

Universidade de Évora

Mestrado em Modelação Estatística e Análise de Dados

Estudo Estatístico dos Fluxos Migratórios em Portugal, no Período de 1991 a 2001

Filipe José Batista Ribeiro

2010

Orientador: Professora Doutora Maria Filomena Mendes

Co-Orientador: Professor Doutor Paulo Infante

Esta dissertação não inclui as críticas e sugestões feitas pelo júri

Universidade de Évora

Mestrado em Modelação Estatística e Análise de Dados

Estudo Estatístico dos Fluxos Migratórios em Portugal, no Período de 1991 a 2001

Filipe José Batista Ribeiro

2010

Orientador: Professora Doutora Maria Filomena Mendes

Co-Orientador: Professor Doutor Paulo Infante

Esta dissertação não inclui as críticas e sugestões feitas pelo júri

Índice

1.	Introdução	6
1.1.	Breve revisão da literatura.....	7
1.1.1.	Contextualização das Migrações em Portugal	9
1.1.2.	Teorias das Migrações	12
1.1.3.	Novas Abordagens ao Estudo das Migrações	15
1.2.	Origem dos dados e contextualização.....	16
2.	Caracterização dos Indivíduos em Estudo.....	18
3.	Alguns Aspectos Teóricos.....	28
3.1.	Comparação de populações a partir de amostras independentes	28
3.2.	Regressão logística e modelos lineares generalizados	34
3.3.	Modelação: Regressão Logística	36
3.4.	O teste de independência do χ^2	39
4.	Estimação Indirecta dos Fluxos Migratórios Internos	41
4.1.	Especificidades Etárias.....	41
4.2.	A Estimação Baseada na Propensão Migratória das Crianças	44
4.3.	A Estimação Baseada na Migração Líquida Específica por Idade	48
4.4.	A Estimação Baseada em duas Distribuições Sucessivas de Stocks Populacionais	52
4.5.	Avaliação Crítica dos Resultados	55
5.	Redução do Erro da Estimativa Indirecta	63
5.1.	<i>“A Estimação da Curva”</i>	63
5.2.	Conjugação Metodológica	68
5.3.	A estrutura Etária.....	72
6.	Análise Estatística das Migrações Internas em Portugal Continental	73
6.1.	A Residência cinco anos antes.....	74
6.2.	A região de nascimento.....	84
6.3.	Interacções regionais	93
7.	Conclusão/Discussão de Resultados	94
	Bibliografia	96

Estudo Estatístico dos Fluxos Migratórios em Portugal, no Período de 1991 a 2001

Resumo:

A falta de um registo sistemático associado à problemática das migrações resulta numa situação bastante complexa, pois torna inexecutável a sua contabilização. A alternativa reside, assim, na aplicação de metodologias de estimação indirecta que nos permitam avaliar e quantificar os fluxos migratórios inter-regionais. Esta temática tem sido alvo de estudo por parte de diversos investigadores, no entanto, *Andrei Rogers* surge como aquele que mais se destaca, desenvolvendo, em conjunto com diversos investigadores, metodologias que nos permitem, além de analisar e quantificar indirectamente os diferentes comportamentos migratórios inter-regionais, identificar os padrões migratórios a eles associados.

Por outro lado, a elaboração de uma análise estatística das migrações internas, permitir-nos-á também, complementar os resultados obtidos através das metodologias de estimação indirecta aplicadas.

Deste modo, durante a elaboração deste estudo procedeu-se a uma breve introdução da temática das migrações, seguida de uma breve revisão da literatura, da estimação indirecta dos movimentos migratórios internos em Portugal Continental, e uma análise estatística das já referidas migrações.

De forma a consolidar esta dissertação foram abordados alguns aspectos teóricos sobre a metodologia aplicada, bem como algumas críticas aos resultados obtidos.

A Statistical Study of Migration Flows in Portugal (1991 – 2001)

Abstract:

The nonexistence of a systematic process of collecting background information about migration results in a quite complex situation, because it makes impossible to estimate these movements accurate and reliably. One alternative is therefore the application of indirect estimation methods that allow us to assess and quantify the inter-regional migration flows. This subject has been studied by several authors; however, *Andrei Rogers* emerges as one of the pioneers, developing, together with other researchers, new methodologies that allow us, to quantify and analyze the different migratory inter-regional behaviors, and simultaneously to identify different migration patterns associated with those flows.

On other hand, a statistical approach to internal migration flows will help us also to complement the results achieved by the indirect estimation methodology.

Thus, during this essay, it was done a brief introduction about the migration topic, followed by a brief revision of literature, the indirect estimation of internal migration within the Portuguese Mainland, and a statistical approach to the migration study.

With the intention of a secure presentation of the used approach, some theoretical aspects were presented and some critics to results were done.

1. Introdução

Contrariamente à fecundidade e mortalidade em que a inércia caracteriza a dinâmica demográfica, no presente e futuro próximo, as migrações assumem-se como a componente mais volátil da evolução demográfica surgindo a necessidade de se realizarem projecções/estimativas demográficas que se ajustem à previsão de fluxos migratórios futuros.

Por outro lado, a realização de uma abordagem estatística, permitir-nos-á, além da sua caracterização, identificar quais os factores que mais “contribuem” para a existência destes movimentos, identificando, entre outros aspectos, quais as regiões que são mais atractivas ou mais repulsivas.

Assim, o objectivo principal desta dissertação reside na caracterização dos fluxos migratórios internos ao nível das NUTS II constituintes do continente português, e os seus objectivos específicos pela caracterização do padrão dos fluxos migratórios, em termos de idade, sexo, e região de origem/destino; caracterizar as migrações efectivas através de outros indicadores adicionais considerados pertinentes; observar e caracterizar a evolução registada entre os dois recenseamentos gerais da população mais recentes em Portugal; aplicação da metodologia de estimação indirecta dos fluxos migratórios inter-regionais a Portugal (Rogers *et al.*, 2003); e por último, comparar, os resultados obtidos com os de Espanha.

Deste modo, os capítulos seguintes desta dissertação corresponderão à aplicação das metodologias de estimação indirecta de Rogers *et al.* (2003), isto após a realização de uma breve revisão da literatura e da apresentação de alguns aspectos teóricos sobre as ferramentas estatísticas por nós utilizadas. De seguida, procedeu-se à elaboração de um modelo de regressão (logística) que nos permitisse identificar quais as variáveis associadas ao acto de migrar, bem como de outras metodologias complementares para a análise.

Por último, seguir-se-á uma conclusão/discussão de resultados, onde iremos referir as ideias principais desta dissertação, e estabelecer a comparação possível com resultados obtidos por outros investigadores.

1.1. Breve revisão da literatura

A origem do fenómeno migratório passará muito para além de algumas vivências mais recentes. Muito pelo contrário, desde os primórdios da raça humana se têm registado diversos movimentos populacionais cujo incitamento se deve às mais variadas questões. Se nos nossos dias, e referindo-nos inicialmente aos países mais desenvolvidos, as migrações se encontram essencialmente associadas a questões de cariz económico (Genereux, 2007), tempos houve em que estas se deveram simplesmente à procura de mais alimentos, ou de terrenos que providenciassem a subsistência familiar.

No entanto, estes são somente alguns dos fundamentos que podemos encontrar na questão migratória, pois não raras vezes estes movimentos se encontram associados a fugas de guerras fomentadas pela sede de poder que alimenta todas as nações e culturas. Dada esta perseverança do fenómeno migratório constatamos que, independentemente das suas diversas formas e motivações, os movimentos populacionais têm sido uma constante ao longo da história da humanidade, e que por detrás de uma história da humanidade se encontra associada uma história das migrações (Marques, 2008).

Os movimentos populacionais podem, assim, ser observados nas suas diferentes formas e tipologias, ou seja, como invasões, conquistas, colonizações, ou simples movimentações entre países e regiões, e de tipologia mais ou menos duradoura.

Apesar da referida perseverança dos movimentos migratórios no decorrer dos tempos, o estudo das migrações é sempre difícil, principalmente devido ao seu carácter conjuntural. A inércia demográfica que caracteriza a mortalidade e a fecundidade (Carrilho e Patrício, 2007) não se observa do mesmo modo em relação aos movimentos migratórios, e a falta de registo das mudanças de residência, sobretudo a nível interno, não permite um conhecimento do comportamento dos migrantes que possibilite sustentar projecções demográficas, sobretudo no respeitante aos movimentos de base a nível regional.

Esta situação coloca-nos perante a problemática da inexistência de dados, o que não nos permite analisar ou estimar de forma “eficiente” os fluxos migratórios, quer a nível nacional quer a nível internacional. Porém, e no que concerne à temática das migrações internas, uma das soluções passa pela utilização dos dados relativos aos recenseamentos gerais da população, pois trata-se de uma base (considerada) mais fidedigna.

