
	:								
	:							L_i	
	I_i					O_{ij}			
	:								
		C_1	C_j					N	

Quadro 4.16 - Tabela de contingência

em que V_I - é a variável-linha V_c - é a variável-coluna
 $I_1, I_2, I_3, \dots, I_i$ estados da variável segundo as linhas
 $C_1, C_2, C_3, \dots, C_j$ estados da variável segundo as colunas
 O_{ij} número de ocorrências com o estado i da variável-linha e com o estado j da variável-coluna
 C_j Total das ocorrências do estado j da variável-coluna
 L_i Total das ocorrências do estado i da variável-linha

$$C_1 = \sum_i O_{i1} \quad C_j = \sum_i O_{ij} \quad L_1 = \sum_j O_{1j} \quad L_i = \sum_j O_{ij}$$

N Total de todas as ocorrências

$$N = \sum_i \sum_j O_{ij}$$

gl Graus de liberdade

$$gl = (l - 1) \times (c - 1)$$

e_{ij} Valor esperado para cada célula

$$e_{ij} = \frac{L_i \times C_j}{N}$$

m Número mínimo de estados nas variáveis

$$m = \min \{l, c\}$$

3.1.1 – Medidas de dependência

A) - Qui quadrado (χ^2)

É uma das medidas mais utilizadas para medir a dependência entre variáveis que se apresentam em tabelas de contingência ($l \times c$).

O valor de qui quadrado χ^2 global é a soma total dos valores de qui quadrado calculados para cada uma das células da tabela de contingência. Calcula-se segundo as expressões

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - e_{ij})^2}{e_{ij}} \quad \text{com} \quad e_{ij} = \frac{L_i \times C_j}{N}$$

Esta medida testa e mede a dependência / independência entre as duas variáveis. O seu valor varia entre 0 (zero) e infinito (∞).

O valor de qui quadrado revela a diferença entre as duas distribuições, permitindo estimar a probabilidade das duas distribuições serem iguais. Um valor elevado mostra a independência das variáveis. Por outro lado, quanto mais próximo de zero é o valor do qui quadrado mais aproximadas são as distribuições.

O valor de qui quadrado é avaliado pela tabela de valores críticos (Anexo 1), geralmente para valores significativos - probabilidade de 0,05, ou muito significativos - probabilidade de 0,01. É necessário calcular os graus de liberdade. A hipótese nula H_0 é a de que não existe diferença entre as variáveis, e deve ser rejeitada, isto é deve considerar-se que existe "diferença" entre elas se o valor é mais elevado que o valor crítico,

geralmente para uma probabilidade de ocorrência por simples acaso de 5%.

Sempre que a probabilidade de ocorrência do χ^2 mostrar dependência entre as variáveis (valores próximos de zero), pode-se utilizar o valor esperado para extrapolar as relações de dependência das variáveis

Utiliza-se para escalas nominais, ou superiores, desde que em quadros de contingência.

B) - Teste de probabilidade G quadrado (G^2)

É uma medida que, tal como o qui quadrado, se baseia nos valores observados e esperados das células da tabela de contingência.

O valor de Guê quadrado G^2 calcula-se segundo as expressões

$$G^2 = 2 \times \sum_i \sum_j \left(o_{ij} \times \ln \left(\frac{o_{ij}}{e_{ij}} \right) \right) \quad \text{com} \quad e_{ij} = \frac{L_i \times C_j}{N}$$

Esta medida testa e mede a dependência / independência entre as duas variáveis. O seu valor varia entre zero (0) e infinito (∞).

Assemelha-se ao qui quadrado. É utilizado para tabelas grandes e muito grandes, pois por ter o somatório do logaritmo numa razão em vez da soma dum quadrado das diferenças, é menos instável face aos arredondamentos do cálculo automático, e por isso, de maior confiança.

Usa-se para escalas nominais ou superiores.

C) - Teste Fi (para matrizes I x c)

É uma medida de associação que se baseia no qui quadrado. Como valores de qui quadrado para conjuntos com diferentes números de elementos não são comparáveis directamente, o teste **Fi** pondera o χ^2 calculado através do número de unidades de análise envolvidas (**N**).

O valor de Fi (ou Phi), que se representa pela letra grega ϕ , calcula-se através da fórmula

$$\phi = \left(\frac{\chi^2}{N} \right)^{\frac{1}{2}} = \sqrt{\left(\frac{\chi^2}{N} \right)}$$

Mede a dependência / independência entre as variáveis. Também se usa para efectuar comparações, por exemplo entre a evolução de duas variáveis no mesmo tempo, ou de uma só em dois períodos diferentes, ou

ainda, entre dois subconjuntos de tipo similar, como quando um diz respeito a homens e outro a mulheres, etc.

Varia entre zero (0) e um valor finito (T). É independente do tamanho da população estudada (ao contrário do χ^2), mas não do grau de liberdade do conjunto. Quanto maior for o grau de liberdade **gl**, maior pode ser o valor de **T**.

Mede a robustez da relação entre variáveis (a robustez encarada no sentido de Tukey).

Serve para escalas nominais ou superiores.

D) - Teste V de Cramer

Tal como o Phi, tem por base o qui quadrado. Calcula-se segundo a fórmula.

$$V = \left(\frac{\chi^2}{N \times (m - 1)} \right)^{\frac{1}{2}} = \sqrt{\left(\frac{\chi^2}{N \times (m - 1)} \right)}$$

Varia entre 0 e 1. Corresponde ao coeficiente fi ponderado também pelo valor **m**, que é o valor mínimo do número de estratos, classes ou categorias existente nas duas variáveis em causa, segundo a expressão

$$m = \min \{ (l), (c) \}$$

É independente do tamanho da população e dos graus de liberdade.

É muito usado para grandes tabelas, e nas circunstâncias em que tanto o número de unidades de análise como o de graus de liberdade (o número de estados nas variáveis) dos conjuntos é diferente, mas em que é interessante, mesmo assim, efectuar comparações entre eles.

Mede a dependência / independência entre as variáveis, mas é especialmente interessante para comparar tabelas, de qualquer dimensão, quando as populações são diferentes.

Serve para escalas nominais e superiores.

Para o caso especial de tabelas do tipo (2 x 2) reduz-se ao coeficiente fi. O mesmo acontece quando uma das variáveis tem somente dois estratos, classes ou categorias.

E) - Coeficiente de Contingência

Calcula-se pela expressão

$$C = \sqrt{\left(\frac{\chi^2}{N + \chi^2} \right)}$$

Mede a intensidade da dependência entre variáveis. Serve também para comparar tabelas.

Só deve ser usado para comparar tabelas com o mesmo número de linhas e de colunas, visto não ter factor correctivo.

Varia entre 0 e 1. Por isso é de interpretação mais clara do que o coeficiente fi.

Pode ser usado para escalas nominais e superiores.

Note-se que ϕ , V e C medem a intensidade da dependência e permitem comparar várias tabelas. Diferem apenas no valor do denominador.

3.1.2 – Medidas de associação

Apresentam-se de seguida três medidas de associação, entre duas variáveis, utilizadas em tabelas de contingência de $I \times c$ estados:

Gama	Tau b de Kendall	Tau c de Stuart

Estas medidas de associação comparam as ordenações dos pares numa tabela, ou em diferentes tabelas, indicando se essa ordem se mantém de uma para outra variável, ou não.

A leitura e interpretação do valor calculado são similares às do coeficiente de correlação linear simples de Pearson utilizado na análise de regressão linear entre duas variáveis de intervalo ou razão.

Quando o valor destas medidas é igual a zero (0) não existe associação (correlação) entre as variáveis; todos os pares concordam ou discordam igualmente.

Quando é igual a menos um (-1) há mais pares discordantes; existe uma relação no sentido da diagonal menor (direcção esquerda/alta direita/baixa);

Quando é igual a mais um (1) há mais pares concordantes; a relação é na direcção esquerda/baixa direita/alta.

Para uma mesma tabela de contingência, os valores calculados para estes três medidas é diferente, e por isso não tem sentido compará-los uns com os outros, mas apenas entre si. De uma maneira geral o

coeficiente gama é o que dá valores mais elevados (em módulo), seguido pelo tau bê de Kendall e só depois pelo tau cê de Stuart

$$|\Gamma_c| < |\Gamma_b| < |\text{Gama}|$$

O coeficiente gama privilegia a ordenação das variáveis, enquanto os taus, para além da ordenação das variáveis, dão uma grande ênfase à ordenação dos pares.

A) - Coeficiente Gama

É um coeficiente cujo valor varia entre -1 e $+1$, com o sentido que é habitual nos coeficientes de correlação, como se disse.

Toma em conta toda a tabela, e não apenas as ocorrências próximas da diagonal.

Utiliza-se para escalas ordinais e superiores.

Calcula-se pela expressão

$$\text{Gama} = \frac{P - Q}{P + Q}$$

em que **P** e **Q** se calculam do seguinte modo:

$$P = \sum_i \sum_j (O_{ij} \times A_{ij}) \quad Q = \sum_i \sum_j (O_{ij} \times D_{ij})$$

$$A_{ij} = \sum_{k>i} \sum_{n>j} O_{kn} + \sum_{k<i} \sum_{n<j} O_{kn}$$

$$D_{ij} = \sum_{k>i} \sum_{n<j} O_{kn} + \sum_{k<i} \sum_{n>j} O_{kn}$$

O valor **A_{ij}** é a das frequências com índices **i** e **c** maiores do que **i** e **j** (as colunas e linhas após a célula de que se trata), com a soma das frequências com índices **i** e **c** menores do que **i** e **j** (as colunas e linhas antes da célula de referência).

As células que se somam em cada caso são as que se encontram à esquerda/acima com as que estão à direita/abaixo da célula de referência.

Na Figura 4.1 pode ver-se graficamente as células que se somam para cada conjunto **i** e **j**, que estão marcadas a sombreado.

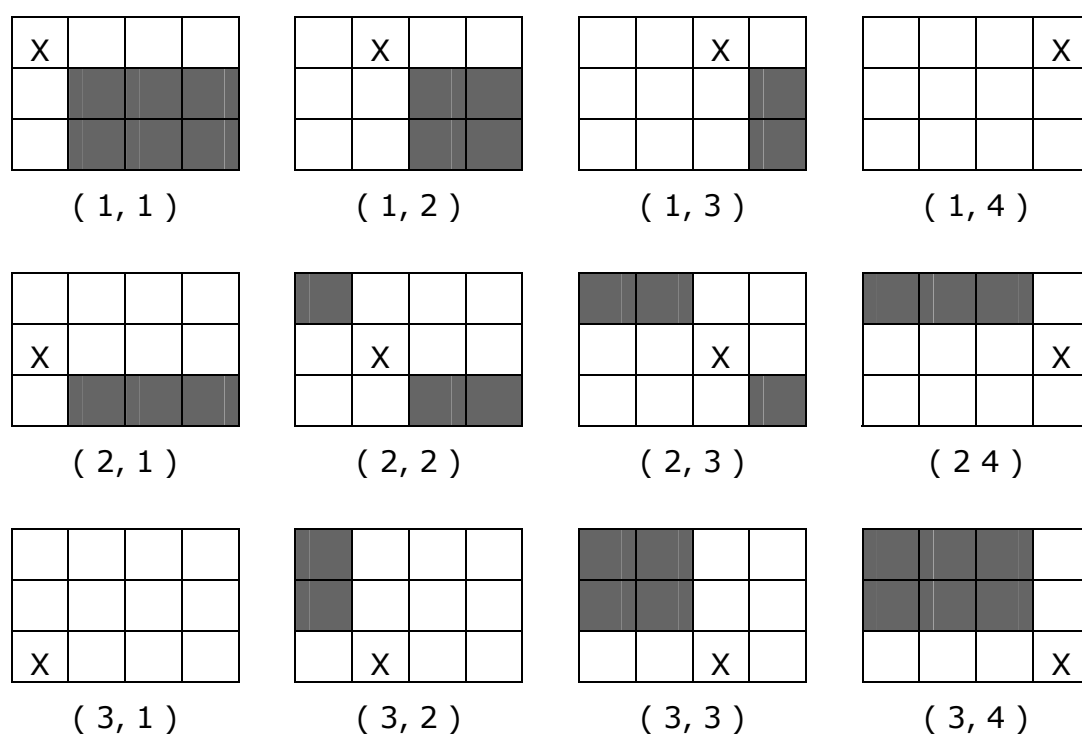


Figura 4.1 - Exemplificação do modo de calcular os valores A_{ij}

O valor D_{ij} é a das frequências com índices i maiores e c menores do que i e j (as colunas e linhas à esquerda e abaixo da célula de referência), com a soma das frequências com índices i menores e c maiores do que i e j (as colunas e linhas à direita e acima). Na Figura 4.2 mostra-se graficamente as células que se somam para cada conjunto de i e de j , marcadas a sombreado.

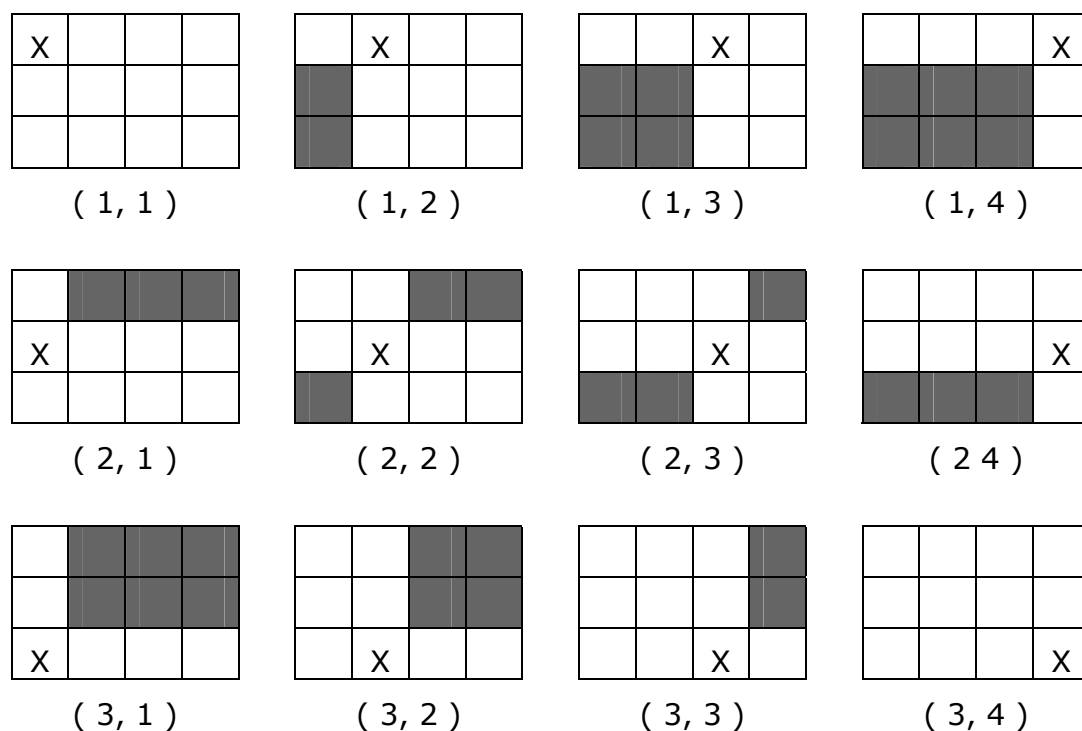


Figura 4.2 - Exemplificação do modo de calcular os valores D_{ij}

B) - Coeficiente Γ_b (de Kendall)

É um coeficiente que utiliza também os valores de P e Q calculados do mesmo modo que para o coeficiente gama.

Também varia entre -1 e $+1$.

Mede a associação das variáveis, e é utilizado para escalas de medida ordinais e superiores.

$$\Gamma_b = \frac{P - Q}{\sqrt{\left[\left(N^2 - \sum_i l_i^2 \right) \times \left(N^2 - \sum_j c_j^2 \right) \right]}}$$

Deve ser usado para tabelas de contingência quadradas, pois dá a mesma importância ao número de linhas do que ao número de colunas.

C) - Coeficiente Γ_c (de Stuart)

Este coeficiente usa também os valores de P e Q,

$$\Gamma_c = \left(\frac{P - Q}{N^2} \right) \times \left(\frac{m}{m - 1} \right)$$

Varia entre -1 e $+1$.

Mede a associação das variáveis.

Usa-se para escalas de medida ordinais e superiores.

É mais usado em tabelas rectangulares, pois corrige os graus de liberdade.

3.1.3 – Medidas de associação e predição

Existem medidas que para além de medirem a associação, também são preditivas. Apresentam-se três medidas que medem o conjunto da variação e os valores previstos, tanto consideradas para o conjunto da tabela, como só para as linhas ou as colunas.

As medidas são o coeficiente D de Somer (simétrico), que toma em conta as duas variáveis, não considerando dependências ou independências, e os coeficientes D assimétricos, com duas variantes, um com as linhas dependentes ($D_{j/i}$) ou com colunas dependentes ($D_{i/j}$).

Deve referir-se que existe uma relação entre estes últimos e o coeficiente tau bê de Kendall

$$(\Gamma_b)^2 = D_{j/i} \times D_{i/j}$$

A) - Coeficiente D de Somer (simétrico)

$$D = \frac{P - Q}{\frac{1}{2} \times \left[\left(N^2 - \sum_j c_j^2 \right) + \left(N^2 - \sum_i l_i^2 \right) \right]}$$

Varia entre - 1 e + 1.

Utiliza-se para escalas ordinais e superiores.

Mede a associação entre as variáveis, mas também tem valor preditivo.

B) - Coeficiente D j/i assimétrico (linhas dependente)

$$D_{j/i} = \frac{P - Q}{N^2 - \sum_i l_i^2}$$

Varia entre - 1 e + 1.

Utiliza-se para escalas ordinais e superiores.

Mede a associação entre as variáveis, e também a predição, considerando as linhas dependentes. Assume que as linhas dependem das colunas.

3.1.3.3 - Coeficiente D i/j assimétrico (colunas dependente)

$$D_{i/j} = \frac{P - Q}{N^2 - \sum_j c_j^2}$$

Varia entre - 1 e + 1.

Utiliza-se para escalas ordinais e superiores.

Mede a associação entre as variáveis, e também a predição, considerando dependentes as colunas.

3.1.3 – Medidas de predição

As medidas preditivas que se vão apresentar são:

O <u>coeficiente de incerteza U</u>	Simétrico
	Assimétrico i/j - colunas dependente j/i - linhas dependente
O <u>Tau assimétrico de Kruskal</u>	i/j - colunas dependente j/i - linhas dependente
O <u>coeficiente lambda λ</u>	

A) - Coeficiente de Incerteza U (simétrico)

É uma medida que se baseia nas probabilidades de ocorrência.

O conceito de incerteza deriva da teoria da informação, e da noção de entropia utilizada em Física e também nas ciências sociais. É uma medida do grau de organização/desorganização de um sistema.

A propósito, lembra-se que o conceito de entropia absoluta é igual

$$\text{Entropia} = -\sum_{i=1}^n (P_{(a_i)} \times \log(P_{(a_i)}))$$

em que $P_{(a_i)}$ é a probabilidade de um objecto pertencer à classe a_i e o logaritmo é de base dois se os dados forem de tipo binário.

O valor do coeficiente de incerteza **U** calcula-se através da expressão abaixo, em que

$$U = \frac{U_j + U_i - U_{ij}}{(U_j + U_i) / 2}$$

o valor dos diferentes termos calcula-se do seguinte modo:

$$U_j = -\sum_j \left(\frac{c_j}{N} \times \log \frac{c_j}{N} \right)$$

$$U_i = -\sum_i \left(\frac{l_i}{N} \times \log \frac{l_i}{N} \right)$$

$$U_{ij} = -\sum_i \sum_j \left(\frac{o_{ij}}{N} \times \log \frac{o_{ij}}{N} \right)$$

O coeficiente de incerteza varia entre 0 e 1. Quando $U=1$ não existe incerteza.

Aplica-se a escalas nominais ou superiores.

É um coeficiente preditivo, no sentido que compara a soma da "organização" das linhas e das colunas menos a "organização" das células, com metade da soma das linhas e das colunas (que é o valor máximo possível de "organização"). Ou dito de outro modo, compara a "organização" que se deve às linhas e às colunas menos a das células, com a "organização" máxima devido às linhas e às colunas apenas.

B) - Coeficiente de Incerteza das linhas devido às colunas (assimétrico)

$$U_{j/i} = \frac{U_j + U_i - U_{ij}}{U_j}$$

Varia também de 0 a 1. Quando $U=1$ não há incerteza.

Aplica-se a variáveis medidas em escalas nominais e superiores.

Pode ser interpretado com o ganho na predição das linhas quando se conhece o valor das colunas, comparativamente com o caso de não se conhecer o seu respectivo valor.

C) - Coeficiente de Incerteza das colunas devido às linhas (assimétrico)

É uma medida coluna dependente.

$$U_{i/j} = \frac{U_j + U_i - U_{ij}}{U_i}$$

O seu valor varia também entre 0 e 1. Quando $U=1$ não há incerteza na relação entre as duas variáveis..

Aplica-se a variáveis nominais e superiores.

Pode interpretar-se como o ganho na predição das colunas quando se sabe o valor das linhas, relativamente a não se conhecer.

D) - Coeficiente Tau j/i de Kruskal (assimétrico)

$$\Gamma_{j/i} = \frac{N \times \sum_i \sum_j \left(\begin{array}{c} o_{ij}^2 \\ \hline l_i \end{array} \right) - \sum_j c_j^2}{N^2 - \sum_j c_j^2}$$

É uma medida linha dependente. É uma medida preditiva. Ao considerar uma variável dependente, a medida de predição informa sobre a aleatoriedade, ou não, da relação.

Varia de 0 a 1. Se $\Gamma_{j/i} = 0$ não pode haver previsão.

Usa-se para variáveis nominais e superiores.

E) - Coeficiente Tau i/j de Kruskal (assimétrico)

$$\Gamma_{i/j} = \frac{N \times \sum_i \sum_j \left(\begin{array}{c} o_{ij}^2 \\ \hline c_j \end{array} \right) - \sum_i l_i^2}{N^2 - \sum_i l_i^2}$$

Como a anterior é uma medida preditiva. Ao considerar uma variável dependente mede na predição a aleatoriedade do processo.

Varia de 0 a 1. Se $\Gamma_{i/j} = 0$ não há previsão.

Pode ser utilizado com todas as escalas (nominais e superiores).

F) - Coeficiente Lambda

É uma medida de previsão que se baseia nas modas, seleccionando como, ou para, melhor estratégia de previsão, a categoria com mais unidades de análise, sujeitos, indivíduos ou casos.

$$\lambda = \frac{\sum_i (\max_j o_{ij}) + \sum_j (\max_i o_{ij}) - \max_j c_j - \max_i l_i}{2 \times N - \max_j c_j - \max_i l_i}$$

em que

$$\sum_j (\max_i o_{ij})$$

é a soma dos **j** elementos que são o máximo (a moda) de cada coluna

e

$$\sum_i (\max_j o_{ij})$$

é a soma dos **i** elementos que são os máximos de cada linha

Varia entre zero (0) caso em que não é possível prever uma com a outra e um (1) em que a previsão de uma a partir da outra, se fosse feita, viria sem erro.

G) - Coeficiente Lambda j/i (assimétrico)

É uma medida linha dependente.

$$\lambda_{j/i} = \frac{\sum_i (\max_j o_{ij}) - \max_j c_j}{N - \max_j c_j}$$

H) - Coeficiente Lambda i/j (assimétrico)

$$\lambda_{i/j} = \frac{\sum_j (\max_i o_{ij}) - \max_i l_i}{N - \max_i l_i}$$

É uma medida coluna dependente.

3.2. – Tabelas de contingência (l x c) a 2 estados

As tabelas (ou quadros) de contingência de duas variáveis, que combinam dois apenas estados (2×2), permitem simplificar bastante algumas das fórmulas já apresentadas.

São tabelas que relacionam 2 estados de uma variável-linha com 2 estados da variável-coluna.

Do ponto de vista da notação matemática tem-se

		V_c		
		c₁	c₂	
V_l	l₁	a	b	L₁
	l₂	c	d	L₂
		C₁	C₂	N

Quadro 4.17 - Tabela de contingência 2 x 2

para os seguintes graus de liberdade $(ln-1) \times (lc-1) = gl$

que neste caso é sempre $(2-1) \times (2-1) = 1$

em que **V_l** - é a variável-linha **V_c** - é a variável-coluna

l₁ e **l₂** os dois estados da variável linha

c₁ e **c₂** os dois estados da variável segundo as colunas

a número de ocorrências dos estados **1** nas variáveis linha e coluna;

b número de ocorrências do estado **1** na variável-linha com o estado 2 na variável-coluna;

c número de ocorrências do estado **2** na variável-linha com o estado 1 na variável-coluna;

d número de ocorrências dos estados **1** nas variáveis linha e coluna;

C_j Total das ocorrências do estado **j** da variável-coluna

L_i Total das ocorrências do estado **i** da variável-linha

$$C_1 = \sum_i O_{i1} \quad C_j = \sum_i O_{ij} \quad L_1 = \sum_j O_{1j} \quad L_i = \sum_j O_{ij}$$

N Total de todas as ocorrências

$$N = \sum_i \sum_j O_{ij}$$

As medidas mais utilizadas na análise de tabelas de contingência com dois estados em cada variável (2×2), são:

3.2.1. - Qui quadrado (χ^2)

$$\chi^2 = \frac{N \times (a \times d - b \times c)^2}{L_1 \times L_2 \times C_1 \times C_2}$$

Varia de 0 (zero) a infinito (∞).

Usa-se para escalas nominais ou superiores.

Mede a dependência/independência das duas variáveis.

3.2.2. - Qui quadrado (χ^2) corrigido por Yates

$$\chi_y^2 = \frac{N \times (|a \times d - b \times c| - N/2)^2}{L_1 \times L_2 \times C_1 \times C_2}$$

Varia de 0 (zero) a infinito (∞). Usa-se para escalas nominais ou superiores.

Mede a dependência/independência das duas variáveis.

Procura melhorar a adaptação à distribuição de χ^2 .

3.2.3. - Coeficiente Fi (ϕ)

$$\phi = \frac{(a \times d - b \times c)}{\sqrt{L_1 \times L_2 \times C_1 \times C_2}}$$

Varia de -1 a +1, com o sentido do coeficiente de correlação.

Utiliza-se para medir a dependência/independência das duas variáveis e para comparar conjuntos diversos de duas variáveis. É aplicável a diferentes tabelas 2×2 , tanto para as mesmas variáveis como para uma só em dois tempos ou localizações diferentes.

Usa-se para escalas nominais ou superiores.

Este coeficiente ϕ é igual ao coeficiente V de Cramer para quadros de contingência do tipo 2 x 2.

3.2.4. - Coeficiente Q de Yule

$$Q_{\text{Yule}} = \frac{(a \times d - b \times c)}{(a \times d + b \times c)}$$

Varia de -1 a $+1$. Serve para comparar conjuntos iguais ou diferentes de duas variáveis (a dois estados). Pode usar-se para comparar a mesma variável em dois tempos ou espaços diferentes.

Para escalas nominais e superiores.

3.2.5. - Coeficiente Y

$$Y = \frac{\sqrt{a \times d} - \sqrt{b \times c}}{\sqrt{a \times d} + \sqrt{b \times c}}$$

Este coeficiente Y deve-se também a Yule.

Varia entre -1 e $+1$.

É similar ao Q , mas compensa o alto valor dos produtos pela respectiva raiz quadrada, de modo que em vez dos produtos dos valores diagonais passa a comparar as suas respectivas médias geométricas.

Serve para comparar tabelas 2 x 2, para escalas nominais e superiores.

3.2.6. - Razão dos produtos cruzados (α)

$$\alpha = \frac{a \times d}{b \times c}$$

Este coeficiente α (alfa) varia entre 0 (zero) e ∞ (infinito).

Usa-se para escalas nominais e superiores.

Serve para comparar tabelas.

É a razão das proporções.

É zero, quando um ou os dois valores da diagonal principal têm o valor de. É infinito quando uma ou as duas frequências na diagonal secundária são iguais a zero. É indeterminado quando existem frequências de zero simultaneamente na diagonal principal e na secundária (cf. Figura 3).

0	f1
f2	0

$\alpha=0$

f1	0
f2	0

α indeterminado

f1	0
0	f2

$\alpha=\infty$

Figura 4.3 - Exemplificação dos valores de alfa

3.2.7. - Logaritmo de α

Como por vezes o coeficiente alfa atinge valores muito elevados, também se usa o seu logaritmo.

O logaritmo dos produtos cruzados (LPC) também varia entre 0 (zero) e ∞ (infinito), mas apresenta sempre valores muito menores

O seu valor calcula-se determinando o logaritmo neperiano do coeficiente alfa.

$$LPC = \ln (\alpha)$$

3.3. – Exemplos

3.3.1- Medidas de dependência para 2 variáveis a n estados

Considere-se como exemplo a seguinte tabela de contingência, em que se apresentam 2430 indivíduos descritos por duas variáveis. Uma é a raça, segundo dois estados ou categorias - *Branca* e *Não-branca* - e a outra o rendimento, em quatro classes - de menos de 4000 dólares, entre 4000 e 8000, de 8000 a 12500 e mais deste valor - conforme Quadro 4.18).

US\$	Branca	Não-branca	
< 4000	396	98	494
4000 - 7999	526	70	596
8000 - 12499	612	64	676
> 12500	624	40	664
	2158	272	2430

Quadro 4.18 - Tabela de contingência do primeiro exemplo

O que se pode concluir deste quadro surge como resultante dos valores das diversas medidas que se podem calcular.

Como não é um quadro (2X2) calculam-se as medidas apropriadas ao tipo de tabelas (l x c).

A) - Qui quadrado (χ^2)

É uma medida que testa a relação de dependência ou independência entre as duas variáveis.

Quanto mais elevado é o seu valor, mais relacionadas entre si estão as variáveis.

Veja-se o processo de cálculo do valor de χ^2

1).- A tabela a tratar é esta,

396	98	494
526	70	596
612	64	676
624	40	664
2158	272	2430

2).- Os graus de liberdade são 3, pois

g. l. = (nº. colunas-1 x nº linhas-1)

g.l. = (2-1) x (4-1) = 1 x 3 = 3

3).- Os valores observados são

396	98	= O_{ij}
526	70	
612	64	
624	40	

4).- E os esperados

E_{ij}	438,70453	55,29547
	529,28724	66,71276
	600,33251	75,66749
	589,67572	74,32428

5).- A diferença entre valores observados e esperados é

- 42,70453	42,70453
- 3,28724	3,28724
11,66749	- 11,66749
34,32428	- 34,32428

6).- E o respectivo quadrado é

1823,67660	1823,67660
10,80597	10,80597
136,13032	136,13032
1178,15619	1178,15619

7).- A razão entre a diferença quadrática e o valor esperado é

4,15696	32,98058
0,02042	0,16198
0,22676	1,79906
1,99797	15,85157

8).- Pelo que o valor do qui quadrado é

$$\chi^2 = 57,195285 \text{ com g.l.} = 3$$

Consultado a tabela de Qui quadrado, do Anexo 1, os valores críticos para 3 graus de liberdade, são

Probabilidade	0.05	0.01	0.005	0.001
χ^2	7,82	1,34	12,84	16,27

Como a probabilidade de as categorias de uma variável pertencerem à mesma população da outra, isto é, terem distribuições semelhantes nas categorias da outra, é muitíssimo baixa, pode-se concluir que, para este conjunto de indivíduos, existe uma forte relação de dependência entre a raça e o rendimento.

B) - Teste de probabilidade G quadrado (G^2)

Como parte do cálculo é semelhante ao do qui quadrado, vão ser omitidos alguns dos passos.

Assim, veja-se

1).- Os valores observados são

396	98	$= o_{ij}$
526	70	
612	64	
624	40	

2).- E os esperados

$e_{ij} =$	438,70453	55,29547
	529,28724	66,71276
	600,33251	75,66749
	589,67572	74,32428

3).- As razões entre observados e esperados são

0,90266	1,77230	$= \frac{o_{ij}}{e_{ij}}$
0,99379	1,04928	
1,01944	0,84581	
1,05821	0,53818	

4).- E os seus logaritmos,

$\ln \frac{o_{ij}}{e_{ij}} =$	-0,10241	0,57228
	-0,00623	0,04810
	0,01925	-0,16747
	0,05658	-0,61956

5).- O passo seguinte é o cálculo de

$$\left[o_{ij} \times \ln \left(\frac{o_{ij}}{e_{ij}} \right) \right]$$

obtendo-se

-40,55512	56,08309
-3,27701	3,36693
11,78014	-10,71779
35,30443	-24,78233

pelo que

$$G^2 = 2 \times 27,2023385 = 54,404677$$

C) - Teste Fi (para matrizes I x c)

Calcula-se a partir dos valores de Qui quadrado e do número total de unidades de análise em estudo (**N**), segundo

$$\phi = \sqrt[2]{\frac{\chi^2}{N}} = \sqrt[2]{\frac{57,19529}{2430}} = \sqrt[2]{0,02353715..} = 0,153418234..$$

obtendo-se, por arredondamento, $\phi = 0,15342$

D) - Teste V de Cramer

Esta medida baseia-se no valor de qui quadrado, no número de elementos em análise (**N**), e no número de linhas e de colunas da tabela de contingência.

É a raiz quadrada da razão entre o qui quadrado e o número de elementos em análise multiplicado pelo valor do grau de liberdade mínimo de linhas e colunas

$$V = \sqrt{\frac{\chi^2}{N \times (\min\{(l), (c)\} - 1)}} = \sqrt{\frac{57,195285}{2430 \times (2 - 1)}} = \sqrt{0,02353715} = 0,153418234$$

o que por arredondamento dá $V = 0,15342$

E) - Coeficiente de Contingência

Do mesmo modo, também se calcula com o qui quadrado e N, o número total de unidades analisadas.

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} = \sqrt{\frac{57,195285}{2430 + 57,195285}} =$$

$$\sqrt{\frac{57,195285}{2487,195285}} = \sqrt{0,022995896 \dots} =$$

$$C = 0,15164398 \dots$$

3.3.2- Medidas de associação para 2 variáveis a n estados**A) - Cálculo de P e Q**

O primeiro passo para o cálculo destas medidas é a determinação dos valores de P e Q. Para isso é necessário calcular as matrizes de valores auxiliares A e D.

$$P = \sum_i \sum_j (O_{ij} \times A_{ij}) \qquad Q = \sum_i \sum_j (O_{ij} \times D_{ij})$$

$$A_{ij} = \sum_{k>i} \sum_{n>j} O_{kn} + \sum_{k<i} \sum_{n<j} O_{kn}$$

$$D_{ij} = \sum_{k>i} \sum_{n<j} O_{kn} + \sum_{k<i} \sum_{n>j} O_{kn}$$

No exemplo que estamos tratando, os valores são:

$$\begin{aligned}
 A_{11} &= 70 + 64 + 40 & A_{12} &= 0 & A_{21} &= 64 + 40 \\
 A_{22} &= 396 & A_{31} &= 40 & A_{32} &= 396 + 526 \\
 A_{41} &= 0 & A_{42} &= 396 + 526 + 612 \\
 \text{e} \quad D_{11} &= 0 & D_{12} &= 526 + 612 + 624 \\
 & D_{21} &= 98 & D_{22} &= 612 + 624 \\
 & D_{31} &= 98 + 70 & D_{32} &= 624 \\
 & D_{41} &= 98 + 70 + 64 & D_{42} &= 0
 \end{aligned}$$

1).- Assim sendo, os valores de A e D são

$$\begin{array}{|c|c|} \hline 174 & 0 \\ \hline 104 & 396 \\ \hline 40 & 922 \\ \hline 0 & 1534 \\ \hline \end{array} = \mathbf{A}_{ij} \quad \mathbf{A}_{ij} \times \mathbf{O}_{ij} = \begin{array}{|c|c|} \hline 68904 & 0 \\ \hline 54704 & 27720 \\ \hline 24480 & 59008 \\ \hline 0 & 61630 \\ \hline \end{array}$$

$$\begin{array}{|c|c|} \hline 0 & 1762 \\ \hline 98 & 1236 \\ \hline 168 & 624 \\ \hline 232 & 0 \\ \hline \end{array} = \mathbf{D}_{ij} \quad \mathbf{D}_{ij} \times \mathbf{O}_{ij} = \begin{array}{|c|c|} \hline 0 & 172676 \\ \hline 51548 & 86520 \\ \hline 102816 & 39936 \\ \hline 144768 & 0 \\ \hline \end{array}$$

2).- E os de $P = 296176$ e $Q = 598264$

Com estes valores é possível calcular os coeficientes que seguem.

B) - Coeficiente Gama

Como se sabe, o Coeficiente Gama é a razão da diferença de P e Q com a sua respectiva soma,

$$\text{Gama} = \frac{P - Q}{P + Q} = \frac{296176 - 598264}{296176 + 598264} = \frac{-302088}{894440} = -0,3377...$$

C) - Coeficiente Γ_b (de Kendall)

$$\Gamma_b = \frac{P - Q}{\sqrt{\left[\left(N^2 - \sum_i l_i^2 \right) \times \left(N^2 - \sum_j c_j^2 \right) \right]}}$$

Calcula-se através da seguinte expressão
que no exemplo que se está tratando, vem a ser igual a

$$\begin{aligned} \Gamma_b &= \frac{296176 - 598264}{\sqrt{(5904900 - 1497124) \times (5904900 - 4730948)}} = \frac{-302088}{\sqrt{4407776 \times 1173952}} \\ &= \frac{-302088}{\sqrt{5174517450752}} = \frac{-302088}{2274756,6...} = -0,1328 \dots \end{aligned}$$

D) - Coeficiente Γ_c (de Stuart)

Calcula-se com a fórmula

$$\Gamma_c = \left(\frac{P - Q}{N^2} \right) \times \left(\frac{m}{m - 1} \right)$$

que assume aqui os seguintes valores

$$\Gamma_c = \frac{-302088}{5904900} \times \frac{2}{1} = -0,102317736 \dots$$

3.3.3- Medidas de associação e predição de 2 variáveis a n estados

A) - Coeficiente D de Somer (simétrico)

$$\begin{aligned} D &= \frac{296176 - 598264}{0,5 \times ((5904900 - 4730948) + (5904900 - 1497124))} = \\ &= \frac{-302088}{0,5 \times (1173952 + 4407776)} = \frac{-302088}{2790864} = -0,1082417 \dots \end{aligned}$$

B) - Coeficiente D j/i assimétrico (dependência entre linhas)

$$D_{j/i} = \frac{296176 - 598264}{5904900 - 1497124} = \frac{-302088}{4407776} = -0,068535 \dots$$

C) - Coeficiente D_{i/j} assimétrico (dependência entre colunas)

$$D_{i/j} = \frac{296176 - 598264}{5904900 - 4730948} = \frac{-302088}{1173952} = -0,2573257 \dots$$

Deve notar-se que $\tau_b^2 = D_{j/i} \times D_{i/j}$, como se verifica dos dados, pois

$$-0,132800 \times -0,132800 = 0,0176358 \quad e$$

$$-0,068535 \times -0,257326 = 0,0176358$$

3.3.4- Medidas de predição para 2 variáveis a n estados**A) - Coeficiente de Incerteza U (simétrico)**

A primeira operação a fazer, é a de calcular os valores de U_i , U_j e U_{ij}

$$U_i = -[(-0,14065) + (-0,14970) + (-0,15457) + (-0,15396)] = -(-0,59889) = 0,59889$$

$$U_j = -[(-0,04578) + (-0,10645)] = -(-0,15223) = 0,15223$$

$$U_{ij} = -[(-0,12840) + (-0,05623) + (-0,14386) + (-0,04437) + (-0,15082) + (-0,04159) + (-0,15161) + (-0,02935)] = -(-0,74625) = 0,74626$$

Determinados estes valores é fácil calcular todos os coeficientes de incerteza.