Embora os dados relativos ao recenseamento façam parte da solução, deparamo-nos, ainda, com a problemática da sua confidencialidade, ou seja, o difícil acesso aos micro-

dados do próprio recenseamento, e com a sua amplitude temporal bastante alargada (10 anos), possibilitando-nos apenas realizar uma análise mais abrangente do que profunda. Esta situação conjuntural direcciona-nos para a realização de estimativas que nesta situação se tornam mais viáveis através da utilização de metodologias de estimação indirecta que nos permitam projectar para o país e desagregar ao nível regional, ou projectar para as regiões e agrupar para o país. No entanto, quer num caso, quer noutro, a estimação das migrações internas é fundamental, não só em termos demográficos, mas também em termos de planeamento, porque as informações sobre a capacidade de atracção das regiões devem ser tidas em conta no desenho de um desenvolvimento futuro das regiões e do país, de forma equilibrada e sustentável.

Torna-se então necessário contrariar as deslocações em massa para as regiões do litoral (Peixoto e Oliveira, 2001), onde se concentra o maior número de pessoas e as cidades são mais desenvolvidas. Estes movimentos migratórios entre o meio rural e o meio urbano resultam, assim, na desertificação do interior e seu consequente envelhecimento populacional, o que cria um afastamento económico e até cultural cada vez maior entre as diversas regiões continentais, permitindo a identificação de uma fronteira clara entre um grupo de regiões com maior e menor mobilidade, em que o Alentejo e o Algarve são as regiões com maior mobilidade e o Norte com menor (Magalhães, 2003).

Além dos volumes dos fluxos migratórios e das suas movimentações (origem/destino), é ainda crucial identificar as causas associadas aos diferentes movimentos migratórios, bem como definir um perfil por idades que nos permita conhecer as grandes tendências em função da idade.

A análise da estrutura espacial leva a que seja mais fácil compreender a influência da migração nas mais variadas regiões, conseguida através da comparação entre a estrutura espacial da migração e a própria estrutura populacional da região sem migrações (Rogers *et al.*, 2002).

Devido à escassez de dados de base para o estudo das migrações internas e internacionais, situação que também se regista em Portugal, têm vindo a ser envidados esforços que permitam ultrapassar esta lacuna, tal como a realização de estimativas indirectas dos fluxos migratórios (Rogers *et al.*, 2003) e a reconstituição dos *stocks* de população imigrante (Bijak e Kupiszewska, 2008).

Com o intuito de contrariar esta tendência, Rogers (1975) desenvolveu uma metodologia de projecção “multistate” e “multiregional”, que parte da região de

residência dos próprios migrantes, permitindo-nos identificar quais são as mais atractivas (ou repulsivas).

No entanto, esta metodologia tem sido alvo de algumas extensões a temáticas como a escolaridade (Lutz *et al.*, 1999; Lutz e Goujon, 2001), a grupos específicos de mão-de-obra (Siegel, 2002; Smith *et al.*, 2000; van Imhoff e Henkens, 1998), ou a regiões de origem ou nascimento (Rogers *et al.*, 1999, 2003; Ledent, 1980).

A. Rogers é um dos principais investigadores entre os que se têm dedicado à temática da estimação indirecta dos fluxos migratórios internos, permitindo, apesar da referida escassez de dados, observar, analisar e prever, os diversos movimentos populacionais dentro de uma determinada região ou país.

1.1.1. Contextualização das Migrações em Portugal

Os estudos relativos às migrações têm vindo a ganhar a atenção dos mais variados investigadores a nível mundial, isto porque desde sempre se verificam elevadas movimentações de pessoas por todo o mundo. Os estudos realizados são, na sua maioria, direccionados para as migrações internacionais, mas nunca deixando totalmente de parte a própria realidade dos vários países. Portugal também não tem sido excepção, pois encontram-se vários trabalhos realizados que fazem alusão a esta temática.

As migrações são normalmente analisadas a nível internacional (Jennissen, 2003), não existindo muitas comparações entre países, mas sim análises próprias da sua realidade. Sem nunca se deixar de parte a vertente internacional, a investigação sobre estes temas centra-se essencialmente nos factores que induzem a migração (Abreu, 2004), e nos próprios indivíduos que migram, de modo a determinar também a sua condição social e as suas qualificações. O estudo da distribuição espacial da população portuguesa tem sido cada vez mais deixado de parte, dando lugar a uma análise mais ao nível imigratório (Machado, 2003), principalmente desde a entrada de países de Leste para a União Europeia (UE) (Pires, 2002) e, ainda que, em menor abundância, para a emigração, analisando os “novos” emigrantes e a sua redistribuição, deixando de parte as migrações internas. As migrações em Portugal sofreram grandes alterações ao longo dos tempos, verificando-se uma transição de um país de Emigração para um país de

Imigração (Magalhães, 2003), marcando agora presença como sendo um destino “apetecível” (Carneiro, 2005).

No entanto, as migrações internas têm sido deixadas de parte e existem relativamente poucas investigações realizadas sobre este assunto, e as que existem não fazem um acompanhamento da evolução destes movimentos. Deste modo, continua-se a verificar uma continuidade da deslocação dos indivíduos para regiões mais atractivas, geralmente situadas no litoral do país (Peixoto e Oliveira, 2001), apesar de por vezes não ser possível determinar a sua intensidade e proporção devido à escassez de dados disponíveis.

A importância atribuída às migrações nos nossos dias é principalmente devida às dinâmicas populacionais serem resultantes do crescimento migratório e não tanto do crescimento natural, pois dado o acesso generalizado dos países mais desenvolvidos a uma estabilização ou recessão demográfica resultante do equilíbrio a baixos níveis de mortalidade e natalidade é, sobretudo, a variação migratória que explica decréscimos ou aumentos populacionais (Magalhães, 2003). Verifica-se a existência de uma fronteira clara entre um grupo de regiões com maior e menor mobilidade, sendo o Alentejo e o Algarve as regiões com maior mobilidade e o Norte com menor (Magalhães, 2003). Verifica-se também, a existência de equilíbrio territorial, contrariando a imagem difundida de que existem regiões claramente atractivas e outras repulsivas em Portugal, e de regiões como Lisboa e Vale do Tejo (L.V.T.), que são consideradas “redistribuidoras”, visto que surgem como um forte pólo emissor e receptor de agentes sociais no Continente, mantendo uma interacção elevada com todas as regiões (Peixoto e Oliveira, 2001).

O crescimento da população portuguesa está, cada vez mais, dependente dos imigrantes, pois a evolução registada na baixa da mortalidade infantil, uma esperança média de vida cada vez mais elevada e uma taxa de fecundidade geral mantida a um nível muito baixo, têm um grande impacto negativo na população, envelhecendo-a cada vez mais (Carrilho e Patrício, 2007). No entanto, este envelhecimento nem sempre é registado da mesma forma e ao mesmo ritmo em todas as regiões ou países, podendo diferenciar-se de região para região (Patrício e Gonçalves, 2007).

As migrações registadas durante a idade activa têm como principal motivo o carácter económico. Como existe uma estreita relação entre a população em idade activa e em idade de procriar, assistimos também a uma transferência dos nascimentos da região

emissora para a região receptora. Por outro lado, os motivos que levam à decisão de migrar depois da entrada na reforma encontram-se essencialmente associados a uma diferente motivação: o clima do país de origem (Carrilho e Patrício, 2002).

As próprias migrações exerceram uma forte influência no crescimento da população portuguesa no período intercensitário 1991-2001, como resultado do aumento dos fluxos imigratórios no decorrer da década de 90 (Magalhães, 2003).

A mobilidade geográfica não age aleatoriamente sobre o conjunto da população, afectando grupos particulares dessa população. A observação dos valores migratórios, por idades e por sexos, indica-nos uma maior diferenciação de comportamento entre os dois sexos nas idades adultas jovens, onde a migração é um facto predominantemente feminino. O nível de instrução dos migrantes é significativamente mais qualificado, em média, que a restante população portuguesa. São os indivíduos mais qualificados e ligados às actividades terciárias que detêm maior mobilidade, sendo os grupos assalariados relativamente mais móveis que os restantes. Existe uma maior repulsão feminina nas regiões rurais do interior e, por outro lado, maior atractividade de L.V.T. (Peixoto, 1998).

Mais uma vez, se verifica que o despovoamento das regiões do interior é um fenómeno com grande importância nos nossos dias, pois a população deixa essas regiões para se dirigir para as cidades do litoral na tentativa de acompanhar as oportunidades económicas que um litoral mais urbanizado e desenvolvido tem para oferecer. Todavia, relacionado com estudos anteriores, verifica-se um decréscimo do volume migratório para os grandes centros de atracção que são Lisboa e Porto (Rees *et al.*, 1998).

Tal situação também se verifica em diversos países, sejam da U.E. ou de outro qualquer Continente, pois os migrantes deixam para trás as zonas com uma taxa de desemprego mais alta e menos desenvolvidas para se dirigirem para zonas mais desenvolvidas que lhes possam proporcionar melhor qualidade de vida. Um exemplo desta situação é o caso da Suécia (Kupiszewski *et al.*, 2001), onde as regiões menos desenvolvidas são abandonadas em detrimento de melhores empregos e consequente melhoria na qualidade de vida. Tal como Portugal, também a Suécia se via a par com uma desertificação do interior do país, procurando-se as regiões costeiras e mais desenvolvidas. A diferença consiste em que, ao contrário de Portugal, a zona mais atractiva seria o sul e não o norte. Mas esta situação está a inverter-se, pois denota-se um ganho de população por parte do norte e um saldo migratório negativo nas regiões

do sul (Kupiszewski *et al.*, 2001), denotando-se também um decréscimo da atracção exercida pelos grandes centros urbanos dos próprios países. É importante comparar a situação portuguesa com a de outros países para melhor poder compreender a sua possível evolução futura.

1.1.2. Teorias das Migrações

Qualquer movimento migratório encontra a si associados os mais variados factores, sendo talvez o factor económico o mais influente, pois um país ou região em que a taxa de desemprego se encontre em níveis mais baixos, exerce uma maior influência nos migrantes do que uma região com elevadas taxas de desemprego (Jennissen, 2003). O nível de desenvolvimento de uma região exerce uma força positiva nos próprios migrantes, visto que estes não estão dispostos a “trocar” a sua região por outra que lhe ofereça menos oportunidades (Abreu, 2004).

No entanto, a relação entre migrações e desenvolvimento acaba por não ser linear, ou seja, nem sempre este movimento segue em direcção aos países ou regiões desenvolvidas, mas sim aquele que reúne as condições consideradas como essenciais para o migrante (Nyberg-Sorensen *et al.*, 2002).

O passado e presente das migrações pode ser visto simultaneamente como sendo manifestação e consequência da Globalização, pois esta leva, entre outras, a grandes movimentações de capital, produção e, consequentemente, de pessoas à procura de melhores condições de vida e de novos desafios. Esta movimentação do capital humano torna-se possível essencialmente devido à utilização das mais modernas tecnologias, quer a nível dos *media*, que divulgam a informação, quer através da facilidade e comodidade em viajar de um lado para o outro (Nyberg-Sorensen *et al.*, 2002). Há que ter também em conta o aumento da migração do sexo feminino, que não necessita de migrar acompanhada pela família, procurando “ela” também melhores condições de vida e formas de singrar na vida profissional, deixando de parte o antigo papel que lhe era atribuído (Nyberg-Sorensen *et al.*, 2002).

A própria migração em si é composta por uma estrutura “espacial” que compreende as suas componentes através da construção de uma matriz. Esta matriz é composta pelos próprios fluxos migratórios; pelos factores repulsivos e atractivos (“push and pull

factors”); e por último, pelos níveis de interacção espaciais entre a origem, destino e especificidade das próprias migrações. A análise da estrutura espacial leva a que seja mais fácil compreender a influência da migração nas mais variadas regiões, conseguido através da comparação entre a estrutura espacial da migração e a própria estrutura populacional da região sem migrações (Rogers *et al.*, 2002).

Através da migração conseguimos apreender informações essenciais sobre uma determinada região, sejam estas informações sobre a própria composição da sua população, ou até transmitir os seus indicadores económicos, sociais e ambientais, pois o nível de atracção que exerce nos indivíduos permitindo indagar sobre as suas potencialidades (Rees e Kupiszewski, 1999).

O tema das migrações nem sempre foi visto com atenção, sendo até ignorado pela maioria dos investigadores clássicos das principais ciências sociais mesmo durante o seu período de hegemonia e consolidação. Os fluxos migratórios já assumiam grande importância, quer sob a forma de movimentos internos intensos que se dirigiam dos campos para as cidades, quer através de migrações transoceânicas que permitiram a povoação de novos continentes, e no entanto a importância dispensada sobre o tema era pouco significativa. No entanto, com o passar dos anos têm surgido as mais variadas teorias explicativas das migrações, que apesar de se poderem aplicar também às migrações internas, estas formaram-se essencialmente tendo em vista o cariz internacional das próprias migrações.

Segundo Peixoto (2004), podemos afirmar a divisão destas em Teorias Micro e Macro-Sociológicas. As teorias micro-sociológicas são aquelas que se reportam à decisão individual de migrar, independentemente das mais variadas condicionantes ou vantagens que possam exercer influência. Dentro das teorias “micro” encontram-se ligados vários modelos que explicam as migrações tendo em conta a decisão individual do migrante. Assim, existem os chamados modelos de *Push-Pull* e do Capital Humano, que tal como a maioria das teorias consideradas como micro-sociológicas têm presente uma raiz económica. A essência do modelo de *Push-Pull* reside na base em informação recolhida acerca das zonas de origem e de destino, levando a considerar qual a melhor opção, exercendo assim as regiões de destino “atração” (*pull*) e as regiões de origem “repulsão” (*push*). No caso da teoria do capital humano, o essencial dos seus argumentos, reside no facto de que o migrante analisa a sua saída da região de origem através de um cálculo económico permitindo-lhe ter em conta a sua vida a longo prazo.

Aqui o migrante tende a calcular se o seu investimento inicial irá ser compensado pelo seu trabalho, explicando o facto de indivíduos com idades mais avançadas não decidirem migrar, pois o período em que o investimento possa ser compensado será menor.

Uma outra perspectiva sociológica apresenta uma natureza diferente da anterior, agora de cariz biográfico, sendo as principais variáveis estudadas a influência do Ciclo da Vida (entrada na vida adulta, casamento, nascimento dos filhos, reforma, etc.) e a Trajectória da Mobilidade Social. Enquanto que no caso da influência do ciclo da vida na decisão de migrar este compreende a etapa da vida em que o migrante se insere, no caso da trajectória da mobilidade social, o que está essencialmente em causa é a realização de um percurso por parte do indivíduo, por diferentes posições sociais.

No caso das teorias “*macro*”, podem distinguir-se como tais as que privilegiam a acção de factores de tipo colectivo. A esta teoria estão também associados vários desenvolvimentos teóricos, tais como o Mercado de Trabalho Segmentado ou “Dual”. No caso do primeiro, este centra-se na vertente receptora da migração uma vez que é o próprio mercado de trabalho que afirma qual o tipo de migrante que procura.

Um outro tipo de teorias que também se enquadra no nível “*macro*” são as “*análises que lidam essencialmente com a variável espaço e procuram enunciar os factores que levam a um desenvolvimento particular dos territórios. Existem assim, mecanismos que levam a uma dada localização dos estabelecimentos humanos em realidades de tipo urbano ou regional, central ou periférico*” (Peixoto, 2004).

A perspectiva “macro” das migrações pode ser dividida em alguns grupos: em primeiro lugar, as teorias que salientam o papel das instituições no desencadear ou desenvolvimento dos fluxos migratórios; em segundo lugar as que defendem o papel das redes migratórias; e por último, aquelas que salientam os laços de natureza social e étnica que une alguma actividades no destino.

Mesmo podendo ser divididas, estas duas perspectivas encontram-se muito interligadas entre si, pois por vezes umas contêm componentes das outras tomadas como diferente perspectiva.

Tal como atrás referi, algumas teorias reportaram-se essencialmente às migrações internacionais, mas que devidamente analisadas e aplicadas a novos contextos também podem contribuir para a sua compreensão a nível interno.

Assim temos, a *teoria económica neo-clássica*, que explica a migração através das diferenças salariais entre os diversos países ou, neste caso regiões, levando a que as regiões onde os salários são mais elevados exerçam um elevado poder de atracção nos próprios migrantes; a *teoria do mercado de trabalho “dual”*, em que a atracção exercida incide em dois tipos de mão-de-obra, a qualificada e a não-qualificada, explicando-se pela divisão da economia num sector primário de uso intensivo de capital, e num sector secundário de uso intensivo de mão-de-obra e de baixa produtividade, dando lugar a um mercado de trabalho segmentado; a *teoria da nova economia da migração laboral*, em que a decisão de migrar é levada a cabo tendo em conta toda a família, migrando por vezes, apenas o “chefe de família” de forma a garantir um melhor nível de vida para os familiares que ficam na região ou país de origem e poupando-os, de certa forma ao incómodo de migrar; e por último, a *teoria do sistema mundial*, que se baseia no pressuposto de o capitalismo ser um sistema social histórico, tendo como principal ideia que as economias altamente desenvolvidas necessitam de mão-de-obra estrangeira para trabalhar em determinados sectores em troca de salários reduzidos (Jennissen, 2004).

1.1.3. Novas Abordagens ao Estudo das Migrações

O estudo sobre as migrações nem sempre se dirige aos padrões migratórios ou ao estudo destas ocorrências nos mais diversos países, tentando-se, por vezes, a realização de uma estimação das migrações de forma a tentar analisar e compreender melhor estes movimentos. No entanto, a realização de uma estimação/projecção sobre esta temática nem sempre é fácil de realizar devido à falta de dados disponíveis ou até de dados correctos, visto isto constata-se que os diversos estudiosos desta temática recorrem a mais diversas técnicas para efectuarem as suas estimação e projecções.