O simétrico é igual a

$$U = \frac{U_j + U_i - U_{ij}}{(U_j + U_i)/2} = \frac{0,15223 + 0,59889 - 0,74626}{(0,15223 + 0,59889)/2} = \frac{0,004861}{0,375565} = 0,012944$$

B) - Coeficiente de Incerteza das linhas devido às colunas (assimétrico)

$$U_{j/i} = \frac{U_j + U_i - U_{ij}}{U_j} = \frac{0,15223 + 0,59889 - 0,74626}{0,15223} = \frac{0,004861}{0,15223} = 0,031935$$

C) - Coeficiente de Incerteza das colunas devido às linhas (assimétrico)

$$U_{i/j} = \frac{U_j + U_i - U_{ij}}{U_i} = \frac{0,15223 + 0,59889 - 0,74626}{0,59889} = \frac{0,004861}{0,59889} = 0,008118$$

D) - Coeficiente Tau j/i de Kruskal (assimétrico)

Como $\sum \sum (o_{ij}^2 / l_i) = 1958,2632$ e $\sum (c_j^2) = 4730948$ tem-se que

$$\Gamma_{j/i} = \frac{2430 \times 1958,2632 - 4730948}{5904900 - 4730948} = \frac{27631,5}{1173952} = 0,023537$$

E) - Coeficiente Tau i/j de Kruskal (assimétrico)

E como $\sum \sum (o_{ij}^2 / c_j) = 629,1358$ e $\sum (l_i^2) = 1497124$

$$\Gamma_{i/j} = \frac{2430 \times 629,1358 - 1497124}{5904900 - 1497124} = \frac{31675,994}{4407776} = 0,007186$$

F) - Coeficiente Lambda

Os coeficientes lambda baseiam-se nas modas, os maiores valores observados segundo as linhas e segundo as colunas, e nos máximos do total de linhas e do total das colunas. No exemplo que estamos tratando, que se relembra é

396	98	494
526	70	596
612	64	676
624	40	664
2158	272	2430

Assim, $\sum (\max_j o_{ij}) = 396 + 526 + 612 + 624 = 2158$

e $\sum (\max_i o_{ij}) = 624 + 98 = 722$.

Como o $\max_j c_j = 2158$ e o $\max_i l_i = 676$

$$\lambda = \frac{2158 + 722 - 2158 - 676}{2 \times 2430 - 2158 - 676} = \frac{46}{2026} = 0,0227048371$$

G) - Coeficiente Lambda j/i (assimétrico)

$$\lambda_{j/i} = \frac{2158 - 2158}{2430 - 2158} = \frac{0}{272} = 0$$

H) - Coeficiente Lambda i/j (assimétrico)

$$\lambda_{i/j} = \frac{722 - 676}{2430 - 676} = \frac{46}{1754} = 0,0262257697$$

I) - Resultados gerais:

MEDIDA		VALOR
Símbolo	Nome	
χ^2	Qui quadrado	57,195285
G^2	G quadrado	54,404677
ϕ	Phi	0,153418
V	V de Cramer	0,153418
C	Coeficiente de contingência	0,151644
Gama	Coeficiente Gama	-0,337740
Γb	Tau b de Kendall	-0,132800
Γc	Tau c de Stuart	-0,102318
D	D simétrico de Somer	-0,108242
D j/i	D dependência de linhas	-0,068535
D i/j	D dependência de colunas	-0,257326
U	Coeficiente de Incerteza	0,129449
U i/j	Coef. Incerteza - col. dependente	0,008118
U j/i	Coef. Incerteza - linh. dependente	0,031935
$\Gamma i/j$	Tau i/j de Kruskal	0,007186
$\Gamma j/i$	Tau j/i de Kruskal	0,023537
λ	Lambda simétrico	0,022705
$\lambda i/j$	Lambda i/j	0,000000
$\lambda j/i$	Lambda j/i	0,026226

3.3.5- Tabelas de Contingência de 2 variáveis a 2 estados

Alguns dos coeficientes têm o seu cálculo simplificado quando as tabelas são do tipo 2 x 2 (só com dois estados na variável linha e na variável coluna). Por outro lado, para estas tabelas podem utilizar-se alguns coeficientes diferentes dos utilizados para tabelas mais complexas.

O exemplo de tabela 2 x 2 que se irá tratar respeita a um conjunto de 117 indivíduos, em que se relaciona a idade e o número de cigarros que fumam por dia.

Nas linhas tem-se na linha 1 - Os que fumam menos de 20 cigarros por dia
na linha 2 - Os que fumam mais de 20 cigarros/dia
Nas colunas na coluna 1 - Os que têm mais de 40 anos de idade
na coluna 2 - Os com menos de 40 anos.

	Mais de 40	Menos de 40	
20 e mais cigarros/dia	3	16	19
Menos de 20 cig/dia	40	58	98
	43	74	117

A). - Qui quadrado (χ^2)

$$\chi^2 = \frac{117 \times (3 \times 58 - 16 \times 40)^2}{19 \times 98 \times 43 \times 74} = \frac{117 \times (174 - 640)^2}{1862 \times 3182} = \frac{117 \times (-466)^2}{5924884} =$$

$$= \frac{117 \times 217156}{5924884} = \frac{25407252}{5924884} = 4,28822775$$

B). - Qui quadrado (χ^2) corrigido por Yates

$$\chi^2 = \frac{117 \times (|3 \times 58 - 16 \times 40| - 117/2)^2}{19 \times 98 \times 43 \times 74} = \frac{117 \times (|174 - 640| - 58,5)^2}{1862 \times 3182} =$$

$$= \frac{117 \times (466 - 58,5)^2}{5924884} = \frac{117 \times (407,5)^2}{5924884} = \frac{117 \times 166056,25}{5924884} =$$

$$= \frac{19428581,25}{5924884} = 3,27914964$$

C). - Coeficiente Fi (ϕ)

$$\phi = \frac{3 \times 58 - 16 \times 40}{\sqrt{19 \times 98 \times 43 \times 74}} = \frac{174 - 640}{\sqrt{1862 \times 3182}} = \frac{-466}{\sqrt{5924884}} =$$

$$= -\frac{466}{2434,1084} = -0,191445865$$

D). - Coeficiente Q de Yule

$$Q = \frac{3 \times 58 - 16 \times 40}{3 \times 58 + 16 \times 40} = \frac{174 - 640}{174 + 640} = \frac{-466}{814} = -0,5724157$$

E). - Coeficiente Y

$$Y = \frac{\sqrt{3 \times 58} - \sqrt{16 \times 40}}{\sqrt{3 \times 58} + \sqrt{16 \times 40}} = \frac{\sqrt{174} - \sqrt{640}}{\sqrt{174} + \sqrt{640}} = \frac{13,190905 - 25,298221}{13,190905 + 25,298221} =$$

$$= \frac{-12,107316}{38,489126} = -0,314564559$$

F). - Razão dos produtos cruzados (α)

$$\alpha = \frac{3 \times 58}{16 \times 40} = \frac{174}{640} = 0,271875$$

G). - Logaritmo de α

$$\ln(\alpha) = \ln 0,271875 = -1,302412877$$

H). - Resultados gerais

MEDIDA		VALOR
Símbolo	Nome	
χ^2	Qui quadrado	4,288228
$\chi^2 Y$	Qui quadrado, corrigido por Yates	3,270150
ϕ	Coeficiente Phi	-0,191446
Q	Q de Youle	-0,572482
Y	Y de Youle	-0,314565
α	Razão alfa	0,271875
$\ln \alpha$	Logaritmo neperiano de alfa	-1,302413

4. – Análise de Tabelas de Burt

4.1. – Análise de χ^2

4.1.1- Exemplo

Considere-se o exemplo já antes referido de um quadro de dados de quatro indivíduos em três variáveis, a idade, o sexo e a pontuação num teste (Quadro 13), medidas respectivamente em escalas de medida de razão, nominal e ordinal. (Quadro 4.19)

i	Idade	Sexo	Exame final
1	35	M	90
2	54	M	110
3	18	F	98
4	45	F	137

Quadro 4.19 - Matriz de atributos inicial

As variáveis têm três, dois e três estados cada, a saber, jovens (Jv) com menos de 20 anos, médios (Md) com idades entre 20 e 50 anos e velhos (Ve) com mais de cinquenta, do sexo masculino (Ma) e do feminino (Fe), e Reprovados (Rep) com menos de 100 no teste, sofrível (Sf) com nota entre 100 e 135 e bom (Bm) com mais de 135.

No quadro 4.20 podem ver-se os valores desta matriz já classificados

i	Idade	Sexo	Exame final
1	Md	Ma	Rep
2	Ve	Ma	Sf
3	Jv	Fe	Rep
4	Md	Fe	Bm

Quadro 4.20 - Matriz de atributos classificada

E com estes dados é possível construir a respectiva matriz, quadro ou tabela de Burt (quadro 4.21)

		Idade			Sexo		Teste		
		Jv	Md	Ve	Ma	Fe	Rep	Sf	Bm
Idade	Jv	1	0	0	0	1	1	0	0
	Md	0	2	0	1	1	1	0	1
	Ve	0	0	1	1	0	0	1	0
Sexo	Ma	0	1	1	2	0	1	1	0
	Fe	1	1	0	0	2	1	0	1
Teste	Rep	1	1	0	1	1	2	0	0
	Sf	0	0	1	1	0	0	1	0
	Bm	0	1	0	0	1	0	0	1

Quadro 4.21 - Tabela de Burt para os dados

A leitura da primeira linha da tabela informa que existe apenas um jovem, que é mulher e reprovou no teste, A segunda linha diz que existem dois sujeitos de meia idade, um deles homem e o outro mulher, que um deles reprovou e o outro teve bom no teste (mas daqui não se sabe se foi o homem ou mulher o que passou). Mas vendo as linhas relativas ao sexo, pode saber-se que existem duas mulheres, uma que reprovou e outra que teve bom. E assim sucessivamente.

A tabela de Burt contem toda a informação existente na matriz de atributos de dados classificados, e a sua análise também se baseará nessa mesma informação total.

Para a analisar utilize-se o método do χ^2 , que é uma medida estatística que compara os desvios entre as frequências observadas e as esperadas (a partir dos respectivos totais marginais).

Da tabela de Burt do quadro 4.21, retiram-se os valores dos totais marginais de linha

3 6 3 6 6 6 3 3

e dos totais marginais de coluna

3 6 3 6 6 6 3 3

que como é óbvio terão de ser iguais. Com estes valores constrói-se uma tabela auxiliar (quadro 4.22) com os valores esperados para cada célula, calculados com a seguinte expressão

$$e_{ij} = \frac{\bar{l}_i \times \bar{c}_j}{Nt}$$

em que se multiplicam os respectivos totais marginais da linha e da coluna que depois se dividem por **Nt** o total geral da tabela.

		Idade			Sexo		Teste		
		Jv	Md	Ve	Ma	Fe	Rep	Sf	Bm
Idade	Jv	0.25	0.50	0.25	0.50	0.50	0.25	0.50	0.25
	Md	0.50	1.00	0.50	1.00	1.00	0.50	1.00	0.50
	Ve	0.25	0.50	0.25	0.50	0.50	0.25	0.50	0.25
Sexo	Ma	0.50	1.00	0.50	1.00	1.00	0.50	1.00	0.50
	Fe	0.50	1.00	0.50	1.00	1.00	0.50	1.00	0.50
Teste	Rep	0.50	1.00	0.50	1.00	1.00	1.00	0.50	0.50
	Sf	0.25	0.50	0.25	0.50	0.50	0.50	0.25	0.25
	Bm	0.25	0.50	0.25	0.50	0.50	0.50	0.25	0.25

Quadro 4.22 - Valores esperados

Com os valores observados da tabela de Burt (no quadro 4.21 e os valores esperados (do quadro 4.22) calculam-se os valores de χ^2 , segundo a expressão

$$\chi^2 = \frac{(f_o - f_e)^2}{f_e}$$

obtendo-se um quadro dos valores de χ^2 (quadro 4.23)

		Idade			Sexo		Teste		
		Jv	Md	Ve	Ma	Fe	Rep	Sf	Bm
Idade	Jv	2.25	0.50	0.25	0.50	0.50	0.50	0.25	0.25
	Md	0.50	1.00	0.50	0.00	0.00	0.00	0.50	0.50
	Ve	0.25	0.50	2.25	0.50	0.50	0.50	2.25	0.25
Sexo	Ma	0.50	0.00	0.50	1.00	1.00	0.00	0.50	0.50
	Fe	0.50	0.00	0.50	1.00	1.00	0.00	0.50	0.50
Teste	Rep	0.50	0.00	0.50	0.00	0.00	1.00	0.50	0.50
	Sf	0.25	0.50	2.25	0.50	0.50	0.50	2.25	0.25
	Bm	0.25	0.50	0.25	0.50	0.50	0.50	0.25	2.25

Quadro 4.23 - Valores de χ^2

A matriz é simétrica. As relações entre as variáveis podem ser vistas na parte superior direita da matriz de χ^2 (a *bold* e sombreado), em que se têm as relações entre as três variáveis, a saber, Idade/Sexo e Idade/Teste na primeira linha e Sexo/Teste na segunda.

A análise tanto pode ser feita agregada (quadro 4.24) como célula a célula (quadro 4.25), e tanto em valor absoluto como em valor percentual do universo que interesse.

No quadro 4.24 b), a percentagem é sobre o total das relações entre variáveis diferentes, ou seja 9,00 (pois 2,00+5,00+2,00=9,00).

	Idade	Sexo	Teste
Idade	8,00	2,00	5,00
Sexo		4,00	2,00
Teste			8,00

Quadro 4.24 a) Valor do χ^2 por variável

	Idade	Sexo	Teste
Idade		22,22	55,56
Sexo			22,22
Teste			

Quadro 4.24 b) Percentagem do total de variáveis diferentes

Dos Quadros 4.24 verifica-se que, de um modo geral, as relações mais importantes são as que relacionam o resultado do teste com a idade. Tanto a relação do sexo com a idade como a do sexo com o resultado do teste têm menos importância.

Analisando célula a célula, observa-se que o resultado mais excepcional (o menos esperado) é o valor de *Sofrível* obtido pelo *Velho*; os doze valores de 0,50 são interessantes, quer no que respeita às ocorrências, quer às não ocorrências de valores. No Quadro 4.25 resume-se as ocorrências dos diversos valores.

Valor de χ^2	Ocorrências	% unitária	% total	% cumulativa
2,25	1	25,00	25,00	25,00
0,50	12	5,56	66,67	91,67
0,25	3	2,78	8,33	100,00
0,00	5	0,00	0,00	100,00

Quadro 4.25 - Resumo do valor de χ^2 das relações entre variáveis

O total das três células é, como sabemos, igual a 9,00. A percentagem unitária é o contributo de cada célula para este total, como por exemplo, $2,25/9,00 = 25\%$, enquanto a percentagem total é a correspondente a esse estado, tomando em conta quantas vezes aparece.

A análise de χ^2 considera tanto as células individuais (relações entre 2 estados de 2 variáveis) como os quadros isolados (relações entre todos os estados de duas variáveis) das tabelas de Burt. Dá uma ideia da quantidade de informação em cada quadro isolado, e em cada célula.

O cálculo dos χ^2 e a sua ordenação (até em percentagem e em percentagem cumulativa), mostra até que ponto é aberrante o quadro isolado (relação entre duas variáveis), ou a célula (relação entre dois estados de variáveis diferentes). O termo aberrante é aqui tomado no sentido de não esperado, e não esperado face aos totais marginais.

Por isso, os valores mais elevados de χ^2 são os mais aberrantes. Tanto podem ser faltas como excessos. São os que trazem mais informação (no sentido de Shannon).

Os valores de zero são aqueles em que os valores que realmente ocorrem são iguais aos que se esperam (sempre relativamente aos totais marginais).

5. –Tabela de valores críticos de Qui quadrado

Graus de liberdade	Níveis de significação								
	0,995	0,99	0,95	0,90	0,10	0,05	0,01	0,005	0,001
1	3.94E-4	1.57E-3	3.93E-2	1.58E-1	2.71	3.84	6.64	7.88	10.83
2	0.0100	0.0201	0.103	0.211	4.60	5.99	9.21	10.60	13.82
3	0.072	0.115	0.352	0.584	6.25	7.82	11.34	12.84	16.27
4	0.207	0.297	0.711	1.064	7.78	9.49	13.28	14.86	18.46
5	0.412	0.554	1.145	1.61	9.24	11.07	15.09	16.75	20.52
6	0.676	0.872	1.64	2.20	10.64	12.59	16.81	18.55	22.46
7	0.989	1.24	2.17	2.83	12.02	14.07	18.48	20.28	24.32
8	1.34	1.65	2.73	3.49	13.36	15.51	20.09	21.96	26.12
9	1.73	2.09	3.33	4.17	14.68	16.92	21.67	23.59	27.88
10	2.16	2.56	3.94	4.87	15.99	18.31	23.21	25.19	29.59
11					17.28	19.68	24.72	26.76	31.26
12					18.55	21.03	26.22	28.30	32.91
13					19.81	22.36	27.69	30.82	34.53
14					21.06	23.68	29.14	31.32	36.12
15					22.31	25.00	30.58	32.80	37.70
16					23.54	26.30	32.00	34.27	39.29
17					24.77	27.59	33.41	35.72	40.75
18					25.99	28.87	34.80	37.16	42.31
19					27.20	30.14	36.19	38.58	43.82
20					28.41	31.41	37.57	40.00	45.32
21					29.62	32.67	38.93	41.40	46.80
22					30.81	33.92	40.29	42.80	48.27
23					32.01	35.17	41.64	44.18	49.73
24					33.20	36.42	42.98	45.56	51.18
25					34.38	37.65	44.31	46.93	52.62
26					35.56	38.88	45.64	48.29	54.05
27					36.74	40.11	46.96	49.65	55.48
28					37.92	41.34	48.28	50.99	56.89
29					39.09	42.56	49.59	52.34	58.30
30					40.26	43.77	50.89	53.67	59.70
40					51.81	55.76	63.69	66.77	73.40
50					63.17	67.51	76.15	79.49	86.66
60					74.40	79.08	88.38	91.95	99.61
70					85.53	90.53	100.43	104.22	112.32
80					96.58	101.88	112.33	116.32	124.84
90					107.57	113.15	124.12	128.30	137.21
100	67.33	70.06	77.93	82.36	118.50	124.34	135.81	140.17	149.45

Rejeita-se a H_0 quando o valor calculado do qui quadrado é maior do que o valor crítico para o nível de significação escolhido.

6 – Bibliografia

BIBLIOGRAFIA

- DAVIS, John C. (1986) – ***Statistics and Data Analysis in Geology***, 2nd edition, New York, John Wiley & Sons, p. 80-86.
- (*) EBDON, David (1996) – ***Statistics in Geology***, 2nd edition, Blackwell, Oxford (UK), p. 65-71.
- (*) ESTABANEZ ALVAREZ, J.; BRADSHAW, R. P. (1978) – ***Técnicas de Quantificación en Geografía***, Editorial Tebar Flores, Madrid, p. 356-358; 365-376.
- PESTANA, Maria Helena; GAGEIRO, João Nunes (1998) – ***Análise de dados para Ciências Sociais. A Complementaridade do SPSS***, Edições Sílabo, Lisboa, p. 91-139.
- VOLLE, Michel (1978) – ***Analyse des Donnees***, Economica, Paris, p. 44-59; 112-143.

Anexo 5

Análise de Variância e Covariância

1. – Introdução

A Análise de Variância faz parte de uma série de métodos que procuram encontrar e esclarecer relações entre variáveis. Em análise de variância, tal como acontece com as técnicas de regressão, assume-se que a variação numa variável está relacionada com a variação de um conjunto de outras variáveis.

A análise de variância pertence pois ao caso geral da análise de dependência, em que se procuram as relações entre uma única variável dependente (a variável y) e n variáveis independentes (as variáveis x).

Y	X _j				
y ₁	x ₁₁	x ₁₂	x ₁₃	...	x _{1m}
y ₂	x ₂₁	x ₂₂	x ₂₃	...	x _{2m}
y ₃					
...	x _{n1}	x _{n2}	x _{n3}	...	x _{nm}

Figura 5.1 Esquema geral das análises de dependência (1 para n)

As técnicas a utilizar nas análises de dependência (de 1 para n) dependem das escalas de medida das variáveis.

Neste capítulo vai-se considerar o caso em que a variável dependente y é uma variável quantitativa, que se encontra medida em escalas de intervalo ou razão.

Se as variáveis independentes, (as diversas variáveis x) também são quantitativas, a técnica de pesquisa de relações a utilizar é a da análise de regressão múltipla. Se as variáveis x são variáveis nominais ou ordinais utiliza-se a análise de variância a um factor (one-way Anova) se a variável independente é só uma e a análise de variância multifactorial (n-way Anova) quando as variáveis independentes são mais do que uma. Finalmente, se as variáveis x são umas quantitativas e outras nominais ou ordinais deve utilizar-se a análise de covariância.

Na análise de variância a um factor os valores de variável independente podem ainda dividir-se em dois casos, um em que a variável apenas assume duas categorias ou estados (análise a **2** grupos) e outro quando se apresenta em mais categorias (análise a **K** grupos).

A base central da análise consiste em saber se as médias da variável nas diferentes categorias ou grupos são ou não iguais.

A hipótese nula a testar é a de que as médias são iguais, e portanto as categorias ou grupos pertencem à mesma população. Se pelo menos uma categoria tiver uma média significativamente diferente a hipótese nula é rejeitada. Testar a igualdade das médias consiste em testar duas variâncias, uma a variância entre as diversas categorias e outra a variância interna às próprias categorias. O teste adequado é o teste **F**, que comparando a variância entre as categorias com a variância dentro das categorias, responde à seguinte questão:

“ Que probabilidade existe de que estes dois valores sejam estimativas da mesma variância de uma mesma população ?”.

2. – Tipos de Análise de Variância

A análise de variância permite conhecer o efeito da variação de variáveis independentes qualitativas numa variável dependente de natureza quantitativa. A variável dependente quantitativa é a variável-critério e as variáveis qualitativas independentes são os factores.

O caso mais simples da análise de variância, aquele em que apenas existe uma variável independente, medida por duas categorias, designado por Análise de Variância a um factor com dois grupos (designado em inglês por two groups one way ANOVA).

2.1. – Análise de Variância a um Factor

Suponha-se que, num dado país, temos informação sobre o rendimento de um conjunto de indivíduos, uns nacionais e outros estrangeiros. Sobre eles temos duas variáveis, uma quantitativa que é o rendimento e uma outra, qualitativa a dois estados, ser ou não estrangeiro, que se supõe influir sobre a primeira.

Uma primeira forma de procurar a relação entre o rendimento e o facto de ser ou não estrangeiro é calcular a média de rendimento do grupo de nacionais e a média de rendimento do grupo de estrangeiros. Comparando-as vê-se se são iguais ou muito diferentes.

Mas há um problema. Se forem feitas outras amostras obter-se-á valores diferentes para as respectivas médias, o que significa que só se pode afirmar que existe realmente influência no rendimento do facto de ser estrangeiro, quando a diferença entre as médias (dos estrangeiros e dos nacionais), for suficientemente grande face às respectivas dispersões, o que equivale a calcular o significado estatístico da diferença das médias.

O problema é então saber até que ponto é significativa a diferença das médias de rendimento dos dois grupos, face à dispersão interna de cada grupo, e é isso precisamente que a análise de variância faz.

O decorrer do processo de análise segue os seguintes passos:

(1) - Cálculo da variância entre grupos ou categorias

$$\sigma_{eg}^2 = \frac{\sum_j^k [N_j \times (\bar{x}_{j.} - \bar{x})^2]}{k - 1}$$

É a soma de quadrados entre os grupos dividida pelo respectivo grau de liberdade (o número de grupos menos um).

N_j é o número de elementos de cada grupo.

(2) - Cálculo da variância nos grupos ou categorias

$$\sigma_{ng}^2 = \frac{\sum_j^k \sum_i^n (x_{ij} - \bar{x})^2}{n - k}$$

É a soma de todas as somas de quadrados dentro dos grupos dividida pelo respectivo grau de liberdade (o número total de elementos menos o de grupos).

(3) - Cálculo da razão F

$$F = \frac{\sigma_{eg}^2}{\sigma_{ng}^2}$$

É a razão entre a variância entre diversos grupos e a variância dentro dos grupos, e é também a razão entre as respectivas somas de quadrados ponderadas pelos respectivos graus de liberdade.

(4) - Avaliação do significado da relação através dos valores de F.

A consulta da tabela pressupõe conhecer o valor de **F_{k-1, n-k, p}** em que k é o número de estados da variável factor, n o número de observações e p o nível de probabilidade desejado, geralmente 0,05. Na tabela de F, k-1 deve ser lido nas colunas e n-k nas linhas.

Para melhor compreensão dos procedimentos e da interpretação, veja-se o seguinte

EXEMPLO: Fez-se um inquérito a 24 empresas, em que se perguntou relação percentual entre receitas e despesas. Procurou-se se haveria relação com o tipo de empresa: cooperativa (A), por acções (B), por quotas (C) ou a título individual (D).

Quadro de dados:

Receitas/despesas	Tipo	Receitas/despesas	Tipo
105	A	95	C
87	B	88	C
118	C	101	C
84	D	95	B
102	A	87	D
113	A	85	D
99	B	97	A
102	B	101	A
98	B	94	C
94	D	96	D
96	A	92	C
86	A	88	D

Com estes dados, procura-se saber se é possível afirmar que existe uma relação entre o sucesso das empresas (medida pela relação entre as respectivas receitas e despesas) e o tipo a que pertencem.

Para isso vai-se testar se existe relação entre as duas variáveis. A hipótese nula (H_0) é que o tipo de empresa não tem a ver com os resultados da exploração, medidos pela relação receitas/despesas.

GRUPOS	A	B	C	D
Relação	105	87	118	84
receitas	102	93	95	85
despesas	113	102	88	87
	96	98	101	96
	86	95	94	88
	97		92	94
	101			

Em cada grupo, calcule-se a média da variável "resultados da exploração",

GRUPOS	A	B	C	D
Total	700	475	588	534
N.º de elementos	7	5	6	6
Média	100	95	98	89

E também a média geral

sendo $K = 4$ e $N = 7 + 5 + 6 + 6 = 24$

virá $(700 + 475 + 588 + 534) / 24 = 2297 / 24 = 95,708333...$

Executando de seguida as restantes operações:

1 - Variância nos grupos

A		B		C		D	
x-média	(x-med)2	x-média	(x-med)2	x-média	(x-med)2	x-média	(x-med)2
5	25	-8	64	20	400	-5	25
2	4	2	4	-3	9	-4	16
13	169	7	49	-10	100	-2	4
-4	16	3	9	3	9	7	49
14	196	0	0	-4	16	-1	1
3	9			-6	36	5	25
1	1						
	420		126		570		120

$$\text{Variância nos grupos} = (420+126+570+120)/(24-4) = 1236/20 = 61,8$$

2 - Variância entre grupos

Grupo	média	Nº.	n x (medx - med geral)
A	100	7	7 x (100 - 95,708333) = 128,9288
B	95	5	5 x (95 - 95,708333) = 2,5087
C	98	6	6 x (98 - 95,708333) = 31,5104
D	89	6	6 x (89 - 95,708333) = 270,0104

$$\text{Variância entre grupos} = (128,929+2,509+31,510+270,010)/3 = 144,319$$

3 - Cálculo de F

$$F = 144,3194 / 61,8 = 2,3353$$

4 - Avaliação do significado

Consultando a tabela de valores críticos de F para os graus de liberdade 3 e 20, com diferentes probabilidades, obtém-se

$$p = 0,05 \quad 3,10 \quad p = 0,01 \quad 4,94 \quad p = 0,005 \quad 5,82 \quad p = 0,001 \quad 8,10$$

Como valor de F que se calculou é de 2,3353, a Hipótese nula (Ho) não pode ser rejeitada,, pois o valor obtido não é superior ao valor crítico da tabela para 0,05 (significativo) e muito menos para 0,01 (muito significativo), o que quer dizer que não é possível afirmar, para estes dados, que o tipo de empresa tem a ver com a relação entre receitas e despesas que apresenta.

A seguir apresenta-se uma listagem produzida, para o mesmo conjunto de dados, por um programa estatístico bastante difundido

O comentário dos resultados é feito através de notas, que devem ser encaradas como exemplos de possível interpretação.

Listagem tipo de um programa de ANOVA *one-way*

*** ANÁLISE DE VARIÂNCIA ***					
<u>Relação Receitas/Despesas por Tipo de Empresa</u>					
<i>Fonte de Variação</i>	<i>Soma de Quadrados</i>	<i>G.L.</i>	<i>Media dos Quadrados</i>	<i>F</i>	<i>Signif. de F</i>
<i>Efeito principal</i>	<i>432.958</i>	<i>3</i>	<i>144.319</i>	<i>2.335</i>	<i>0.105</i>
<i>TIPO</i>	<i>432.958</i>	<i>3</i>	<i>144.319</i>	<i>2.335</i>	<i>0.105</i>
<i>Var. Explicada</i>	<i>432.958</i>	<i>3</i>	<i>144.319</i>	<i>2.335</i>	<i>0.105</i>
<i>Var. Residual</i>	<i>1236.000</i>	<i>20</i>	<i>61.800</i>		
<i>Total</i>	<i>1668.958</i>	<i>23</i>	<i>72.563</i>		

Notas: 1).- Como se pode observar, os dados apresentam dois tipos de variação da "variável razão entre receitas/despesas": a que é explicada pela variável "tipo de empresa" e a residual. Como se trata de uma análise a um factor, a variação explicada é devida apenas a um único efeito, o "TIPO de empresa" (com 432,958 de SQ e 144,319 MQ).

2).- O valor de F, igual a 2,335, é todo devido à variável TIPO (a única independente que existe na análise). O nível de significado da hipótese nula é de 0,105, ou seja, a probabilidade de que as diferentes categorias tenham a mesma média é de 10,5 %. Como é superior a 5 %, não pode ser rejeitada, pelo que não se pode afirmar que o tipo de empresa influi na relação receitas/despesas.

2.2. – Análise de Variância Multifactorial

Nas Análises multifactoriais existe mais de um factor, ou seja, existe mais de uma variável independente. Porém, devido à crescente complexidade de interpretação que ocorre com o aumento do número de factores, se utilizem geralmente apenas dois ou no máximo três factores.

A interpretação é similar à anterior, com a diferença que agora é possível distinguir entre os efeitos de cada um dos factores (as variáveis independentes), quantificando o seu respectivo contributo para a variação do critério (a variável dependente).

No caso mais simples da variação de dois factores A e B, pode considerar-se que a variação total, traduzida pela soma de quadrados (SQ), é igual à variação devida a A (SQ_a), mais a variação devida a B (SQ_b), mais a interacção entre A e B (SQ_{ab}) e ainda mais um valor residual que pode ser interpretado como a variação básica do conjunto (e que corresponde à variação nos grupos da análise a um factor).

$$SQ_t = SQ_a + SQ_b + SQ_{ab} + SQ_r$$

Não interessa conhecer o método manual de determinar as diferentes somas de quadrados, mas apenas interpretar os resultados, calculados por um qualquer pacote estatístico. Os interessados no cálculo manual podem encontrar os pormenores do cálculo em vários manuais (Stein et al, 1996: 528 e sgs).

O cálculo dos valores F é feito ponderado cada soma de quadrados pelos respectivos graus de liberdade, para obter a média de quadrados respectiva, e calculando os respectivos valores de F dividindo cada média de quadrados pela média de quadrados residual, segundo

$$\mathbf{MQa = SQa / (r-1),}$$

em que r é o número de categorias do factor A (linhas)

$$\mathbf{MQb = SQb / (k-1),}$$

em que k é o número de categorias do factor B (colunas)

$$\mathbf{MQab = SQab / [(r-1) \times (k-1)]}$$

$$\mathbf{MQr = SQR / (n-r.k)}$$

com

$$\mathbf{F\ a = MQa / MQr}$$

$$\mathbf{F\ b = MQb / MQr}$$

$$\mathbf{F\ ab = MQab / MQr}$$

Na análise a um factor existem duas fontes de variação, a que se deve ao factor e surge como variação entre as diferentes categorias do factor, e a outra, que existe dentro dos grupos e não tem a ver com o factor, designando-se por variação entre grupos, residual ou errática.

Na análise multifactorial a dois factores existe uma parcela de variação exclusivamente associada a cada factor, uma parte que está associada às combinações de factores e outra parte que não tem a ver com eles, a variação entre grupos, residual e errática.

A variação residual é muito importante, por que o valor de F é calculado pela comparação das outras fontes de variação com a variação residual, após a ponderação pelos respectivos graus de liberdade. Portanto, é fácil ver que quanto menor é a variação residual, maiores são os valores de F e por isso mais provável é que as células pertençam a populações diferentes.

A variação que está associada à combinação dos factores é uma parcela que produz efeitos na variável critério que não seriam esperados se os dois factores tivessem sido considerados isoladamente.

Na análise multifactorial a três factores a situação é mais complexa: existe uma variação exclusiva devida a cada um dos factores (a A, a B e a C), uma variação interactiva de primeira ordem em que os factores se combinam dois a dois (a AB, a AC e a BC) e uma variação interactiva de segunda ordem com os factores combinados três a três (que só pode ser a de ABC). E todas contribuem de forma independente para a variação total, o que significa que os efeitos três a três não surgiriam das combinações dois a dois nem dos factores independentemente.

É evidente verificar que a inclusão de mais factores torna a análise mais complexa e a interpretação vez mais difícil, numa forma exponencial em relação ao número de factores.

Ilustrando o que se disse, considere-se o seguinte exemplo: Numa aldeia queria saber-se até que ponto é que da localização das machambas (nos vales aluvionais ou nas vertentes, mais rochosas) e a intensidade da rega tinham a ver com o rendimento. Note-se que se diz *tinham a ver com o rendimento* e não *provocavam no rendimento*, pois a análise só pode identificar a existência de variações simultâneas, e não a de relações causais, cuja existência, intensidade e direcção só o conhecimento da teoria pode esclarecer.

O problema consiste então em relacionar as variações quantitativas do rendimento (a variável-critério) com as categorias das outras duas variáveis, ambas qualitativas, sendo uma nominal, o tipo de localização (o factor A) e a outra ordinal, a intensidade da rega (o factor B). O factor A tem duas categorias, vale e vertente, enquanto o Factor B tem quatro, rega regularmente, bastante, às vezes e nunca. Foram observadas 32 machambas, quatro em cada uma das 8 células (combinação dos estados dos factores - 2 linhas x 4 colunas).

Quadro de dados:

	Frequência da rega			
	Regular	Bastante	Pouco	Nunca
Na vertente	93	84	73	69
	78	65	76	64
	85	79	77	58
	81	81	71	61
No vale	103	67	80	49
	108	73	78	44
	107	84	85	38
	111	82	81	41

Listagem tipo de um programa de ANOVA *n-way*

Análise de Variância									
32 casos aceites.					8 células não-vazias.				
NÚMERO DAS CÉLULAS									
Variável	1	2	3	4	5	6	7	8	
REGA	1	1	2	2	3	3	4	4	
LOC	1	2	1	2	1	2	1	2	
Fonte da Variação	SQ				GL		MQ		F Sig de F
DENTRO + RESIDUAL	742.00				24		30.92		
REGA	7378.63				3		2459.54		79.55 0.000
LOC	40.50				1		40.50		1.31 0.264
REGA x LOC	1909.75				3		636.58		20.59 0.000
Modelo	9328.88				7		1332.70		43.11 0.000
Total	10070.88				31		324.87		
R-quadrado =	0.926				R quadrado ajustado =		0.905-		

Notas (1). - Verifica-se que o modelo proposto, isto é a pretensão de explicar a variação quantitativa do rendimento das machambas por via da variação qualitativa do factor A - Localização e do factor B - Rega, (com uma divisão em oito células), tem grande capacidade explicativa. O valor de F é 43.11 e a probabilidade das diferenças encontradas nas médias das células acontecer por acaso é de 0,000, ou seja inferior a 0,0005.

(2). - A capacidade explicativa do modelo ronda os 90 %, pois o r quadrado ajustado é de 0.905

(3). - A capacidade explicativa do modelo não decorre igualmente de cada um dos factores considerados. A rega é o factor predominante, tendo associado exclusivamente a si 73,26 % da soma de quadrados (SQ) e em combinação com a localização 18,96 %. A localização apenas tem associada a si própria 0,4 % da soma de quadrados. Note-se que a combinação da rega com a localização produz efeitos no rendimento que nada têm a ver com os que se devem a cada um dos dois factores considerados isoladamente.

(4). - Apenas a interacção com o factor localização não é significativa (p de ocorrência de 26,4%, muito inferior pois ao nível habitual de aceitação de 5%).

3 – Análise de Covariância

A análise de covariância, abreviadamente designada por alguns autores também por ANCOVA, vem complementar as análises possíveis de efectuar com a técnica da comparação de soma de quadrados.

É uma análise que surge da necessidade de considerar como variáveis independentes também variáveis quantitativas, combinando a regressão múltipla com a análise de variância. Permite conhecer o efeito da variação simultânea de variáveis independentes qualitativas e quantitativas, numa variável dependente quantitativa.

A variável dependente quantitativa é a variável-critério. As variáveis independentes chamam-se factores. quando são qualitativas (medidas em escalas nominais ou ordinais) e covariates ou concomitantes quando são quantitativas.

Siga-se a análise e a interpretação de resultados com o seguinte

EXEMPLO: Numa cidade procurou-se conhecer a influência da localização dos apartamentos (quatro bairros: 1, 2, 3 e 4) no preço de venda. Como na amostra se encontravam apartamentos de tamanhos muito diferentes, incluiu-se também a dimensão dos apartamentos na análise.

O problema coloca-se em saber se existe uma relação entre o preço dos apartamentos (a variável-critério) e a localização (variável nominal, factor,) e o tamanho (variável quantitativa, concomitante). A variável localização foi designada por LOCAL e a variável tamanho por METRO2.