Uma metodologia relativamente recente e bastante utilizada pelos investigadores que se dedicaram com mais atenção ao tema das migrações baseia-se na obra de Andrei Rogers (2002), e claro, de todos os seus colaboradores directos, quer através da realização de trabalhos em conjunto, ou apenas de simples contribuições.

Podemos observar um trabalho bastante interessante realizado por Cláudio Caetano Machado (2004) que, ao analisar os fluxos migratórios no Brasil empregou estes

mesmos métodos, encontrando-se aqui o que poderíamos chamar de um “resumo” sobre as suas técnicas de projecção. Neste seu trabalho é utilizado o método da *Migração Infantil*, que utiliza a informação sobre o local de nascimento das crianças entre os 0-4 anos como parâmetro de estimativa indirecta das taxas específicas da estimativa por idade. Este método assenta em três pressupostos essenciais, que são, “*as taxas de migração têm um padrão por idade muito semelhante, independentemente do seu volume e o lugar de ocorrência; a questão sobre o nascimento encontra-se sempre presente em quase todos os censos para as crianças pertencentes ao grupo etário dos 0-4 anos, tornando-se numa aproximação à análise migratória; e por último, visto que estas crianças migram sempre (ou na sua grande maioria) com os pais, jovens adultos, a sua migração pode ser vista como um espelho da migração dos adultos*” (Machado, 2004).

Este método utiliza a regressão linear simples entre a proporção de crianças dos 0-4 anos que nasceram em i e viviam em j para se obter para cada grupo etário, a proporção de pessoas que viviam em i há cinco anos e hoje vivem em j .

Um outro método é a utilização do *Saldo Líquido Migratório* em conjunto com a *Migração Infantil*, mas aqui há que ter em conta os inconvenientes apresentados pela migração líquida, uma vez que consideramos um conjunto de regiões que interage entre si, e que a soma dos SML deverá ser no fim igual a 0.

O método da *Migração Infantil* também pode ser utilizado de uma outra forma, através da sua *separação entre o nível e a estrutura etária*. Pode-se assim, separar a estimativa indirecta de taxas específicas de emigração na necessidade de descobrir o nível da migração, onde interessa a intensidade da migração e o seu padrão por idade; bem como a sua distribuição por grupo etário (Machado, 2004).

1.2. Origem dos dados e contextualização

A já referida dificuldade devida à inexistência de dados que permitam a realização de uma análise detalhada das migrações, leva-nos a utilizar dois conjuntos de dados de provenientes de duas fontes distintas. O primeiro, proveniente do Instituto Nacional de Estatística (INE), correspondente ao universo de todas as observações, mas que por

estarem agrupados não nos permite realizar uma análise descritiva, permitindo somente a aplicação das duas primeiras metodologias de estimação indirecta desenvolvidas por Rogers *et al.* (2002).

Enquanto o segundo conjunto, proveniente do IPUMS – International Database (Integrated Public Use Microdata Series¹), correspondente a uma amostra de 5% da população portuguesa (em cada ano de observação), foi utilizado quer para a aplicação da terceira metodologia de estimação indirecta, bem como para a realização de uma análise descritiva.

Os dados, referentes a este estudo foram, assim, recolhidos através dos Recenseamentos da População de 1991 e 2001 e disponibilizados pelo IPUMS/INE. Tal como na maioria dos países, no questionário utilizado para o recenseamento em Portugal consta uma pergunta sobre o local (região) de residência no momento actual, isto é, à data do recenseamento, e num dado período de tempo no passado (neste caso 5 anos antes da data censitária). As migrações foram avaliadas comparando, para cada indivíduo, se existem diferenças entre o local de residência, à data do recenseamento, e o local de residência no período anterior. Usámos para o efeito dados relativos à população residente, por grupo etário e sexo, ao nível da área geográfica de residência segundo as NUTS II, em 31 de Dezembro de 1985 e 1995 e população residente, por grupo etário e sexo, segundo a NUT II de residência da mãe à data do nascimento.

A análise, que incidiu nas regiões portuguesas correspondentes às NUTS II, excluiu quer os indivíduos residentes, quer os nascidos no estrangeiro (estes últimos só foram considerados relevantes para a análise dos indivíduos pertencentes ao grupo etário 0-4 anos).

Importa ainda referir o facto de também termos deixado de parte as regiões insulares (Região Autónoma dos Açores e Região Autónoma da Madeira), pois os movimentos migratórios *de* e *para* estas regiões são bastante reduzidos, o que não nos permitiu elaborar estimativas coerentes.

¹ O autor agradece ao Instituto Nacional de Estatística, produtor dos dados disponibilizados pelo IPUMS-International Database para Portugal.Minnesota Population Center. Integrated Public Use Microdata Series — International: Version 4.0. Minneapolis: University of Minnesota, 2008.

2. Caracterização dos Indivíduos em Estudo

Para uma melhor compreensão dos dados que utilizamos, elaborou-se uma análise descritiva onde, além de podermos caracterizar os indivíduos em estudo, podemos também observar a evolução registada entre os dois recenseamentos em questão.

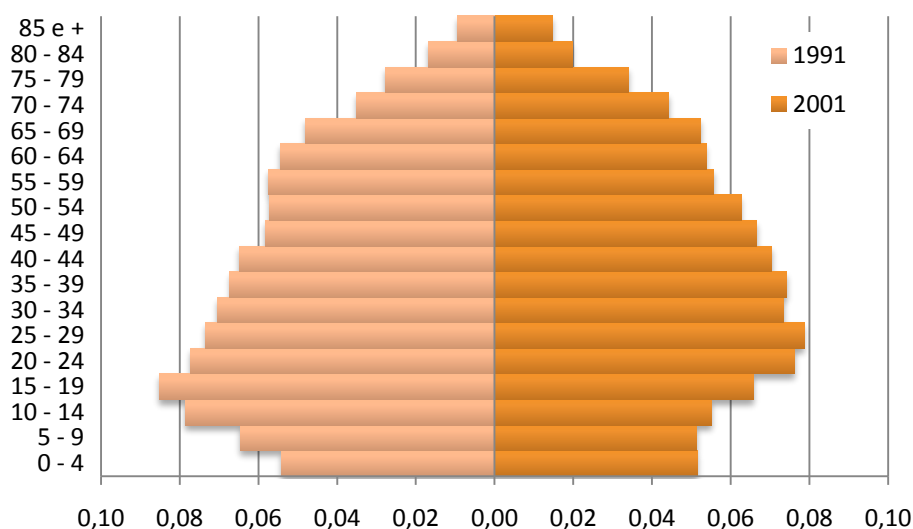
Numa primeira fase, e com o intuito de compreendermos da melhor forma a evolução registada, representámos graficamente, através de uma *pirâmide de idades* (figura 1), a estrutura populacional de Portugal Continental nos dois anos em estudo. Para o efeito, foi utilizada informação utilizada para a aplicação das metodologias de estimação indirecta.

Deste modo, observando então a estrutura da população registada nos dois recenseamentos gerais da população de 1991 e 2001, podemos afirmar que, neste intervalo inter-censitário (10 anos), se registaram importantes alterações. Aqui, o estreitamento da *base* em relação ao *topo* é um forte indicador de envelhecimento, pois o facto de as classes anteriores serem mais pequenas que as posteriores, sugerem a não renovação de gerações (devido aos baixos níveis de fecundidade), o que aliado a uma baixa mortalidade resulta no já referido envelhecimento da estrutura populacional.

Importa também destacar que, apesar destas diferenças estruturais, os níveis de fecundidade acabaram por não diferir muito nestes 10 anos, situação que pode ser corroborada através da observação dos três primeiros grupos de idades. Em 1991, a população com idades até aos 4 anos completos perfazia apenas cerca de 5,4% da população total, enquanto em 2001, esse valor desceu, ainda que ligeiramente, para cerca de 5,2%.

O comportamento registado neste hiato de tempo, a manter-se, resultará numa população envelhecida, com um alto índice de dependência, quer dos próprios jovens, quer dos idosos, e numa incapacidade de se auto-sustentar, pois o peso do topo acaba por ser muito superior ao dos adultos activos na população.

Figura 1 – Estrutura Populacional de Portugal Continental em 1991 e 2001



Uma vez que os dados referentes ao total da população residente em Portugal Continental apenas nos permitem observar a sua distribuição em termos estruturais iremos, de seguida, recorrer aos dados correspondentes às amostras disponíveis para os dois anos em observação.

Analisando a tabela 1, podemos reforçar o que acabamos de referir anteriormente, ou seja, de que a população residente em Portugal Continental em 2001 registou algum envelhecimento comparativamente a 1991, situação que se pode observar através da média de idades registada nos dois anos, passando de 36,56 anos em 1991, para 39,21 em 2001. Não variando muito mesmo quando *aparada a 5%* (5% Trimmed Mean), descendo ligeiramente para 36,03 e 38,88 anos, em 1991 e 2001, respectivamente, o que significa que mesmo retirando os casos mais extremos não se regista uma grande variação.