Quadro de valores

Critério PREÇO	Factor LOCAL	Concom. METRO2	Critério PREÇO	Factor LOCAL	Concom. METRO2
320	1	387	160	3	188
430	2	390	120	1	166
515	3	376	165	2	167
160	1	160	205	3	168
105	2	160	170	1	212
185	3	160	215	2	212
195	1	163	240	3	210
170	2	162	120	1	175
95	3	184	159	2	174
165	1	134	305	3	174
160	2	134	300	1	183
185	3	134	150	2	182
300	1	183	240	3	180
206	2	183	210	1	214
220	3	185	320	2	213
315	1	190	235	3	213
190	2	189			

Listagem-tipo de um programa de ANCOVA

*** ANÁLISE DE COVARIÂNCIA ***					
33 casos aceites					
3 células não-vazias					

*** Análise de Variância -- 1 ***					
Fonte da Variação	SQ	GL	MQ	F	Sig de F
DENTRO + RESIDUA L	109904.62	29	3789.81		
REGRESSÃO	150432.83	1	150432.83	39.69	0.000
LOCAL	4481.65	2	2240.82	0.59	0.560
Modelo	155110.10	3	51703.37	13.64	0.000
Total	265014.73	32	8281.71		

R-quadrado	0.585	R-quadrado ajustado =		0.542	

Análise de Regressão para o termo DENTRO + RESIDUAL					
Variável Dependente (Y) -- VALOR					
Variáveis concomitantes (covariates)	B	Beta	Erro Pdr.	Valor t	Sig. de t
METRO2	1.07138	0.75342	0.170	6.300	0.000
Intervalos de confiança a 0.9500 das variáveis concomitantes (covariates)					
METRO2	[0.724	1.419]			

Notas: (1). - O modelo proposto é que a variação do preço dos apartamentos depende de um factor locativo, a variável LOCAL (situação em um dos quatro bairros) e do tamanho dos apartamentos (pois não foi possível garantir uma amostra com apartamentos de igual tamanho). A listagem de resultados informa que o modelo geral é significativo (o valor F

geral do modelo é de 13,64 que corresponde a uma probabilidade de acontecer por simples acaso praticamente nula pois $p=0,000^4$.

(2). - A capacidade explicativa do modelo não é porém muito grande, pois as duas variáveis independentes: Localização e Tamanho apenas “explicam” 58,52% da variância total (o valor de R-quadrado, que é a razão entre a soma de quadrados total e a do modelo. Aliás a explicação ajustada para os graus de liberdade em causa é ainda ligeiramente inferior (54,2%).

(3). - Como o modelo tem duas componentes explicativas, uma ligada à regressão com a variável concomitante quantitativa e outra relacionada com o factor qualitativo, é necessário conhecer a importância relativa de cada uma. Os resultados da análise mostram que:

a)- a relação com o factor qualitativo - a Localização - é muito fraca, e sem qualquer significado estatístico, pois a probabilidade de acontecer por simples acaso é de 56 % ! Assim sendo tem de se considerar que a localização não tem a ver com o preço dos apartamentos na amostra que se analisou, e, por ser uma amostra representativa, também não tem a ver com a realidade estudada.

b)- a capacidade explicativa do modelo é quase exclusivamente devida à variável concomitante quantitativa - o Tamanho dos apartamentos . Esta variável explica por si só 56,86% da variação total.

(4). - Esta análise deve continuar com a relação de regressão. Aconselha-se aos alunos que a façam, a seguir à leitura do capítulo sobre regressão múltipla.

4 – Considerações Gerais

Não é muito o que se pode dizer sobre a análise de variância e covariância que não tenha sido já referido.

O mais importante é reafirmar que estas análises são poderosos instrumentos de compreensão da realidade por permitirem comprovar a existência de relações entre variáveis e quantificar a sua importância.

Fazendo parte do conjunto de técnicas de análise de dependência, são especialmente importantes no caso de inquéritos ou levantamentos por observação pessoal em que predominem variáveis nominais ou ordinais.

A análise de variância deve ser utilizada quando se buscam as relações entre uma variável dependente de natureza quantitativa com estados de uma ou mais variáveis qualitativas, nominais ou ordinais, a que se chamam factores.

Deve ter-se em conta que embora seja possível analisar muitos factores simultaneamente, não é conveniente usar mais de duas ou no

⁴ - O valor 0,000 pode não ser matematicamente zero. Significa apenas que o valor arredonda para zero, e por isso, que a probabilidade é inferior a metade de um milésimo, ou seja, é inferior a 0,0005.

máximo três variáveis, pois o estudo de todas as possibilidades combinatórias da origem da variância, torna muito complexos os resultados e impossível a sua compreensão e interpretação.

Tal facto exige um grande cuidado na selecção dos factores a analisar, que devem ser os mais adequados, os necessários e também os suficientes para o problema em estudo que será sempre um modelo de relação e dependência.

A análise de covariância usa-se para estudar as relações entre uma variável dependente de natureza quantitativa e os estados de uma ou mais variáveis qualitativas e valores de uma ou várias variáveis quantitativas.

Mas a análise de variância não deve ser utilizada para muitas variáveis independentes simultaneamente.

Associa a análise de regressão múltipla com a de variância, e é uma das formas de introduzir variáveis nominais ou ordinais na regressão múltipla. A outra forma de o fazer é pela conversão dos estados das variáveis qualitativas em variáveis "fingidas" ou "mudas" (*dummy variables*, na literatura de língua inglesa) de tipo dicotómico, ausência/presença.

5 – Bibliografia

- (*) ESTEBANEZ ALVAREZ, J.; BRADSHAW, R.. P. (1978) – ***Técnicas de Quantificación en Geografía***, Editorial Tebar Flores, Madrid, p. 381-386.
- IVERSEN, G. R.; NORPOTH, H. (1976) – ***Analysis of Variance***, Sage Publications, London, 95 p..
- (*) KIM, Jae-On; KOHOUT, F. J. (1975) - "***Analysis of Variance and Covariance***" in NIE, Norman H. *et alia* - ***SPSS -Statistical Package for Social Sciences, 2th edition***, McGraw-Hill Book Company, New York, p. 398-432.
- (*) PESTANA, M. H.; GAGEIRO, J. N. (1998) - ***Análise de Dados para Ciências Sociais. A Complementaridade do SPSS***, Edições Sílabo, Lisboa, p. 177-243.
- STEYN, A.G.W. *et alia* (1996) – ***Modern Statistics in Prattice***, J. L. van Schaik, Pretoria ,p. 508-546.

Anexo 6

Regressão Linear Múltipla

A maior parte dos sistemas estudados pelos geógrafos - sistemas ambientais, sociais, económicos - envolvem variáveis numéricas, que se apresentam sob forma nominal, ordinal ou paramétrica (de intervalo e razão).

A busca das relações existentes entre as variáveis, da sua força e do seu significado é um dos aspectos fundamentais da pesquisa geográfica.

O conhecimento destas relações entre variáveis é importante por três motivos:

- 1).- Confirmar ou refutar noções teóricas sobre os processos de causa-efeito no sistema em estudo;
- 2).- Ilustrar e realçar aspectos não contemplados pelas teorias e *corpus* teóricos existentes, e por essa via, estimular novas ideias;
- 3).- Predizer a resposta do sistema a alterações futuras nas condições do sistema, ou mesmo, no estado actual, completar a informação existente sobre ele.

Quando se conhece, ou se julga conhecer, o sentido das dependências, isto é, sempre que é possível estabelecer que uma variável (a variável dependente), depende do valor de outras variáveis (as independentes), temos um problema de **regressão**.

Os matemáticos desenvolveram metodologias complexas de regressão para funções muito complexas, mas em geografia, como noutras ciências sociais, o interesse reside, na grande maioria dos casos, nas relações proporcionais, o mesmo é dizer, nas regressões lineares.

É claro que existem excepções importantes a esta regra, que geralmente são contempladas por **relações não lineares**. A mais importante excepção que merece ser aqui referida é a relação com distância. A sua relação com a expressão espacial de diversos fenómenos é quase sempre quadrática, pois de facto a relação não ocorre com a distância, mas sim com a respectiva área (que é proporcional ao quadrado da distância).

1. - Revisão da Regressão Linear Simples

A regressão linear simples, que relaciona duas variáveis, **x** e **y**, apresenta-se sob a seguinte forma geral

$$y = a + b \times x$$

e portanto para cada unidade de análise **i**, tem-se

$$y_i = a + b \times x_i + e_i$$

em que o termo **e_i** é a diferença entre o valor de **y** que se calcula aplicando a fórmula (o **y_i** previsto) e o valor de **y_i** realmente observado. Os matemáticos chamam-lhe **erro**, e admitem que é provocado por inúmeras causas aleatórias de pequena importância individual e não relacionadas entre si, mas para a generalidade dos sistemas sociais isto não é bem assim. Este valor **e_i**, é uma parte da variação da variável dependente **y** que não é explicada pela ou pelas variáveis independentes **x** que se incluíram no modelo, e é devida à falta de conhecimento teórico ou à falta da possibilidade prática, de incluir no modelo regressivo variáveis que trariam contributos importantes para a variação da variável dependente. Dá-se-lhe então nome de **resíduo**, e o seu estudo constitui muitas vezes a parte analítica mais interessante de muitos trabalhos de investigação.

2. - O modelo de Regressão Linear Múltipla

2.1. - Coeficientes de Regressão

O modelo de regressão linear geral (ou múltipla) é apenas a extensão do modelo de regressão linear simples, segundo a expressão:

$$y = b_0 + b_1 \times x_1 + b_2 \times x_2 + b_3 \times x_3 + \dots + b_n \times x_n$$

onde

y é a variável dependente:

b₀ é o intercepto, isto é, o valor que a variável dependente assume, quando todas as independentes são iguais a zero;

b_i são os coeficientes de regressão (os **coeficientes bê**), que indicam a variação de **y** associada à variação de cada variável **x**. Deve notar-se que estes coeficientes, contêm ainda o efeito de escala de cada **x**, excepto quando as variáveis foram todas previamente estandardizadas, caso em que estes coeficientes **β_i**'s medem a importância relativa de cada variável independente na variação de **y**,

e se chamam coeficientes estandardizados ou coeficientes beta, conforme a seguinte equação

$$y = \beta_0 + \beta_1 \times z_1 + \beta_2 \times z_2 + \beta_3 \times z_3 + \dots + \beta_n \times z_n$$

Os coeficientes **beta** (estandardizados) estão relacionados com os coeficientes **bê**, através da multiplicação pela razão entre o desvio-padrão da variável independente em causa, e o desvio-padrão da variável dependente, segundo a expressão

$$\beta_i = b_i \times \frac{\sigma_i}{\sigma_y}$$

E de um modo similar ao que sucede para a regressão simples, para cada unidade de análise **i**, em que se tem

$$y_i = b_0 + b_1 \times x_{1i} + b_2 \times x_{2i} + b_3 \times x_{3i} + \dots + b_n \times x_{ni} + e_i$$

em que o termo **e_i** é a diferença entre o valor de **y** que se calcula aplicando a fórmula (**y_i** previsto) e o valor de **y_i** realmente observado nessa unidade de análise. É o **resíduo** da regressão múltipla para cada unidade de análise.

O objectivo principal da análise de regressão múltipla é o de conhecer o valor dos coeficientes de regressão, o que se consegue geralmente utilizando um programa (*package*) estatístico ou econométrico, e muito raramente, por cálculo manual. Embora seja ainda relativamente fácil calcular o seu valor para 2 (duas) variáveis independentes, e o cálculo matricial para mais variáveis, não seja muito difícil, o geógrafo, como os outros investigadores de ciências sociais, deve utilizar programas de computador profissionais.

2.2. - Coeficiente de Correlação Múltipla

Mas se é importante o conhecimento da tendência geral da relação entre a variável dependente e as variáveis independentes, tal facto é, como no caso da regressão linear simples, insuficiente, pois falta saber a importância relativa dessa relação, ou seja, a parcela da variação de **y** que é devida às variações das variáveis independentes, **x_i**. Como se sabe, esta parcela é medida pelo valor do coeficiente de correlação, que no caso, é o coeficiente de correlação múltipla, de **y** com todos os **x** que entram na regressão múltipla, e que é representado pela letra **r**. Do mesmo modo, se define um coeficiente de determinação múltiplo, cujo valor é o de **r²** (érré ao quadrado), e que corresponde à parcela da variação da variável dependente que é explicada pela variação das variáveis independentes que entram na regressão.

2.3. - Significado da regressão

Os valores calculados respeitam a um conjunto de unidades de análise. Se constituem o total da população em estudo têm um significado perfeito - são a relação que existe entre as variáveis naquele caso concreto. Porém, se respeitam a uma amostra, fica por saber o grau de confiança que se pode ter na equação encontrada. Pois outra amostra poderia dar valores de correlação e de regressão ligeira ou muito diferentes. Assim para se poder utilizar a equação na previsão de outros valores, tem de se conhecer o seu significado.

E mesmo no caso em que os resultados respeitam a uma população, o seu sentido ou significado será diferente consoante o valor do coeficiente de correlação múltipla (quanto maior for, mais ajustada a relação), e o número de unidades de análise, que quanto maior é menos provável torna o aparecimento por acaso, de valores elevados do coeficiente de correlação.

Nestes casos são dois os testes de significação que é habitual fazer:

- um teste para o coeficiente de correlação (ou de determinação), com a hipótese nula (H_0) de que o coeficiente de correlação da população de onde esta amostra foi retirada era igual a zero;
- um teste para os coeficientes de regressão, com a hipótese nula (H_0) de que os coeficientes de regressão das variáveis independentes da população a que esta amostra pertence, são iguais a zero.

2.3.1 - Teste F

O primeiro teste de significação é feito calculando a razão **F**, que exprime a relação entre a variável dependente e o conjunto das variáveis independentes. A variância total da variável dependente y encontra-se dividida em duas partes, uma que está ligada à variância das estimativas da regressão, e outra, associada aos resíduos da regressão,

Variância Total de y = Variância dos resíduos + Variância da regressão

Sendo a Variância total de **Y**, igual a

$$\text{Var}_y = \sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

Uma forma de calcular o valor de **F** é determinando a razão entre a variância da regressão e a dos resíduos, como se disse,

$$F = \frac{\sum (y_i - \hat{y}_i)^2}{\sum (\hat{y}_i - \bar{y})^2}$$

Outra é usando a seguinte expressão:

$$F = \frac{r^2}{1 - r^2} \times \frac{n - k - 1}{k}$$

em que **r** é o coeficiente de correlação múltipla, **n** o número de unidades de análise e **k** o número de variáveis que entram na equação de regressão.

O valor de **F** pode também ser determinado dividindo as somas de quadrados da regressão e dos resíduos, ponderadas pelos respectivos graus de liberdade

$$F = \frac{\frac{(\text{SQ Re gressão})}{k}}{\frac{(\text{SQ Total}) - (\text{SQ Re gressão})}{n - k - 1}} = \frac{\frac{(\text{SQ Re gressão})}{k}}{\frac{(\text{SQ Re síduos})}{n - k - 1}}$$

Para avaliar se a equação encontrada para a amostra é uma imagem fiel (a 95% geralmente) da população, ou se tem sentido efectuar estimativas de novos valores a partir da equação, compara-se com o valor crítico da tabela respectiva. Se o valor calculado for superior ao valor crítico rejeita-se a Hipótese nula (H_0), de que não havia relações entre as variáveis, sempre para uma dada probabilidade, geralmente de 95%.

2.3.2 - Testes t

Este segundo tipo de teste mede a significação dos coeficientes de regressão e do intercepto e efectua-se através do cálculo do valor **t** para cada coeficiente de regressão e para o intercepto, associando-lhe um teste bilateral de significação.

A hipótese nula (H_0) é a de que tanto o intercepto como os coeficientes de regressão são iguais a zero, ou seja, que os valores estatísticos encontrados pelo cálculo não têm significado. O teste é feito com os valores de **t** e ou a probabilidade que lhes está associada. Rejeita-se a hipótese nula, e portanto consideram-se significativos valores dos parâmetros calculados quando a probabilidade associada a **t** é inferior a 0,05. Então pode dizer-se que a variável em causa aumenta significativamente a capacidade predictiva do modelo.

Por outro lado, tem de se aceitar a hipótese nula, quando a probabilidade daquele valor de **t** aparecer por simples acaso for superior a 0,05%, e nesse caso, não se pode afirmar que o coeficiente em que tal facto se passa seja realmente diferente de zero, pelo que não tem valor que permita utilizá-lo na regressão.

Os programas de regressão múltipla mais usuais fornecem os valores de **t** e das probabilidades que estão associadas a esses valores, geralmente sob a designação de **P**, ou de **Sign**.

Quando assim é, conhece-se imediatamente o significado dos coeficientes e do intercepto. Mas nem sempre assim acontece: umas vezes só é fornecido o valor de **t** e noutras nem isso.

Neste último caso, é possível calcular o valor de **t** dos coeficientes, a partir do coeficiente de regressão de cada variável (**b_i**) e do seu respectivo erro-padrão, (**se_i**), do seguinte modo

$$t_i = \frac{b_i}{se_i}$$

E para o intercepto, o cálculo é similar

$$t_0 = \frac{b_0}{se_0}$$

Sabendo-se os valores de **t** mas não as probabilidades que lhes estão associadas, é necessário verificar se esses valores permitem ou não rejeitar a hipótese nula.

Para tal, procura-se na tabela dos valores críticos de **t** (na região crítica bilateral) o valor crítico para a probabilidade de 0,05 e para o grau de liberdade respectivo (**n-k-1**), em que **n** é o número de observações e **k** o de variáveis independentes.

Sempre que um dos valores de **t** calculados for superior ao valor crítico da tabela rejeita-se a hipótese nula, o que significa que o respectivo coeficiente trás um significativo aumento da capacidade explicativa da equação de regressão. Se for inferior, indica que não introduz informação significativa para o modelo, e por isso, o melhor é não tomar em conta essa variável.

Importa salientar que, no caso de se estar a trabalhar com populações, os valores encontrados para os coeficientes de regressão, mesmo que a partir de um pequeno número de unidades de análise, e estatisticamente não significativos, no sentido de que pouco ou nada acrescentarem à capacidade explicativa, têm sempre significado. Por exemplo, se nos concelhos de um distrito, se encontra que a relação entre os votos do PCP e do CDS não é significativa, isso pode ser uma importante conclusão, especialmente quando se esperava uma forte relação negativa.

No caso de se estar a trabalhar com uma amostra, é diferente. Os valores obtidos (os valores mostrais) são apenas uma estimativa dos valores da população que desconhecemos. Então o teste t permite-nos saber se devemos ou não aceitar os resultado (para um dado grau de significação, geralmente 95%) e por outro lado calcular o intervalo de confiança dos coeficientes.

2.4. - Erro-padrão da estimativa

Tem a ver com a adequação da equação de regressão aos dados. Se os dados respeitam a uma população, o valor é descritivo e tem todo o significado.

Mas no caso de se estar a trabalhar com uma amostra, o erro-padrão da estimativa é muito importante, pois mede a variação (ao nível de probabilidade de 95 %) que está associada às estimativas dos valores de **Y**. É o que se passa quando queremos calcular valores futuros ou para outras situações de **Y**. Mesmo que os valores originalmente pertencessem a uma mesma população, estamos a assumir que os valores que temos são os de uma amostra, a partir da qual se fazem estimativas dos valores de **Y** para outras combinações de **Xs**.

Se não for dado directamente pelo programa, divide-se o valor estimado para o intercepto (o termo independente) pelo erro-padrão geral da estimativa, obtendo-se o respectivo valor de **t**.

Consultada a tabela dos valores críticos de **t**, para os graus de liberdade do caso em estudo, que é igual a **u = n-k-1**, verifica-se se os resultados são ou não aceitáveis, através da consulta da tabela. Aceita-se a hipótese nula **H₀** (de que não há relação significativa) se o valor de **t** for inferior ao valor crítico da tabela, e rejeita-se quando for superior.

Com o valor do erro padrão é também possível calcular os intervalos de confiança das estimativas de **Y**, segundo a fórmula

$$Y_i^{\text{real}} = Y_i^{\text{estim.}} \pm se_i \times t_p^v$$

em que o valor do intervalo de confiança de **Y** para uma dada probabilidade **p** se situa entre o valor de **Y** estimado mais ou menos o resultado da multiplicação do erro-padrão da estimativa **se_i** pelo valor crítico de **t** (para a probabilidade **p** e graus de liberdade **u**).

2.5. - Erro-padrão dos coeficientes de regressão

O erro-padrão de cada coeficiente de regressão reflecte a precisão que esse coeficiente traz para a equação.

No caso duma amostra, se o coeficiente é **b_i** e o erro-padrão **se_i**, pode-se ter 95 % de certeza de que o coeficiente de regressão real se situará entre

$$b_i^{\text{real}} = b_i^{\text{estim.}} \pm se_i \times t_p^v$$

o valor crítico de **t**, vê-se na tabela dos valores críticos, para a probabilidade desejada (geralmente de 95%) e os respectivos graus de liberdade. Este intervalo de confiança indica a precisão associada ao valor de cada coeficiente, no caso de amostras, e o valor que se pode dar às

previsões que se fizerem com eles, mesmo quando se trate originalmente de populações.

2.6. - Tipos de correlação múltipla

Existem várias maneiras de executar regressão múltipla:

2.6.1 - Simultânea

É o mais comum, em que se considera um conjunto de variáveis independentes simultaneamente. Definida a variável dependente, escolhe-se uma série de variáveis independentes, cuja relação com a primeira se procura determinar. As variáveis são tratadas todas ao mesmo tempo, e obtém-se uma equação de regressão, com os coeficientes de regressão de cada variável independente e o termo independente. Ao mesmo tempo conhece-se o valor dos coeficientes de correlação múltipla e o seu quadrado, o coeficiente de determinação.

Para o caso de se estar a tratar duma amostra, ou se pretender usar a equação para prever valores da variável dependente, são fornecidos também os valores do erro-padrão da estimativa e dos coeficientes de regressão, e geralmente ainda é indicada o respectivo grau de significado estatístico, ou seja, a probabilidade de o valor obtido não ter ocorrido por acaso, mas sim pela existência de uma relação significativa. Este tipo de **regressão múltipla simultânea**, é designado por **Enter** no SPSS, e constitui uma opção no package ESTAT.

Também é possível remover variáveis da regressão, sendo o modo como se faz dependente das especificidades de cada programa.

2.6.2 - Passo-a-passo

Outro modo de fazer regressões múltiplas, é passo-a-passo, com as variantes Stepwise, e Backward. Na primeira, Stepwise, as variáveis entram uma a uma, de acordo com um modelo que se escolhe (geralmente o ganho ou a perda na probabilidade associada a **F** ou no próprio valor de **F**, como no SPSS, ou a redução na soma de quadrados no caso do ESTAT). Quanto maior for a probabilidade de **F** ou menor for o valor de **F** mais variáveis entram na regressão e vice-versa. Para eliminar da equação as variáveis menos importantes, aquelas que trazem menos informação, deve-se diminuir o valor da probabilidade de **F** ou aumentar o valor de **F**, na opção do programa.

O processo pode iniciar-se com uma regressão simples, em que é seleccionada a variável que melhor satisfaz o critério utilizado (maior valor de **F**, da probabilidade de **F** ou da soma de quadrados). No passo seguinte, o mesmo se passa com a variável independente que agora melhor satisfaz o critério, deduzidos obviamente, os efeitos das variáveis já entradas na equação de regressão. A entrada de cada variável, corresponde à parcela de variação que ela reduz, retiradas os efeitos das que já entraram em passos anteriores, o que vem eliminar os efeitos de auto-correlação entre as variáveis independentes, e pode provocar que uma variável que tem uma correlação simples elevada com a variável dependente, entre muito tarde ou nem chegue mesmo a entrar na equação de regressão, devido à sua variância ter, entretanto, já entrado na equação de regressão múltipla,

associada à variação de outras variáveis independentes, com que esta muito se auto-correlaciona.

Uma forma de avaliar a auto-correlação na equação de regressão deve-se efectuar o teste de Durbin-Watson. É um teste para a auto-correlação (a existência de correlação entre os elementos de uma série, ou seja, a existência de um efeito de vizinhança). Um dos pressupostos da análise de regressão é a de que os consecutivos resíduos de regressão não estão relacionados entre si. Se isto é verdade, o valor de teste Durbin-Watson é 2, se for menor, existe uma auto-correlação positiva, se for maior a auto-correlação é negativa.

Neste tipo de análise é necessário definir muito bem os critérios a utilizar. No SPSS o valor de entrada da probabilidade de **F** tem de ser sempre inferior ao da remoção, e quanto maiores ambos forem mais variáveis entram na regressão, e quanto menores, menos entram. No *ESTAT* quanto menor o critério, mais variáveis entram. Para quem começa, no *SPSS* podem escolher entre 0.45 de entrada e 0.5 de remoção até 0.05 de entrada e 0.10 de remoção, e no *ESTAT* um valor de 0.01.

No método Backward, as variáveis entram primeiro todas e depois são eliminadas uma a uma. Quanto maior é o valor de **F** ou menor a probabilidade que lhe está associada, maior será número de variáveis excluídas da selecção final.

Utilizando os valores mais adequados a cada caso concreto, assim se eliminam ou agregam as variáveis de modo a obter o equilíbrio (difícil) na relação de regressão entre a capacidade de explicação do modelo (que deve ser a maior) e o número de variáveis (que quanto menor, melhor será).

2.6.3 - Outros

Outros processos de selecção ou remoção de variáveis são possíveis em diversos packages, como seja a definição de blocos, e a utilização de diversos métodos para fazer entrar ou retirar variáveis da selecção. Devem ser utilizados com cuidado, e depois de saber bem qual o seu efeito, pelo que não se aconselham a quem se inicia nestes trabalhos.

2.7. – Extensões do Modelo

Em regressão múltipla as variáveis independentes são variáveis que foram medidas em escalas de intervalo ou razão. Uma forma de introduzir variáveis de tipo nominal, é através do conceito de variáveis fingidas ou mudas (variáveis "dummy", em inglês).

Para tal, cada um dos estados da variável nominal que se ache interessante introduzir na análise deve ser transformada em uma variável dicotómica do tipo presença – ausência.

Efectuada a respectiva regressão, o coeficiente associado a cada variável fingida, quantifica na variável dependente (**y**) o efeito da ocorrência do referido estado da variável independente.

3. - Problemas

3.1. – Enunciados

1).- A partir dos seguintes dados:

Y	10	12	15	17	19	22	24	27	29	30
X₁	1	1	2	2	3	4	4	5	5	6
X₂	10	9	8	7	6	5	4	3	2	1

- Calcule a relação proporcional entre **Y** e **X₁**.
- Teste se os coeficientes encontrados são significativos.
- Calcule a relação proporcional entre **Y** e **X₂**.
- Teste se são significativos os coeficientes encontrados nesta relação.
- Com os mesmos dados, calcule a equação de regressão múltipla de **Y** com **X₁** e **X₂**.
- Para esta última equação estabeleça a hipótese nula e verifique se a pode rejeitar.
- Qual o valor do coeficiente de correlação múltipla. Diga o que significa.

3.2. – Soluções

1) a) $Y = 6,550 + 4,650 \times X_1$

b) Como o valor de **t** do intercepto é igual a 6,664 e o do coeficiente é 15,692, ambos com a probabilidade associada inferior a 0,000 os valores são significativos.

c) $Y = 33,267 - 2,321 \times X_2$

d) Os valores de **t** do intercepto e do coeficiente são respectivamente de 97,691 e de -42,295 (o sinal não interessa), ambos com a probabilidade associada também inferior a 0,000 os valores são significativos.

e) $Y = 31,850 + 0,250 \times X_1 - 2,200 \times X_2$

f) A **H₀** é que o intercepto e os coeficientes de regressão de **X₁** e **X₂** são na realidade iguais a zero (não influem na Variação de **Y**). Sendo

	Valor de t	Sign.
Intercepto	8,254	0,000
Coeficiente de X₁	-6,591	0,000
Coeficiente de X₂	0,369	0,723

pode afirmar-se que a entrada da variável **X₂** na regressão não introduziu uma significativa melhoria na capacidade explicativa / predictiva da equação, o que se deve à elevada correlação existente entre as duas variáveis **X₁** e **X₂**.

g) $r = 0,998$ logo o coeficiente de correlação múltipla é de 0,996.

Este valor significa que a variação do conjunto das duas variáveis **X₁** e **X₂** "explica" 99,6% da variação da variável **Y**.

A marcha de uma análise de regressão com alguns exemplos de interpretação dos resultados obtidos, é seguidamente apresentada.

Considere-se, por exemplo, a seguinte relação entre as produções de algumas explorações agrícolas, e algumas técnicas de cultivo, com a seguinte matriz de valores, extraída de ESTEBANEZ & BRADSHAW (1978)

A Adubação e o Regadio em cada exploração agrícola, são as variáveis independentes, e o Rendimento respectivo a variável dependente. Com os valores para estas seis unidades de análise, calculem-se os valores da regressão. No cálculo foram utilizados os packages estatísticos *ESTAT* e *SPSS*.

```

*****
*****                                     *****
*****                                     *****
***** DIOGO & ROQUETTE SOFTWARE          *****
*****                                     *****
*****                                     *****
*****                                     *****
***** REGRESSÃO MÚLTIPLA PASSO-A-PASSO ****

NOME DO TRABALHO:           Teste para as explorações agrícolas

***** RESULTADOS DO PROGRAMA *****
No. de Variáveis:            3      No. de Observações:        6
No. de Selecções             1      Constante para limitar variáveis:    0.00000

VARIABEL       MEDIA         DESV-PADRAO
   1           1.75000       0.85391
   2          10.16667       4.01732
   3          44.83333       8.19383

*** MATRIZ DE CORRELACAO DAS VARIABEIS
Linha  1              1.00000    0.88667    0.93495
Linha  2              0.88667    1.00000    0.98817
Linha  3              0.93495    0.98817    1.00000

***** SELECCAO..... 1               Variável Dependente:                3
No. de Variáveis obrigadas:           0      No. de Variáveis abandonadas:0

PASSO 1
Variável entrada..... 2
Soma de quadrados reduzida neste passo     393.356500   Proporção reduz. neste passo:    0.976475
Soma de quadrados reduzida acumulada       393.356500   Proporção reduz. acumulada:.....0.976475 de 402.8330

Para 1 variáveis entradas

```

Coeficiente de correlação múltipla....		0.988167	(Ajustado para os G.L.).....0.988167
Valor de F para análise da variância.....		166.029100	
Erro-padrão da estimativa.....		1.539221	(Ajustado para os G.L.).....1.539221
No. da variável	Coefficientes de Regressão	Erro-Padrão do Coef. Regres.	Valor de t calculado
2	2.01549	0.15642	12.885
Intercepção:		24.34251	
PASSO 2			
Variável entrada..... 1			
Soma de quadrados reduzida neste passo:		6.507617	Proporção reduz. neste passo: 0.016165
Soma de quadrados reduzida acumulada.		399.864100	Proporção reduz. acumulada:.....0.992629 de 402.83330
Para 2 variáveis entradas			
Coeficiente de correlação múltipla....		0.996308	(Ajustado para os G.L.).....0.995383
Valor de F para análise da variância.....		202.007500	
Erro-padrão da estimativa.....		0.994850	(Ajustado para os G.L.).....1.112276
No. da variável	Coefficientes de Regressão	Erro-Padrão do Coef. Regres.	Valor de t calculado
2	1.51840	0.21864	6.945
1	2.63754	1.02860	2.564
Intercepção:		24.78059	
TABELA DE RESÍDUOS			
Obs. no.	Valor	Estimativa	Resíduo
1	52.00	52.63	-0.63
2	42.00	42.60	-0.60
3	49.00	48.28	0.72
4	33.00	32.17	0.83
5	56.00	55.47	0.53
6	37.00	37.85	-0.85

A equação de regressão calculada é:

$$Y = 2,63754 \times X_1 + 1,51840 \times X_2 + 24,78059$$

para um coeficiente de correlação múltipla de $r = 0,99631$

correspondente ao coeficiente de determinação de $r^2 = 0,99263$

Interpretação: O ajuste da regressão é muito elevado, o que significa que as variáveis independentes "*explicam*" quase a totalidade (99,26 %) da variação da variável dependente. É um valor enorme, e por isso, sabe-se que existe uma forte relação entre as variáveis em estudo. Não se sabe se a relação é de causa-efeito, embora seja previsível que sim, nem se é devida a outra qualquer variável que afecte as três.

O intercepto tem o valor de 24,78, e isso significa o rendimento médio que é de esperar obter nestas explorações, quando não se utiliza nem adubos, nem regas;

O valor do coeficiente b da variável 1 - adubação - é 2,6375; é o acréscimo médio de rendimento (Qts/ha), que se obtém por cada quintal de adubo por ha empregue no cultivo destas explorações. Do mesmo modo, 1,5184 é o número médio de rendimento, em quintais/ha, que se produz a mais por cada hora de rega semanal.

O SPSS fornece o valor dos coeficientes beta, que no *ESTAT* podem ser calculados com os desvios-padrões da variável dependente e das variáveis independentes. Os coeficientes beta são iguais a

$$X1 = 0,2749$$

$$X2 = 0,74445$$

e destes valores se pode avaliar a importância relativa de cada variável independente na variação da variável dependente, que, no caso concreto em análise, é que a variável Regadio tem 2,7084 mais importância na variação do Rendimento, que a Adubação.

A equação descreve a relação encontrada nestas seis explorações. Pode ser usada para prever o rendimento noutras explorações, desde que essa nova exploração seja similar às seis em que a relação foi estudada. Quer isto dizer que o uso da regressão múltipla, assim como o da regressão simples, obriga a uma destas duas situações:

1).- as unidades de análise para que se faz a previsão são elementos duma população (em sentido estatístico), da qual as unidades de análise com que se fez a regressão são amostra (subconjunto representativo);

2).- as unidades de análise para que se faz a previsão são semelhantes (ou pelo menos não são significativamente diferentes), daquelas para que se fez a regressão, nem em dimensão, nem em formas de cultivo, nem em localização, nem em tipos de solo, etc., etc..

Então, estando satisfeitas uma destas condições, é possível conhecer o rendimento que seria de esperar numa outra exploração semelhante, a partir do conhecimento dos valores de adubação e regadio que utiliza, ou prever que rendimentos seriam de esperar se nalguma das estudadas se alterasse um ou ambos destes valores.

Se por exemplo, uma exploração agrícola (de tipo similar, como se disse) utilizar uma Adubação de **2,75** Qts/ha e um Regadio de **11** horas de rega/semana

o Rendimento que se espera é de **48,73** Quintais/ha

pois **$48,73 = 24,7806 + 2,75 \times 2,6375 + 11 \times 1,5184$** .

Mas, mesmo considerando as duas condições acima referidas, o valor que se obtém não é um valor nem exacto, nem seguro, isto é, tem um grau de confiança associado e os valores dos coeficientes e do termo independente (o intercepto), podem não ser exactamente os que se determinaram para a amostra, mas situarem-se entre determinados intervalos de variação.

O erro-padrão da estimativa é de 0,9948 Quintais/ha, ou melhor de 1,1123 Quintais/ha ajustado para os graus de liberdade (aqui, ajustado, significa que sendo uma amostra estará estatisticamente mais próximo do valor real da população).

Como os graus de liberdade são 3 (6 unidades de análise, 2 variáveis independentes menos 1) o valor crítico de t, para a probabilidade de 95% é, na tabela bilateral, de 3.182. Como o erro padrão da estimativa é de 1,1123, os intervalos de confiança, a 95% são de 3,558515, o que significaria, neste exemplo, que para o valor previsto de 48,73 Quintais/ha, o que realmente se esperaria encontrar ao fazerem-se várias experiências, seria um resultado variável, mas em que 95% dos valores obtidos deveriam estar compreendidos entre 45,17 e 52,29 Quintais/ha.

O significado da regressão é dada pelo valor de *F*, que ambos os programas fornecem, sendo **F = 202,007**, pois se o quadrado do

coeficiente de correlação múltiplo para a relação com duas variáveis é = **0,99263**

$$F = \frac{0,99263}{1 - 0,99263} \times \frac{6 - 2 - 1}{2} = \frac{0,99263}{0,00737} \times \frac{3}{2} = 134,6852 \times 1,5 = 202,027$$

onde a diferença é devido a arredondamentos.

O valor de **F** calculado é de 202,007, e o valor crítico da tabela respectiva, para o nível de probabilidade de 0.05, e para o **k** das variáveis (colunas) e para o **n-k-1** das unidades de análise (linhas) que neste caso são, respectivamente, 2 e 3, é de 9,55, pelo que a relação encontrada é muitíssimo significativa, pois 202,007 é muitíssimo maior que 9,55. Na realidade, e como o *SPSS* fornece o nível de significação respectivo, verifica-se que é de 0,0006, ou seja, a probabilidade deste valor ter sido fruto do acaso, é inferior a um por mil. E deve tomar-se em conta que a regressão apenas respeita a seis unidades de análise, que é de facto, muito pouco.

A entrada das variáveis na regressão passo-a-passo *Steepwise* foi:

Passo	Variável	Coef. correlação múltiplo	Perc. de explicação (%)
1	V2 - Regadio	0,988167	97,65
2	V1 - Adubação	0,995383	99,08

o que significa que com apenas uma variável se encontra já uma elevadíssima relação (de quase 98%). A entrada da segunda variável, no passo dois, acrescenta muito pouca explicação à regressão pelo que a sua inclusão será discutível. O que acontece é que as duas variáveis independentes são muito pouco independentes entre si, apresentando uma elevadíssima correlação entre si.

Deve notar-se que como estamos a tratar de uma estimativa dos valores de uma população a partir de uma amostra (de seis elementos), se utilizaram os valores do coeficiente de correlação múltipla ajustados para os graus de liberdade.