Por outro lado, observando a *variância* e o *desvio-padrão* verifica-se a existência de uma ligeira dispersão de valores em torno da média de idades, registando-se um desvio-padrão de 22,38 e 22,48, em 1991 e 2001, respectivamente. Neste sentido, e apesar de se ter registado um desvio padrão em 2001 ligeiramente mais elevado relativamente a 1991, o *coeficiente de variação* indica-nos a existência de uma maior dispersão de idades em 1991 (61,23%) comparativamente a 2001 (57,34%).

Observando agora o valor correspondente à *mediana*, constatamos que em 1991, 50% dos indivíduos tinha 34 ou menos anos de idade, contrapondo com os 38 anos registados em 2001, sendo mais um indicador do envelhecimento ocorrido entre recenseamentos.

No entanto, podemos ainda reforçar que, em 1991, 25% dos residentes em Portugal Continental tinham 18 anos ou menos, valor que aumentou para os 21 anos em 2001. No outro quartil, podemos referir que 25% dos indivíduos registam idades superiores ou iguais a 54 e 57 anos em 1991 e 2001, respectivamente.

A *amplitude interquartílica* indica-nos ainda que, 50% dos indivíduos residentes em Portugal Continental no ano de 1991, tinham idades entre os 18 e os 54 anos, perfazendo um intervalo de 36 anos entre eles, enquanto que no ano de 2001, 50% dos residentes registava idades entre os 21 e os 57 anos, mantendo assim o intervalo em 36 anos.

O *máximo* e o *mínimo* indicam-nos que as idades dos inquiridos variam entre os 0, ou seja, entre as crianças que a ainda não completaram o primeiro ano de idade e os idosos com 100 anos, tendo assim, uma amplitude (*range*) de 100 anos de intervalo entre a idade mais baixa e a mais elevada.

Por último, no respeitante à assimetria e ao achatamento da distribuição das idades, podemos constatar que em 1991 o coeficiente entre o coeficiente de assimetria obtido no *software* utilizado (*SPSS*) e o seu desvio-padrão é de 81,34, e em 2001 de 53,82, o que, sendo muito superior a 1,96, nos indica que estamos perante duas distribuições muito assimétricas positivas. No entanto, o facto de este valor ter descido em 2001, pode ser mais um indicador de envelhecimento, pois se a tendência se mantiver, observaremos uma grande “deslocação” de indivíduos para idades mais elevadas.

Quanto ao coeficiente obtido entre o coeficiente de achatamento e o seu desvio-padrão ($Kurtosis/SE_K$), este é de -129,56 e -127,71 em 1991 e 2001, respectivamente. Uma vez que estamos perante valores muito inferiores a -1,96, estes indicam-nos que estamos perante duas distribuições claramente *platicúrticas*.

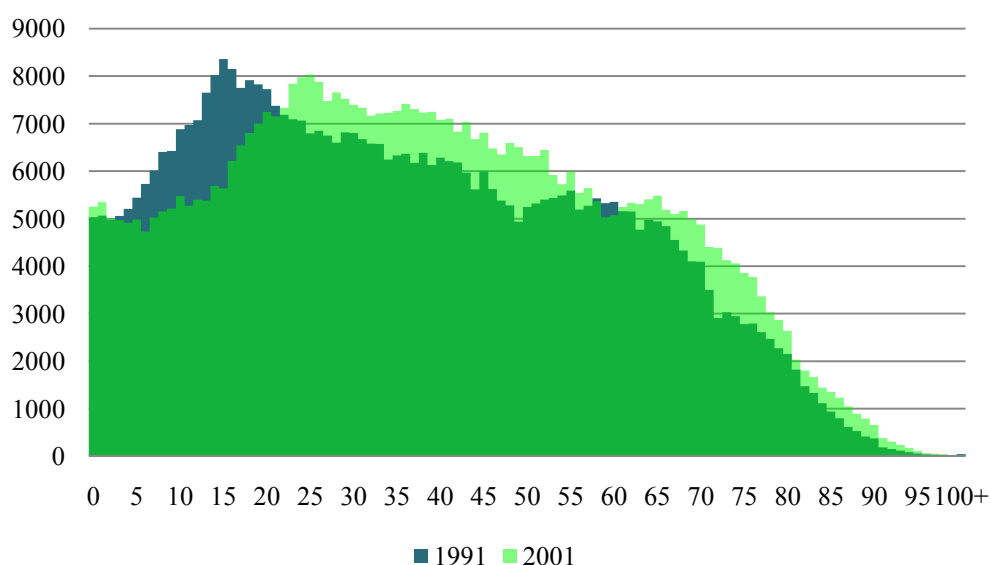
Tabela 1 – Idade (Portugal Continental)

		1991		2001	
		Estatística	Erro Padrão	Estatística	Erro Padrão
Média		36,56	0,0327	39,21	0,0320
95% I. C. Média	Lim. Sup.	36,49	-	39,15	-
	Lim. Inf.	36,62	-	39,28	-
Média Aparada a 5%		36,03	-	38,88	-
Variância		500,97	-	505,52	-
Desvio Padrão		22,38	-	22,48	-
Mínimo		0,00	-	0,00	-
Máximo		100,00	-	100,00	-
Amplitude		100,00	-	100,00	-
Quartis	25	18,00	-	21,00	-
	50	34,00	-	38,00	-
	75	54,00	-	57,00	-
Amplitude Interquartilica		36,00	-	36,00	-
Enviesamento – “ <i>Skewness</i> ”		0,2914	0,0036	0,1878	0,0035
Achatamento – “ <i>Kurtosis</i> ”		-0,9282	0,0072	-0,8913	0,0070

Tal como foi possível verificar na análise da *figura 1* e foi descrito através dos resultados apresentados na tabela 1, denota-se um envelhecimento da população portuguesa na sua totalidade, no entanto, o facto de os dados disponíveis para a *população* serem agrupados em grupos de idade quinquenais, não nos permitiu ter uma ideia clara da evolução ocorrida entre os dois anos em questão.

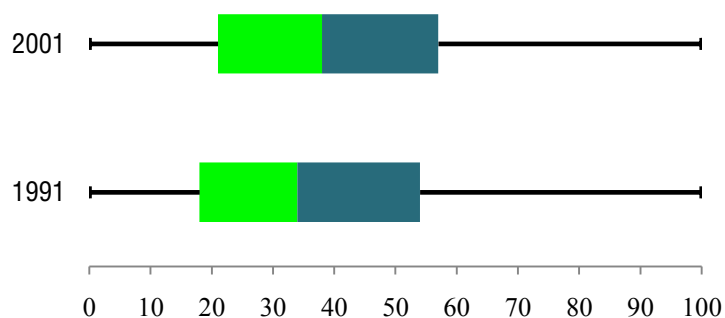
Desta forma, e baseado na *amostra* disponível, elaborou-se um gráfico de barras que realça a diferença resultante do contraste entre a população portuguesa entre 1991 e 2001 (*gráfico 1*) idade a idade.

Gráfico 1 – Evolução da estrutura populacional de Portugal Continental entre 1991 e 2001



Através da *figura 3*, e utilizando uma *caixa de bigodes*, podemos constatar que se verifica um deslocamento da população para idades mais avançadas, resultando num envelhecimento populacional. De uma forma clara, a *figura 2* apresenta e reforça o aumento na idade mediana da população portuguesa já atrás descrito.

Figura 2 – Quartis da evolução da estrutura Populacional de Portugal Continental entre 1991 e 2001



No entanto, será que este envelhecimento populacional se verifica em todo o continente e de uma forma uniforme? Para que seja encontrada uma resposta para esta pergunta foi elaborada uma tabela e os valores obtidos referentes à *amostra* populacional distribuída pelas diferentes regiões portuguesas foram realçados através da utilização de cores (tabela 2).

Com esta tabela é possível verificar que o envelhecimento populacional é distinto consoante as diferentes regiões continentais. Como seria de esperar, são as regiões menos povoadas (Alentejo e Algarve) que denotam um maior envelhecimento, pois além de se observar um decréscimo no número de indivíduos com idades até aos 19 anos, denotam um aumento populacional em idades superiores aos 70 anos.

No entanto, mesmo entre estas semelhanças, as duas regiões em foco, reflectem ainda comportamentos distintos, tal como por exemplo o aumento populacional na região Algarvia em idades activas, uma vez que é uma região que atrai muitos migrantes à procura de um emprego na área do turismo.

Para terminar, e como era de esperar, as regiões do Norte, Lisboa e Vale do Tejo e Centro, são aquelas que apresentam um comportamento mais semelhante ao nível do envelhecimento populacional.

Tabela 2 – Distribuição da População pelas diferentes regiões Continentais Portuguesas

	Norte		Centro		L.V.T.		Alentejo		Algarve	
	1991	2001	1991	2001	1991	2001	1991	2001	1991	2001
0 - 4	10713	10338	4458	4172	8016	8900	1273	1151	873	933
5 - 9	12473	10506	5369	4268	9592	8348	1554	1088	1028	898
10 - 14	14711	11429	6583	4817	12258	8617	1849	1424	1203	932
15 - 19	16324	13224	6953	5912	13556	10265	1893	1617	1291	1171
20 - 24	15140	14666	6376	6552	12053	13213	1736	1757	1139	1361
25 - 29	13690	14957	5658	6321	11651	14114	1674	1696	1144	1490
30 - 34	12664	14116	5647	6082	11656	12912	1763	1791	1149	1443
35 - 39	11630	14748	5331	6360	11624	12196	1607	1752	1197	1424
40 - 44	10477	13447	5121	6139	12058	11945	1515	1832	1093	1356
45 - 49	9029	12334	4836	5879	10945	11538	1475	1679	927	1316
50 - 54	8799	10732	4936	5369	10424	11865	1652	1523	1078	1237
55 - 59	8679	9210	5226	4989	9940	10765	1914	1457	1054	1156
60 - 64	8242	8713	5048	5171	9202	9728	1902	1693	1014	1068
65 - 69	7166	8465	4832	5219	8024	9292	1821	1787	925	1155
70 - 74	5133	6918	3721	4572	5477	7720	1424	1681	727	953
75 - 79	3934	5121	3035	3753	4197	5837	1141	1357	625	820
80 e +	3707	5192	2993	3886	4001	5634	1084	1360	529	892

A evolução da proporção dos residentes distribuídos pelas cinco NUTS II em análise, é outro dos factores que nos pode ajudar a compreender as migrações, pois a sua proporção de residentes pode, além da mortalidade e da natalidade, ser um bom indicador de atractividade regional, gerando um maior número de movimentos migratórios na sua direcção

A tabela 3 indica-nos que são as regiões do Norte e de Lisboa e Vale do Tejo que detêm maior proporção de residentes. Por outro lado, são as duas regiões do Sul do país Alentejo e Algarve, que menos proporções de residentes registam, denotando-se uma grande diferença entre o Norte, onde se concentra a grande maioria da população, e o Sul do Continente português.

Tabela 3 – Residentes em Portugal Continental

	1991			2001		
	n	%	Cum.	n	%	Cum.
Norte	172511	36,9	36,9	184116	37,4	37,4
L.V.T.	164674	35,2	72,1	172889	35,1	72,5
Centro	86123	18,4	90,5	89461	18,2	90,6
Alentejo	27277	5,8	96,4	26645	5,4	96,0
Algarve	16996	3,6	100,0	19605	4,0	100,0
Total	467581	100,0	-	492716	100,0	-

De seguida, e uma vez que, nem homens, nem mulheres, registam o mesmo comportamento demográfico, será bastante interessante observar as diversas diferenças entre os dois sexos. Assim, e segundo a tabela 4, constatamos que, independentemente do ano de observação, a média de idades correspondente ao sexo masculino é sempre inferior à do sexo feminino, apesar de, nos dois sexos, ambos terem registado um aumento entre 1991 e 2001.

Embora sejam os indivíduos do sexo masculino que registam uma menor média de idades, e um desvio-padrão igualmente mais reduzido quando comparados com o sexo feminino, estes são os que registam um coeficiente de variação mais elevado nos dois anos (62,07% e 58,04%). No respeitante ao sexo feminino, estes valores descem para 60,25% e 56,50%, em 1991 e 2001, respectivamente. Neste sentido, e em ambos os sexos, apesar de se registar um aumento da idade média dos indivíduos constituintes da amostra entre 1991 e 2001, o desvio-padrão associado apenas aumentou ligeiramente, o que resultou num decréscimo do coeficiente de variação associado.

Por outro lado, e começando novamente pelo sexo masculino, constatamos que, em 1991, 50% dos residentes tinha pelo menos 33 anos de idade, valor que subiu para os 36 anos em 2001, enquanto que no caso do sexo feminino, este valor passou dos 36 anos para os 39.

Tabela 4 – Idade segundo o sexo (Portugal Continental)

		1991		2001	
		Homens	Mulheres	Homens	Mulheres
Média		35,12	37,90	37,80	40,53
95% I. C. Média	Lim. Sup.	35,03	37,81	37,71	40,44
	Lim. Inf.	35,21	37,99	37,89	40,62
Desvio Padrão		21,80	22,83	21,94	22,90
Quartis	25	17,00	18,00	20,00	22,00
	50	33,00	36,00	36,00	39,00
	75	52,00	56,00	55,00	59,00

Como se sabe, nem sempre a idade média de uma população corresponde à idade média das diferentes subpopulações que a compõem. Neste caso, e entendendo por subpopulações, a região de residência e o sexo dos indivíduos, podemos verificar que de uma forma geral são as mulheres que valores mais elevados quando comparados com os homens (tabela 5). Estes valores podem ser explicados pelo facto de a esperança de vida do sexo feminino ser mais elevada comparativamente ao masculino.

Neste sentido, e ainda relativamente ao sexo feminino, verifica-se que é a região Alentejo que apresenta valores mais elevados, passando de uma média de 42 anos de idade em 1991 para os 44, já em 2001. No extremo oposto, encontra-se a região Norte, que em 1991 registou uma média de 35 anos para o sexo feminino, e em 2001 esse valor subiu para os 38 anos de idade.

No que diz respeito ao sexo masculino, verifica-se que é mais uma vez a região norte que apresenta a idade média mais baixa das cinco regiões, e o Alentejo (1991: 33; 2001: 36) a mais elevada (1991: 40; 2001: 42).

Com a análise da tabela 5, podemos verificar que de uma forma geral a idade média está a aumentar, muito provavelmente fruto dos avanços medicinais e na qualidade de vida que resultam num melhor estado de saúde. Em 2001, e no que diz respeito ao sexo feminino, verifica-se ainda que apenas a região norte se encontra com uma idade média abaixo dos 40 anos, enquanto que todas as outras já ultrapassaram esta idade.

Tabela 5 – Idade média segundo o sexo e região de residência

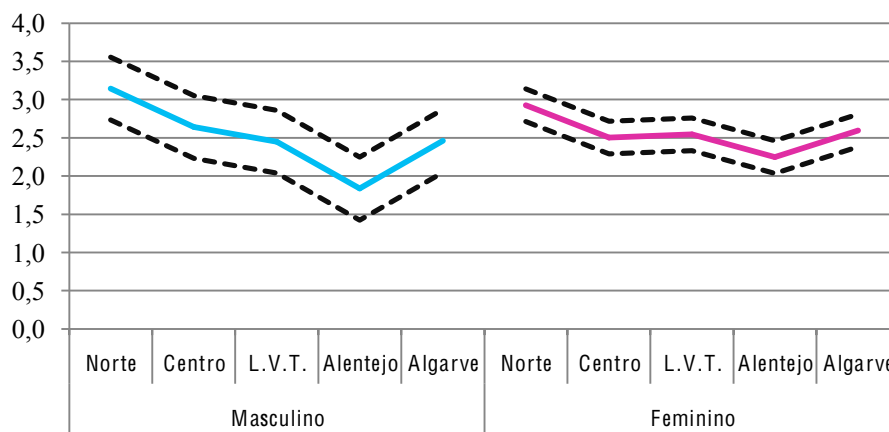
	1991		2001	
	Masc.	Fem.	Masc.	Fem.
Norte	33	35	36	38
Centro	37	40	39	42
L.V.T.	36	39	38	41
Alentejo	40	42	42	44
Algarve	38	40	40	42

No entanto, e apesar de se terem analisado as evoluções etárias ocorridas entre os dois anos em observação, será interessante observar graficamente as diferenças registadas entre 1991 e 2001, segundo as regiões de residência, e uma vez mais segundo o sexo.

Deste modo, observando o *gráfico 2*, verifica-se que foi a região do Alentejo que menos viu aumentar a idade média dos seus residentes, quer para o sexo masculino, quer para o feminino. Por outro lado, é a região Norte que maiores alterações sofreu, especialmente no caso do sexo masculino.

Por outro lado ainda, verifica-se que no caso do sexo feminino a variabilidade é mais homogénea entre regiões, enquanto que no sexo masculino as oscilações nas diferenças acumuladas são mais díspares. O resultado desta dissemelhança regional encontra-se reflectido no cálculo dos intervalos de confiança obtidos com o nível de significância de 5%, onde a maior amplitude se encontra conectada ao sexo masculino.

Gráfico 2 – Diferenças entre as idades médias registadas nos dois anos, segundo o sexo e a região de residência



Apesar de verificarmos a existência de diferenças na idade média entre as diferentes regiões em estudo, serão estas diferenças significativas? Neste sentido, e com o intuito de se obter uma comparação entre as diferentes regiões de Portugal Continental, pensou-se em utilizar a análise de variância, ou seja, a ANOVA. No entanto, esta metodologia necessita de ver validados os seus pressupostos, que são a normalidade e a homocedasticidade, pressupostos que não foram passíveis de validar. Assim, optou-se pela sua alternativa não-paramétrica para este tipo de amostra, o teste de Kruskal-Wallis (tabela 6).