4.2. - Criminalidade nos Estados Unidos, 1978

Considere-se agora a relação entre a criminalidade violenta e algumas variáveis, potencialmente explicativas (ou relacionadas), nos Estados Unidos, em 1978. Os valores são os constantes do quadro abaixo, para os 48 estados contíguos do território continental dos Estados Unidos da América do Norte,

Unidades de Análise		Variáveis						
ESTADOS		Y	X1	X2	X3	X4	X5	X6
1	Maine	207.7	26.4	0.9	6.9	11.8	318	6.1
2	New Hampshire	119.1	22.2	0.4	6.1	14.3	315	3.8
3	Vermont	166.3	26.0	1.3	6.3	12.4	359	5.7
4	Massachusetts	461.6	26.4	4.5	7.9	15.5	426	6.1
5	Rhode Island	348.2	26.6	5.7	7.2	14.5	346	6.6
6	Connecticut	315.2	25.9	8.4	3.6	16.2	378	5.2
7	New York	841.5	28.5	15.4	7.4	15.3	458	7.7
8	New Jersey	423.6	27.1	12.8	6.0	16.4	426	7.2
9	Pennsylvania	301.6	27.1	10.4	7.3	14.2	383	6.9
10	Ohio	412.9	28.7	10.7	6.0	14.8	321	5.4
11	Indiana	323.6	28.3	8.4	6.4	14.4	322	5.7
12	Illinois	465.7	28.5	17.7	6.8	16.1	367	6.1
13	Michigan	577.5	28.8	12.1	5.6	15.4	460	6.9
14	Wisconsin	131.9	26.7	4.2	6.2	15.1	382	5.1
15	Minnesota	190.0	26.5	3.7	7.9	14.7	445	3.8
16	Iowa	161.9	27.6	2.7	7.5	14.5	381	4.0
17	Missouri	468.4	26.3	13.5	7.4	13.0	287	5.0
18	North Dakota	66.7	27.8	5.2	9.1	13.6	329	4.6
19	South Dakota	163.7	29.0	5.6	8.4	12.1	295	3.1
20	Nebraska	190.2	27.2	5.0	7.6	14.2	359	2.9
21	Kansas	318.4	24.6	8.8	7.9	13.4	311	3.1
22	Delaware	440.6	28.0	14.7	3.6	15.7	424	7.6
23	Maryland	732.4	27.1	23.1	6.1	17.6	422	5.6
24	Virginia	284.6	24.7	18.4	6.4	14.6	330	5.4
25	West Virginia	167.8	25.7	2.8	7.9	12.0	319	6.3
26	North Carolina	413.1	26.5	25.5	6.2	11.8	300	4.3
27	South Carolina	637.2	28.0	32.9	6.0	12.2	306	5.7
28	Georgia	482.6	27.7	28.2	6.4	12.4	298	5.7
29	Florida	766.1	19.6	13.7	5.2	12.2	292	6.6
30	Kentucky	223.0	26.7	9.7	5.5	11.0	257	5.2
31	Tennessee	382.4	25.0	16.5	7.4	11.3	285	5.8
32	Alabama	418.9	26.8	27.6	7.0	11.8	285	6.3
33	Mississippi	321.6	30.9	36.6	6.9	10.0	276	7.1
34	Arkansas	344.2	27.0	17.3	4.9	10.1	271	6.3
35	Louisiana	585.3	28.9	30.5	6.6	12.6	316	7.0
36	Oklahoma	352.8	23.9	11.7	6.4	12.2	323	3.9
37	Texas	435.5	24.4	12.2	6.4	12.7	395	4.8
38	Montana	237.4	26.2	3.5	7.6	13.6	449	6.0
39	Idaho	236.6	24.9	0.1	4.5	12.8	348	5.7
40	Wyoming	280.8	17.2	3.4	6.6	14.8	520	3.3
41	Colorado	498.1	24.4	4.8	4.4	15.0	451	5.5
42	New Mexico	533.0	27.1	11.3	5.7	11.8	371	5.8
43	Arizona	552.5	22.9	4.8	4.0	13.6	401	6.1
44	Utah	272.1	28.5	0.6	4.2	14.3	405	3.8
45	Nevada	780.9	16.1	7.6	5.4	15.0	403	4.4
46	Washington	404.8	20.6	6.3	4.5	15.0	417	6.8
47	Oregon	502.3	20.2	3.5	5.1	13.9	429	6.0
48	California	742.6	22.3	12.9	3.9	15.1	409	7.1

Fonte: Bureau of the Census(ed.) - *Statistical Abstract of the United States*, 1980, 101 th edition, Washington DC

As variáveis são as seguintes:

Dependente:

Y - Crimes Violentos por 100.000 habitantes (1978)
(consideraram-se crimes violentos os assassinios, as violações, os roubos e os assaltos agravados)

Independentes:

X1 - Percentagem de População com menos de 18 anos (1979)
X2 - Percentagem de População não branca (1976)
X3 - Camas hospital/mil hab (1978)
X4 - Rendimento médio familiar (1975) 1000 US\$
X5 - Gastos na educação/capita (1978)
X6 - Percentagem de Desempregados (1978)

Neste exercício de interpretação, procura-se "explicar" a criminalidade violenta nos diversos estados dos U.S.A. numa perspectiva de ecologia social. Parte-se da hipótese de que poderá ser provocada pelas deficientes condições de vida, traduzidas de por esta série de variáveis. Por exemplo, de acordo com esta perspectiva, é de esperar mais crime, onde a pobreza é maior (variáveis 4 e 6), ou onde existe pouco apoio social, o que pode ser medido por baixos níveis de assistência médica (variável 3) e educativa (variável 5). Por outro lado, pode assumir-se uma posição conservadora e racista, esperando mais crime nas áreas em que existem menos população *white, male, protestant* (variável 2) e assumindo que a delinquência é fundamentalmente juvenil (variável 1). É claro que estas considerações são pouco consistentes e as relações entre as variáveis muito mais complexas, mas vamos ver o que se pode concluir desta correlação.

Vamos executar desta vez uma regressão múltipla passo-a-passo, cuja listagem (simplificada para os primeiros passos e mais detalhada para o último) a seguir se lista:

**** MULTIPLE REGRESSION ****						
Variable(s) Entered on Step Number 1.. X6						
Multiple R	0.46692		R Square	0.21802		
Adjusted R Square	0.20102		Standard Error			
169.17329						
F =	12.82471		Signif F =	0.0008		
Equation -----						
Variable	B	SE B	Beta	T	Sig T	
X6	70.944383	19.810443	0.466922	3.581	0.0008	
(Constant)	-2.551167		112.103086		-0.023	0.9819
Variable(s) Entered on Step Number 2.. X2						
Multiple R	0.56425		R Square	0.31837		
Adjusted R Square	0.28808		Standard Error			
159.69021						
F =	10.50934		Signif F =	0.0002		
Equation -----						
Variable	B	SE B	Beta	T	Sig T	
X2	7.027578	2.730198	0.338637	2.574	0.0134	
X6	52.764659	19.989283	0.347271	2.640		0.0114
(Constant)	19.145061	106.154271		0.180	0.8577	

Variable(s) Entered on Step Number 3.. X1

Multiple R 0.70858 R Square 0.50209
 Adjusted R Square 0.46814 Standard Error
 138.02600

F = 14.78978 Signif F = .0000

Equation -----

Variable	B	SE B	Beta	T	Sig T
X1	-28.402815	7.049174	-0.457936	-4.029	0.0002
X2	10.052845	2.476376	0.484415	4.059	0.0002
X6	58.001995	17.326283	0.381741	3.348	0.0017
(Constant)	688.596783	189.799597		3.628	0.0007

Variable(s) Entered on Step Number 4.. X4

Multiple R 0.76742 R Square 0.58894
 Adjusted R Square 0.55070 Standard Error
 126.86140

F = 15.40197 Signif F = .0000

Equation -----

Variable	B	SE B	Beta	T	Sig T
X1	-27.740926		6.482703	-0.447264	-4.279 0.0001
X2	12.146366		2.379683	0.585296	5.104 0.0000
X4	34.490479		11.442705	0.310899	3.014 0.0043
X6	50.184831		16.134598	0.330292	3.110 0.0033
(Constant)	219.170138	233.851288		0.937	0.3539

Variable(s) Entered on Step Number 5.. X5

Multiple R 0.77499 R Square 0.60061
 Adjusted R Square 0.55306 Standard Error
 126.52763

F = 12.63210 Signif F = .0000

Equation -----

Variable	B	SE B	Beta	T	Sig T
X1	-25.953297	6.663983	-0.418443	-3.895	0.0003
X2	12.889067	2.466299	0.621084	5.226	0.0000
X4	22.568231	15.686808	0.203431	1.439	0.1577
X5	0.517210	0.466892	0.170042	1.108	0.2743
X6	46.026872	16.524091	0.302927	2.785	0.0080
(Constant)	163.749168	238.541338		0.686	0.4962

Variable(s) Entered on Step Number 6.. X3

Multiple R 0.78097 R Square 0.60992
 Adjusted R Square 0.55283 Standard Error
 126.56052

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	6	1026819.68391	171136.61399
Residual	41	656720.13921	16017.56437

F = 10.68431 Signif F = .0000

Equation -----

Variable	B	SE B	Beta	T	Sig T
X1	-23.475371		7.121016	-0.378491	-3.297 0.0020
X2	12.816398	2.468034	0.617582	5.193	0.0000
X3	-15.778191		15.953241	-0.109954	-0.989 0.3285
X4	20.960155		15.774900	0.188936	1.329 0.1913
X5	0.517612	0.467013	0.170174	1.108	0.2742
X6	40.382772		17.485828	0.265780	2.309 0.0260
(Constant)	252.427987	254.893993		0.990	0.3278

5.3. - Evapotranspiração em Quelimane (Moçambique)

Conhecem-se os valores normais de sete variáveis climatológicas da estação de Quelimane (Moçambique), para os doze meses do ano.

As normais são as do período 1930-1960 e as variáveis são

- V1 - Precipitação mensal (em milímetros)
- V2 - Temperatura média do mês
- V3 - Temperatura máxima
- V4 - Temperatura mínima
- V5 - Pressão de vapor
- V6 - Insolação (em percentagem)
- V7 - Evapotranspiração (em milímetros)

Meses	V ₁ R (mm)	V ₂ T °C	V ₃ T Max	V ₄ T Min	V ₅ Pressão	V ₆ Insol.	V ₇ Evapot.
J	237	28,0	32,5	23,3	28,5	53	160
F	238	28,0	32,2	23,6	29,1	55	139
M	248	27,3	31,5	23,2	28,3	55	139
A	114	26,2	30,7	21,7	26,3	63	115
M	77	24,0	29,0	19,0	23,1	63	93
J	61	21,7	26,7	16,6	20,6	60	72
J	61	21,1	26,5	15,8	19,8	57	75
A	30	22,1	27,7	16,5	20,5	69	97
S	16	24,1	29,8	18,5	21,8	71	127
O	17	26,5	32,0	21,0	23,5	74	169
N	84	27,5	32,7	22,5	25,2	64	168
D	90	28,0	33,0	23,1	27,5	57	166
Anual	1373	21,4	30,4	20,4	24,5	61	1520

- Considere a Evapotranspiração a variável dependente;
- Calcule o coeficiente de correlação simples entre ela e todas as outras. Interprete.
- Efectue uma regressão múltipla passo-a-passo.
- Interprete o valor do coeficiente de correlação múltipla, os coeficientes, o valor de F e os erros-padrão.

Considerando os coeficientes de correlação simples entre a Evapotranspiração (variável sete) e as outras variáveis

Variáveis	Coef. correlação	Coef. Determinação
V1 - Precipitação	0,40212	16,17 %
V2 - Temperatura média	0,89514	80,13 %
V3 - Temperatura máxima	0,95444	91,10 %
V4 - Temperatura mínima	0,83586	69,87 %
V5 - Pressão	0,69693	48,57 %
V6 - Insolação	0,01442	0,02 %

A evapotranspiração numericamente está mais relacionada com a temperatura máxima de cada mês ($r = 0,95444$) "*explicando*" esta variável a variação de 91,1 % da evotranspiração ($r^2 = 0,9110$). Existem também relações fortes entre a evapotranspiração e outras variáveis; o que importa conhecer é em que medida essas relações são independentes entre si ou não. Salienta-se o valor da variável seis, a insolação, que parece não ter nada a ver com as evapotranspiração, o que sugere que os valores de temperatura e de humidade são gerais, independentes das condições locais de nebulosidade.

6. - Bibliografia

(*) CURWIN, Jon; SLATER, Roger (1996) – **Quantitative Methods for Business Decisions**, London, International Thomson Business Press, pp. 378 420, especialmente 397 a 420.

É um bom manual de introdução às técnicas de regressão simples e múltipla, exercícios trabalhados e problemas resolvidos, embora sejam quase todos de gestão e de economia.

DAVIS, John C. (1986) – **Statistics and Data Analysis in Geology, 2nd edition**, New York, John Wiley & Sons, pags. 468-478.

Um capítulo mais fraco numa obra de referência.

FERGUSON, Rob (1979) – **Linear Regression in Geography**, London, Institute of British Geographers, 44 pags. (CATMOG, 15)

Uma apresentação da aplicação das técnicas de regressão múltipla a Geografia. Não sendo muito simples, pode dar boas sugestões de trabalhos. Não é fácil de encontrar.

LEWIS-BECK, Michael (1980) – **Applied Regression. An Introduction**, London, Sage Publications, 77 pags. (SAGE, 22)

Uma apresentação simples das técnicas de regressão múltipla. Sobre variáveis "dummy" ver pags 66 a 71.

Anexo 7

Análise Fatorial

1. - Introdução

A maioria dos sistemas geográficos, sejam naturais ou socio-económicos, são descritos por um grande número de variáveis que estabelecem complexas relações entre si.

As Análises Factoriais são uma técnica de Análise Multidimensional de Dados que procuram tornar mais simples e mais evidentes estas relações.

O que as análises factoriais fazem é reescrever a informação contida nos dados. É um tipo de análise em que se estudam as relações entre o conjunto das variáveis iniciais, que foram consideradas como descritoras do sistema em estudo, e um outro conjunto de variáveis, secundárias, geradas a partir do primeiro conjunto e consideradas "melhores" descritores do sistema. É "melhor" não por que fornece mais informação, mas por que a fornece mais arrumada, ou seja, menos redundante, mais hierarquizada e mais facilmente interpretável.

No conjunto das variáveis iniciais, cada uma é simultaneamente considerada dependente e independente, com uma parte da variância que se relaciona com as outras variáveis e uma outra parte que lhe é própria ou específica. As variáveis secundárias do conjunto de variáveis que resulta da análise factorial são geralmente independentes entre si.

As Análises Factoriais usam-se com os seguintes objectivos:

1.- **Identificação de grupos de variáveis muito relacionadas entre si**

O que permite pesquisar de forma sistemática as "proximidades" entre as variáveis e agrupá-las, o que é difícil de fazer manualmente para conjuntos de dados muito grandes. Note-se que só a busca de semelhanças de comportamento entre 80 variáveis, exige a comparação memorizada de 3160 coeficientes de correlação.

2.- **Redução do número de variáveis em estudo**

Pode colocar-se a questão se todas as variáveis iniciais são mesmo necessárias. por exemplo, se todas as oitenta variáveis do exemplo anterior são mesmo necessárias. Se algumas delas estiverem contidas em outras, isto é, se uma parte da sua respectiva variância estiver contida noutras, é possível simplificar, reduzindo o seu número.

3.- **Descrição do conjunto de dados numa forma alternativa**

Tendo-se identificado "grupos" e reduzido o número de variáveis o conjunto de variáveis secundárias descreve o sistema em estudo de uma forma mais clara e mais simples do que o conjunto das variáveis iniciais.

2. – Tipos de Análise Factorial

Debaixo da designação comum de análise factorial encontra-se uma variedade de técnicas semelhantes, mas diferentes entre si, que se podem agrupar em três grandes tipos:

- A Análise em Componentes Principais (ACP);
- A Análise Factorial propriamente dita (AF);
- A Análise Factorial de Correspondências Múltiplas (AFC)

Do ponto de vista do tipo de dados as Análises em Componentes Principais e as Análise Factorial propriamente ditas são apropriadas para dados que se apresentem sob a forma de matrizes de atributos, medidos em escalas de intervalo ou razão, enquanto a Análise Factorial de Correspondência é especialmente adaptada a dados sob a forma de tabelas de contingência (quadros de frequência).

As diferenças entre Análise de Componentes Principais e Análise Factorial são de três tipos:

- do modelo matemático
- da utilização
- de procedimento

1) - Diferenças de utilização

A Análise de Componentes Principais é óptima na eliminação da redundância, razoável na selecção de grupos de variáveis e ineficaz na sua identificação.

A Análise Factorial é muito conveniente em todos estes aspectos, mas é mais complexa e morosa, o que nestes tempos de computadores rápidos já não tem importância.

2) - Diferenças do modelo matemático

Em Análise de Componentes Principais nada se presume sobre a estrutura subjacente às variáveis. Apenas se procuram as melhores combinações lineares das variáveis pré-existentes que irão definir as componentes.

Haverá tantas componentes principais (variáveis secundárias) quantas as variáveis iniciais.

A solução obtida é só uma.

Em Análise Factorial presume-se que as correlações observadas são devidas principalmente a "alguma" regularidade nos dados, considera-se que cada variável é influenciada por vários determinantes alguns dos quais são partilhados por outras variáveis., e outros não o são. A variância de cada variável inicial é considerada constituída por duas partes, uma comum, a comunalidade, que é a parcela da variação que é partilhada com outras variáveis e vai ser objecto de análise e uma outra específica apenas dessa variável, a unicidade, que lhe é própria e nada tem a ver com a estrutura de relações comuns entre variáveis e portanto ficará de fora da análise.

O número de factores é menor que o das variáveis iniciais.

Teoricamente há uma infinidade de soluções, mas delas escolhem-se apenas uma ou poucas, de acordo com critérios suplementares que visam dar uma melhor interpretabilidade aos factores.

3) - Diferenças do procedimento

Em Análise de Componentes Principais calcula-se a matriz de correlação a partir da matriz de dados com os valores das variáveis (colunas) nas unidades de análise (linhas). O traço da matriz de correlação é n , o número de elementos a factorializar (em regra as variáveis).

Em Análise Factorial a diagonal da matriz de correlação (que só tem 1's), é substituída pela comunalidade estimada para cada uma das variáveis. O traço da matriz é inferior a n , e igual ao total da comunalidade estimada.

A factorialização de cada uma destas matrizes é que dá a solução de Análise de Componentes Principais ou Análise Factorial.

Em Análise de Componentes Principais existe uma solução única. Em Análise Factorial existem infinitudes de soluções e muitos métodos diferentes para as obter. O principal problema nos métodos de Análise Factorial é que para a fazer é necessário conhecer as comunalidades, como acima vimos, e isso só se sabe após efectuar a análise.

Um dos processos consiste na determinação de valores aproximados para as comunalidades (por exemplo, considerar como comunalidade em vez do valor 1, o do coeficiente de correlação múltipla dessa variável com todas as outras) e depois recalcular novas comunalidades finais, e repetir o processo até que não haja modificações nas soluções.

2.1. – Preparação da Matriz de Correlação

Uma das alternativas que se colocam ao executar análises factoriais é a do cálculo de uma medida apropriada de associação a partir da matriz de dados, uma matriz de atributos.

Em geral as análises factoriais requerem como medida de associação, a utilização do coeficiente de correlação linear de Pearson embora possam ser utilizadas também matrizes de variância/covariância. Nas primeiras assume-se que a variância de todas e cada uma das variáveis é igual a um, ou seja, que têm igual importância (e o traço da matriz é igual a n), enquanto nas segundas se assume uma variância diferente para cada variável, proporcional a qualquer elemento mensurável que tenha significado. Por exemplo, nos resultados de uma eleição, tanto se pode considerar que todos os partidos têm igual importância, como assumir que a importância que têm é proporcional aos resultados percentuais que obtiveram.

Sendo os dados do tipo atributo, tanto se pode medir a associação entre as variáveis (modo R) como entre as unidades de análise (modo Q). Embora seja mais comum o modo R que dá origem a matrizes de correlação entre as variáveis, também é possível calcular matrizes de correlação entre as unidades de análise, em modo Q.

2.2. – A Extração dos Factores

Outra das alternativas que se colocam ao investigador é a escolha dos métodos de extração dos factores, isto é de identificar combinações das variáveis iniciais que, respeitando certas condições permitam reduzir e melhorar a interpretação da variância total existente na matriz de dados.

São duas as abordagens fundamentais à selecção / identificação dos factores: a do método das componentes principais e o método clássico do factor comum.

No método das componentes principais, que é o que se aconselha os alunos a seguir, o processo inicia-se por uma simples análise em componentes principais, em que se calculam os vectores e valores próprios da matriz de dados, que podem ser identificados respectivamente como as dimensões estruturantes da matriz e a sua capacidade de “explicação” da variância. Observando os valores próprios, verifica-se que, em geral, um pequeno número de componentes tem associado a si a maior parte da variância dos dados; retêm-se esses na análise e eliminam-se as restantes componentes, em grande número mas com pouca variância associada.

Como se eliminaram componentes, o conjunto das relações entre as variáveis iniciais e os componentes ficou indeterminado, o que permite introduzir mais restrições para a sua determinação, o que é efectuado através de rotações das componentes, obtendo-se factores.

No método clássico do factor comum, admite-se que nem toda a variância das variáveis é partilhada pelos factores. Deste modo, na diagonal da matriz de associação não deve estar 1 (a unidade, pois a variância da variável não é partilhada pelas outras), mas sim o valor da variância partilhada), que é menor. E é aí que reside o problema: como estimar a parcela de variância que cada variável partilha com todas as outras ? Têm sido propostas várias metodologias, umas directas outras iterativas. Dentro muitos outros salientam-se:

1) - R múltiplo ao quadrado. Assume-se que a comunalidade de cada variável é igual ao quadrado do seu coeficiente de correlação múltipla com todas as outras;

2) - R múltiplo ao quadrado melhorado. Foi proposto por Harman e Jones em 1966. Inicialmente também se assume, como no método anterior, que a comunalidade de cada variável é igual ao quadrado do seu coeficiente de correlação múltipla com todas as outras, mas após a extracção inicial dos factores a comunalidade é reajustada ao longo de várias iterações até que não haja melhoria no processo de ajustamento;

3) - Método da máxima verosimilhança. Proposto por Lawley em 1940 nele se considera que existe um número fixo de factores, sendo então estimados os "loadings" e as comunalidades (deles derivadas) que maximizam a probabilidade da matriz de correlação observada ocorrer.

4) - Método do eixo principal. Neste método em cada iteração os valores próprios são calculados a partir das comunalidades e depois as comunalidades são recalculadas a partir dos valores e vectores próprios extraídos anteriormente, e assim sucessivamente, até que se alcancem valores estabilizados.

5) - Método do centróide. Apresentado por Thurstone em 1931, hoje tem principalmente um interesse histórico. É um método geométrico, de formulação matemática complicada mas de aplicação simples que dava soluções aproximadas.

São ainda regularmente apresentados como opções alternativas em programas comerciais de análise factorial, os métodos Alfa e Imagem e a factorialização canónica de Rao, mas os alunos e investigadores, na fase inicial, devem utilizar o método das componentes principais.

2.3. – A Rotação dos Factores

Os factores extraídos pelo método de componentes principais, ou por alguns dos outros, podem dar origem, ou não, a um padrão significativo de variáveis, isto é, podem estar estreitamente associados a um grupo de variáveis (com quem têm elevadas correlações), ou estar divididos entre dois ou mais grupos de variáveis. Se na investigação em curso não era necessário que os factores tivessem significado, ambas as situações são aceitáveis, mas em geral espera-se que sejam significativos, e nesse caso deve fazer-se uma rotação para tentar melhorar a interpretabilidade dos factos, o mesmo é dizer, aumentar, tanto quanto possível, a sua relação com as variáveis.

Existem também vários métodos de rotação de factores, tanto ortogonais, isto é, mantendo a independência dos factores:

1) - Varimax. Simplifica as colunas da matriz de factores. Procura que a variância por coluna seja máxima, isto é, que cada variável tenha, tanto quanto possível, uma grande correlação com um factor.

2) - Quartimax. Simplifica as linhas da matriz de factores. Procura que a variância por linha seja máxima, isto é, que cada factor tenha, tanto quanto possível, uma elevada correlação com uma variável.

3) – Equimax. É uma solução de compromisso entre as duas anteriores.

4) – Biquartix. É também uma solução de compromisso entre as duas anteriores, mas com peso igual para as duas.

como oblíquas, em que os factores, procurando ajustar-se melhor aos grupos de factores, admitem alguma correlação entre si, em geral pequena, como é o caso das rotações Oblimin, Oblimax e Quartimin.

Os factores rodados estão sempre muito mais associados às variáveis do que os iniciais, razão pela qual, quando se fazem rotação de factores, deixa de ser interessante a análise das características das componentes ou dos factores iniciais.

A principal dificuldade e também o principal desafio que se coloca ao investigador é a de compreender o que representa cada factor. Na realidade, se no conjunto de dados, que se considera ser o necessário e suficiente para compreender o fenómeno ou processo em análise, as variáveis se tendem a juntar em grupos (em torno de cada um dos factores) então essas associações têm sentido, que compete ao investigador decifrar.

Este processo de identificação dos factores, ou seja, dos grupos estruturados de variáveis, é a parte que coloca mais desafios mas também é a mais estimulante.

2.4. – O Cálculo do Valor dos Factores

Identificados os factores, que são, recorda-se, combinações lineares das variáveis iniciais (e portanto índices que resumem o grupo de variáveis a que estão associados), resta apenas redescrever as unidades de análise pelos os factores, calculando os seus valores.

Ao valor dos factores nas unidades de análise chama-se *score*, pontuação, ou peso. E a maioria dos programas permite calculá-los.

A sua importância deriva de capacidade de identificação dos factores. Se foi possível identificar um factor da análise com um factor estruturante da realidade ou do processo em estudo, então os valores desse factor da análise servem como um indicador ou índice do factor estruturante da realidade.

3. – Marchas de Análise

As listagens de resultados que os programas de análise factorial disponíveis no comércio, integrados em grandes “packages” como o Statistica, SPSS (Statistical Package for Social Sciences), SPAD, BMD, SAS, entre muitos outros são longas e difíceis de interpretar.

O que vem dificultar a interpretação é a existência de muita informação, uma grande parte da qual não tem de ser utilizada, nas utilizações mais habituais.

Para facilitar a leitura e a interpretação apresenta-se de seguida um exemplo das várias matrizes que podem surgir nas listagens e o que são. As mais importantes para a interpretação estão assinaladas a negrito (*bold*).

3.1. – Sequência de uma Análise em Componentes Principais

- [D] (m,n) - **Matriz de Dados**, com m linhas (os objectos) e n colunas (as variáveis)
- [S] (m,n) - Matriz de dados padronizados
- [C] (n,n) - **Matriz de correlação** entre as variáveis (modo R). Representa a proximidade entre as variáveis.
- [λ] (l,n) - Matriz linha de **Valores Próprios** da matriz de correlação. Representa a parte da variação 'explicado' por cada componente. Dividindo por n e multiplicando por 100 obtém-se a percentagem de explicação de cada componente.
- [I] (n,n) - Matriz de Vectors próprios, que por vezes não é dada como *output*. Dela se obtém
- [L] (n,n) - **Component loading matrix**. Matriz de 'loadings'.; cada loading representa simultaneamente a correlação entre a componente e a variável (cosseno do ângulo que fazem entre si) e os coeficientes de regressão para calcular as variáveis a partir dos componentes.
- [CS] (m,n) - **Component score matrix**. Matriz de 'scores'. Cada 'score' representa o valor (peso ou carga) que a componente tem em cada caso ou unidade de análise. Aparece com as variâncias de cada componente proporcionais à importância que o componente tinha nos dados.
- [CSS] (m,n) - **Component score matrix** padronizada, com variâncias unitárias nos factores.

3.2. – Sequência de uma Análise Factorial

- [D] (m,n) - **Matriz de dados**. m casos, n variáveis.
- [S] (m,n) - Matriz de dados padronizados
- [C] (n,n) - **Matriz de correlação** das variáveis (em modo R)
- [λ] (l,p) - Matriz linha dos **Valores Próprios**. Representa a 'potência' explicativa de cada factor inicial medida em termos da variação comum. O número de factores (p) é menor que o número de variáveis (n)
- [L] (n,p) - Matriz inicial ou não rodada de loadings. Representa a correlação entre n variáveis e os p factores iniciais e simultaneamente os coeficientes de regressão para estimar as variáveis a partir dos factores iniciais.

Como se sabe, existem vários métodos para as obter, consoante a técnica de estimação das comunalidades empregue: Componentes principais, Common factor, Image factor, etc.

E como existem várias soluções factoriais possíveis, importa, a partir desta estrutura de factores iniciais, procurar outras estruturas de factores que permitam uma melhor correspondência com as variáveis. Para isso roda-se a estrutura de factores no espaço multidimensional por elas definido, procurando uma melhor solução, ou seja, uma maior "proximidade" entre os factores e grupos de variáveis. Existem dois tipos de rotações, as ortogonais e as oblíquas, com diversas variantes: Salientam-se as rotações QUARTIMAX e VARIMAX de entre as ortogonais e a OBLIMIN nas oblíquas.

Obtém-se assim, conforme o tipo de rotação:

- [LR] (n,p) - **Rotated Loading Matrix**. Matriz de loadings rodada. Contém os valores dos coeficientes de correlação entre as variáveis e os factores rodados, que são ao mesmo tempo os coeficientes de regressão para estimar aquelas a partir destes. Esta matriz só existe nas rotações ortogonais, pois nas oblíquas é desdobrada em:
- [RP] (n,p) - Rotated Pattern Matrix, dos coeficientes de regressão para estimar as variáveis a partir dos factores rodados.
- [SM] (n,p) - Structure Matrix, das correlações entre os factores rodados e as variáveis.

Nas rotações oblíquas existe ainda a

- [FC] (p,p) - Factor Correlation Matrix, da correlação dos factores rodados obliquamente, entre si. Obviamente que nas rotações ortogonais é igual à matriz identidade.

Nalguns 'packages' por vezes poderá aparecer como output a seguinte matriz:

- [FT] (p,p) - Factor Transformation Matrix, das correlações entre os factores iniciais e os rodados. É a matriz dos ângulos de rotação respectivos. A sua análise não tem interesse, a não ser que se tenha decido analisar a matriz inicial.
- [CS] (n,p) - Factor Score Coefficients Matrix. Matriz dos coeficientes dos scores factoriais. Contém os coeficientes de regressão para estimar os factores rodados a partir das variáveis. É uma estimativa e não um valor exactamente calculado, devido aos factores terem a ver, não com a totalidade de variação das variáveis, mas apenas com a respectiva comunalidade.
- [SF] (m,p) - **Factor Score Matrix**. Matriz de scores factoriais, em que cada valor representa o 'score', peso ou carga que se estima que o factor tem em cada caso. Tanto pode aparecer com as variâncias de cada factor proporcionais à importância que tinham na análise,
- [SFS] (m,p) - **Factor Score Matrix** padronizada, com variâncias unitárias.

4 – Exemplo

Na parte final de um folheto de informação estatística *Moçambique em Números 1997*, vem um quadro informativo com comparações internacionais.

Para exemplificar a utilização da metodologia de análise factorial, irá proceder-se a uma análise factorial da matriz de dados do Quadro 7.1, em que estão os valores de sete variáveis em quinze países

- V1 - Densidade de população (hab/km²)**
V2 - % de variação da população
V3 - Taxa de fecundidade total
V4 - Esperança de vida
V5 - Mortalidade infantil (por 1000 nados vivos)
V6 - PIB/capita em dólares americanos de 1994
V7 - Variação percentual do PIB/cap.

Vai procede-se à análise e interpretar os resultados, seguindo um listagem-tipo de resultados.

País	V1	V2	V3	V4	V5	V6	V7
Angola	9	3.0	7.0	46	124	384	8.6
Botswana	3	2.0	4.0	51	56	2666	4.1
Brasil	19	1.0	2.0	67	36	3787	5.7
Cabo Verde	97	2.0	4.0	66	43	900	2.6
Lesotho	67	2.0	5.0	58	74	424	-1.4
Malawi	106	3.0	6.0	43	133	130	-9.8
Namíbia	2	3.0	5.0	56	61	1924	7.3
Portugal	109	0.0	1.0	75	7	8832	1.0
S.Tomé Príncipe	141	2.0	5.0	64	50	120	1.5
Africa do Sul	31	2.0	3.0	65	49	2835	2.1
Swazilandia	54	3.0	5.0	57	67	1311	0.3
Tanzania	30	2.9	5.8	51	84	119	3.0
Zambia	12	3.0	6.0	44	112	367	1.0
Zimbabwe	29	2.0	4.0	56	56	568	4.6
Moçambique	21	2.6	5.6	45	135	217	9.9

Quadro 7.1 – Comparação internacional de alguns indicadores de desenvolvimento

LISTAGEM-TIPO DE RESULTADOS

***** ***** DIOGO & ROQUETTE SOFTWARE ***** ***** ***** ANÁLISE FACTORIAL ***** Nome do Trabalho: Comparações internacionais - INE 97 ***** RESULTADOS DO PROGRAMA ***** No. de Variáveis 7 No. de Observações 15				INTERPRETAÇÃO O programa utilizado nesta demonstração é o de análise factorial do conjunto de programas ESTAT, dada a facilidade com que se pode processar a sua utilização. Mais tarde, aconselha-se a utilização de pacotes mais profissionais. O número de variáveis são sete e o de

*** Médias das Variáveis ***

48.67	2.23	4.56	56.27	72.47	1638.93	2.70

*** Desvios-Padrões das Variáveis ***

44.74	0.86	1.61	9.69	37.96	2303.82	4.70

*** Matriz de Correlação das Variáveis ***

1)	2)	3)	4)	5)	6)	7)
1.000	-0.343	-0.189	0.435	-0.265	0.130	-0.610
-0.343	1.000	0.912	-0.815	0.766	-0.823	-0.020
-0.189	0.912	1.000	-0.853	0.856	-0.853	-0.019
0.435	-0.815	-0.853	1.000	-0.927	0.686	0.016
-0.265	0.766	0.856	-0.927	1.000	-0.684	-0.048
0.130	-0.823	-0.853	0.686	-0.684	1.000	0.032
-0.610	-0.020	-0.019	0.016	-0.048	0.032	1.000

*** Valores Próprios ***

4.3855	1.5782

*** Percentagem de explicação ***

62.65	22.55

*** Perc. cumulativa de explicação ***

62.65	85.20

*** Matriz de Loadings (2 factores) ***

	1)	2)
(1)	-0.3792	-0.8392
(2)	0.9365	-0.0380
(3)	0.9564	-0.1319
(4)	-0.9384	-0.0325
(5)	0.9129	-0.0803
(6)	-0.8573	0.1662
(7)	0.0319	0.9055

Iteração Variâncias

0	0.941235
1	0.962314
2	0.962314
3	0.962314
4	0.962314
5	0.962314

*** ROTACAO VARIMAX ***

*** Matriz de Loadings rodada (2 factores) ***

	1)	2)
(1)	-0.2476	-0.8870
(2)	0.9314	0.1043
(3)	0.9654	0.0146
(4)	-0.9226	-0.1743
(5)	0.9145	0.0589
(6)	-0.8726	0.0343
(7)	-0.1057	0.8999

*** Nova Variância explicada ***

4.3210	1.6427

*** Percentagem de explicação ***

61.73	23.47

*** Perc. cumulativa de explicação ***

61.73	85.20

observações quinze. O número de observações deve ser maior que o de variáveis, para que não haja dependência linear, que traz problemas de cálculo, e também por que não vale a pena medir muito em universos pequenos.

As médias e os desvios-padrões valem fundamentalmente para confirmar a correcção dos dados entrados, não sendo geralmente objecto de análise especial.

A matriz de coeficientes de correlação deve ser bem observada. É aqui que se revelam as relações entre as diversas variáveis. Convém inventariar os grupos de variáveis com maiores correlações (negativas ou positivas). Pode até pintar-se os coeficientes mais elevados como forma de salientar os grupos. Aqui identifica-se facilmente um com as variáveis V2, V3, V4, V5 e V6.

A seguir procede-se à análise factorial propriamente dita.

Inicia-se com a extracção dos valores próprios. Existem tantos quantas as variáveis (sete neste caso), mas vão diminuindo de valor. Aqui escolheu-se como critério de selecção o do coeficiente de Kaiser igual a um (significando que só se querem factores que “expliquem” mais variância do que uma das variáveis iniciais isoladamente). Se o coeficiente de Kaiser fosse zero, seriam seleccionados todos valores próprios, não poderia haver qualquer rotação de factores e estaríamos perante uma análise de componentes principais.

Verifica-se que os dois factores seleccionados explicam 82,5 % da variação total das sete variáveis iniciais, e que os outros cinco factores cuja variância é desprezada corresponderiam apenas a 14,8% da variância.

Mas estes primeiros factores são geralmente pouco interessantes e difíceis de interpretar, pelo que se procedeu a uma rotação de tipo Varimax.

A matriz rodada é a chave de toda a análise factorial. Contém as correlações entre as variáveis iniciais e os factores, que simplificam e estruturam a variação total do conjunto.

São dois os factores. A parcela total da variância que explicam é a mesma da matriz de factores não rodada, mas a repartição entre cada um é diversa.

O primeiro factor está muito associado à variável 3 (taxa de fecundidade total), à 2 (crescimento da população), à 4, negativamente (esperança de vida), à 5 (mortalidade infantil) e à 6, também de forma negativa (o Pib/capita).

Estas relações, e o facto da variável 6 estar bem correlacionada negativamente

*** Conferência das comunalidades ***

Variável	Original	Final	Diferença
1	0.84809	0.84809	0.00000
2	0.87839	0.87839	0.00000
3	0.93219	0.93219	0.00000
4	0.88162	0.88162	0.00000
5	0.83984	0.83984	0.00000
6	0.76263	0.76263	0.00000
7	0.82091	0.82091	0.00000

*** Matrizes de scores - 2 factores ***

** Coeficientes dos scores

	1)	2)
(1)	-0.02104	-0.88497
(2)	0.92782	0.01401
(3)	0.98623	-0.08141
(4)	-0.90042	-0.08672
(5)	0.92242	-0.03083
(6)	-0.90392	0.12228
(7)	-0.34468	0.93339

Scores proporcionais à importância dos factores

	1)	2)
(1)	4.61	1.83
(2)	-0.99	1.32
(3)	-5.84	1.34
(4)	-1.94	-1.05
(5)	0.66	-1.28
(6)	5.90	-3.69
(7)	0.42	1.85
(8)	-10.65	-1.12
(9)	-0.61	-2.22
(10)	-3.01	0.31
(11)	1.20	-0.61
(12)	2.83	0.33
(13)	4.45	0.34
(14)	-0.68	0.75
(15)	3.64	1.91

Scores padronizados

	1)	2)
(1)	1.07	1.11
(2)	-0.23	0.80
(3)	-1.35	0.82
(4)	-0.45	-0.64
(5)	0.15	-0.78
(6)	1.36	-2.25
(7)	0.10	1.13
(8)	-2.47	-0.68
(9)	-0.14	-1.35
(10)	-0.70	0.19
(11)	0.28	-0.37
(12)	0.66	0.20
(13)	1.03	0.21
(14)	-0.16	0.46
(15)	0.84	1.16

***** Fim do trabalho *****

com a densidade e o crescimento da população, sugere que o factor está ligado ao desenvolvimento, ou melhor, ao não desenvolvimento humano, que neste conjunto também está muito relacionado com o facto de se ser rico ou pobre.

Como hipótese de trabalho irá chamar-se a este factor o de desenvolvimento, especialmente humano, mas também económico. Explica 61,73% da variação total.

O factor 2 relaciona-se com as variáveis 1 (densidade de população) e 7 negativamente (variação do PIB), com valores próximos de 0,9, embora estas duas variáveis não tenham uma correlação muito forte entre si (-0.61).

Este factor é mais difícil de caracterizar. Agrupa uma variável conjuntural, a variação do produto, com a densidade, ambas com poucas relações com as outras.

Poderá ser um indicador do dinamismo recente dos países, que parece não ter a ver com o grau de desenvolvimento (o PIB/capita) mas sim com característica dos territórios, .

Temos assim as sete variáveis agrupadas em dois factores, o primeiro relativo ao desenvolvimento humano e económico, e o segundo ao dinamismo recente.

No quadro da conferência das comunalidades descreve-se a parte de cada variável que está incluída na variância que foi desprezada aquando da selecção dos valores próprios.

Com a identificação do significado dos factores, que se faz, como se viu, através das relações que mantêm com as variáveis iniciais, é possível descrever grupos de variáveis associadas entre si, que são as grandes dimensões subjacentes à organização dos dados

A matriz de coeficientes dos scores é uma matriz de trabalho que geralmente não tem interesse analisar. Os scores, pontuações ou pesos, vêm mostrar a expressão dos factores nas unidades de análise, que neste caso são os países. Estes pesos factoriais permitam hierarquizar e agrupar os países segundo os dois factores estruturantes anteriormente identificados, o nível de desenvolvimento humano e económico (Factor I) e o dinamismo recente (Factor II).

As análises a fazer a seguir são a cartografia dos factores e a classificação das unidades de análise. Os países vêm organizados pelo peso dos factores segundo duas formas, uma proporcional à importância que cada factor tinha nas variáveis em estudo, outra com os factores padronizados, com média nula e desvio-padrão unitário

Existem formas muito expressivas de apresentar alguns destes quadros. Apresenta-se de seguida dois gráficos, um com a projecção das variáveis iniciais no espaço dos factores rodados (com os valores da matriz de "loadings" rodada) e outro com a projecção dos países nesse mesmo espaço (Figura 7.1).

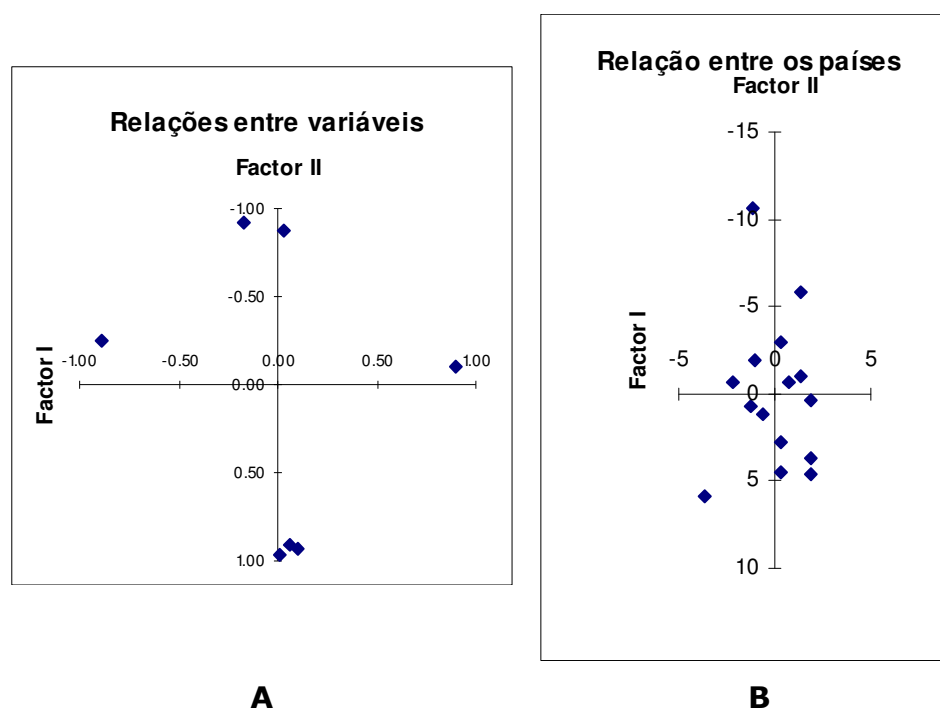


Figura 7.1 - Representação das variáveis e dos países no espaço dos factores.

Na figura A vê-se bem como as variáveis se organizam: no Factor II à esquerda a densidade (V1), e à direita a variação do produto (V7), enquanto no Factor I se tem em cima, com valores negativos a V4 e a V6, a esperança de vida e a riqueza, opostas à variação da população, taxa de fecundidade e mortalidade infantil. Quanto aos países é Portugal o mais desenvolvido no conjunto, logo seguida pelo Brasil, com valores elevados nas variáveis homólogas do respectivo factor. O Malawi destaca-se com valores preocupantes nos dois factores. Angola, com um nível de desenvolvimento similar tem muito mais dinamismo. E qual é a posição de Moçambique ?

5 – Bibliografia

BRYMAN, Alan; CRAMER, Duncan (1993) – **Análise de Dados em Ciências Sociais. Introdução às técnicas utilizando o SPSS**, Celta Editores, Oeiras, p. 307-323.

(*) CICERI, M.F.; MARCHANT, B.; RIMBERT, S. (1977) – **Introduction a l'Analyse de l'Espace**, Masson, Paris, p. 13-42.

- DAVIS, John C. (1986) – ***Statistics and Data Analysis in Geology, 2nd edition***, New York, John Wiley & Sons, p. 515-573.
- (*) KIM, Jae-On (1975) - “***Factor Analysis***” in NIE, Norman H. et alia - ***SPSS -Statistical Package for Social Sciences, 2th edition***, McGraw-Hill Book Company, New York, p. 468-514.
- MORENO JIMÉNEZ, Antonio (1994) – “***Análisis Factorial***” in BOSQUE SENDRA, Joaquín; MORENO JIMÉNEZ, Antonio – ***Prácticas de Análisis Exploratório y Multivariante de Datos***, Oikos-Rau, Barcelona, p. 87-131.
- MATHER, Paul (1976) – ***Computational Methods of Multivariate Analysis in Physical Geography***, John Wiley & Sons, London, p. 215-306.
- MATHER, P.M. (1981) “***Factor Analysis***”, in WRIGLEY, N.; BENNETT, R.J. (Eds) – ***Quantitative Geography***, Routledge & Krgan Paul, London, p. 144-150.
- MELO, F. Galvão de (1982) – ***Modelos e Métodos Estatísticos em Psicologia***, Escola Nacional de Saúde Pública, Lisboa, p. 54-112.
- MORRISON, Donald F. (1976) – ***Multivariate Statistical Methods***, McGraw-Hill Kogakusha, Ltd., Tóquio, p. 226-345.
- (*) PESTANA, Maria Helena; GAGEIRO, João Nunes (1998) – ***Análise de Dados para Ciências Sociais. A complementaridade do SPSS***, Lisboa, Edições Sílabo, p. 321-344.
- (*) REIS, Elizabeth (1997) – ***Estatística Multivariada Aplicada***, Lisboa, Edições Sílabo, p. 251-284.
- TORRENS-IBERN, J. (1972) – ***Modèles et Méthodes de l’Analyse Factorielle***, Dunod, Paris, 202 p.
- VOLLE, Michel (1978) – ***Analyse des Donnees***, Economica, Paris, p. 61-194 p.
- YEATES, Maurice (1974) – ***An Introduction ro Quantitative Analysis in Human Geography***, MacGraw-Hil, New York, p. 207-238.

Anexo 8

Classificação Numérica

1.- APRESENTAÇÃO DO PROBLEMA

As Taxionomias numéricas, também chamadas Metodologias de Classificação Automática, são técnicas gerais de agrupamento ou de classificação de indivíduos, casos, elementos ou unidades de análise. Dizem-se numéricas porque a base do agrupamento ou classificação destes indivíduos ou unidades de análise se baseia nos valores numéricos que cada um apresenta em uma ou várias variáveis.

Existem diversos métodos de classificação automática, mas todos eles procuram organizar as unidades de análise em grupos ou classes da melhor maneira possível. O que em cada caso se entende por "melhor" é uma das principais dificuldades práticas que se encontram na utilização destas metodologias.

A marcha geral que se segue neste tipo de análise é simples. Partindo-se duma *matriz de atributos* (Figura 8.1), em que existem uma série de valores a_{ij} correspondentes aos valores da variável j no caso ou unidade de análise i , constrói-se uma *matriz de proximidade* (ou uma *matriz de distâncias*, conforme os casos), entre os casos ou unidades de análise.

De seguida agrupam-se estas, com base em critérios de semelhança, considerando que estão mais próximas e por isso devem pertencer ao mesmo grupo ou classe os casos ou unidades de

análise que têm valores mais aproximados nas diversas variáveis.

	V_1	V_2	V_3	V_4	...	V_j
C_1	a_{11}	a_{12}	a_{13}	a_{14}	...	a_{1j}
C_2	a_{21}	a_{22}	a_{23}	a_{24}	...	a_{2j}
C_3	a_{31}	a_{32}	a_{33}	a_{34}	...	a_{3j}
C_4	a_{41}	a_{42}	a_{43}	a_{44}	...	a_{4j}
...	...	...	...	...	...	...
C_i	a_{i1}	a_{i2}	a_{i3}	a_{i4}	...	a_{ij}

Figura 8.1 – Matriz de Atributos

Existem dois grandes tipos de taxionomias numéricas, as *classificações hierárquicas* e as *partições*. Nas primeiras, cada vez que se forma um grupo, ele constitui-se por junção (no caso das *classificações hierárquicas ascendentes*) ou por segmentação (no das *hierárquicas descendentes*) de grupos pré-existent, ao passo que nas partições, os grupos ou classes de unidades de análise existentes em etapas anteriores do método não constituem restrições à formação de grupos.

Assim, as classificações hierárquicas formam uma sequência de agrupamentos ou separações de grupos pré-existent, numa série que vai de **n** grupos de **1** elemento a **1** único grupo final com **n** elementos. Esta sequência de classes pode ser visualizada por um gráfico esquemático, a que se dá o nome de árvore de ligação, dendograma, ou fenograma (segundo a denominação proposta por SOKAL).

As árvores de ligação evidenciam, muito facilmente, o modo como se efectuam as “fusões” entre indivíduos ou grupos e o nível de proximidade ou de semelhança a que se efectuam. No gráfico, a proporção com a proximidade é dada pela distância vertical entre a base do gráfico, onde estão os nomes, letras ou números que identificam as unidades de análise e a barra horizontal da união dos respectivos grupos. Deve notar-se que a disposição das unidades de análise no eixo dos **XX's** não tem significado especial, decorrendo apenas das restrições da classificação. Alguns autores procuram uma disposição que “seja lógica”, o mesmo é dizer, tanto quanto possível ordenadas por uma, ou várias, variáveis de controle, que por vezes é o eixo de máxima variância da nuvem de pontos que a matriz de atributos representa. Geralmente os algoritmos disponíveis não apresentam essa ordenação, pelo que, nesses casos, a leitura da disposição das variáveis segundo a horizontal carece de sentido. Com efeito, toda a informação do gráfico respeita aos grupos que se constituem e ao diferente grau de proximidade a que se efectua cada junção (figura 8.2).

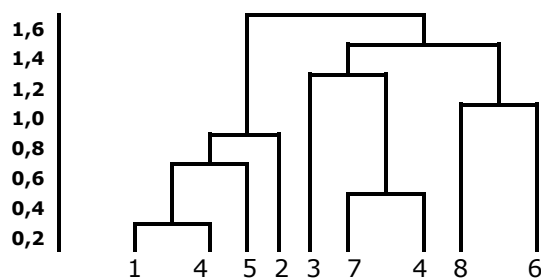


Figura 8.2 – Esquema de uma árvore de ligação

Quando a classificação ou agrupamento é não-hierárquica está-se perante uma partição. O resultado é um conjunto de **k** grupos, ou classes, com a “melhor” atribuição a cada grupo das **n** unidades de análise, face ao critério de semelhança em causa.

Se porventura se pede uma sequência de partições, por exemplo, de **1** a **n** grupos, obtêm-se sucessivas partições, com um número crescente de classes, como nas classificações hierárquicas, mas em que os grupos de cada nível não se formam por agregação ou desagregação de grupos do nível anterior, ao invés do que acontece com estas.

Por isso, não existe qualquer representação gráfica, do tipo de árvore.

Em taxionomia numérica são três os problemas que se colocam ao investigador:

1.- A escolha da metodologia de agrupamento:

- partição ou classificação hierárquica ?
- classificação hierárquica ascendente ou descendente ?

Quando se opta por uma partição, geralmente é necessário indicar o número de grupos a formar. Uma alternativa possível, que poucos programas permitem, é construir todos os conjuntos de grupos, desde um único, com todos os elementos, até n , em que cada elemento está isolado.

Quando se escolhe a classificação hierárquica, é preciso decidir se será ascendente (de longe a que mais se utiliza) ou descendente (que só muito raramente se utiliza). Assim, obtém-se uma sequência de k grupos, desde $k=1$ até $k=n$ (sendo n o número de elementos iniciais), de tal modo que cada grupo do nível $k=i+1$ é formado pela junção de dois grupos existentes ao nível $k=i$.

2.- Definição da medida de distância ou de proximidade a usar:

É a distância que se utiliza na passagem da matriz de atributos para a matriz de distância ou de proximidade.

Deve obviamente ser aquela que mais se adapte à estrutura de semelhança contida nos dados iniciais. A título de exemplo, se se tratar de valores absolutos de variáveis socio-económicas de vários países, para estudar situações sociais e económicas, deve-se ser usada uma medida de distância que não seja sensível ao tamanho ou dimensão dos países; e o mesmo acontece quando se trabalha com dados que já se encontram ponderados, como índices ou percentagens; se os dados de base forem dados em escalas ordinais obviamente que as medidas de perfil, que acima eram as mais adequadas, agora já não são interessantes.

Deve-se, por isso, ensaiar as medidas que parecem ser as mais adequadas às diversas utilizações, especialmente nos "limites", junto aos valores extremos, verificando como essa distância se comporta nas maiores e menores diferenças, se mantém homogeneidade nos valores elevados e nos valores baixos, etc.

3.- Definir a estratégia ou critério de agrupamento (a agregação de grupos) mais adequada no caso das classificações ascendentes, que são as mais frequentes. Se se optar por classificações hierárquicas descendentes, então importa definir o critério de separação a utilizar.

Consiste em definir como se mede a distância entre um elemento isolado e um grupo, ou entre dois grupos, pois a medida de distância que escolhida inicialmente já informa de qual a distancia entre dois elementos isolados.

2.- MEDIDAS DE SEMELHANÇA E DISTÂNCIAS A UTILIZAR

Existem muitas medidas de semelhança e de proximidade ou distância, que são utilizadas nas Metodologias de Classificação Automática. Procuram adaptar-se ao tipo de dados de que se dispõe e ao tipo de similitude que se procura encontrar entre as unidades de análise. Descreveremos de seguida as mais usuais.

2.1.- Para dados nominais

São muitas as medidas que têm sido apresentadas para medir a semelhança entre duas unidades de análise, a partir dos valores que ambos têm num conjunto de variáveis.

No caso das variáveis binárias ou dicotómicas, os valores das variáveis apresentam-se sob a forma de **0** ou **1**. No caso de variáveis nominais com várias classes, é conveniente transformá-las, construindo uma variável secundária de ausência/presença para cada classe.

A base do cálculo das medidas de semelhança para este tipo de informação é o quadro de frequências dos valores das diversas variáveis para as duas unidades de análise **k** e **l**.

Dada uma matriz atributos com dados que foram medidos em escalas nominais dicotómicas ou alternativas, isto é, que apenas assumem os valores de **0** ou **1**, (Figura 8.3), pode construir-se um

quadro de frequências (Figura 8.4), em que **a**, **b**, **c** e **d** são as frequências (número de vezes), em que ocorrem simultaneamente nas mesmas unidades de análise, os valores das variáveis (**0** e **1**)

	V ₁	V ₂	V ₃	V ₄	...	V _j
C ₁	a ₁₁	a ₁₂	a ₁₃	a ₁₄	...	a _{1j}
C ₂	a ₂₁	a ₂₂	a ₂₃	a ₂₄	...	a _{2j}
C ₃	a ₃₁	a ₃₂	a ₃₃	a ₃₄	...	a _{3j}
C ₄	a ₄₁	a ₄₂	a ₄₃	a ₄₄	...	a _{4j}
...	...	...	...	...	...	...
C _l	a _{l1}	a _{l2}	a _{l3}	a _{l4}	...	a _{lj}

Figura 8.3 – Matriz de Atributos

		Caso l		Σ
		1	0	
Caso k	1	a	b	a+b
	0	c	d	c+d
Σ		a+c	b+d	a+b+c+d

Figura 8.4 – Quadro de frequência dos valores das variáveis nas unidades de análise k e l.

Assim se vê, por exemplo, que **a** é o número de vezes em que os valores das variáveis que descrevem as unidades de análise **k** e **l** têm ao mesmo tempo, o valor de **1**, e **c** é o número de vezes que a unidade de análise **k** tem o valor **0** e a unidade de análise **l** tem o valor de **1**.

2.1.1.- Para dados binários dicotômicos

As dicotomias são tipos de dados binários de tipo presença/ausência, em que a ocorrência simultânea dos dois caracteres (**a**) é importante na definição da noção de semelhança e a sua não ocorrência ao mesmo tempo (**d**) não tem ou tem pouco significado nessa definição. Por exemplo, ao analisar semelhanças entre cidades, o facto de terem ambas um aeroporto internacional é uma medida de semelhança razoável, mas não o terem, pode corresponder a situações muito distintas, que vão desde uma cidade média a um pequeno lugarejo.

2.1.1.1.- Coeficiente de JACCARD (1908)

$$J_{kl} = \frac{a}{(a + b + c)}$$

Varia entre o mínimo de **0** e o máximo de **1**.

2.1.1.2.- Coeficiente de RUSSEL & RAO (1940)

$$R_{kl} = \frac{a}{(a + b + c + d)}$$

Também varia entre **0** e **1**.

2.1.1.3.- Coeficiente de DICE (1915)

$$D_{kl} = \frac{2a}{(2a + b + c)}$$

Varia entre **0** e **1**. Reforça o valor das presenças simultâneas, ainda mais que os dois anteriores. Também é conhecido por coeficiente de Sørensen (1948)

2.1.2.- Para dados binários alternativos

As alternativas são tipos de dados binários, em que quer a ocorrência do estado **1** quer a ocorrência do estado **0**, tem aproximadamente o mesmo valor na construção da noção de semelhança e assim tanto **a** como **d** têm importância nessa construção.

2.1.2.1.- Coeficiente de Concordância Simples

$$Cs_{kl} = \frac{(a + d)}{(a + b + c + d)}$$

Pode assumir valores entre **0** e **1**. Em inglês chama-se *simple matching*. É também referido por alguns autores como o Coeficiente de Sokal & Michener (1958).

2.1.2.2.- Coeficiente de SNEATH & SOKAL

$$S_{kl} = \frac{2(a + d)}{2(a + d) + (b + c)}$$

Varia entre **0** e **1**. Reforça o valor das ocorrências simultâneas, tanto do estado 1 como do estado 0.

2.1.2.3.- Coeficiente de YOULE

$$Y_{kl} = \frac{(a \times d - b \times c)}{(a \times d + b \times c)}$$

Este coeficiente varia entre **- 1** e **+ 1**.

2.1.2.4.- Coeficiente correlação de PEARSON (dados nominais)

$$\Phi_{kl} = \frac{(a \times d - b \times c)}{\sqrt{(a + b) \times (a + c) \times (c + d) \times (b + d)}}$$

Tal como o anterior, também este coeficiente varia entre **- 1** e **+ 1**. É também conhecido por Coeficiente Fi.

Os coeficientes de semelhança que são mais utilizados em trabalhos de tipo geral com os dados de tipo nominal, são o Coeficiente de Jaccard, para dicotomias e o Coeficiente de Concordância Simples, para o caso das alternativas. São estes que se devem utilizar se não houver outras razões para sustentar a escolha. Para comparar com estudos feitos noutros contextos, devem usar-se os mesmos coeficientes, pois só assim os valores determinados são comparáveis.

2.2.- Para dados de intervalo e razão

Para dados de intervalo e razão e para a generalidade dos dados ordinais, a questão prévia que se coloca é a da comparabilidade das diversas colunas da matriz, ou seja, como ponderar a diferença de valores entre as unidades de análise, que é diferente em cada variável e respeita na maioria das vezes a unidades de medida e a intervalos de variação muito diferentes.

Dois casos se podem apresentar.

No primeiro, as unidades de medida são as mesmas e o que é muito variável é o intervalo de variação de cada uma das variáveis; importa verificar se essa variação existe ou não na realidade do mundo que o modelo de dados pretende descrever. Por exemplo, no caso dos dados serem resultados eleitorais, a diferença na amplitude dos valores é um dado importante e significativo no comportamento eleitoral das populações e nesse sentido deve ser mantido.

Se porventura as unidades de medida não são as mesmas, ou sendo-o, as diferenças na amplitude de variação não têm o mesmo peso, como poderá ser o caso de uma série de percentagens, uma de votos, outra de distribuição da população, outras ainda da ocorrência de vários fenómenos como a natalidade, a emigração, etc., o caso é diferente. E torna-se necessário ponderar previamente as variáveis (ver *Comparação de variáveis e Mudanças de Escala*).

Resolvido o problema da comparabilidade importa considerar as medidas de semelhança que se podem utilizar.

O espaço das variáveis de intervalo e razão é sempre um espaço métrico e o das variáveis ordinais, por extensão analógica, é-o às vezes. Nos espaços métricos é sempre possível definir relações de semelhança e distâncias.

As distâncias que se calculam variam entre zero (no caso da similitude completa entre duas unidades de análise) e um número que não tem limite superior máximo, que será tanto maior quanto mais distintas foram as unidades de análise.

O sentido dos valores numéricos das medidas de distância é inverso do das medidas de semelhança ou de proximidade. Nestas, o valor mínimo é o da maior diferenciação e o máximo, o da maior identificação. Quando existem simultaneamente medidas de distância e semelhança, convém estar atento a este facto, ao interpretar os resultados.

O grupo mais utilizado de distâncias é o das distâncias métricas. Uma métrica é uma relação que se estabelece entre um conjunto de três unidades de análise (**a,b,c**), satisfazendo as seguintes propriedades:

elemento neutro	$D(a,a) = 0,$	(1)

não-negatividade	$D(a,b) \geq 0$	(2)

reflexividade	$D(a,b) = D(b,a)$	(3)

transitividade	$D(a,c) \leq D(a,b) + D(b,c)$	(4)

Quando apenas se verificam as três primeiras, estamos perante uma *semimétrica* e temos uma *ultramétrica* quando a quarta condição é substituído pela seguinte expressão:

$$D(a,c) \leq \max (D(a,b), D(b,c)) \quad (5).$$

2.2.1.- Métricas de Minkowski

As métricas de MINKOWSKI são as que obedecem à seguinte fórmula geral:

$$D^{r_{(k,l)}} = \left(\sum_j^n |a_{k,j} - a_{l,j}|^r \right)^{1/r}$$

Pode definir-se um número infinito de métricas de Minkowski, fazendo variar o parâmetro **r** na fórmula. As distâncias de Minkowski são também sensíveis ao número de variáveis, **j**, pois sendo uma soma de módulos, tenderão a ser maiores quando existem mais parcelas a somar, razão por que, para fazer com que a distância seja independente do número de variáveis, alguns autores apresentaram o coeficiente de distância de Minkowski, distância de Minkowski média ou distância de Minkowski ponderada, da seguinte forma geral:

$$D^{r_{(k,l)}} = \frac{1}{n} \left(\sum_j^n |a_{k,j} - a_{l,j}|^r \right)^{1/r}$$

Nota: O uso desta distância ponderada é muito conveniente se existem valores omissos, pois quando não se conhecem **a_{kj}**, **a_{lj}** ou ambos, a diferença não pode entrar no somatório e o **n** não incrementa e, deste modo, as distâncias obtidas para valores de **n** diversos são comparáveis, mesmo quando respeitam a um número diferente de pares de valores.

As distâncias derivadas de Minkowski que mais se utilizam são as de **r=1** e **r=2**, conhecidas respectivamente por distância de Manhattan e distância euclidiana.

2.2.1.1.- Distância de Manhattan

Também conhecida por distância dos quarteirões, distância *city-block* ou simplesmente *Block*. Sob a forma ponderada é igual a:

$$D_{(k,l)} = \frac{1}{n} \sum_j^n |a_{k,j} - a_{l,j}|$$

Do ponto de vista formal é a média das diferenças entre os valores das coordenadas (variáveis) das duas unidades de análise entre as quais se está a medir a distância.

É muito fácil de interpretar. Tem a vantagem de ser insensível às translações dos eixos (soma ou diminuição de constantes às variáveis) e o inconveniente de ser sensível à rotação do referencial cartesiano de referência. Apresenta uma série de diferentes caminhos possíveis em

estruturas rectangulares, todos com igual distância. As pequenas e as grandes diferenças são tratadas de igual modo

2.2.1.2.- Distância Euclidiana

É a medida de distância que se utiliza quando se refere a “distância” na linguagem corrente. Tem a seguinte forma, na versão ponderada:

$$D^2_{(k,l)} = \frac{1}{n} \left(\sum_j^n |a_{k,j} - a_{l,j}|^2 \right)^{1/2}$$

É invariante, tanto às translações (soma e diminuição de constantes aos valores das variáveis), como às rotações (multiplicação das duas variáveis por um valor entre **-1** e **+1**) dos eixos de coordenadas do referencial cartesiano.

É também muito fácil de interpretar: é a mais curta distância entre dois pontos num espaço métrico.

Ao contrário da distância de Manhattan, aqui uma grande diferença numa das variáveis tem mais importância do que muitas pequenas diferenças. Por exemplo, para a distância de Manhattan, 10 casos de diferenças de 10 unidades equivalem a uma diferença única de 100 unidades, enquanto para a euclidiana, esta última é 10 vezes mais importante, pois se:

$$10 \cdot 10 = 1 \cdot 100 \quad \text{temos que } 10 \cdot (10^2) = 1.000, \text{ mas } 100^2 = 10.000.$$

2.2.1.3.- Coeficiente de Chebychev

É a diferença absoluta máxima existente nos valores da variável.

$$D^C_{(k,l)} = \text{Max} (|a_{k,j} - a_{l,j}|)$$

Como se poderá deduzir da expressão, com o aumento do expoente, as distâncias de Minkowski vão privilegiando cada vez mais a importância das maiores diferenças. Demonstra-se que esta distância de Chebychev é o limite para que tende a distância de Minkowski, quando o r cresce, tendendo para infinito.

$$D^C_{(k,l)} = \lim_{r \rightarrow \infty} \left(\frac{1}{n} \left(\sum_j^n |a_{k,j} - a_{l,j}|^r \right)^{1/r} \right)$$

2.2.2.- Coeficiente de Correlação de Pearson

Esta medida é utilizada tanto como medida de semelhança entre variáveis, como entre unidades de análise. No primeiro caso é insensível às unidades de medida e às escalas de variação utilizadas, mas no segundo é necessário que todas as variáveis sejam comparáveis entre si, ou que sejam ponderadas previamente (ver *Comparação de Variáveis e Mudanças de Escala*).

$$r_{(k,l)} = \frac{\sum_{j=1}^n [(a_{k,j} - \bar{a}_k) \times (a_{l,j} - \bar{a}_l)]}{\sqrt{\sum_{j=1}^n (a_{k,j} - \bar{a}_k)^2 \times \sum_{j=1}^n (a_{l,j} - \bar{a}_l)^2}}$$

Como se sabe, varia entre **-1** e **+1**. Quando é utilizada neste contexto, comparando linhas, não tem sentido estatístico. É tão somente uma medida que “mede” a correspondência entre o perfil (ou forma) dos valores de todas as variáveis numa unidade de análise com o mesmo perfil das variáveis da outra unidade de análise. Assim sendo, o valor de **+1** é o da maior semelhança nos padrões de variação e o **-1** o mais diferenciado. Deve-se ainda tomar em conta que as médias de **a_k** e **a_l** são a média dos valores segundo as linhas (calculadas para cada unidade de análise, com os valores de todas as variáveis).

Neste contexto não é uma medida de distância, mas apenas um índice de semelhança parcial.

Esta medida não é uma métrica, podendo ocorrer que nem sempre se verifique a relação de transitividade:

$$D(a,c) \leq D(a,b) + D(b,c) \quad (4)$$

E ignora as diferenças proporcionais e aditivas nos valores das variáveis. O que às vezes é uma vantagem, pois torna-se invariante face ao efeito de escala.

2.2.3.- Cosseno

Mede o cosseno do ângulo formado pelos vectores que representam as duas unidades de análise, assumindo que os valores nas variáveis são coordenadas num espaço multidimensional.

$$\cos \theta_{(k,l)} = \frac{\sum_{j=1}^n (a_{k,j} \times a_{l,j})}{\sqrt{\sum_{j=1}^n a_{k,j}^2 \times \sum_{j=1}^n a_{l,j}^2}}$$

É uma medida que também compara a forma ou perfil dos valores nas variáveis de duas unidades de análise. Varia também entre **-1** e **+1**. O valor de **0** indica que os vectores são ortogonais.

Assemelha-se, no cálculo e na interpretação, ao Coeficiente de Correlação de Pearson¹. Se as linhas estiverem normalizadas (com média zero e variância unitária), o valor calculado para o cosseno é igual ao do coeficiente de correlação (BENZÉCRI *et al* (1973) e MATHER (1976)).

2.2.4.- Distância de χ^2 (Qui quadrado)

Esta é uma medida de semelhança, especialmente aplicável a matrizes de medição conjunta. Nem sempre se encontra disponível para dados de intervalo ou razão. Como necessita do valor dos somatórios marginais de colunas e linhas, é necessário que os valores dessas somas na matriz tenham verdadeiro significado (quer devido à natureza das variáveis usadas, quer aos esforços de ponderação aplicados).

$$d\chi^2_{(k,l)} = \sum_{j=1} \left[\frac{a_{..}}{a_{.j}} \times \left(\frac{a_{kj}}{a_{k.}} - \frac{a_{lj}}{a_{l.}} \right)^2 \right]$$

em que **a_{..}** é o total geral da matriz, **a_{.j}** a soma da coluna **j**, **a_{k.}** a soma da linha **k** e **a_{l.}** a da linha **l**.

Esta distância compara o quadrado da diferença nas importâncias relativas (face ao total da unidade de análise) dos valores de duas variáveis, ponderadas pelo inverso da importância da variável no total do quadro, o que vem a corresponder a torná-las equivalentes.

2.2.5.- Exemplos de cálculo

Dada a seguinte matriz de três unidades de análise descritas por

	V1	V2	V3	V4	
C1	11	14	107	75	207
C2	1	10	60	31	102
C3	23	2	16	40	81
	35	26	183	146	390

quatro variáveis, vai exemplificar-se o cálculo das diversas distâncias:

¹ - Que os franceses denominam de Coeficiente de Correlação de Bravais-Pearson.

1.- Distância de Manhattann (ponderada)

$$D_{12} = 1/4(|11-1|+|14-10|+|107-60|+|75-31|) = 1/4(10 + 4 + 47 + 44) = 105/4 = 26,25$$

$$D_{13} = 1/4(|11-23|+|14-2|+|107-16|+|75-40|) = 1/4(12 + 12 + 91 + 35) = 150/4 = 37,50$$

$$D_{23} = 83 / 4 = 20,75$$

	C1	C2	C3
C1	0,00	26,25	37,50
C2	26,25	0,00	20,75
C3	37,50	20,75	0,00

2.- Distância Euclidiana (ponderada)

$$D_{12} = \frac{1}{4} \times \sqrt{(11-1)^2 + (14-10)^2 + (107-60)^2 + (75-31)^2} = \frac{1}{4} \times \sqrt{10^2 + 4^2 + 47^2 + 44^2} =$$

$$D_{12} = \frac{1}{4} \times \sqrt{100 + 16 + 2209 + 1936} = \frac{1}{4} \times \sqrt{4261} = \frac{65,276...}{4} = 16,319...$$

$$D_{13} = 98,965.../4 = 24,741...$$

$$D_{23} = 50,646.../4 = 12,661...$$

	C1	C2	C3
C1	0,000	16,319	24,741
C2	16,319	0,000	12,661
C3	24,741	12,661	0,000

3.- Distância Chebychev

$$D_{12} = \text{Max}(|11-1|, |14-10|, |107-60|, |75-31|) = \text{Max}(10, 4, 47, 44) = 47$$

$$D_{13} = \text{Max}(12, 12, 91, 35) = 91$$

$$D_{23} = \text{Max}(22, 8, 44, 9) = 44$$

	C1	C2	C3
C1	0	47	91
C2	47	0	44
C3	91	44	0

4- Coeficiente de Correlação de Pearson

$$D_{12} = \frac{(-40,75 \times -24,5) + (-37,75 \times -15,5) + (55,25 \times 34,5) + (23,25 \times 5,5)}{\sqrt{(-40,75^2 - 37,75^2 + 55,25^2 + 23,25^2) \times (-24,5^2 - 15,5^2 + 34,5^2 + 5,5^2)}} = \frac{3617,5}{\sqrt{(6678,75 \times 2061)}} = \frac{3617,5}{3710,108} = 0,975039$$

$$D_{12} = \frac{[(11 - 51,75) \times (1 - 25,5) + (14 - 51,75) \times (10 - 25,5) + (107 - 51,75) \times (60 - 25,5) + (75 - 51,75) \times (31 - 25,5)]}{\sqrt{[(11 - 51,75)^2 + (14 - 51,75)^2 + (107 - 51,75)^2 + (75 - 51,75)^2] \times [(1 - 25,5)^2 + (10 - 25,5)^2 + (60 - 25,5)^2 + (31 - 25,5)^2]}} =$$

$$D_{13} = 0,358304...$$

$$D_{23} = 0,142886...$$

	C1	C2	C3
C1	0,0000	0,9750	0,3583
C2	0,9750	0,0000	0,1429
C3	0,3538	0,1429	0,0000

5- Cosseno

$$D_{12} = \frac{[11 \times 1 + 14 \times 10 + 107 \times 60 + 75 \times 31]}{\sqrt{[(11)^2 + (14)^2 + (107)^2 + (75)^2] \times [(1)^2 + (10)^2 + (60)^2 + (31)^2]}} =$$

$$= \frac{8896}{\sqrt{(17391) \times (4662)}} = \frac{8896}{9004,2678} = 0,9880 ...$$

$$D_{13} = 0,7746...$$

$$D_{23} = 0,6721...$$

	C1	C2	C3
C1	0,0000	0,9880	0,7746
C2	0,9880	0,0000	0,6721
C3	0,7746	0,6721	0,0000

6- Distância de Qui-quadrado

$$D_{12} = \frac{390}{35} \times \left(\frac{11}{207} - \frac{1}{102} \right)^2 + \frac{390}{26} \times \left(\frac{14}{207} - \frac{10}{102} \right)^2 + \frac{390}{183} \times \left(\frac{107}{207} - \frac{60}{102} \right)^2 + \frac{390}{146} \times \left(\frac{75}{207} - \frac{31}{102} \right)^2$$

$$D_{12} = 0,0209262 + 0,0138675 + 0,0108422 + 0,0091094 = 0,0547453...$$

$$D_{13} = 0,884857...$$

$$D_{23} = 1,339811...$$

	C1	C2	C3
C1	0,0000	0,0547	0,8849
C2	0,0547	0,0000	1,3398
C3	0,8849	1,3398	0,0000

2.3.- Para dados nominais, ordinais, de intervalo e razão

São situações complexas, que não são fáceis de tratar e que exigem procedimentos especiais.

Parker (1969), citado por MATHER, sugeriu que se utilizassem todas as variáveis transformadas para o intervalo entre zero e um. As variáveis nominais teriam dois valores para cada estado: **0** seria a ausência desse estado e **1** a presença. As variáveis ordinais seriam divididas em classe equidistantes, como por exemplo em **0** (ausente), **0,5** (escasso) e **1** (abundante) ou em mais classes, sempre de modo proporcional entre **0** e **1**. As variáveis de intervalo/razão seriam codificadas no intervalo de 0 a 1 segundo a expressão:

$$t_i = \frac{(a_i - \min(a_i))}{(\text{Max}(a_i) - \min(a_i))}$$

Feitas estas transformações, utiliza-se como medida de semelhança a distância euclidiana calculada entre as unidades de análise.

Gower, em 1971, propôs uma forma semelhante para resolver o mesmo problema de comparar dados medidos em escalas de medida de diversa natureza, apresentando um coeficiente de semelhança, **S_g**, que compara duas unidades de análise através dos valores que têm em todas as variáveis.

O coeficiente não considera as variáveis em que existem valores omissos em uma ou nas duas unidades de análise e pondera o resultado final obtido, achando a média ponderada dos três tipos diferentes de medidas de semelhança que calcula.

Para o conjunto de todas as **p** variáveis de tipo dicotômico computa-se o respectivo Coeficiente de Jaccard (**S_j**).

Para as **m** variáveis nominais com mais de dois estados, calcula-se um índice de semelhança **Se**, tal que:

$$Se_{k,l} = \sum_{i=1}^m s_{i,k,l} / m$$

onde **S_{ikl}** assume o valor **1** se as variáveis estão no mesmo estado e o valor **0** quando o não estão.

Para o conjunto das variáveis ordinais, de intervalo e razão (**n**), computa-se **S_i**, de acordo com a seguinte expressão:

$$Si_{k,l} = \sum_{i=1}^n (1 - (|a_{k,i} - a_{l,i}| / (\text{Amp}(a_i)))) / n$$

O Coeficiente de Gower **S_g** é a média ponderada de **S_j**, **S_e** e **S_i**, conforme:

$$S_{g_{k,l}} = \frac{p \times S_j + m \times S_e + n \times S_i}{(p + m + n)}$$

3.- ESTRATÉGIAS DE AGRUPAMENTO

3.1.- Para classificações não-hierárquicas (partições)

Existem muitos métodos. De seguida iremos referir-nos sucintamente a alguns dos mais usuais.

3.1.1.- Nuvens Dinâmicas

Este método deve-se a DIDAY. Esquemáticamente consiste no seguinte:

- 1.- Parte-se da matriz de atributos **A**, que contém os valores das variáveis nas unidades de análise;
 - 2.- Normaliza-se **A**, por colunas, diminuindo da média e dividindo pelo desvio-padrão, de modo a tornar comparáveis os valores segundo as linhas;
 - 3.- Escolhe-se os centros das **k** “nuvens”, tantos centros quantos os grupos a formar, por inspecção da matriz (método teórico) ou de modo aleatório (método empírico). Este último pode provocar alguns problemas: é preciso garantir que a sequência de procedimentos não conduza para um máximo local e que os pontos escolhidos como centros se encontrem no interior da nuvem, o que é fácil de fazer a duas dimensões, mas muito mais difícil a **n** dimensões (um dos melhores meios é usar pontos já existentes);
 - 4.- Constrói-se uma partição inicial em **k** grupos, afectando cada unidade de análise ao centro que lhe estiver mais próximo, segundo a medida de distância que se tiver escolhido e, que no caso do Diday, é a distância euclidiana;
 - 5.- Calculam-se de novo os centros das **k** “nuvens”, agora como resultado das unidades de análise que correspondem a cada uma;
 - 6.- Constrói-se nova partição em **k** grupos, calculando outra vez as distâncias das unidades de análise aos novos centros das “nuvens” e afectando-as aos centros mais próximos;
- e assim sucessivamente, até que se encontrem duas partições sucessivas iguais, em que os indivíduos afectos às “nuvens” e as

respectivas distâncias são as mesmas, ou seja, em que a solução não se altere.

3.1.2.- K - médias

Este método minimiza as variâncias inter grupos e maximiza as variâncias entre os grupos.

3.1.3.- K - centróides

Minimiza a distância entre os grupos.

3.2.- Para classificações hierárquicas ascendentes

Depois de escolher uma medida de distância, constrói-se uma matriz de distâncias entre as unidades de análise.

Deve ter-se em conta se as variáveis necessitam ou não de ser padronizadas.

Como estamos numa metodologia de agrupamento hierárquico ascendente, o primeiro passo consiste em determinar quais as duas unidades mais próximas e constituir com elas um grupo.