Assim, e recorrendo ao *software* estatístico *R*, verificou-se que, com um $\alpha = 0,05$ existem diferenças significativas nas idades registadas entre as diferentes regiões. No entanto, será que todas as regiões diferem entre si? Como resposta a esta questão foram efectuadas comparações múltiplas, e cujos resultados serão apresentados de seguida.

Tabela 6 – Kruskal-Wallis

<i>Chi-Squared</i>	3716.142
<i>Df</i>	4
<i>P-Value</i>	2.2e-16

Observando deste modo os resultados provenientes das comparações múltiplas (tabela 7) podemos verificar que ao nível de significância de 5%, apenas não existem diferenças significativas entre as regiões do Algarve e do Centro.

Tabela 7 – Comparações múltiplas *a posteriori*

Multiple comparison test after Kruskal-Wallis			
p.value: 0.05			
Comparisons			
	obs.dif	critical.dif	difference
Alentejo-Algarve	12450.434	3702.646	TRUE
Alentejo-Centro	15256.975	2632.467	TRUE
Alentejo-L.V.T.	19925.000	2476.844	TRUE
Alentejo-Norte	39359.553	2468.837	TRUE
Algarve-Centro	2806.541	3180.173	FALSE
Algarve-L.V.T.	7474.566	3052.601	TRUE
Algarve-Norte	26909.119	3046.107	TRUE
Centro-L.V.T.	4668.025	1593.321	TRUE
Centro-Norte	24102.578	1580.845	TRUE
L.V.T.-Norte	19434.552	1305.352	TRUE

3. Alguns Aspectos Teóricos

Para a elaboração desta dissertação de mestrado, além das metodologias de estimação indirecta, utilizaram-se diferentes ferramentas estatísticas, desde o próprio *software* às diferentes abordagens utilizadas para atingir os nossos objectivos.

Assim, para além do *software* SPSS, utilizou-se também o *R*, entre os quais fomos alternando consoante a abordagem a realizar. No entanto, foi sobre o *R* que mais recaiu a nossa preferência, principalmente por ser uma ferramenta que nos permite ter uma maior liberdade de acção.

No respeitante à própria modelação em si, e uma vez que um dos nossos objectivos passa pela modelação das relações entre diversas variáveis consideradas e predizer o valor de uma variável dependente a partir de um conjunto de variáveis independentes, em que a nossa variável resposta é binária (migrou/não migrou), a metodologia utilizada foi a da regressão categórica, mais especificamente, a regressão logística.

3.1. Comparação de populações a partir de amostras independentes

Durante a análise descritiva dos dados utilizados ao longo desta dissertação, foram utilizadas diferentes técnicas estatísticas simples, que em conjunto com algumas representações gráficas, permitiu uma clara e cuidada análise descritiva.

No entanto, e uma vez que os dados disponíveis o permitiam, pensou-se em realizar uma comparação das idades médias ao nível regional (NUTS II), com o intuito de se perceber se existiam diferenças significativas entre as cinco regiões em estudo ao nível da idade.

Deste modo, pretendeu-se utilizar a análise de variâncias, ou seja, a ANOVA, que é um teste paramétrico que nos permite determinar, dado um determinado nível de significância, se existem diferenças significativas entre as idades médias nas respectivas regiões. A escolha desta metodologia recaiu sobre o facto de nos encontrarmos perante a necessidade de comparar duas ou mais médias populacionais.

Apesar de existirem diferentes tipos de ANOVA, a aqui descrita corresponde à ANOVA tipo I ou *one way*, em que o modelo de análise de variância aqui utilizado assenta apenas num factor de classificação, neste caso a região de residência.

Pretendemos assim testar as seguintes hipóteses:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k \text{ Vs } H_1: \exists i, j: \mu_i \neq \mu_j (i \neq j; i, j = 1, \dots, k)$$

Onde a hipótese nula corresponde à igualdade entre todas as médias populacionais, e H_1 corresponde à existência de pelo menos duas dessas médias serem diferentes.

Neste sentido, e tomando \bar{X}_i como sendo a média correspondente a cada amostra (região), temos que:

$$\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$$

Em que X_{ij} representa o elemento j da amostra i (sendo $i = 1, \dots, k$ e $j = 1, \dots, n_i$).

Enquanto que a média para todas as observações é dada por:

$$\bar{X} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij}}{N}$$

em que N é dado por $N = \sum_{i=1}^k n_i$.

As diferentes k amostras extraídas das populações em estudo permitem assim analisar a variância quer entre amostras, quer dentro das próprias amostras. Assim, é a variância total calculada para cada amostra que nos vai permitir identificar a existência ou não de desvios significativos entre a média observada entre as diferentes amostras e dentro delas próprias.

É a variação entre as médias das amostras em relação à média total e os desvios observados entre as médias de cada uma das amostras que resultam na variação total, e que nos permitem fazer a referida distinção.

Deste modo, se a variação total corresponde a $X_{ij} - \bar{X}$, os desvios em relação à média total a $\bar{X}_i - \bar{X}$, e os desvios em relação à média de cada uma das amostras por $X_{ij} - \bar{X}_i$, temos:

$$(X_{ij} - \bar{X}) = (\bar{X}_i - \bar{X}) + (X_{ij} - \bar{X}_i)$$

Que elevando ao quadrado resulta na denominada *equação de análise de variâncias*:

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 = \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

O que significa que a soma dos quadrados total é resultado da soma entre a adição da soma dos quadrados factorial e da soma dos quadrados do erro, e simplificando temos:

$$SQT = SQF + SQE$$

Assim, podemos então quantificar a variação total, que é resultado da razão entre as somas dos quadrados e a os respectivos graus de liberdade, designada por quadrados médios, e obtidos por:

$$QMT = \frac{SQT}{N-1} \quad QMF = \frac{SQF}{k-1} \quad QME = \frac{SQE}{N-k}$$

Deste modo, é o quociente entre a variação factorial e a variação residual que permitem rejeitar ou não a hipótese nula, baseando-se na estatística de teste dada por:

$$F = \frac{QMF}{QME} = \frac{\frac{SQF}{(k-1)}}{\frac{SQE}{(N-k)}}$$

Seguindo uma distribuição F de Snedecor com $(k-1, N-k)$ graus de liberdade, rejeitando-se a H_0 quando:

$$F \geq f_{1-\alpha; (k-1, N-k)}$$

Com este teste, ficamos assim a saber se pelo menos uma das médias populacional é diferente das outras, não nos indicando qual ou quais são os pares de médias que são diferentes. Para esse efeito, terão se ser aplicados alguns testes *a posteriori*, testando:

$$H_0: \mu_i = \mu_j \text{ Vs } H_1: \exists i, j: \mu_i \neq \mu_j (i, j = 1, \dots, k)$$

Para todos os pares de médias possíveis dado por ${}^kC_2 = (k^2 - k)/2$.

Existem no entanto, varias testes que possibilitam esta análise, mas não um consenso, e desta forma optámos por apresentar o teste de *Tukey*, que considerado um dos mais robustos aos desvios à normalidade e à homogeneidade de variâncias. Sendo a sua estatística de teste dada por:

$$Q = \frac{|\bar{X}_i - \bar{X}_j|}{\sqrt{\frac{QME}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}}$$

Rejeitando-se a hipótese nula sempre que Q for maior que o valor crítico do teste de *Tukey* $q_{\alpha, v, k}$, onde v é o número de graus de liberdade dos erros na ANOVA.

No entanto, para que este modelo de análise de variância possa ser utilizado, três pressupostos têm de se verificar:

- ✓ Dados provenientes de amostras simples e independentes;
- ✓ Homogeneidade de variâncias entre as populações em estudo: $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$;
- ✓ E, populações com distribuição normal: $X_i \sim N(\mu_i, \sigma)$ ($i = 1, \dots, k$).

Dois destes pressupostos podem ser avaliados estatisticamente, como é o caso da homogeneidade de variâncias e a normalidade dos dados. De seguida serão apresentadas duas das formas que podem ser usadas.

Começando pela homogeneidade de variâncias, temos o teste de Levene, que é um teste robusto a desvios da normalidade e um dos considerados mais potentes para testar a homogeneidade de variâncias.

Onde a estatística de teste é dada por:

$$W = \frac{(N - k)}{(k - 1)} \times \frac{\sum_{i=1}^k n_i (\bar{Z} - \bar{Z}_i)^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_i)^2}$$

Em que n_i é a dimensão de cada uma das k amostras ($i = 1, \dots, k$) e N é a dimensão total da amostra, onde Z se pode definir-se como $Z_{ij} = |\bar{X}_{ij} - \bar{X}_i|$ ($i = 1, \dots, k$ e $j = 1, \dots, n_i$). Rejeita-se assim a hipótese nula se $W \geq f_{1-\alpha; (k-1, N-k)}$, onde as hipóteses a testar são:

$$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 \quad \text{Vs} \quad H_1: \exists i, j: \sigma_i^2 \neq \sigma_j^2 (i \neq j; i, j = 1, \dots, k)$$

No que diz respeito à normalidade, apresentaremos o exemplo do teste de ajustamento de Kolmogorov-Smirnov, que é o que mais se ajusta à dimensão amostral em causa. Deste modo, o teste de Kolmogorov-Smirnov baseia-se numa comparação da função de distribuição dos valores observados com a função de distribuição teórica esperada de acordo com H_0 .