Depois refaz-se a matriz e assim sucessivamente. O problema está em como se mede a distância de entre um elemento e um grupo, ou entre dois grupos.

3.2.1.- Método do vizinho mais próximo

É também denominado de ligação simples e método mínimo (em inglês, *nearest neighbor*, *single linkage*, e *minimum method*).

Define-se como segue: dadas duas classes **a** e **b**, a distância entre elas é dada pelo valor mínimo de todas as distâncias entre todos os elementos de **a** e todos os de **b**.

$$d_{(a,b)} = \sup \{d_{(i,i')}; i \in a, i' \in b\}$$

A expressão pode ser lida como a distância entre **a** e **b**; é a menor das distâncias entre todo e qualquer elemento de **a** e todo e qualquer elemento de **b**.

Gera grupos bastante heterogêneos, com tendências para formar cadeias. Privilegia o facto de fazer classes ou grupos tão distintos internamente quanto os atributos o permitem.

3.2.2.- Método do vizinho mais distante

Método da ligação completa ou método máximo (em inglês, é o *complete linkage*, o *maximum method*, ou o *furthest neighbor*).

É definido do seguinte modo:

A distância entre duas classes **a** e **b**, é o valor máximo de todas as distâncias entre todos os elementos de **a** e todos os de **b**.

$$d_{(a,b)} = \sup \{d_{(i,i')}; i \in a, i' \in b\}$$

A expressão pode ser lida como a distância entre **a** e **b**; é a maior das distâncias entre todo e qualquer elemento de **a** e todo e qualquer elemento de **b**.

Quando se juntam dois grupos, primeiro calculam-se todas as distâncias deste modo e depois agregam-se os que se encontram à menor distância.

Cria grupos muito compactos, que com dificuldade se agregam a outros. Privilegia o facto de fazer grupos tão homogêneos quanto os atributos o permitem.

Estes dois métodos dão resultados muito contrastados. Foram por isso propostos alguns que procuram a sua conciliação. Em vez de utilizar a menor ou a maior das distâncias de um conjunto, usam-se médias dessas distâncias, o que vem a dar uma solução, de compromisso, entre as duas. Dentre eles destaca-se o Método da média do grupo.

3.2.3.- Método da média do grupo

Em inglês, é o *group average* ou o *between-groups linkage*. Também é muitas vezes designado por UPGMA, abreviatura de *unweighted pair-group method with average*.

A distância entre as duas classes **a** e **b** é a média de todas as distâncias entre todos os elementos de **a** e todos os de **b**.

$$d_{(a,b)} = \frac{\sum (d_{(i,i')}; i \in a, i' \in b)}{(\text{card}(a) \times \text{card}(b))}$$

A expressão significa que a distância entre **a** e **b** é a média das distâncias entre todo e qualquer elemento de **a** e todo e qualquer

elemento de **b** (é a soma das distâncias dividida pelo produto do número de elementos de **a** com o número de elementos de **b**).

Devido à clareza do seu significado, este é o método mais utilizado em taxinomias numéricas.

3.2.4.- Método da média simples

Em inglês é o *simple average*, que se conhece por WPGMA, de *weighted pair-group method with average*.

A distância de um elemento **j** ao grupo formado por dois grupos **a** e **b**, é a média simples das distâncias entre **j** e o grupo **a** e **j** e o grupo **b**.

$$d_{[(a,b),j]} = \frac{d_{(j,a)} + d_{(j,b)}}{2}$$

A expressão significa que a distância entre **j** e os grupos **a** e **b** é a média simples das distâncias a **a** e a **b** (é a soma das distâncias dividida pelo produto do número de elementos de **a** com o número de elementos de **b**). Embora ajude a preservar a individualidade dos pequenos grupos, a sequência de uniões deste tipo leva muitas vezes a grandes enviesamentos em benefício das últimas agregações efectuadas, desvalorizando grandes concentrações que se efectuaram em estágios muito iniciais do processo ascendente de agregação.

3.2.5.- Método do Centróide

É o *centroid clustering* da bibliografia de língua inglesa. É designado por UPGC, derivado de *unweighted pair-group centroid*. A distância entre as duas classes **a** e **b** é a distância entre os centróides de **a** e de **b**. Os centróides são o centro de gravidade das classes, determinadas pelas médias em cada variável. A localização do centróide da nova classe resulta da média dos dois centróides, ponderada pelo número de unidades de análise existentes em cada classe.

Como não é difícil de entender o que significa, é também um método de agregação muito utilizado.

3.2.6.- Método da Mediana

Chama-se, em inglês, *median average linkage*. Designa-se também por WPGC, de *weighted pair-group centroid*.

Neste método a distância entre as duas classes a e b é a distância entre os centróides de a e de b, mas, ao contrário do anterior, no cálculo do centróide do novo grupo não se toma em consideração o número de elementos que existe em cada um, sendo as suas coordenadas a média simples das coordenadas dos centros dos dois grupos que se uniram.

Apresenta uma certa tendência para formar cadeias, embora menos forte do que o método do vizinho mais próximo.

3.2.7.- Método de Ward

Também designado da mínima variância.

Baseia-se no pressuposto de que classes devem ter a máxima homogeneidade interna possível. Ward propôs uma técnica de minimização da soma dos quadrados das distâncias de cada ponto ao respectivo centro de classe.

Em cada nível, o novo grupo constitui-se pela combinação dos dois grupos cuja junção provoca o menor aumento da soma do quadrado das destas distâncias.

No método de Ward usa-se geralmente como distância, a distância euclidiana.

3.2.8.- Método das Distâncias internas

É o método designado em inglês por *within-group linkage* e está disponível em alguns programas de análise, como por exemplo no SPSS.

Faz grupos de tal modo que, para cada nível de agregação, o total das distâncias internas (as entre todos os elementos de cada grupo) seja mínima. Assim, a distância entre dois grupos considera-se que é a média das distâncias entre todos os pares possíveis de constituir no novo grupo que resultaria da combinação daqueles dois.

3.2.9.- A fórmula geral de agregação

Foi proposta já há algum tempo uma fórmula geral de agregação, que simplifica os processos de cálculo automático das estratégias de constituição hierárquica de grupos. Embora actualmente a disponibilidade de programas e de pacotes (*packages*) lhe retire a importância prática que já teve, é importante conhecer que todas estas estratégias de agregação podem ser consideradas como derivadas de uma única fórmula geral:

$$d_{[(j,k),l]} = a_j \times d_{(j,l)} + a_k \times d_{(k,l)} + b \times d_{(j,k)} + c \times |d_{(j,l)} - d_{(k,l)}|$$

em que os valores de a_j , a_k , b e c assumem valores diferentes para cada uma das estratégias de agregação.

Esta fórmula foi proposta em 1967 por dois investigadores, Lance e Williams² e nela os diferentes parâmetros para cada estratégia de agrupamento são os seguintes:

Ligação	a_j	a_k	b	c
Simples	0,5	0,5	0	- 0,5
Completa	0,5	0,5	0	0,5
UPGMA	$n_j/n(j+k)$	$n_k/n(j+k)$	0	0
WPGMA	0,5	0,5	0	0
UPGC	$n_j/n(j+k)$	$n_k/n(j+k)$	$- a_j \times a_k$	0
WPGC	0,5	0,5	- 0,25	0
Ward	$n(m+j)/n(m+j+k)$	$n(m+k)/n(m+j+k)$	$- nm/n(m+j+k)$	0

3.2.10.- Exemplificação da agregação

Para que se compreendam melhor as estratégias de agrupamento, apresentam-se alguns exemplos gráficos da maneira como é medida a distância de um elemento a um grupo (este constituído por dois subconjuntos, um com dois e outro com quatro elementos). Em todas as figuras estão representados os mesmos pontos e as mesmas distâncias. A preto, mais grosso, está indicada a distância associada a cada estratégia de agrupamento.

Na Figura 8.5 A está a distância do vizinho mais próximo (SL), em B a do vizinho mais distante (CL) e em C a da média do grupo (UPGMA).

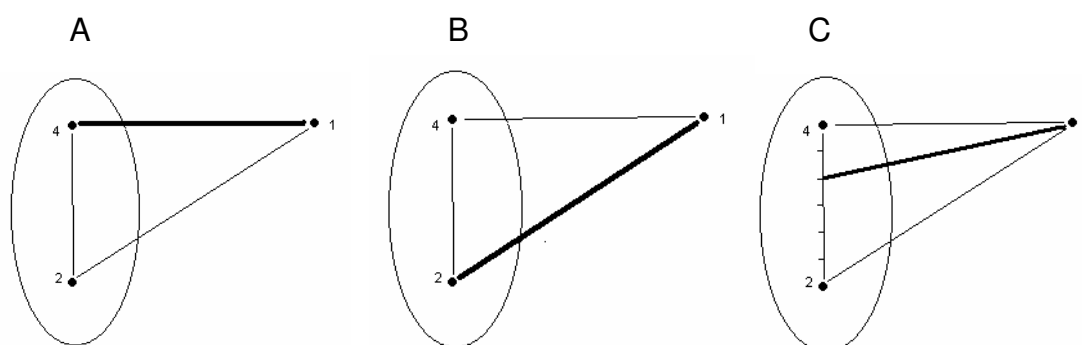


Figura 8.5 - Distâncias entre um elemento e um grupo constituído por dois subconjuntos de 2 e 4 elementos.

² - Num artigo fundamental: Lance, G. N. & Williams, W. T. (1967) – A general theory of classificatory sorting strategies. I: Hierarchical systems. *Computer Journal*, 9, 373-380.

Na Figura 8.4 D está a distância da média simples (WPGMA), em E a do centróide (UPGC) e finalmente em F a da mediana, ou do centróide mediano (WPGC).

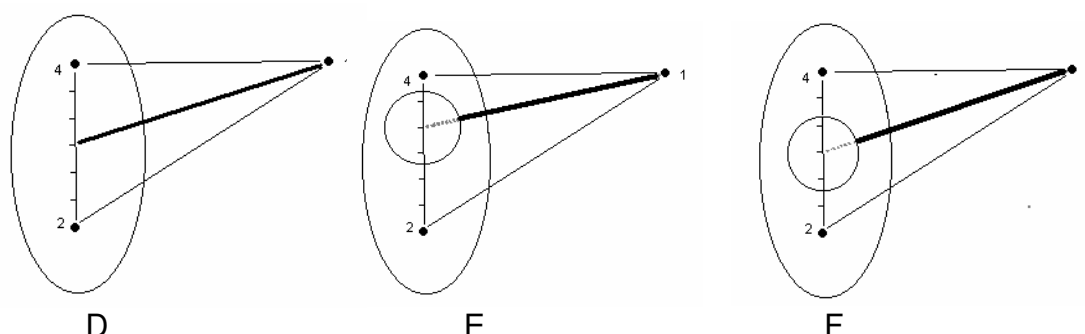


Figura 8.4 (Continuação)- Distâncias entre um elemento e um grupo constituído por dois subconjuntos de 2 e 4 elementos.

Como já dissemos, a mais utilizada é a média do grupo (UPGMA), que é a que deve ser utilizada quando não se tiver evidência de que é necessário ou conveniente utilizar uma das outras.

3.2.11.- Exemplos

Para ilustrar o modo como funcionam alguns destes modos de agregação, considere-se como exemplo a matriz de dados do Quadro 8.1.

NUT's II	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆
Norte	165.38	.45	12.09	8.52	48.23	45.17
Centro	72.41	-.01	9.97	11.39	50.15	46.36
L.V.Tejo	277.33	.30	10.45	10.00	48.62	45.53
Alentejo	19.63	-.80	8.28	13.51	43.94	42.36
Algarve	69.13	.28	10.57	12.71	44.40	41.41
Açores	103.26	.59	15.19	10.95	39.24	33.90
Madeira	329.40	.50	13.00	9.66	42.63	45.61

Quadro 8.1 – Variáveis exemplificativas. Portugal, NUT's II, 1993/95

As variáveis referem-se todas ao ano de 1995, excepto a segunda, que trata da variação entre 1993 e 1995. A variável V1 é a densidade de população, a V2 a variação de população entre 93 e 95, a V3 a taxa de natalidade, a V4 a de mortalidade, a V5 a de actividade e a V7 a percentagem da participação feminina nos activos.

As distâncias entre as sete NUT's II (os sete casos C_i) foram calculadas pela distância euclidiana, após prévia ponderação dos valores, por variável (por coluna), como forma de dar a cada uma o mesmo peso na análise, independentemente das escalas e unidades em que foram medidas e são apresentadas sob a forma de uma matriz de relação quadrada de sete casos.

	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆	C ₇
C ₁		2,3348	1,5167	4,5929	2,9564	4,0289	2,1710
C ₂	2,3348		2,0988	2,9065	2,1162	4,8054	3,5546
C ₃	1,5167	2,0988		4,1479	2,7769	4,5021	2,0213
C ₄	4,5929	2,9065	4,1479		2,5701	5,0729	4,9480
C ₅	2,9564	2,1162	2,7769	2,5701		3,2302	3,2573
C ₆	4,0289	4,8054	4,5021	5,0729	3,2302		3,6433
C ₇	2,1710	3,5546	2,0213	4,9480	3,2573	3,6433	

Cada valor pode ser interpretado como a diferença acumulada, em desvios padrões, nas seis variáveis dos valores de cada par de casos ou de unidades de análise.

1).- Vizinho mais próximo

O valor mais baixo da matriz acima é o de 1,5167, entre as unidades de análise 1 e 3, que se juntam, calculando-se depois os valores deste grupo às outras unidades de análise, obtendo-se:

	C _{1.3}	C ₂	C ₄	C ₅	C ₆	C ₇
C _{1.3}		2,0988	4,1479	2,7769	4,0289	2,0213
C ₂	2,0988		2,9065	2,1162	4,8054	3,5546
C ₄	4,1479	2,9065		2,5701	5,0729	4,9480
C ₅	2,7769	2,1162	2,5701		3,2302	3,2573
C ₆	4,0289	4,8054	5,0729	3,2302		3,6433
C ₇	2,0213	3,5546	4,9480	3,2573	3,6433	

A menor distância é agora 2,0213 (entre 1/3 e 7), que se unem, dando origem a:

	C _{1.37}	C ₂	C ₄	C ₅	C ₆
C _{1.37}		2,0988	4,1479	2,7769	3,6433
C ₂	2,0988		2,9065	2,1162	4,8054
C ₄	4,1479	2,9065		2,5701	5,0729
C ₅	2,7769	2,1162	2,5701		3,2302
C ₆	3,6433	4,8054	5,0729	3,2302	

Agora une-se o grupo de 1/3/7 com o elemento 2, à distância de 2,0988 e repetindo o processo,

	C _{12.37}	C ₄	C ₅	C ₆
C _{1.237}		2,9065	2,1162	3,6433
C ₄	2,9065		2,5701	5,0729
C ₅	2,1162	2,5701		3,2302
C ₆	3,6433	5,0729	3,2302	

verifica-se agora que a menor distância é a do grupo 1/2/3/7 com o elemento 5, no valor de 2,1162.

Repetindo o processo, obtém-se a união com o elemento 4 e depois com o 6.

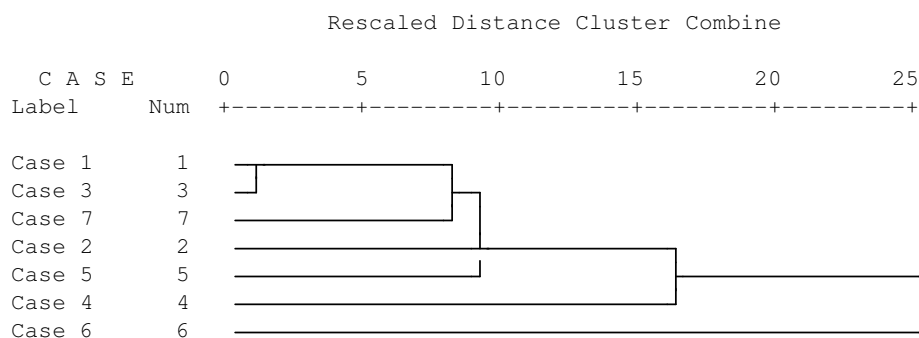
	C_{12.357}	C₄	C₆
C_{1.2357}		2,5701	3,2302
C₄	2,5701		5,0729
C₆	3,2302	5,0729	

	C_{12.357}	C₆
C_{1.2357}		3,2302
C₆	3,2302	

Recapitulando, fizeram-se as seguintes junções, que permitiram fazer a respectiva árvore de ligação:

Novo grupo	Grupo anterior 1	Grupo anterior 2	Distância
8	1	3	1,516653
9	8	7	2,021251
10	9	2	2,098766
11	10	5	2,11620
12	11	4	2,5700
13	12	6	3,23023

Dendrogram using Single Linkage



2).- Vizinho mais distante

De modo similar, vê-se que o valor mais baixo da matriz inicial de distâncias é o de 1,5167, entre as unidades de análise 1 e 3, mas agora o cálculo das distâncias entre este grupo 1/3 e os outros elementos é diferente:

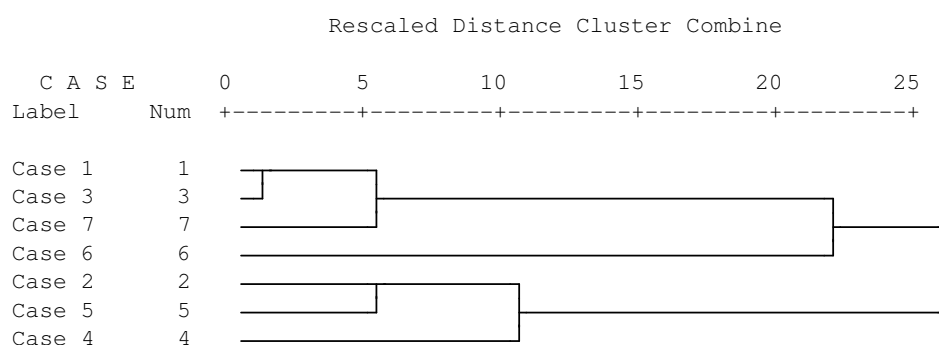
	C_{1.3}	C₂	C₄	C₅	C₆	C₇
C_{1.3}		2,3348	4,5929	2,9564	4,5021	2,1710
C₂	2,3348		2,9065	2,1162	4,8054	3,5546
C₄	4,5929	2,9065		2,5701	5,0729	4,9480
C₅	2,9546	2,1162	2,5701		3,2302	3,2573
C₆	4,5021	4,8054	5,0729	3,2302		3,6433
C₇	2,1710	3,5546	4,9480	3,2573	3,6433	

A união seguinte é a do grupo 1/3 com 2 (à distância de 2,1162). E, repetindo o processo, obtêm-se as seguintes junções:

Novo grupo	Grupo anterior 1	Grupo anterior 2	Distância
8	1	3	1,5166
9	2	5	2,1162
10	8	7	2,1710
11	9	4	2,9065
12	10	8	4,5021
13	12	11	5,0729

e correspondente árvore de ligação:

Dendrogram using Complete Linkage

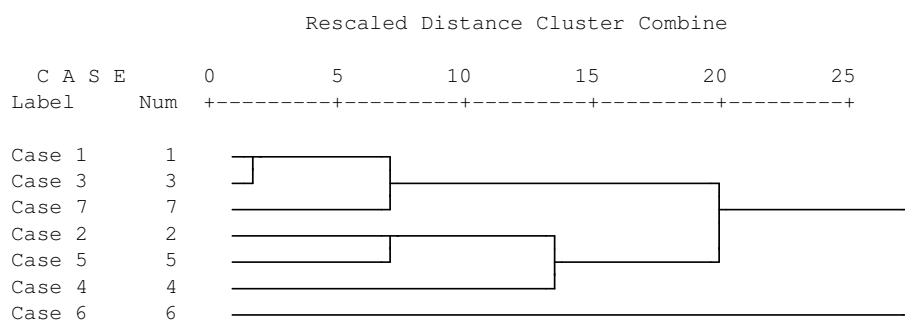


3).- UPGMA

Com esta estratégia de agrupamento é o seguinte o esquema de junção e a árvore de ligação:

Novo grupo	Grupo anterior 1	Grupo anterior 2	Distância
8	1	3	1,5167
9	8	7	2,0961
10	2	5	2,1162
11	10	4	2,7383
12	11	9	3,4075
13	6	12	4,2138

Dendrogram using Average Linkage (Between Groups)



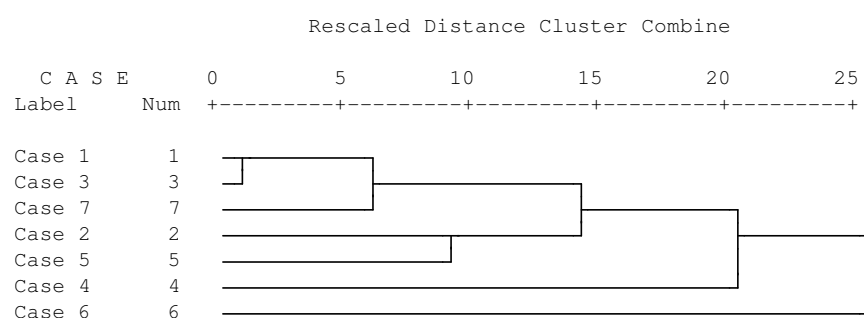
4).- WPGMA

O esquema de junção desta metodologia é:

Novo grupo	Grupo anterior 1	Grupo anterior 2	Distância
8	1	3	1,5167
9	8	7	2,0961
10	2	5	2,1162
11	10	4	2,7383
12	11	9	3,8166
13	6	12	4,2879

e a árvore de ligação

Dendrogram using Average Linkage (Within Group)

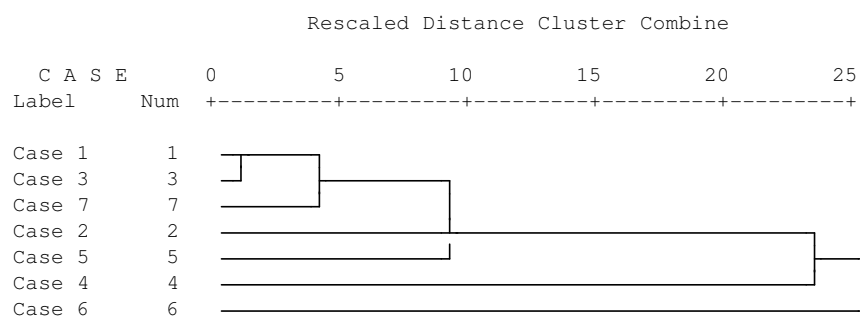


5).- UPGC

O esquema de junção desta metodologia e a árvore de ligação são:

Novo grupo	Grupo anterior 1	Grupo anterior 2	Distância
8	1	3	1,5167
9	8	7	1,7169
10	9	2	2,02842
11	10	5	1,92063
12	11	4	2,84092
13	6	12	2,99242

Dendrogram using Centroid Method

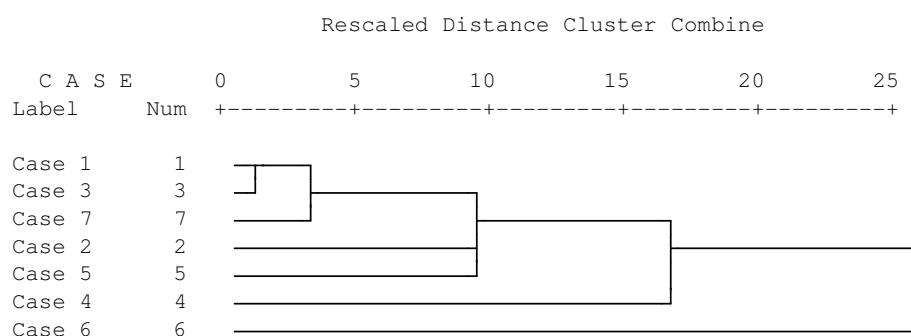


6).- WPGC

O esquema de junção desta metodologia e a árvore de ligação são:

Novo grupo	Grupo anterior 1	Grupo anterior 2	Distância
8	1	3	1,5167
9	8	7	1,7169
10	2	5	2,11623
11	9	10	1,825963
12	11	4	2,668308
13	12	6	3,347181

Dendrogram using Median Method

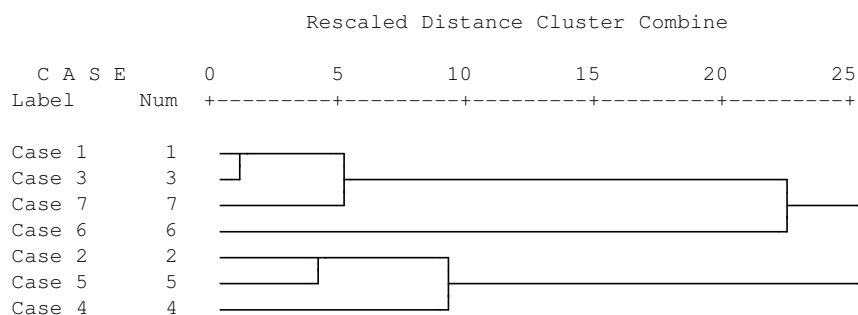


7).- Método de Ward

O esquema de junção desta metodologia e a árvore de ligação são:

Novo grupo	Grupo anterior 1	Grupo anterior 2	Distância
8	1	3	0,758326
9	2	5	1,816442
10	8	7	2,961088
11	9	4	4,433907
12	10	6	7,001731
13	12	11	9,893149

Dendrogram using Ward Method



4.- AVALIAÇÃO DO AJUSTAMENTO

Os dois principais problemas dos métodos de classificação automática são a escolha da distância a utilizar e da estratégia de agrupamento. Tanto para um como para o outro, o fundamental é a adequação teórica entre o problema que se está a estudar no mundo real e as suas características e as metodologias que se utilizam.

Porém, para a estratégia de agregação é possível utilizar um método empírico auxiliar, o da correlação cofenética.

Para tal é necessário construir uma matriz de valores cofenéticos, que é uma matriz em que constam os valores das distâncias que as unidades de análise apresentam na árvore de ligação. O ajustamento entre os valores desta matriz e o da matriz de distâncias inicial (também entre todas as unidades de análise) mede a qualidade da classificação efectuada.

Podem comparar-se os valores das duas matrizes através de um diagrama de dispersão, geralmente com as distâncias da matriz original em abcissas e as da matriz cofenética nas ordenadas, que se chama diagrama cofenético.

Também é habitual calcular o coeficiente de correlação linear de Pearson entre os valores da matriz inicial e os da matriz cofenética, obtendo-se o Coeficiente de Correlação Cofenética.

Exemplificando com o caso anterior, as distâncias calculadas entre os elementos são:

	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆	C ₇
C ₁		2,3348	1,5167	4,5929	2,9564	4,0289	2,1710
C ₂	2,3348		2,0988	2,9065	2,1162	4,8054	3,5546
C ₃	1,5167	2,0988		4,1479	2,7769	4,5021	2,0213
C ₄	4,5929	2,9065	4,1479		2,5701	5,0729	4,9480
C ₅	2,9564	2,1162	2,7769	2,5701		3,2302	3,2573
C ₆	4,0289	4,8054	4,5021	5,0729	3,2302		3,6433
C ₇	2,1710	3,5546	2,0213	4,9480	3,2573	3,6433	

e as matrizes de distâncias nas diversas árvores de ligação, para os métodos do vizinho mais próximo, vizinho mais distante, UPGMA e WPGMA são, sucessivamente, as que se seguem.

1) Vizinho mais próximo

	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆	C ₇
C ₁		2,0988	1,5167	2,5700	2,1162	3,2302	2,0213
C ₂	2,0988		2,0988	2,5700	2,1162	3,2302	2,0988
C ₃	1,5167	2,0988		2,5700	2,1162	3,2302	2,0213
C ₄	2,5700	2,5700	2,5700		2,5700	3,2302	2,5700
C ₅	2,1162	2,1162	2,1162	2,5700		3,2302	2,1162
C ₆	3,2302	3,2302	3,2302	3,2302	3,2302		3,2302
C ₇	2,0213	2,0988	2,0213	2,5700	2,1162	3,2302	

2) Vizinho mais distante

	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆	C ₇
C ₁		5,0729	1,5167	5,0729	5,0729	4,5021	2,1710
C ₂	5,0729		5,0729	2,9065	2,1162	5,0729	5,0729
C ₃	1,5167	5,0729		5,0729	5,0729	4,5021	2,1710
C ₄	5,0729	2,9065	5,0729		2,9065	5,0729	5,0729
C ₅	5,0729	2,1162	5,0729	2,9065		5,0729	5,0729
C ₆	4,5021	5,0729	4,5021	5,0729	5,0729		4,5021
C ₇	2,1710	5,0729	2,1710	5,0729	5,0729	4,5021	

3) UPGMA

	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆	C ₇
C ₁		3,4075	1,5167	3,4075	3,4075	4,2138	2,0961
C ₂	3,4075		3,4075	2,7383	2,1162	4,2138	3,4075
C ₃	1,5167	3,4075		3,4075	3,4075	4,2138	2,0961
C ₄	3,4075	2,7383	3,4075		2,7383	4,2138	3,4075
C ₅	3,4075	2,1162	3,4075	2,7383		4,2138	3,4075
C ₆	4,2138	4,2138	4,2138	4,2138	4,2138		4,2138
C ₇	2,0961	3,4075	2,0961	3,4075	3,4075	4,2138	

etc..

6. - BIBLIOGRAFIA

- BENZÉCRI, J.-P. et collaborateurs (1973) – ***L'Analyse des Donnees. 1 – La Taxinomie***, Dunod, Paris.
- (*) BROCHADO, Ana Margarida (2002) – “**A análise de clusters: técnica de classificação na análise espacial**” in COSTA, José da Silva (Coord.), ***Compêndio de Economia Regional***, Coimbra, A.P.D.R., pp 743 - 766.
- MATHER, Paul (1976) – ***Computational Methods of Multivariate Analysis in Physical Geography***, John Wiley & Sons, London, p. 309-329, 342-351.
- (**) REIS, Elizabeth (1997) – ***Estatística Multivariada Aplicada***, Lisboa, Edições Sílabo, pp 285 – 336.
- (*) PESTANA, Maria Helena; GAGEIRO, João Nunes (1998) – ***Análise de Dados para Ciências Sociais. A complementaridade do SPSS***, Lisboa, Edições Sílabo, pp 345 – 354.

Anexo 9

Análise Discriminante

1.- INTRODUÇÃO

A análise discriminante é uma técnica de análise de dados que tem algumas semelhanças com as taxionomias numéricas. Com efeito, enquanto nas taxionomias numéricas se parte de uma matriz de atributos a várias variáveis para obter os grupos mais homogêneos que ali se podem encontrar, isto é, grupos em que os valores das variáveis correspondem a uma regra de semelhança, nas análises discriminantes conhecem-se os grupos e o que se pretende é construir um modelo predictivo, baseado nos valores das variáveis que nos permita distinguir entre os diversos grupos.

2.-O DECURSO DA ANÁLISE

O objectivo é, pois, construir uma função discriminante para dois ou mais grupos, baseada na combinação linear das variáveis observadas, que permita a melhor discriminação ou distinção das unidades de análise nos grupos iniciais.

Os grupos iniciais são definidos pela situação de investigação específica; por exemplo, ao analisar uma série de concelhos sabemos *a priori* que uns são rurais e outros são urbanos, estudando sedes de concelho sabemos que umas são cidades e outras não, ou ainda num estudo sobre empresas sabemos que umas faliram e outras não.

Neste tipo de análises, o que interessa ao investigador não é se uma dada unidade de análise pertence ou não a um dado grupo, pois isso é um resultado da mera observação, mas sim as razões por que pertence, expressas pelas variáveis e tendo em conta as relações médias que a matriz de atributos expressa.

Um exemplo clássico em ciência política e geografia eleitoral é a procura das razões (o mesmo é dizer, dos valores das variáveis) que “explicam” por que as pessoas votam em determinados partidos. Ou se alguma característica locativa de países ou regiões a que variáveis está associada e com que intensidade.

2.1.- As funções discriminantes

Matematicamente o problema da análise discriminante consiste em encontrar uma função linear das variáveis, cujo valor calculado permite atribuir, com um certo grau de probabilidade, a que grupo pertence cada unidade de análise e, portanto, numa perspectiva de estatística multidimensional indutiva, prever a que grupo pertencerão os casos novos, desde que se conheça o valor que neles assumem as variáveis independentes.

Quando existem muitos grupos (**K**) é possível definir várias funções discriminantes (**K-1**), cuja combinação é que atribui as unidades de análise aos grupos, mas geralmente as funções discriminantes são em muito menor número.

A função discriminante assume então a seguinte forma geral:

$$D_i = d_{i1} \times V_1 + d_{i2} \times V_2 + d_{i3} \times V_3 + \dots + d_{ip} \times V_p$$

em que **D_i** é o valor da função discriminante **i**, os **d** são os coeficientes ponderados, os **V** os valores das **p** variáveis utilizadas na análise.

Quando se utilizam procedimentos passo a passo, as participações de cada variável são analisadas e são eliminadas as que não satisfazem aos critérios de selecção (geralmente associados à melhoria da capacidade de explicação da função) e as funções discriminantes vão sendo construídas

A ordem de entrada das variáveis assim como o seu respectivo contributo e a identificação das que não têm poder discriminatório são o primeiro passo da análise.

De seguida analisam-se as funções discriminantes, quantas são seleccionadas (através da avaliação do o seu significado) e qual o valor e percentagem explicativa de cada uma (pelo valor do respectivo valor próprio).

2.2.- As coordenadas e os centróides dos grupos

Pode-se, então, calcular os valores das funções para cada ponto, assim como os centróides dos grupos, que se poderão *plotar*. Quanto mais distintos estiverem os centróides mais discriminantes as funções e quanto mais próximos menos discriminantes.

Por último e ainda a partir dos valores dos coeficientes das funções discriminantes, pode atribuir-se um grupo a uma nova unidade de análise de que se conheçam os valores das variáveis independentes. A unidade de análise pertencerá ao grupo que corresponder ao maior valor.

Obviamente, se houverem dois valores muito aproximados, isso indica que não existe a certeza do grupo a que a unidade de análise pertence e que esta se encontra na região da fronteira multidimensional entre os grupos

3.- EXEMPLO

Seja dado o exemplo seguinte de nove clientes de um banco que foram agrupados pelo tipo de risco que apresentaram para a instituição (Reis, 1997:246). As variáveis a ter em conta são:

a variável de agrupamento RISCO

- 0 – sem problemas nos pagamentos
- 1 - com algumas dificuldades
- 2 - não pagaram a tempo

e as quatro variáveis independentes:

CREDIT – montante de crédito concedido em milhares

PROF - anos de experiência profissional

IDADE - número de anos

SALAR - salário mensal

RISCO	CREDIT	PROF	IDADE	SALAR
0	17,5	20	45	400
0	15,0	15	38	340
0	13,0	13	36	230
1	18,0	16	37	200
1	20,0	25	46	300
1	14,0	30	54	160
2	10,0	9	29	150
2	9,0	18	37	180

Utilizando um método discriminante passo a passo (*stepwise*) e com o critério de selecção Lambda de Wilks com valores de entrada de $F = 3,84$ e saída de $F = 2,71$, obtém-se uma listagem de resultados semelhante a esta:

Discriminant				
Analysis Case Processing Summary			Valid	9
Analysis 1				
Stepwise Statistics				
Step	Entred	Lambda	F	Signif de F
1	CREDIT	0,303	6,900	0,280
2	SALAR	0,052	8,460	0,003
1	PROF	0,011	11,604	0,001

Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
1	CREDIT	1,000	6,900	
2	CREDIT	,240	18,537	,438
	SALAR	,240	12,059	,303
3	CREDIT	,079	36,925	,207
	SALAR	,075	33,718	,190
	PROF	,304	7,798	,052

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
0	CREDIT	1,000	1,000	6,900	,303
	PROF	1,000	1,000	2,944	,505
	IDADE	1,000	1,000	2,893	,509
	SALAR	1,000	1,000	3,852	,438
1	PROF	,981	,981	1,494	,190
	IDADE	,979	,979	1,193	,205
	SALAR	,240	,240	12,059	,052
2	PROF	,304	,075	7,798	,011
	IDADE	,411	,101	4,989	,015
3	IDADE	,024	,018	3,406	,003

**Summary of Canonical Discriminant Functions
Eigenvalues**

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	41,189	97,1	97,1	,988
2	1,232	2,9	100,0	,743

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,011	22,724	6	,001
2	,448	4,014	2	,134

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
CREDIT	3,493	,197
PROF	1,638	-,010
SALAR	-3,548	,825

Functions at Group Centroids

RISCO	Function	
	1	2
0	-3,684	1,112
1	7,411	-,004
2	-3,727	-1,108

Unstandardized canonical discriminant functions evaluated at group means

Classification Function Coefficients

	RISCO		
	0	1	2
CREDIT	15,298	31,607	15,048
PROF	3,204	6,646	3,195
SALAR	-,473	-1,078	-,498
(Constant)	-66,323	-235,035	-56,696

Fisher's linear discriminant functions

Do primeiro quadro conclui-se que a variável IDADE tem um poder discriminatório inferior ao critério utilizado, que a variável CREDIT é a que maior poder discriminatório, com $\lambda = 0,303$, seguida por SALAR (λ total igual a 0,052) e PROF (0,011). As três variáveis em conjunto têm um Λ igual a 0,011, altamente significativo, portanto.

Sendo três grupos, poderia ter-se duas funções discriminantes. Mas o quadro dos valores próprios (eigenvalues) mostra que a primeira função discriminante explica 97,1 % e a segunda apenas 2,9 % da variação dos grupos. Por outro lado, pela última coluna do quadro de Wilks- λ vê-se que, depois de removido o efeito discriminante da primeira função, a segunda deixa de ser significativa (probabilidade de acontecer por acaso de $p = 0,134$, muito superior ao limite do significado $p = 0,05$).

Os centróides de cada um dos três grupos são os do quadro *Functions at group centroids* e os respectivos coeficientes das funções discriminantes os da última tabela. Com eles é possível atribuir RISCO a novas unidades de análise. Por exemplo, se um cliente fizesse um pedido de 25 unidades monetárias, com 5 anos de experiência e 320 de salário, deveria ser classificado como um cliente de risco médio, porque o valor máximo das equações de classificação se verifica para o grupo 2:

$$C1 = 180,787 \quad C2 = 243,41 \quad C3 = 176,119$$

4.- BIBLIOGRAFIA

- DAVIS, John C. (1986) – ***Statistics and Data Analysis in Geology, 2nd edition***, New York, John Wiley & Sons, p. 602-606.
- (*) KLECKA, William R. (1975) - “***Discriminant Analysis***” in NIE, Norman H. *et alia* - ***SPSS -Statistical Package for Social Sciences, 2th edition***, McGraw-Hill Book Company, New York, p. 434-448.
- MATHER, Paul (1976) – ***Computational Methods of Multivariate Analysis in Physical Geography***, John Wiley & Sons, London, p. 420- 459.
- (*) REIS, Elizabeth (1997) – ***Estatística Multivariada Aplicada***, Lisboa, Edições Sílabo, p. 199-249.

Anexo 10

Redes e Grafos

Os fenómenos de interacção são dos mais relevantes fenómenos com expressão geográfica. Decorrentes das relações que as sociedades humanas estabelecem entre si, constituem uma das chaves da explicação em Geografia Humana.

Em Geografia, as interacções podem ser de produtos, de indivíduos e de informação. Efectuam-se segundo redes, que podem estar fortemente implantadas no território - as redes de estradas, as vias férreas e as linhas telefónicas -, que podem estar apenas baseadas em nós ou pontos localizados no espaço como é o caso das redes marítimas e aéreas -, ou que até podem ser em grande parte virtuais - como a InterNet e as redes rádio.

Em todos estes casos a interacção pode ser expressa por redes. As redes constituem, pois, um sistema de ligações entre lugares (simplificadamente assumidos como pontos) através das quais se podem processar movimentos.

1. – A Linguagem Topológica: Redes, Grafos e Matrizes

Como as interacções e as redes são relações que se estabelecem entre um ponto de origem e um ponto de destino, podem ser representadas matematicamente por uma matriz e graficamente por um grafo.

A matriz é uma entidade matemática constituída por números dispostos em linhas e colunas. Nas interacções o número representa uma qualquer característica dessa interacção (existência ou não, intensidade, etc.), a linha o lugar de origem e a coluna o de destino. No caso das redes, o número representa a existência ou não de ligação entre os lugares ou a distância a que se encontram.

Grafo é uma estrutura topológica constituída por um conjunto de elementos, os nós e pelas ligações entre os nós, os arcos. Os nós, que também são chamados vértices, representam os lugares, ao passo que os arcos representam as ligações.

Nos grafos, nem os nós nem os arcos assumem hierarquia, direcção, extensão ou forma. A estrutura é definida apenas pelo conjunto das relações entre as entidades, ou seja, pelos arcos e nós. Este conjunto representa uma rede topológica, isto é, uma estrutura existente num espaço em que importa a relação entre os lugares – os nós. Nos grafos, a localização exacta e as distâncias não são tomados em conta.

Nessa estrutura topológica, a localização e a distância passam de absolutas a relativas e como tal são descritas e analisadas.

Existem outros tipos de representações gráficas em que são considerados as localizações e as distâncias absolutas, assim como as

formas e as hierarquias, que são os mapas, amplamente utilizados em Geografia, com as suas regras e técnicas de utilização específicas. Mas não são grafos, nem se lhe aplicam as técnicas de análise de grafos.

1.1. – Grafos

Relativamente aos grafos, considera-se geralmente que:

- uma rede possui um número finito de lugares, isto é, de nós ou vértices;
- cada arco liga apenas dois nós diferentes;
- cada par de nós é ligado apenas por um único arco;
- a existência de uma ligação (de um arco) tanto pode ser unidireccional como pressupor a existência de tráfego nos dois sentidos. No primeiro caso o grafo diz-se orientado.

Qualquer grafo, orientado ou não, pode ser representado por uma matriz, em que as linhas são os nós de origem e as colunas os de destino. Como é natural, os grafos não orientados são representados por matrizes simétricas.

1.2. – Matriz de Conectividade

O tipo de grafo mais simples é o de ligação directa ou de conectividade, em que se indica graficamente se os nós têm ou não ligações directas entre si.

Ao grafo de ligação directa corresponde a matriz de ligação directa ou de conectividade, que é uma matriz binária onde se relacionam pares de lugares, indicando-se estes se encontram ligados ou não através de um arco, ou seja, se possuem ou não uma ligação directa.

Em caso afirmativo, a relação é representada pela unidade; caso contrário, por um zero.

Esta matriz de conectividade, a que chamaremos por agora matriz C1, é simétrica para todos os gráficos não-orientados.

C1	1	2	3	4	5	6
1	0	1	0	0	0	0
2	1	0	1	0	0	0
3	0	1	0	1	1	0
4	0	0	1	0	0	0
5	0	0	1	0	0	1
6	0	0	0	0	1	0

Figura 10.1 – Matriz de Conectividade

É importante fazer a distinção entre gráfico planar e não planar. Um grafo diz-se planar quando as relações entre lugares podem estabelecer-se apenas a duas dimensões e diz-se não planar quando essas relações apenas podem estabelecer-se a mais de duas dimensões.

Num desenho de um grafo planar numa folha de papel, os arcos não se interseccionam necessariamente. Pelo contrário, na representação, no plano, de um grafo não planar, alguns dos arcos têm de representar-se cruzados, como resultado da projecção tridimensional (ou outra superior) nesse mesmo plano.

1.3. – Conceitos

Os conceitos mais importantes associados à teoria dos grafos são:

- **Distância topológica entre dois nós:** é a distância entre esses nós na rede, medida pelo número de arcos que os separam;
- **Circuito:** é um caminho que se inicia e termina no mesmo nó, sem passar duas vezes pelo mesmo arco;
- **Região:** área intersticial ou envolvente a um grafo;
- **Rede Nula:** rede sem arcos;
- **Rede em árvore:** uma rede em que não existem circuitos;
- **Número Associado:** (ou Índice de Köning) de um nó, é a distância topológica desse nó ao nó que se encontra mais afastado topologicamente na rede, considerando-se o trajecto mais curto;
- **Nó Central:** o nó da rede que possui o número associado de valor mais baixo;
- **Diâmetro da Rede:** o diâmetro de uma rede é igual ao número associado mais elevado que nela se encontra.

2. – Medidas de Ligação e Conectividade

O estudo dos grafos inicia-se pelo estudo do conceito de ligação, ou de conexão.

Numa rede qualquer o número **mínimo** de arcos necessários para que todos os nós estejam ligados entre si é de **$n-1$** , sendo **n** é o número de nós ou vértices.

Assim, o número mínimo de ligações (**l**) será:

$$l = n - 1$$

O número **máximo** de ligações que se podem estabelecer num grafo depende do facto de se querer que o grafo seja ou não planar.

Sendo planar, o número de ligações máximo possível (**L**) é:

$$L = 3 \times (n - 2)$$

No caso de ser não-planar, ter-se-á então:

$$L = \frac{1}{2} \times (n^2 - n)$$

Pode definir-se uma série de medidas que relacionam o número de arcos com o número de nós, sendo que, para um mesmo número de nós, quanto maior for o número de arcos maior será o grau de interligação ou de conectividade da rede.

O **Índice Beta** é o mais simples. Calcula-se dividindo o número de arcos (**a**) pelo número de nós (**n**) e corresponde à densidade de ligações. Se for inferior a um estamos perante uma rede geral em árvore, ou redes em árvores, desconexas. Se for igual a um existe pelo menos um circuito e se for maior do que um, é uma rede complexa, com mais de um circuito.

$$\beta = \frac{a}{n}$$

O **Número Ciclomático** mede o número de circuitos fundamentais que existem na rede.

$$Nc = a - n + sg$$

Quando é igual a zero, todos os arcos estão ligados entre si e não existem circuitos. Se for negativo, há nós isolados. Quando é positivo existem circuitos, tanto mais quanto maior for o seu valor.

Estes dois índices informam sobre a maior ou menor conectividade de uma rede. São porém difíceis de interpretar, pois o seu valor máximo está dependente do número de nós que existem na rede, pelo que é muito difícil comparar redes com um número de elementos diferente.

Para obstar a este inconveniente foram desenvolvidos índices comparativos, que comparam um valor qualquer calculado com o valor que existiria se a ligação entre os nós fosse máxima. Os seus valores variam entre 0 e 1, sendo 0 a inexistência de ligações e 1 a conectividade máxima.

O número máximo de circuitos numa rede é o número máximo de ligações (L) menos o número mínimo de ligações necessário para ligar todos os nós (I), ou seja:

Para grafos planares: $L - I = 3 \times (n - 2) - (n - 1) = 2 \times n - 5$

Para grafos não planares: $L - I = \frac{n \times (n - 1)}{2} - (n - 1) = \frac{n^2 - 3n + 2}{2}$

Assim sendo, o **Índice Alfa**, que compara o número de circuitos existente na rede (o número ciclomático) com o número máximo de circuitos calculado para a rede em questão, calcula-se da seguinte maneira

Para grafos planares: $\alpha = \frac{a - n + sg}{2 \times n - 5}$

Para grafos não planares: $\alpha = \frac{a - n + sg}{\frac{n^2 - 3 \times n + 2}{2}}$

O **Índice Gama**, que mede a conectividade relativa da rede, compara o número de arcos existente na rede com o número máximo de arcos que existe na rede em análise, segundo as expressões:

Para grafos planares: $\gamma = \frac{a}{3 \times (n - 2)}$

Para grafos não planares: $\gamma = \frac{2 \times a}{n^2 - n}$

O índice Gama é a medida de conectividade mais utilizada. Quando num texto aparecer referência ao índice de conectividade, sem especial referência, trata-se do Índice Gama.

A fim de compreender melhor o cálculo e a interpretação de resultados, vai tratar-se do exemplo da rede que está representada, juntamente com a respectiva matriz de conectividade, na Figura 10.2.

Analisando este grafo, vê-se que tem 10 nós e 9 arcos.

O Índice Beta (de Kansky) é 0,9, o que significa que existe uma média de 0,9 por cada nó. O Número Ciclomático é 0, o que quer dizer que a rede não tem circuitos (é uma rede tipo árvore).

O Índice Alfa (também devido a Kansky) tem o valor de 0% e indica a percentagem entre o nº de circuitos existentes e o máximo possível, enquanto o Índice Gama (que também foi proposto por Kansky) é igual a 37,5% para um grafo planar e 20% para um não-planar, com o significado de qual é o peso do número de arcos existentes, face ao máximo possível.

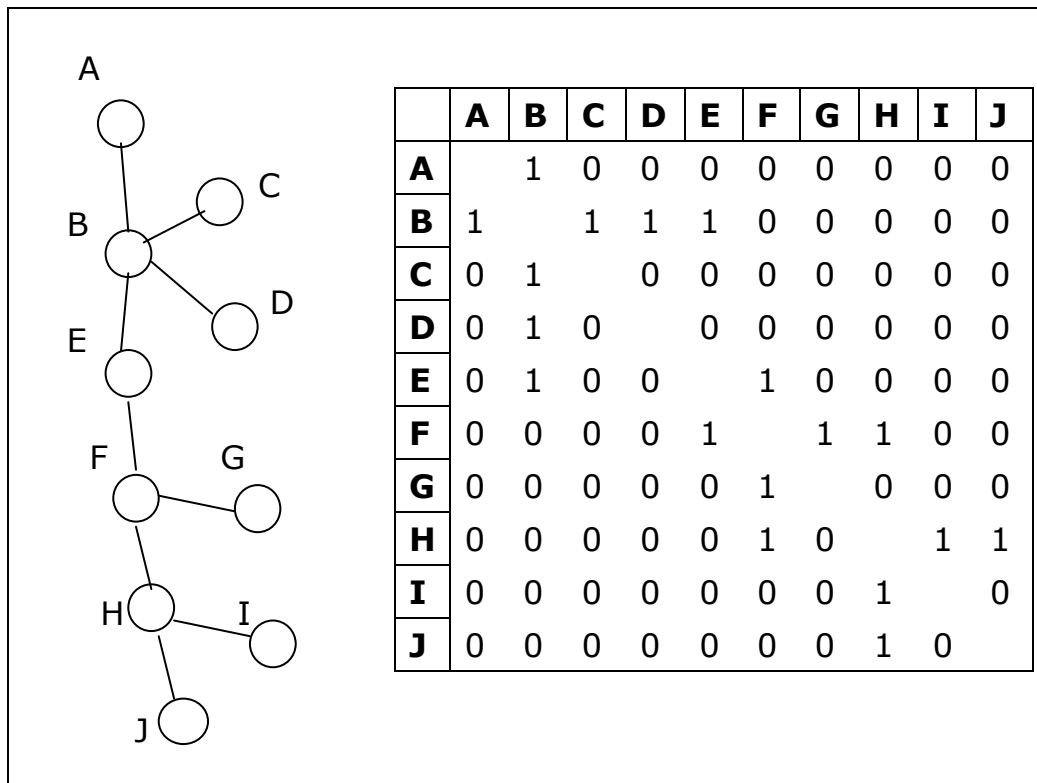


Figura 10.2 – Grafo da rede e respectiva matriz de conectividade

3. – Medidas de Acessibilidade e Centralidade

As medidas de conectividade referem-se, fundamentalmente, às características da rede no seu todo. Os índices de centralidade ou de acessibilidade permitem, para além de descrever algumas das características da rede, diferenciar os nós uns dos outros.

Nas redes mais simples, com poucos nós, a centralidade ou acessibilidade de um nó será tanto maior quanto maior for o número de ligações directas a outros nós. Deste modo, o nó que possuir maior número de ligações directas será o mais acessível na rede. Mas nas redes mais complexas, a simples observação não basta. É necessário construir a matriz de distâncias topológicas da rede e depois analisá-la.

Veja-se a figura 10.3, em que está representado um grafo e a respectiva matriz de distâncias topológicas, isto é, o número de arcos entre cada par de nós, seguindo sempre o caminho mais curto.

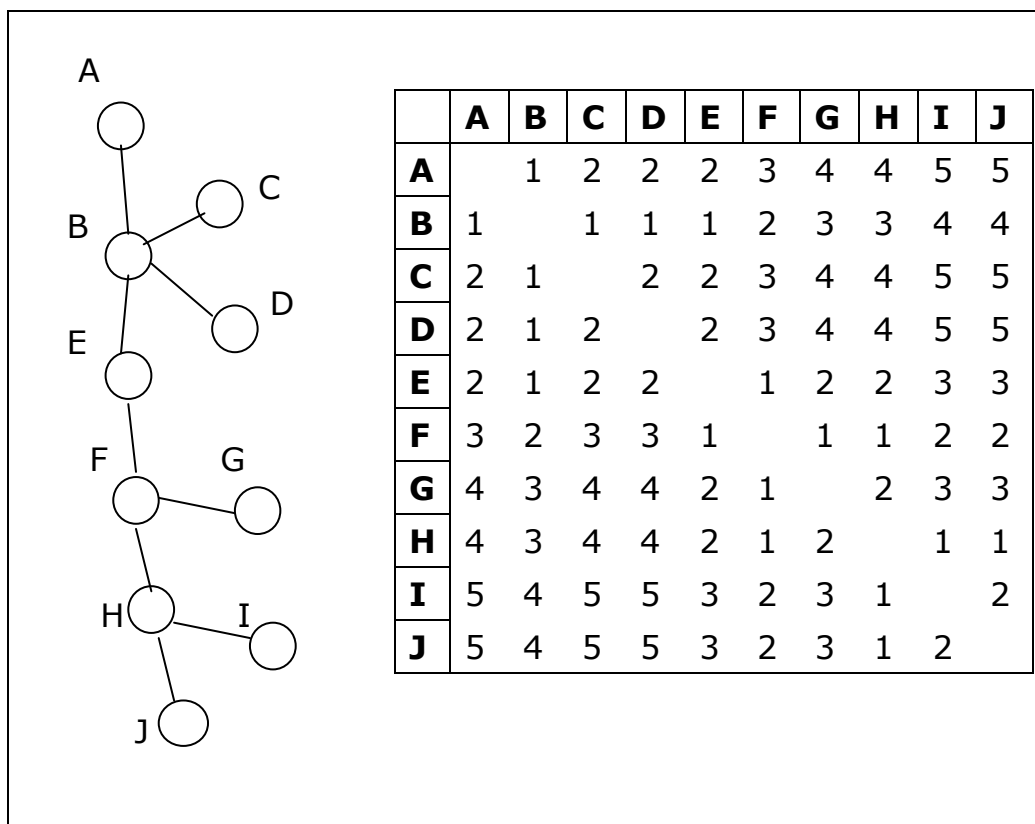


Figura 10.3 – Grafo e respectiva matriz de distâncias topológicas

Para o conjunto da rede pode definir-se o diâmetro, que é a maior distância topológica que nela existe.

Como medida da acessibilidade topológica de cada nó pode usar-se o Índice de Acessibilidade Shimbel (**ac**), que é a soma dum nó a todos os outros, conforme:

$$ac_i = \sum_{j=1}^n d_{ij}$$

ou o Índice de Acessibilidade média de Shimbel de cada nó (**AC**), que é a média das distâncias topológicas desse mesmo nó a todos os outros:

$$AC_i = \frac{\sum_{j=1}^n d_{ij}}{n-1}$$

A acessibilidade geral da rede avalia-se pelo Índice de Dispersão, total ou médio. O Índice de Dispersão total (**d**), calcula-se da seguinte forma:

$$d = \sum_{i=1}^n \sum_{j=1}^n d_{ij}$$

e o Índice de Dispersão médio (**D**), através da seguinte fórmula:

$$D = \frac{\sum_{i=1}^n \sum_{j=1}^n d_{ij}}{(n-1) \times n}$$

4. – Matrizes Derivadas

A partir do matriz de conectividade (**C1**), pode gerar-se uma série de matrizes:

- as matrizes de ligação **Ci**, sendo $i=1, n$, em que se encontram o número de percursos diferentes que é possível de efectuar entre dois nós, usando um número de arcos igual a **i**;

Matriz **C1** – matriz de conectividade
ligações directas entre os nós

Matriz **C2** = $C1 \times C1$
nº de ligações diferentes possíveis entre 2 nós usando 2 arcos

Matriz **C3** = $C2 \times C1 = C1 \times C1 \times C1$
nº de ligações diferentes possíveis entre 2 nós usando 3 arcos

.....

Matriz **Cn** = $C1 \times C1 \times \dots \times C1$ (n vezes)
nº de ligações diferentes possíveis entre 2 nós usando **n** arcos

- as matrizes **Ti**, que indicam quantas ligações diferentes existem entre pares de nós, utilizando percursos com **i** ou menos arcos;

Matriz **T1** = $C1$ – matriz de conectividade
ligações directas entre os nós

Matriz **T2** = $C1 + C2$ -
nº de ligações diferentes possíveis entre 2 nós usando 2 ou 1 arcos

Matriz **T3** = $C1 + C2 + C3$
nº de ligações diferentes possíveis entre 2 nós usando 3 ou menos arcos

.....

Matriz **Tn** = $C1 + C2 + \dots + Cn$
nº de ligações diferentes possíveis entre 2 nós usando **n** ou menos arcos

Nos grafos em que só existe um subgrafo ($sg = 1$) a matriz T_n , em que já não existem zeros, é uma matriz muito importante, pois dá uma medida de acessibilidades dos nós (através da soma das linhas) e da acessibilidade geral da rede (pela soma de todos os elementos da matriz), que será tanto mais acessível, quanto maior for.

- como o número de ligações possíveis cresce muito rapidamente com **Cn**, Garrison propôs uma matriz **TGi**, em que as sucessivas matrizes **Ci** são multiplicadas por um factor-escala S (com valor entre 0 e 1) que vai diminuindo a importância das múltiplas ligações mais longas e torna os valores manuseáveis.

Matriz **TGn** = $S.C1 + S^2.C2 + S^3.C3 + \dots + S^n.Cn$
 proporcional ao n^0 de ligações diferentes possíveis entre 2 nós usando **n** ou menos arcos, multiplicados por em factor **S**, geralmente igual a 0,3.

- e as matrizes de Shimbel, **Di**, em que se expressa o número mínimo de arcos necessários para ligar dois nós. É a matriz de distâncias topológicas.

Enquanto a matriz **Cn** representa o n^0 de ligações diferentes possíveis entre 2 nós usando **n** arcos, a matriz de Shimbel (**Dn**) indica o número de arcos em que existe a primeira ligação entre dois nós.

Chama-se a atenção para que as análises de acessibilidade e centralidade de Shimbel consideram a matriz **Dn** para o valor de **n** em que deixa de haver zeros na matriz **Cn**.

Na Figura 10.4 representa-se um grafo, a respectiva matriz de distâncias topológicas ou de Shimbel, os Índices de Shimbel (**IS**) e o n^0 associado ou Índice de König (**NA**) de cada nó.

	A	B	C	D	E	F	G	H	IS	NA
A	-	1	1	2	2	3	4	4	17	4
B	1	-	2	1	2	3	4	4	17	4
C	1	2	-	2	1	2	3	3	14	3
D	2	1	2	-	1	2	3	3	14	3
E	2	2	1	1	-	1	2	2	11	2
F	3	3	2	2	2	-	1	1	13	3
G	4	4	3	3	2	1	-	2	19	4
H	4	4	3	3	2	1	2	-	19	4

Figura 10.4 – Grafo, matriz de Shimbel e índices de Shimbel e König

De seguida apresenta-se a parte mais importante de um "output" do programa de análise de redes utilizado no C.E.G..

***** RESULTADOS DO PROGRAMA *****

No. de Lugares i (linhas e colunas): 10

MATRIZ DE CONECTIVIDADE - C 1

	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)
1)	0	1	0	0	0	0	0	0	0	0
2)	1	0	1	1	1	0	0	0	0	0
3)	0	1	0	0	0	0	0	0	0	0
4)	0	1	0	0	0	0	0	0	0	0
5)	0	1	0	0	0	1	0	0	0	0
6)	0	0	0	0	1	0	1	1	0	0
7)	0	0	0	0	0	1	0	0	0	0
8)	0	0	0	0	0	1	0	0	1	1
9)	0	0	0	0	0	0	0	1	0	0
10)	0	0	0	0	0	0	0	1	0	0

MATRIZ C 5

	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)
1)	0	17	0	0	0	7	0	0	1	1
2)	17	0	17	17	24	0	7	9	0	0
3)	0	17	0	0	0	7	0	0	1	1
4)	0	17	0	0	0	7	0	0	1	1
5)	0	24	0	0	0	19	0	0	6	6
6)	7	0	7	7	19	0	12	22	0	0
7)	0	7	0	0	0	12	0	0	5	5
8)	0	9	0	0	0	22	0	0	11	11
9)	1	0	1	1	6	0	5	11	0	0
10)	1	0	1	1	6	0	5	11	0	0

MATRIZ T 5

	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)
1)	5	22	5	5	6	8	1	1	1	1
2)	22	21	22	22	30	8	8	10	1	1
3)	5	22	5	5	6	8	1	1	1	1
4)	5	22	5	5	6	8	1	1	1	1
5)	6	30	6	6	11	24	5	7	7	7
6)	8	8	8	8	24	15	16	28	6	6
7)	1	8	1	1	5	16	4	6	6	6
8)	1	10	1	1	7	28	6	14	15	15
9)	1	1	1	1	7	6	6	15	4	4
10)	1	1	1	1	7	6	6	15	4	4

MATRIZ D 5

	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)
1)	0	1	2	2	2	3	4	4	5	5
2)	1	0	1	1	1	2	3	3	4	4
3)	2	1	0	2	2	3	4	4	5	5
4)	2	1	2	0	2	3	4	4	5	5
5)	2	1	2	2	0	1	2	2	3	3
6)	3	2	3	3	1	0	1	1	2	2
7)	4	3	4	4	2	1	0	2	3	3
8)	4	3	4	4	2	1	2	0	1	1
9)	5	4	5	5	3	2	3	1	0	2
10)	5	4	5	5	3	2	3	1	2	0

ANALISE DO GRAFO

Nº de nós.... 9

Nº de arcos..	10
INDICES	
Nº Ciclomático.....	0
Indice Beta de Kansky.....	.90000
Ind. Alfa de Kansky planar (%)....	.00000
Ind. Alfa de Kansky n/planar (%)..	.00000
Ind. Gama de Kansky planar (%)....	37.50000
Ind. Gama de Kansky n/planar (%)..	20.00000
A MATRIZ TOTAL DE LIGAÇÕES É A - T 5	
DIAMETRO DA REDE:.....	5
ACESSIBILIDADE TOTAL DOS NÓS:	
	Absoluta Média
1) -	55 6.11111
2) -	145 16.11111
3) -	55 6.11111
4) -	55 6.11111
5) -	109 12.11111
6) -	127 14.11111
7) -	54 6.00000
8) -	98 10.88889
9) -	46 5.11111
10) -	46 5.11111
Acessibilidade geral da Rede:	
	790 8.77778
MATRIZ DE SHIMBEL - D 5	
Acessibilidades topológicas	
(ligações de um nó a todos)	
	Nº. mínimo de arcos Nº médio
1) -	28 3.11111
2) -	20 2.22222
3) -	28 3.11111
4) -	28 3.11111
5) -	18 2.00000
6) -	18 2.00000
7) -	26 2.88889
8) -	22 2.44444
9) -	30 3.33333
10) -	30 3.33333
Nº. mínimo de arcos p/se chegar a todos os nós:	
	248
Nº. médio de arcos de cada nó: 2.75556	

A análise topológica de redes tem sido bastante criticada, mesmo por aqueles que mais a têm praticado.

Segundo Taaffe, as principais críticas podem resumir-se aos seguintes aspectos:

- Equivalência – Todos os nós são considerados equivalentes, assim como os arcos e, quanto a estes, não se toma em conta a sua localização, direcção e importância;
- Localização – a análise topológica de redes preocupa-se com o número de nós e as suas relações, mas não se sabe exactamente onde se localizam;
- Atenuação – as diferenças entre ligações directas e indirectas (e entre os diversos graus de indirectas) podem ser muito importantes, mas não são assumidas como tal;
- Redundância – não se consegue eliminar os círculos, as viagens de ida e volta, no todo ou em parte das ligações, que não têm interesse;
- Ligações – embora na análise sejam assumidas como equivalentes, são muitas vezes desiguais (distância, intensidade, etc.)

Uma das formas propostas para eliminar a atenuação e a redundância é a utilização da análise com matrizes de Garrison.

5. – Matrizes Ponderadas

Outra forma de ultrapassar as limitações da análise topológica de redes é a de utilizar matrizes que representem grafos ponderados.

São as matrizes **L_i** , em que os números representam uma qualquer característica dos arcos. Esta característica pode ser aditiva, como acontece, por exemplo, quando se consideram as distâncias ou multiplicativa, de que pode ser exemplo de um coeficiente de dificuldade tipo cadeia de probabilidades.

	1	2	3	4	5	6
1	0	10				
2	10	0	20			
3		20	0	10	30	
4			10	0		
5			30		0	5
6					5	0

Figura 10.5 – Matriz de cargas (**L_1**) de um grafo

As matrizes **L_i** começam pela matriz **L₁**, a matriz de cargas do grafo, em que cada célula representa um valor associado ao arco respectivo.

A matriz **L₂** é construída da seguinte forma:

$$l_{2\ ij} = \min(l_{1\ ik} (k = 1, \dots, n) + l_{1\ jk} (k = 1, \dots, n))$$

e do mesmo modo, os valores das matrizes **L_n** são também o mínimo da soma dos respectivos valores nas matrizes de ordem **n-1**.

Matriz L2							Matriz L3						
	1	2	3	4	5	6		1	2	3	4	5	6
1	0	10	30				1	0	10	30	40	60	
2	10	0	20	30	50		2	10	0	20	30	50	55
3	30	20	0	10	30	35	3	30	20	0	10	30	35
4		30	10	0	40		4	40	30	10	0	40	45
5		50	30	40	0	5	5	60	50	30	40	0	5
6			35		5	0	6		55	35	45	5	0

O processo continua-se até que fiquem preenchidos todos os elementos da matriz situados fora da diagonal.

No exemplo que se está a tratar, isso acontece para $n=4$, pois com a matriz **L₄** ficam preenchidos todos os elementos da matriz que se encontram fora da diagonal.

Matriz L4							Σ
	1	2	3	4	5	6	
1	0	10	30	40	60	65	205
2	10	0	20	30	50	55	165
3	30	20	0	10	30	35	125
4	40	30	10	0	40	45	165
5	60	50	30	40	0	5	185
6	65	55	35	45	5	0	205

Hipóteses de interpretação:

- Se a matriz de pesos **L1** for uma matriz de distâncias, **L4** será a matriz que representa as distâncias entre todos os nós (note-se que a distância total é aditiva em relação aos arcos iniciais).
- Se porventura a matriz de pesos **L1** for um coeficiente de dificuldade de tráfego, a fórmula **Ln** deveria ser de tipo multiplicativo e **L4** representaria a dificuldade de tráfego total no trajecto (pois se estaria num tipo multiplicativo de relação).

6. – Bibliografia

- (*) CICERI, M.F.; MARCHANT, B.; RIMBERT, S. (1977) – ***Introduction a l'Analyse de l'Espace***, Masson, Paris, p. 45-53.
- CLIFF, A.D.; HAGGETT, P. (1981) "***Graph Theory***", in WRIGLEY, N.; BENNETT, R.J. (Eds) – ***Quantitative Geography***, Routledge & Kegan Paul, London, p. 225-234.
- SMITH, David M.(1977) – "***Patterns in Human Geography***", Penguin Books, Harmondsworth, p. 279-285.
- HAGGETT, P.; CHORLEY, R. J. (1969) – ***Network Analysis in Geography***, Edward Arnold, London. 348 p.
- POTTS, R.B. (1978) "***Networks***" in LIGHTHILL – ***Newer Uses of Mathematics***, Penguin Books, Harmondsworth, 431 p.
- TAAFFE, E. J.; GAUTHIER, H. L. (1973) – ***Geography of Transportation***, Prentice-Hall, Englewood Cliffs, p. 249-283.

Anexo 11

Geoestatísticas

1. - INTRODUÇÃO

O objecto dos estudos dos geógrafos é o território. Não é de admirar, por isso, que se tenham desenvolvido técnicas especiais de estatística que tomem em conta, de forma explícita, o factor espaço.

São técnicas analíticas que melhoram a análise estatística mais habitual, ao considerarem cada sujeito analisado um "**lugar**", com uma localização específica e mantendo com os outros uma rede de relações ou, ao menos, de proximidades, numa perspectiva teórica completamente diversa do ponto de vista da estatística clássica que considera cada caso um "**evento**" independente.

Com efeito, os factos geográficos, os valores das variáveis com suporte territorial que interessam os geógrafos, não têm as características que os estatísticos gostam de postular para as técnicas que propõem. Os valores das variáveis com suporte geográfico geralmente não são independentes e nem são repetíveis, nem equivalentes (de igual importância). Muito pelo contrário, são conjuntos de dados em que existem e são muito importantes as relações de distância, de proximidade ou de vizinhança e os conceitos de centralidade, periferia e acessibilidade.

A análise espacial, também designada por alguns por **Geoestatística**, toma em consideração a posição que os valores têm no espaço e as relações que estabelecem uns com os outros.

Tendo sempre em conta que uma das melhores formas de analisar e representar distribuições espaciais é o **mapa**, em muitos casos há vantagens em determinar valores resumo, através de medidas de Geoestatística, que permitam com facilidade e rapidez a caracterização geral individual dos lugares, não só através dos valores das variáveis estudadas, mas também através das relações de proximidade que estabelecem entre si.

2. – MEDIDAS ESPACIAIS DE CENTRALIDADE E DISPERSÃO

2.1.- Centro de Gravidade

O centro de gravidade ou centro médio de uma distribuição de localizações de pontos no espaço é um conceito similar ao da média aritmética, adaptado ao caso particular das distribuições espaciais.

Quando se tem uma distribuição de pontos descrita pelas suas coordenadas locativas, x e y , a coordenada x_c do centro médio é a média das coordenadas x_i de todos os pontos. É o mesmo sucede para a coordenada y .

É claro que a média é simples se todos os pontos têm o mesmo peso e ponderada se têm pesos diferentes.

Formalmente, dado um conjunto de n pontos, descritos pelas suas coordenadas x_i e y_i e pelo seu peso p_i , as coordenadas do centro médio ou de gravidade são:

$$x_c = \frac{\sum_{i=1}^n (x_i \times p_i)}{n} \quad y_c = \frac{\sum_{i=1}^n (y_i \times p_i)}{n}$$

As coordenadas x e y das localizações dos pontos são as coordenadas cartesianas dos pontos que resultam da implantação de uma rede gaussiana, ou cartesiana, sobre um mapa. São coordenadas rectangulares, porque os eixos fazem um ângulo recto entre si.

A maior parte dos mapas tem já implantada uma malha quadrangular de referência, que geralmente é quilométrica. Existe um sistema mundial, o UTM (Universal Transverso de Mercator) e em cada país existem um ou vários outros (em Portugal o militar M/P e o X/Y do Instituto Geográfico e Cadastral). O uso dos valores **x** e **y** (ou "**eastings**" e "**northings**") dessa rede facilita a recolha das coordenadas dos pontos e a implantação dos valores no próprio mapa. Por outro lado, como a rede implantada no mapa tem correspondência quilométrica no terreno, a interpretação é muito facilitada.

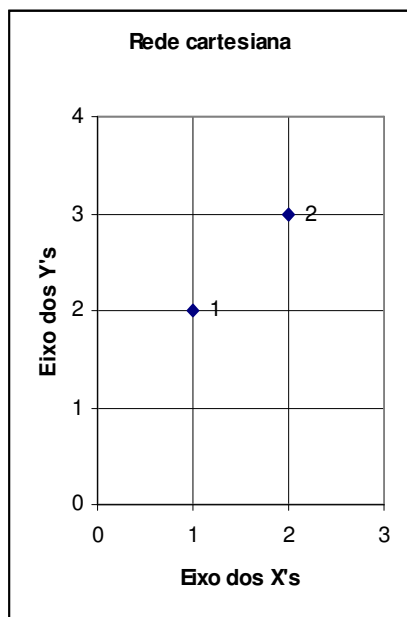


Figura 11.1 - Coordenadas cartesianas dos pontos 1 e 2.

Quando o mapa tem originalmente implantada apenas a rede de coordenadas geográficas - latitude e longitude - ou não tem nenhuma, é necessário que o investigador coloque nele uma rede cartesiana ou gaussiana. Para isso escolhe-se uma direcção de referência, geralmente o lado esquerdo do mapa, que será o eixo dos **y** 's e a partir da base do mapa traça-se uma perpendicular, que será o nosso eixo dos **x** 's. Utiliza-se uma medida qualquer de comprimento (centímetros, milímetros, polegadas, etc.) para medir as coordenadas dos pontos.

Na Figura 11.1 pode também ver-se como se medem as coordenadas dos pontos numa rede gaussiana implantada. O ponto 1 tem como coordenada x o valor 1 e como coordenada y o valor 2. Por outro lado, o ponto 2 tem como coordenada x o valor 2 e como coordenada y o valor de 3.

Sempre que possível, devem ser utilizadas as coordenadas quilométricas existentes no mapa, pois tal facto facilita a localização no próprio mapa dos valores calculados para o centro de gravidade e outros valores de geostatísticas que adiante se apresentam. Pela mesma razão, o uso destas redes quilométricas vem facilitar os cálculos e a interpretação dos resultados.

Quanto não existir uma rede rectangular já implantada nos mapas (ou quando se utilizam fotocópias de escala diferente da original) constrói-se uma rede cartesiana com dois eixos perpendiculares na qual se marcam e posteriormente se medem os valores das coordenadas (a distância em centímetros ou milímetros a cada eixos). A orientação dos eixos perpendiculares é indiferente, mas é conveniente serem paralelos aos limites do mapa.

O Centro de Gravidade de uma distribuição de uma variável pontual é a localização que teria o total dessa variável se todos os valores estivessem concentrados num único ponto. É um conceito bastante utilizado em Geografia, especialmente em termos comparativos. Como exemplo pode referir-se a análise da mudança do centro de gravidade de uma variável ao longo do tempo (população, população urbana, rendimento, etc.). Refira-se que tão cedo como 1872 já J. E. Hilgard empregou este conceito de centro de gravidade para ilustrar a deslocação para Oeste da População dos Estados Unidos da América do Norte.

Também se utiliza para estudar diferentes distribuições num território, seja um país, uma província, um concelho ou qualquer outra unidade regional, comparando os centros de gravidade de variáveis relacionadas, como por exemplo a população total e a activa, os activos no primário, no secundário e no terciário, a população e o rendimento, os jovens e os idosos, o emprego, as actividades e/ou os recursos, etc.. Refira-se que ESTEBANEZ & BRADSHAW (1978, 402 e seguintes) referem um estudo semelhante do conhecido químico russo Mendeleiev sobre os centros de diversos recursos na Rússia.

O centro de gravidade, ou centro médio, é também o ponto de distância mínima dos lugares (distância ponderada, se os lugares tiverem pesos diferentes, como é evidente), o que significa que, se todos os elementos de todos os lugares se deslocarem a um único ponto, de modo a minimizar a deslocação, esse ponto é o centro de gravidade.

Exemplo: Dados cinco pontos (A a E), com as coordenadas e pesos que se indicam no quadro a seguir:

i	x	Y	Peso
A	34	12	3
B	30	25	10
C	15	5	4
D	20	30	5
E	25	10	8

Quadro 11.1. – Exemplo 7.1

o cálculo das coordenadas do centro de gravidade ou centro médio faz-se do seguinte modo:

$$x_c = \frac{34 \times 3 + 30 \times 10 + 15 \times 4 + 20 \times 5 + 25 \times 8}{3 + 10 + 4 + 5 + 8} = \frac{102 + 300 + 60 + 100 + 200}{30} = \frac{762}{30} = 25,4$$

$$y_c = \frac{12 \times 3 + 25 \times 10 + 5 \times 4 + 30 \times 5 + 10 \times 8}{3 + 10 + 4 + 5 + 8} = \frac{36 + 250 + 20 + 150 + 80}{30} = \frac{536}{30} = 17,8666(6)$$

2.2.- Distância – média

É um conceito pouco utilizado.

Consiste na distância média de todos os pontos ao ponto de gravidade. Calcula-se através do somatório das distâncias entre todos os pontos e o centro de gravidade, de acordo com a expressão abaixo, que dá uma ideia da dispersão (ponderada, quando é caso disso) da distribuição dos lugares.

$$dm = \frac{\sum_{i=1}^n \left(\sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \times p_i \right)}{\sum (p_i)}$$

Considerando os dados do exemplo anterior (Quadro 1) e sabendo que a distância do ponto A ao Centro de Gravidade (**d_{ac}**) é igual a:

$$d_{AC} = \sqrt{(34 - 25,4)^2 + (12 - 17,8666(6))^2} = 10,41046...$$

tem-se do mesmo modo que a **distância-média** (**dm**) do exemplo acima tratado é:

$$dm = \frac{10,41046 \times 3 + 8,48790 \times 10 + 16,54422 \times 4 + 13,28073 \times 5 + 7,87683 \times 8}{3 + 10 + 4 + 5 + 8} =$$

$$dm = \frac{311,7055..}{30} = 10,39018..$$

2.3.- Distância – padrão

A mais importante medida de dispersão da distribuição espacial de uma variável não é a distância média, mas sim a distância padrão (**DP**), principalmente devido à sua semelhança com o conceito de desvio padrão, utilizado na estatística clássica.

Formalmente é a raiz quadrada da média do quadrado das distâncias de todos os pontos ao centro de gravidade e por isso é também, ao mesmo tempo, o desvio padrão das distâncias entre o centro de gravidade e todos os pontos ou lugares da distribuição em análise.

O processo de cálculo da Distância padrão é o seguinte:

- 1º - Calcular o centro de gravidade;
- 2º - Calcular a distância do centro de gravidade a todos os pontos;
- 3º - Elevar as distâncias ao quadrado;
- 4º - Achar a respectiva média
- 5º - Determinar a raiz quadrada desta média.

como se vê pelas expressões abaixo, respectivamente para dados simples e ponderados:

$$DP = \sqrt{\frac{\sum_{i=1}^n ((x_i - x_c)^2 + (y_i - y_c)^2)}{n}}$$

$$DP = \sqrt{\frac{\sum_{i=1}^n \{[(x_i - x_c)^2 + (y_i - y_c)^2] \times p_i\}}{\sum_{i=1}^n (p_i)}}$$

A Distância padrão (DP) constitui uma circunferência em torno do centro de gravidade que dá ideia da dispersão das localizações e, de acordo com a regra das distribuições estatísticas, se a distribuição de pontos for normal segundo x e y (com iguais desvios padrões), então 66% dos pontos estarão dentro do círculo a uma distância padrão, 95 % a duas distâncias-padrão e 99% a três.

Considerando o exemplo que se vem tratando (o do Quadro 11.1) para determinar o valor da Distância padrão é conveniente utilizar um quadro auxiliar, como a seguir se exemplifica:

i	x	y	Peso	Dist. a CG	Dist ²	Dist ² x peso
A	34	12	3	10,41046	108,37778	325,13333
B	30	25	10	8,48790	72,04444	720,44444
C	15	5	4	16,54422	273,71111	1094,84444
D	20	30	5	13,28073	176,37778	881,88889
E	25	10	8	7,87683	62,04444	496,35556
Σ	-----	-----	30	-----	-----	3518,66667

Quadro 11.2 – Quadro auxiliar para o cálculo da distância padrão

A Distância padrão (DP) será então a raiz quadrada da média das distâncias quadráticas (ponderadas, se for caso disso, como o é neste exemplo):

$$DP = \sqrt{\frac{3518,6666(6)}{30}} = \sqrt{117,28888(8)} = 10,82999949...$$

2.4.- Raio dinâmico

Até agora vimos valores únicos para todo o conjunto da distribuição, como é o caso do centro de gravidade, da distância média e da distância padrão.

O Raio Dinâmico, pelo contrário, é uma característica de cada ponto da distribuição espacial. É uma medida da proximidade relativa a que se encontram os outros lugares.

O Raio Dinâmico de um dado ponto **k** (**R_k**) é raiz quadrada da média dos quadrados da distância entre esse ponto **k** e todos os outros pontos **i**.

Sendo então:

$$d_{ki} = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2}$$

o Raio Dinâmico do ponto k calcula-se pelas expressões seguintes:

$$R_k = \sqrt{\frac{\sum_{i=1}^{n-1} (d_{ki}^2)}{n-1}} \quad R_k = \sqrt{\frac{\sum_{i=1}^{n-1} (p_i \times d_{ki}^2)}{\sum_{i=1}^{n-1} (p_i)}}$$

respectivamente para dados simples a primeira e para dados ponderados a segunda.

Seguindo o mesmo exemplo, vamos agora proceder ao cálculo do Raio Dinâmico (RD) do lugar A.

i	x	y	Peso	Dist ² a A	Peso que entra	Dist ² x peso
A	34	12	3	0	0	0
B	30	25	10	185	10	1850
C	15	5	4	410	4	1640
D	20	30	5	520	5	2600
E	25	10	8	85	8	680
Σ	----	----	30	-----	27	6770

Quadro 11.3 – Quadro auxiliar para o cálculo do Raio-Dinâmico do lugar A

O Raio Dinâmico do lugar A (**RD_A**) será, então, a raiz quadrada da média das distâncias que o lugar A estabelece com todos os outros, constituindo um total de 27 ligações.

$$RD_A = \sqrt{\frac{6770}{27}} = \sqrt{250,74074} = 15,83480...$$

Os Raios Dinâmicos dos outros lugares são calculados de maneira semelhante e o seu cálculo pode constituir um bom exercício. Abaixo apresentam-se os resultados para todos os lugares do exemplo que se está a trabalhar e para o Centro de Gravidade.

Lugares	Raios Dinâmicos
A	15,83480
B	16,85230
C	21,24038
D	18,77232
E	15,63795
Centro de Gravidade	10,83000

Quadro 11.4 – Raios-Dinâmicos dos lugares e do centro de gravidade

Como se vê o Raio Dinâmico do centro de gravidade é o menor de todos e igual ao valor da Distância padrão.

O valor do Raio Dinâmico de cada lugar é uma medida da centralidade desse lugar, pois quanto mais próximo estiver da Distância padrão (o raio dinâmico do centro de gravidade) mais central será o lugar e vice-versa.

O Raio Dinâmico do Centro de Gravidade (ou Distância padrão) é uma característica do conjunto, da totalidade da distribuição de lugares.

2.5.- Elipse de dispersão

Como se viu, a Distância padrão (DP) constitui uma circunferência em torno do centro de gravidade que dá ideia da dispersão das localizações. É a forma mais correcta de estimar a densidade de probabilidade da distribuição espacial quando esta distribuição tem iguais desvios padrões segundo X e Y.

Mas a dispersão raramente é regular no espaço, podendo apresentar desvios padrões muito diferentes segundo as diversas direcções. Neste caso é preferível considerar a densidade de probabilidade da distribuição como uma elipse em vez de uma circunferência.

O problema é que, enquanto a circunferência é definida apenas pela localização do seu centro (centro de gravidade) e pelo valor do seu raio (distância padrão), para definir uma elipse são necessários outros parâmetros mais difíceis de calcular, a saber:

- o centro (que é também o centro de gravidade ou centro médio da distribuição);
- os valores do eixo maior e do eixo menor (ou de um deles e da excentricidade da elipse);
- a orientação, (a inclinação dos eixos, ou de apenas um deles, pois são perpendiculares).

Deve procurar-se um programa que calcule estas valores, embora se deva confessar que são poucos os disponíveis. Para os alunos interessados em executar os cálculos à mão, dão-se os elementos que podem solucionar o problema.

Facilitará muito os cálculos a fazer considerar-se as coordenadas X e Y expressas em valores transpostos, em vez dos valores habituais das variáveis. Por transpostos quer significar-se que os dados, tanto do conjunto de coordenadas X como do de coordenadas Y, devem ser expressos, em vez do seu valor, por esse valor subtraído da respectiva média.

Para evitar confusões denotaremos estas variáveis estandardizadas por **M** (a correspondente a **X**) e por **P** (a correspondente a **Y**).

2.5.1- Ângulo de Inclinação

O ângulo de rotação ou inclinação (Alfa) é medido a partir do eixo dos Y (vertical ou norte) no sentido dos ponteiros do relógio. O ângulo Alfa é, então, o ângulo entre a vertical e o eixo maior da elipse, sendo o seu valor dado pela expressão:

$$\operatorname{tg} \alpha = \frac{(\sum m^2 - \sum p^2) + \sqrt{(\sum m^2 - \sum p^2)^2 + 4 \times (\sum (m \times p))^2}}{2 \times \sum (m \times p)}$$

2.5.2- Dimensão dos Eixos

No caso da elipse, em vez de uma só Distância Padrão, é necessário conhecer os desvios padrões segundo os eixos da elipse. O desvio padrão segundo o eixo M é igual a:

$$\sigma_M = \sqrt{2 \times \frac{(\sum m^2 \times \cos^2 \alpha) - 2 \times (\sum mp) \times \sin \alpha \times \cos \alpha + (\sum p^2 \times \sin^2 \alpha)}{n}}$$

O desvio padrão segundo cada eixo é o equivalente, na direcção desse eixo, da distância padrão do círculo. Como é óbvio, quando a distribuição é próxima da circular, os desvios padrões segundo os dois eixos assumem valores muito aproximados entre si e à distância padrão.

E segundo o eixo P vem então:

$$\sigma_P = \sqrt{2 \times \frac{(\sum m^2 \times \sin^2 \alpha) + 2 \times (\sum mp) \times \sin \alpha \times \cos \alpha + (\sum p^2 \times \cos^2 \alpha)}{n}}$$

Sendo estes valores dos desvios padrão respectivamente iguais ao respectivo Raio maior (**RM**) e menor (**Rm**), o comprimento dos eixos da elipse eixos será obviamente o dobro.

2.5.3- Excentricidade

É a razão entre o eixo maior e o eixo menor (ou entre o raio maior e o raio menor) anteriormente calculados.

2.5.4- Área da elipse

A área de cada elipse de dispersão para um dado valor de desvio padrão calcula-se multiplicando o produto do respectivo número de desvios padrão segundo M (raio maior) com o mesmo número de desvios padrão segundo P (raio menor) pelo valor de π (3,141592654).

Para se obter a área da elipse a um desvio padrão, tem-se então:

$$\text{Área}_{(a \ 1 \ \sigma)} = 3,141592654 \times \sigma_m \times \sigma_p$$

Se a distribuição em causa for normal, dentro da elipse a um desvio padrão estarão cerca de dois terços dos pontos da distribuição espacial em estudo, na elipse a dois desvios padrão cerca de 95% dos pontos e na de três desvios padrão mais de 99 por cento.

Exemplifiquemos o cálculo com o caso que estamos a tratar. Primeiro calcula-se a média das coordenadas X e Y; depois com as variáveis M e P, respectivamente a diferença entre as variáveis X e Y e as suas médias; de seguida, os produtos MP e M e P ao quadrado, segundo o quadro 11.5, abaixo. Convém calcular todos os valores com máximo de casas decimais possível e arredondar apenas no final dos cálculos, a fim de minimizar os erros e ter em conta os valores ponderados, quando for o caso.

i	X	Y	Peso	M	P	M ²	P ²	MP	M ² x Peso	P ² x Peso	MP x Peso
A	34	12	3	8,6	-5,87	73,96	34,42	-50,45	221,88	103,25	-151,36
B	30	25	10	4,6	7,13	21,16	50,88	32,81	211,60	508,84	328,13
C	15	5	4	-10,4	-12,87	108,16	165,55	133,81	432,64	662,20	535,25
D	20	30	5	-5,4	12,13	29,16	147,22	-65,52	145,80	736,09	-327,60
E	25	10	8	-0,4	-7,87	0,16	61,88	3,15	1,28	495,08	25,17
Σ	25,4	17,87				232,60	459,96	53,80	1013,20	2505,46	409,59

Quadro 11.5. – Tabela auxiliar para o Cálculo da Elipse de Dispersão

Aplicando as fórmulas que acima se apresentaram pode-se calcular todos os elementos da elipse de dispersão desta distribuição de cinco pontos com pesos diferentes.

Sendo:

$$(\sum (M^2 \times \text{peso}) - \sum (P^2 \times \text{peso})) = (1013,2 - 2505,47..) = -1492,27...$$

então:

$$\text{tg}\alpha = \frac{-1492,27 + \sqrt{(-1492,67)^2 + 4 \times (409,6)^2}}{2 \times 409,6} =$$

$$\text{tg}\alpha = \frac{-1492,27 + 1702,336}{819,200} = 0,2564325...$$

e sendo a tangente de α igual a 0,2564325, o ângulo α é igual a 0,25102356 radianos. E como cada radiano corresponde a 57,29577 graus é, então, fácil determinar o valor do ângulo α em graus, que no caso é:

$$\alpha = 14,38259^\circ$$

O ângulo é medido a partir do eixo dos **Y** no sentido dos ponteiros do relógio. Se a equação da tangente de α der como resultado um valor negativo, primeiro determina-se o ângulo a que corresponde, ignorando o sinal negativo e depois subtrai-se esse ângulo de 90° de modo a obter o verdadeiro ângulo de rotação.

Os desvios padrão segundo os eixos são calculados pelas fórmulas respectivas:

$$\sigma_x = \sqrt{2 \times \frac{1013,2 \times 0,9383 - 2 \times 409,6 \times 0,96866 \times 0,2484 + 2505,467 \times 0,0617004}{30}}$$

$$\sigma_y = \sqrt{2 \times \frac{1013,2 \times 0,617004 + 2 \times 409,6 \times 0,96866 \times 0,2484 + 2505,467 \times 0,9383}{30}}$$

Obtendo-se os valores de 7,78102485 para o desvio padrão segundo X e de 13,19217265 segundo Y. A elipse definida por estes eixos tem uma excentricidade de 1,6954 e uma área de 322,480200 e, se a distribuição espacial for normal segundo os dois eixos, então cerca de dois terços dos pontos analisados estarão no seu interior.

3. – POTENCIAIS

Uma das mais interessantes técnicas de análise em Geografia é a do cálculo dos **potenciais**. Ao invés de considerar os valores das variáveis afectos exclusivamente às áreas ou pontos a que dizem respeito, a noção de potencial considera que eles também afectam as áreas vizinhas/próximas, de acordo com uma função qualquer da distância (geralmente muito simples).

Isto vem tornar a análise geográfica que se faz mais próxima da realidade, pois os pontos ou áreas em que se medem as diferentes variáveis da geografia humana não são contentores fechados e isolados uns dos outros, mas sim espaços de interacção e de partilha dos elementos que neles existem.

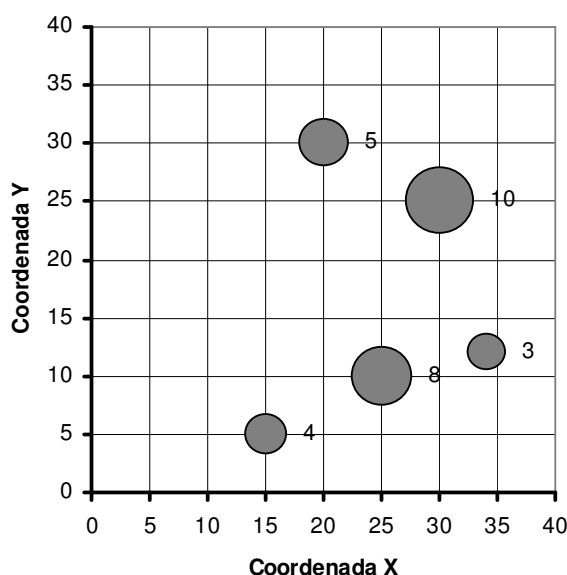


Figura 11.2 – Localização dos lugares do Quadro 11.1

O cálculo dos potenciais vai ser exemplificado com os dados da variável população por lugares, tomando por base a distribuição espacial de lugares do exemplo que se tem vindo a tratar (o exemplo da Figura 11.2, acima e do Quadro 11.1 que a seguir se repete no Quadro 11. 6, para melhor se seguir o exemplo).

i	X	y	Peso
A	34	12	3
B	30	25	10
C	15	5	4
D	20	30	5
E	25	10	8

Quadro 11.6 – Repetição do exemplo do Quadro 11.1

Neste quadro constam cinco lugares (A, B, C, D e E) em que a respectiva população está representada pela variável Peso, em milhares de habitantes e a localização é indicada pelas correspondentes coordenadas (X e Y).

Assim, por exemplo, pode ver-se que o lugar A com 3.000 habitantes tem a coordenada X de 34 e Y de 12.

A sua distribuição pelo espaço será, então, algo similar ao que se mostra na Figura 11.2, onde, para além da posição relativa dos lugares, se indica também a sua população em milhares de habitantes.

Nele se vê que é pouco provável que as populações não se afectem umas às outras: Pode-se identificar dois conjuntos, um a Norte com dois lugares e outro a Sul com três e é mesmo provável que todos tenham alguma interacção entre si, tanto menor quanto maior é a distância.

O conceito de Potencial de uma variável num lugar é geralmente mais adequado em Geografia Humana do que o simples valor da variável nesse lugar, ao considerar não apenas o peso (no exemplo a população) que existe nesse lugar mas, para além dessa, a que existe na vizinhança e que com ela interage mais ou menos, consoante a distância entre os lugares e também a função de que se está tratando³.

O Potencial de uma variável do lugar K (P_k), calcula-se da seguinte forma:

$$P_k = \sum_{i=1}^n \left(\frac{p_i}{d_{ki}^\alpha} \right)$$

em que d_{ki} é a distância entre o lugar de referência k e os outros (incluindo ele próprio) e o parâmetro α (alfa) pode ser muito complexo, mas geralmente considera-se com o valor de um ou dois, o que significa, no primeiro caso, que se entende que a passagem da distância para o dobro se traduzirá na da interacção para metade, numa forma proporcional à distância e no segundo caso, que a interacção passará a um quarto, de um modo proporcional à área do círculo cujo raio é a distância.

Como se poderá verificar pela fórmula de cálculo, este conceito está intimamente relacionado com os modelos gravíticos e é um bom indicador da acessibilidade relativa dos lugares.

Tem tido desenvolvimentos recentes, devido à sua melhor adaptação à análise dos fenómenos em Geografia Humana, como se disse. Como exemplo, refira-se que no estudo do EDEC (Esquema de Desenvolvimento do Espaço Comunitário), efectuado recentemente por uma série de organismos internacionais para a EU e de que o representante português foi o Centro de Estudos Geográficos da Universidade de Lisboa, a maioria das variáveis analisadas foram-no em termos de potenciais.

³ - Veja-se Christaller e a sua teoria dos lugares centrais sobre as diferentes funções e a respectiva facilidade de deslocação.

Deve ainda referir-se que é um dos poucos métodos que é relativamente insensível às mudanças da base territorial das variáveis, ficando como alternativa única para comparar distribuições com bases espaciais diferentes.

Cabe agora referir que em toda a análise espacial que temos vindo a desenvolver se considera que os valores das variáveis das áreas elementares (ou lugares, ou ainda pontos) em que se baseia a análise, estão representados por pontos (no caso das áreas o respectivo centro).

As distâncias medidas também podem ser variadas (veja-se a propósito o que escrevemos sobre taxionomias numéricas).

Assumiu-se que a distância de que se faria uso seria a Euclidiana, tal como decorre do teorema de Pitágoras:

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

É preciso ainda considerar o valor da distância quando $j = i$, isto é, determinar qual é a distância de uma área a si mesma. Como as áreas estão representadas por pontos, a distância entre duas áreas é a distância entre os seus pontos representativos. Matematicamente a distância de um ponto a si mesmo é zero, o que equivaleria a ter uma interacção de valor infinito. Mas de facto, a distância entre dois pontos é a simplificação média do conjunto de distâncias que se podem efectuar entre uma área e a outra.

Considerando assim que a distância quando $j = i$, é a que resulta das distâncias de todos os elementos da área entre si, pode dizer-se que esta distância existe, não é nula e não pode ultrapassar a dimensão da área. Geralmente é assumida como uma função do espaçamento dos lugares (pontos ou áreas em análise). Para estimar o seu valor têm sido propostas várias soluções, a mais simples das quais é a que aqui se aceita, de considerar que é igual a metade da menor distância entre esse lugar e cada um dos outros, tal como:

$$d_{i,i} = \frac{1}{2} \times \min(d_{i,j})$$

Exemplifica-se de seguida a sequência de cálculos, que se aconselha a fazer, sempre que possível, usando uma folha de cálculo (Excel ou outras), com os dados que se tem vindo a tratar, constantes do Quadro 11.5.

Começemos por elaborar a matriz das distâncias euclidianas entre todos os cinco lugares (Quadro 11.7).

	A	B	C	D	E
A	0,00000	13,60147	20,24846	22,80351	9,21954
B	13,60147	0,00000	25,00000	11,18034	15,81139
C	20,24846	25,00000	0,00000	25,49510	11,18034
D	22,80351	11,18034	25,49510	0,00000	20,61553
E	9,21954	15,81139	11,18034	20,61553	0,00000

Quadro 11.7 – Distâncias euclidianas entre os pontos

De seguida estima-se, do modo descrito, os valores da distância de cada ponto a si mesmo, os valores da diagonal, que vão ser metade da distância mais pequena em cada coluna ou linha, pois a matriz é simétrica (Quadro 11.8):

	A	B	C	D	E
A	4,60977	13,60147	20,24846	22,80351	9,21954
B	13,60147	5,59017	25,00000	11,18034	15,81139
C	20,24846	25,00000	5,59017	25,49510	11,18034
D	22,80351	11,18034	25,49510	5,59017	20,61553
E	9,21954	15,81139	11,18034	20,61553	4,60977

Quadro 11.8 – Distâncias euclidianas entre os pontos

Falta agora assumir um valor para o parâmetro α (alfa). Como se disse, geralmente utiliza-se o valor **2** (dois) por ser o que corresponde à relação de aumento da área de interacção com o raio. Se houver razões teóricas que aconselhem outro valor, deve então ser esse o utilizado.

Exemplifica-se o cálculo do potencial do lugar A para o valor de alfa igual a dois:

$$P_A = \frac{3}{4,60977^2} + \frac{10}{13,60147^2} + \frac{4}{20,24846^2} + \frac{5}{22,80351^2} + \frac{8}{9,21954^2}$$

$$P_A = \frac{3}{21,25} + \frac{10}{185} + \frac{4}{410} + \frac{5}{520} + \frac{8}{85} =$$

$$P_A = 0,14118 + 0,05405 + 0,00976 + 0,00962 + 0,09412 = 0,30872$$

Como é evidente, os potenciais dos outros quatro lugares calculam-se de maneira semelhante, obtendo-se os resultados que estão no primeiro exemplo do anexo.

O conhecimento do valor dos potenciais da variável, calculado com o alfa adequado (habitualmente dois, como se disse), é geralmente um valor mais correcto para a variável do que os valores iniciais. Na realidade, o valor de alfa faz entrar a interacção possível entre todos os lugares, oscilando entre os valores extremos de alfa igual a 0 – o que significa que a distância é completa e absolutamente irrelevante para o fenómeno em causa e todos os lugares têm um potencial igual ao total da variável em causa – e o valor de alfa infinito – o que tem o significado de que um só milímetro a mais impede toda a interacção e cada lugar tem o valor de potencial igual ao da sua própria variável. Os valores intermédios privilegiam mais ou menos cada uma das situações extremas, conforme estão mais ou menos distanciados de cada uma delas.

4. – ANÁLISE DO VIZINHO MAIS PRÓXIMO

A análise de distribuições pontuais numa dada superfície é uma das mais antigas e fecundas tarefas em Geografia.

O estudo de fotografias aéreas ou de imagens de satélites e a construção de mapas com a distribuição de casas, árvores, povoações, cidades, fábricas ou, a outra escala, a observação de entidades biológicas em jardins, florestas etc., é um trabalho analítico típico dos estudos geográficos.

Uma das técnicas elementares que permite analisar a vizinhança é a **Análise do Vizinho mais Próximo**.

A vizinhança (de um ponto qualquer, de referência) é um conceito complexo, mas numa distribuição espacial de pontos não é difícil de calcular, pois terá sempre a ver com uma qualquer função das distâncias entre o ponto de referência e os que lhe estão próximos.

A forma mais simples é considerar apenas um valor, o da distância entre o ponto de referência e o ponto que se encontra mais próximo.

A média das distâncias ao vizinho mais próximo de todos os pontos de uma área é uma característica dessa área, que vai dos valores mais pequenos – indicativo de concentração, – até aos valores mais elevados – que indicam espaçamento máximo e, portanto, uma tendência para a dispersão regular.

O método do vizinho mais próximo baseia-se na comparação entre a distância média ao vizinho mais próximo efectivamente verificada na realidade e diferentes distâncias teóricas correspondentes a distribuições concentradas, aleatórias e ordenadas.

4.1.- O Índice R

Para efeitos de cálculo considera-se cada localização como um ponto, independentemente do seu peso (o valor da variável que lhe está associado), pois esta técnica é um método de análise de distribuição de pontos e não de uma variável localizada. Se estamos a analisar uma

variável localizada, passamos a uma distribuição pontual, dando a cada localização o mesmo peso (a unidade) e, se for caso disso, eliminando os pontos que correspondam a valores da variável que se situem abaixo de um dado parâmetro de corte.

A distância média ao vizinho mais próximo de uma distribuição totalmente concentrada de pontos é, obviamente, igual a zero.

A distância média ao vizinho mais próximo de uma distribuição aleatória de pontos pode calcular-se, segundo a distribuição de probabilidade de Poisson, como:

$$\bar{d}_A = \frac{1}{2 \times \sqrt{\frac{n}{A}}}$$

em que ***n*** é o número de pontos na distribuição e ***A*** é a área em estudo.

O Índice ***R***, que se expressa pela fórmula

$$R = \frac{\bar{d}_o}{\bar{d}_A}$$

compara a média das distâncias ao vizinho mais próximo observadas na distribuição em estudo ao valor médio teórico da distribuição aleatório do mesmo número de pontos em idêntica área, obtido pela expressão anterior.

O seu valor varia entre 0 (zero) e cerca de 2,149, com o significado que se indica no Quadro 11.9.

Valor de R	Padrão
0	concentrado
1	aleatório
2,149	regular

Quadro 11.9 – Valores de R e respectivo significado

O valor de R igual a 2,149 é o que corresponde à distância observada numa rede de lugares ordenada de tipo hexagonal ilimitada. A rede hexagonal é a rede regular mais densa que se pode ter num plano e, por isso, este valor é o máximo teórico da dispersão regular (se a rede é ilimitada ou a área muito grande relativamente ao número de pontos). Porém, se a maioria dos pontos tenderem a estar localizados próximo da fronteira da área em estudo (mesmo que de forma regular), o valor obtido para o índice R pode ser maior de que este máximo teórico, por a área ser “artificialmente” reduzida face à distribuição dos pontos como mais adiante se refere, ao tratarmos dos problemas que podem afectar os resultados e a interpretação desta técnica analítica.

4.2.- Teste de Significado

Aos valores obtidos pelo índice R pode aplicar-se um teste de significado estatístico, a fim de avaliar a probabilidade do valor R que se calculou ser devida a um simples acaso.

A hipótese nula (**H₀**) é a de que o valor do Índice R resulta dos pontos estarem localizados ao acaso na área analisada. O teste estatístico utilizado é o do Coeficiente C.

A forma de calcular o valor do Coeficiente C é a seguinte:

$$C = \frac{\bar{d}_O - \bar{d}_A}{SE_{do}} \qquad SE_{do} = \frac{0,26136}{\sqrt{n^2/A}}$$

com as distâncias que já se conhecem e em que **n** é o número de pontos na distribuição e **A** a área em estudo.

A distribuição do Coeficiente C segue uma distribuição normal. Valores positivos indicam um padrão para a dispersão e valores negativos um padrão concentrado. Por outro lado, pode calcular-se a probabilidade do valor C se situar entre determinados valores, a partir da consulta da tabela de distribuição normal. Como para qualquer distribuição normal, existe uma probabilidade de 0,682 de o valor de C estar entre -1,0 e +1,0 e uma probabilidade de 0,954 de se obter um valor de C situado entre -2,0 e +2,0 relativamente à hipótese nula.

4.3.- Considerações finais

A análise do vizinho mais próximo é uma técnica muito útil para estudar distribuições variadas, como seja de casas, lugares, povoações, plantas, ocorrência de infra-estruturas do mesmo tipo, etc., embora relativamente pouco utilizada em Portugal (cf. Abreu, 1978).

É certo que existem alguns problemas na sua aplicação:

1.- Os pesos diferentes de cada localização. Como acima referimos a técnica é exclusivamente para comparação de localizações e não de variáveis com localização espacial, pelo que só se pode utilizar para localizações definidas por um par de coordenadas (X e Y), todas com o mesmo peso.

Alguns programas de Geoestatística que tratam de distribuições espaciais e variáveis localizadas, fazem automaticamente a passagem dos diferentes pesos para um único valor unitário.

2.- O valor teórico da distância média numa distribuição aleatória não considera que, para qualquer área, os pontos que se localizam junto à

fronteira da área em estudo têm menores vizinhanças que os que se localizam no centro. Com efeito, o valor que se obtém para a distância média teórica da distribuição aleatória é o que resultaria de uma infinidade de pontos numa área infinita.

Uma forma de ultrapassar esta restrição é a de considerar que para os pontos junto à fronteira o vizinho mais próximo pode situar-se fora da área em análise (o que equivale a considerar em torna da nossa área de referência como que uma área de protecção, cujos elementos não são pontos que pertençam à área em análise, mas contam para a determinação das distâncias mínimas referidas aos que se situam no interior). É uma forma aceitável em alguns estudos, por exemplo, na consideração dos vizinhos mais próximos de sedes de freguesia, ou lugares onde existe um dado bem ou serviço em que a área de referência é a do concelho, mas a proximidade concorrente do bem pode e deve fazer-se tomando em conta os pontos exteriores, mas que nada significa noutras circunstâncias, como seja a de análise de ilhas ou outros territórios isolados.

3.- O outro problema é de que o valor calculado para o Índice R vem afectado pela área de referência (A) e pela sua forma. Geralmente é uma área administrativa e deve ser fornecido externamente pelo operador ao programa de Geostatística utilizado. Quando se desconhece, muitas vezes assume-se que é o rectângulo de coordenadas que circunscreve a distribuição dos pontos avaliados. Como este rectângulo não é invariante com a eventual rotação dos eixos de referência X e Y, os valores calculados podem vir ser afectados por erros com esta origem.

4.4.- Exemplo de cálculo

Considere-se como de costume os valores do Quadro 11.5, que representam uma distribuição de cinco lugares A, B, C, D e E (todos com o mesmo peso unitário).

1.- A área do rectângulo circunscrito calcula-se através do valor máximo e mínimo de cada uma das coordenadas dos pontos (Quadro 11.1 e 11.5) que são, respectivamente, para a coordenada X de 34 e 15 e para a coordenada Y de 30 e 5. Assim sendo, a área do rectângulo circunscrito dos cinco pontos é igual a 475, o produto de 19 por 25.

2.- A distância média teórica dos cinco pontos com os seus vizinhos mais próximos, se a sua distribuição fosse aleatória, seria:

$$\bar{d}_{\text{aleat}} = \frac{1}{2 \times \sqrt{\frac{5}{475}}} = \frac{1}{2 \times \sqrt{0,010526316}} = \frac{1}{2 \times 0,102597835} = \frac{1}{0,20519567} = 4,87340..$$

enquanto a distância média observada dos cinco pontos com os seus vizinhos mais próximos é igual a 10,39602, conforme resulta dos valores constantes do Quadro 11.10, abaixo:

Pontos	Distância ao vizinho mais próximo
A	9,21954
B	11,18034
C	11,18034
D	11,18034
E	9,21954

Quadro 11.10 – Distâncias ao Vizinho mais próximo

E, assim sendo, o valor do Índice R = 2,13322. Este valor indica que a distribuição de pontos tende para a regularidade.

3.- O valor do Coeficiente C (em desvios padrão) mede a probabilidade da hipótese nula – a distribuição ser aleatória – se verificar, no sentido de quanto mais perto de zero, mais provável é ser aleatória, sendo o desvio positivo indicativo do número de desvios padrão relativamente à distribuição regular e o desvio negativo o mesmo relativamente à distribuição concentrada.

$$C = \frac{10,39602 - 4,87340}{\frac{0,26136}{\sqrt{5^2/475}}} = \frac{5,52262}{1,139241828} = 4,84763..$$

No caso em apreciação, o valor do Coeficiente C, que é de 4,84763, permite rejeitar completamente a hipótese nula, de que a distribuição é aleatória. Com o valor de C a quase cinco desvios padrão, mais precisamente 4,84763, a probabilidade de ocorrência dessa hipótese é inferior a 0,6 por milhão. Como ocorre no sentido positivo, reforça a ideia desta distribuição de pontos se encontrar mais no sentido da dispersão.

5.- LEITURA DO “Output”

1. – Exemplo de dados e de resultados de um programa de Geoestatística:

i	x	Y	Peso
A	34	12	3
B	30	25	10
C	15	5	4
D	20	30	5
E	25	10	8

 ***** DIOGO & ROQUETTE SOFTWARE *****

**** PACOTE GEOSTATISTICO ****

NOME DO TRABALHO: Teste 1

 Coordenadas do centro gravidade

X = 25.40000000
 Y = 17.86666667

Distância-média = 10.39018484
 Distância padrão (DP) = 10.82999949
 Área do Círculo = 368.47390000

 Raios Dinâmicos dos pontos

No.	Coord. X	Coord. Y	Peso	Valor	Relação c/DP
1	34.0000	12.0000	3.0000	15.834795	1.462123
2	30.0000	25.0000	10.0000	16.852300	1.556076
3	15.0000	5.0000	4.0000	21.240382	1.961254
4	20.0000	30.0000	5.0000	18.772320	1.733363
5	25.0000	10.0000	8.0000	15.637949	1.443947
Raio dinâmico do centro médio =				10.829999	1.000000

ELIPSE DE DISPERSÃO

Ângulo de rotação do eixo dos Y = 14.38259 graus
 (sentido ponteiros do relógio)

Desvio padrão segundo novo X = 7.78102485
 Desvio padrão segundo novo Y = 13.19217265

Comprimento do eixo X = 15.562050
 Comprimento do eixo Y = 26.384345

Área da elipse = 322.480200
 Razão eixo maior/menor = 1.6954

POTENCIAIS (alfa = 2.000)

No.	Coord. X	Coord. Y	Peso	Potencial
1	34.00000	12.00000	3.00000	.30872
2	30.00000	25.00000	10.00000	.41462
3	15.00000	5.00000	4.00000	.22301
4	20.00000	30.00000	5.00000	.27075
5	25.00000	10.00000	8.00000	.49553

 ANÁLISE DO VIZINHO MAIS PRÓXIMO

Área = 475.000000
 Índice R = 2.133219
 Coeficiente C = 4.847816

6.- EXERCÍCIOS

Problema 1 – Para o exemplo do quadro 11.1, calcule os potenciais de população assumindo o valor de alfa igual a 1.

Resposta: 2,671; 3,123; 2,175; 2,465; 3,294

Problema 2 – Considere o seguinte exemplo de 10 localizações (todas com o mesmo peso):

Lugar	Coordenada X	Coordenada Y
A	5	9
B	2	8
C	3	8
D	1	7
E	4	7
F	2	6
G	8	4
H	7	2
I	9	2
J	6	1

- Calcule as coordenadas **X** e **Y** do centro de gravidade.
- Calcule a distância média e a distância padrão.
- Consultando a tabela da distribuição normal, determine qual o raio e qual a área da circunferência correspondente à probabilidade de 95% dos pontos estarem no seu interior. Note que o valor acumulado da distribuição normal 0,950000000 corresponde a 1,644853 desvios padrão.
- Calcule os parâmetros da elipse de dispersão: ângulo de rotação de Y e desvios padrão segundo X e Y.
- Para a mesma probabilidade da alínea c), determine os raios maior e menor e a área da elipse de dispersão correspondente.
- Calcule os raios dinâmicos dos dez pontos.
- Calcule também os potenciais para alfa igual a um e igual a dois.
- Calcule os valores da análise do vizinho mais próximo: área do rectângulo circunscrito, Índice R e Coeficiente C.
- Comente e interprete os valores que calculou.

Respostas:

- X=4,7; Y=5,4
- Dm = 3,67700541 DP = 3,80131556
- Para 95% Raio = 6,252607127 Área = 122,820866040
- Ângulo=47,29025° DPy = 2,01415118 DPx = 4,98429477
- Raio Maior = 8,198434598 Raio menor = 3,312983578
Área = 85,329675150

f) e g)

Lugar	Raios dinâmicos	Potencial Alfa = 1	Potencial Alfa = 2
A	5,5277	3,121	1,485
B	5,6273	5,470	6,113
C	5,1747	5,593	6,175
D	5,8405	4,326	3,422
E	4,4096	4,467	3,326
F	4,9554	4,346	3,295
G	5,5076	3,022	1,432
H	5,8973	3,928	3,075
I	7,0317	2,985	1,638
J	6,2805	3,537	2,791

h) Área = 64 R = 1,228798 C = 1,384203

Problema 3 – Dados os três exemplos do quadro abaixo, em que estão as respectivas coordenadas X e Y, calcule para cada um dos casos o centro de gravidade, a distância padrão e a área do respectivo círculo e os parâmetros das respectivas elipses de dispersão.

EXEMPLO A	
1	4
1	3
2	1
3	2
4	3
5	1
5	0

Exemplo B	
2	3
2	6
2	7
3	4
4	2
4	1
4	5

Exemplo C	
2	2
4	2
4	3
6	3
6	4
8	4

Respostas: Exemplo A Centro de gravidade $X = 3$ $Y = 2$
 Distância padrão = 2,07019668; Área do círculo = 13,46397
 Elipse Alfa = $36,65037^\circ$ $DP_x = 2,69604502$ $DP_y = 1,14138930$
 Área da elipse a um desvio-padrão = 9,667425

Exemplo B Centro de gravidade $X = 3$ $Y = 4$
 Distância padrão = 2,20389226; Área do círculo = 15,25916
 Elipse Alfa = $71,98630^\circ$ $DP_x = 2,95690328$ $DP_y = 0,98539758$
 Área da elipse a um desvio-padrão = 9,153738

Exemplo C Centro de gravidade $X = 5$ $Y = 3$
 Distância padrão = 2,081666; Área do círculo = 13,61357
 Elipse Alfa = $69,18322^\circ$ $DP_x = 0,56521541$ $DP_y = 2,88915176$
 Área da elipse a um desvio-padrão = 5,130199

Problema 4 – Calcule e cartografe os raios dinâmicos da população para os dezoito distritos do continente em 1991 e 2001. Compare os valores e saliente as conclusões mais importantes.

Problema 5 – Considere em diversos censos (1900, 1920, 1940, 1960, 1981 e 2001) a população dos dezoito distritos do continente. Para

cada caso compare os respectivos centros de gravidade a distância padrão e as elipses de dispersão. Compare e interprete.

Problema 6 – A partir de duas variáveis por NUT's 3, a população com menos de 15 anos e a com mais de 65, em 2001, faça um programa em Excel ou utilize um já feito, que lhe permita calcular os centros de gravidade e os potenciais das variáveis. Compare e interprete os resultados obtidos para os centros de gravidade. Para os potenciais compare os resultados para alfa igual a um e dois com os valores das variáveis iniciais. Compare e conclua.

Problema 7 – Para Portugal continental, faça uma análise do vizinho mais próximo para as cidades existentes em 2001. Calcule o Índice R e o Coeficiente C e realce as conclusões mais significativas.

Problema 8 – Faça duas análises do vizinho mais próximo das sedes de concelho da ilha da Madeira, utilizando as coordenadas quilométricas UTM, numa com a área do rectângulo circunscrito, noutra com a área real da ilha (em Km²). Interprete os resultados. E para Portugal continental, faça uma análise do vizinho mais próximo para as capitais de distrito, nas mesmas condições.

7.- BIBLIOGRAFIA

- ABREU, Diogo de (1978) – ***Estudo Estatístico de Algumas Características do Povoamento em Portugal***. Faculdade de Letras de Lisboa, Lisboa, 121 p.
- (*) EBDON, David (1996) – ***Statistics in Geography. 2th edition***, Blackwell, Oxford (UK), p. 128-169.
- (*) ESTEBANEZ ALVAREZ, J.; BRADSHAW, R.. P. (1978) – ***Técnicas de Quantificación en Geografía***, Editorial Tebar Flores, Madrid, p. 391-465.
- BRONSTEIN, I.; SEMENDIAEV, K. (1979) – ***Manual de Matemática para Engenheiros e Estudantes***, 2ª edição, Editorial Mir, Moscovo, p.271.

Impressão e acabamento
David J.B. Barreira
Faculdade de Letras da Universidade de Lisboa
Alameda da Universidade
1600-214 Lisboa
Tel./Fax: 21 797 20 49
Email: davidjbbarreira@sapo.pt