Pretendemos assim, testar se a distribuição amostral em estudo $[F(X)]$ segue ou não uma distribuição específica $[F_0(X)]$, neste caso a normal, onde:

$$H_0: X \sim N(\mu, \sigma) \quad \text{Vs} \quad H_1: X \not\sim N(\mu, \sigma)$$

E a estatística de teste é dada por:

$$D_n = \max\{|F(x_i) - F_0(x_i)|, |F(x_{i-1}) - F_0(x_i)|\}$$

em que $F_0(X) \sim N(\mu, \sigma)$.

Por outro lado, e uma vez que os pressupostos podem não se verificar, a alternativa não-paramétrica à ANOVA *one way*, o teste de Kruskal-Wallis, é uma boa solução.

Este teste, ao invés da ANOVA *one way*, baseia-se nas medianas populacionais e não nas médias, e tem como hipóteses a testar:

$$H_0: \theta_1 = \theta_2 = \dots = \theta_k \quad \text{Vs} \quad H_1: \exists i, j: \theta_i \neq \theta_j (i \neq j; i, j = 1, \dots, k)$$

E cuja estatística de teste, apenas aplicada após a ordenação por ordem crescente de todas as observações das diferentes amostras em conjunto, e atribuindo a cada observação a sua ordem na amostra global e mantendo a origem da observação (perante

empates, a ordem dessas observações é dada pela média aritmética das ordens que essas observações teriam se não fossem empatadas), é dada por:

$$H = \frac{\frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(N+1)}{1 - \frac{\sum_{i=1}^g (t_i^3 - t_i)}{N^3 - N}}$$

onde $R_j = \sum_{i=1}^{n_i} r_{ij}$ representa a soma das ordens de cada uma das j ($j=1, \dots, k$) amostras e N a dimensão da amostra global. Deste modo, rejeita-se H_0 se $H \geq \chi_{1-\alpha, (k-1)}^2$.

Tal como no caso da ANOVA, o teste de Kruskal-Wallis também não permite por si só observar quais são os pares de populações que diferem entre si, e torna-se necessário, mais uma vez, recorrer à aplicação de testes *a posteriori* que permitam que estes sejam identificados.

Assim, as hipóteses a testar para as combinações possíveis (duas a duas) são dadas por:

$$H_0: F(X_1) = F(X_2) \text{ Vs } H_1: F(X_1) \neq F(X_2) (i \neq j; i, j = 1, \dots, k)$$

E a estatística de teste é dada por:

$$T_r = \frac{|\bar{R}_i - \bar{R}_j|}{\sqrt{S_r^2 \left(\frac{N-1-H}{N-k} \right) \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}}$$

em que as médias das ordens das amostras i e j são dadas por:

$$\bar{R}_i = \frac{1}{n_i} \sum_{l=1}^{n_i} r_{il} \text{ e } \bar{R}_j = \frac{1}{n_j} \sum_{l=1}^{n_j} r_{jl}$$

e a variância de todas as ordens é:

$$S_r^2 = \frac{1}{N-1} \left(\sum_{j=1}^k \sum_{i=1}^{n_j} r_{ij}^2 - \frac{N(N+1)^2}{4} \right)$$

3.2. Regressão logística e modelos lineares generalizados

A análise de regressão compreende um conjunto de técnicas estatísticas que permitem a modelação e investigação de possíveis relações entre diversas variáveis bem como a previsão de valores referentes à variável dependente para valores não observados da variável independente.

Uma variável pode ser determinada em virtude da magnitude de outra, sem que o inverso seja também válido, no entanto, algumas variáveis encontram-se apenas correlacionadas, não dependendo umas das outras. Todavia, uma análise de correlação não é equivalente a uma análise de regressão, em que a primeira analisa o grau de associação entre as variáveis, e a segunda estuda a relação funcional entre duas variáveis.

No entanto, e dependendo do tipo de variáveis disponíveis, podemos distinguir entre regressão *linear* e *categórica*, em que a segunda é usada em detrimento da primeira quando estamos perante uma variável dependente qualitativa que assume apenas valores de classes discretas e mutuamente exclusivas. Mais precisamente, a técnica de regressão logística é utilizada quando temos uma variável dependente do tipo nominal dicotómico (Hosmer e Lemeshow, 1989), tal como é o nosso caso.

Uma vez que um dos nossos objectivos passa pela modelação das relações entre diversas variáveis consideradas e predizer o valor de uma variável dependente a partir de um conjunto de variáveis independentes, em que a nossa variável resposta é binária (migrou/não migrou), a metodologia utilizada foi a da regressão categórica, mais especificamente, a regressão logística.

De uma forma geral, a função aqui usada para que consigamos estimar a probabilidade de uma dada realização j ($j = 1, \dots, n$) da variável dependente ser o “sucesso”, em que $P\{Y_j = 1\} = \hat{\pi}_j$, é:

$$\hat{\pi}_j = \frac{e^{\beta_0 + \beta_1 X_j}}{1 + e^{\beta_0 + \beta_1 X_j}}$$

Isto com uma só variável independente, pois mais do que uma será:

$$\hat{\pi}_j = \frac{e^{\beta_0 + \beta_1 X_{1j} + \dots + \beta_p X_{pj}}}{1 + e^{\beta_0 + \beta_1 X_{1j} + \dots + \beta_p X_{pj}}}$$

No entanto, será ainda necessário realizar uma transformação de $\hat{\pi}_j$ por forma a obtermos o resultados final, e neste caso é a transformação *logit*.

$$g(x) = \ln \left[\frac{\hat{\pi}_j}{1 - \hat{\pi}_j} \right]$$

Por outro lado, o facto dos erros do modelo de regressão logística não seguirem uma distribuição normal nem apresentarem uma variância constante, a obtenção das estimativas não pode ser efectuada através da aplicação do método dos mínimos quadrados, como na regressão linear, mas sim através do método da máxima verosimilhança, em que a função de verosimilhança é dada por:

$$LL = \sum_{j=1}^n \{y_j \ln(\hat{\pi}_j) - (1 - y_j) \ln(1 - \hat{\pi}_j)\}$$

Tal como vimos, o que distingue entre um modelo de regressão linear e um modelo de regressão logística, é o facto de neste último a variável dependente ser uma variável binária ou dicotómica (Hosmer e Lemeshow, 1989), estando todavia, conectado com a família exponencial, o que faz com que a regressão logística seja parte integrante dos Modelos Lineares Generalizados (MLG). Sendo uma síntese de vários outros modelos da dita família exponencial com o intuito de unificar tanto teórica como conceptualmente a técnica da modelação estatística.

Os MLG surgem assim, como sendo uma extensão do modelo linear clássico, em que a distribuição considerada pertence à família exponencial.

No entanto, há que entender o que se pretende afirmar quando se fala da família exponencial, deste modo, uma determinada variável aleatória pertence à família exponencial quando a sua função densidade de probabilidade (*f. d. p.*) é definida por:

$$f(y|\theta, \phi) = \exp \left\{ \frac{y\theta - b\theta}{a(\phi)} + c(y, \phi) \right\}$$

em que θ e ϕ são parâmetros escalares, $a(\cdot)$, $b(\cdot)$ e $c(\cdot, \cdot)$ são funções reais conhecidas.

Por outro lado, e como já se afirmou anteriormente, os MLG são uma extensão do modelo linear clássico:

$$Y = Z\beta + \varepsilon$$

em que Z é uma matriz de dimensão $n \times p$, associada a um vector $\beta = (\beta_1, \dots, \beta_p)^T$ de parâmetros, e ε vector de erros aleatórios com uma distribuição que se presume $N_n(0, \sigma^2 I)$. O que implica que o valor esperado da variável resposta seja função linear das covariáveis.

Neste sentido, a distribuição não é obrigatoriamente uma distribuição normal, mas sim, e como já foi referido, uma distribuição da família exponencial, e ainda que se mantenha a estrutura da linearidade a função que relaciona o valor esperado e o vector de covariáveis pode assumir qualquer função diferenciável

Segundo Turkmann e Silva (2000), os MLG são assim, caracterizados por uma componente aleatória e por uma componente estrutural ou sistemática.

3.3. Modelação: Regressão Logística

Durante a construção do modelo de regressão logística seguiram-se diferentes etapas com vista a chegar ao melhor modelo possível e que mas capacidade explicativa apresentasse. No entanto, e como veremos adiante algumas dificuldades foram aparecendo, cremos que principalmente pela dimensão da amostra com que trabalhamos, pois tem uma dimensão considerável.

Neste caso, criaram-se duas bases distintas às quais foram aplicados a regressão logística. Uma primeira em que o acto de migrar foi baseado na região de residência cinco anos antes dos recenseamentos em questão e uma segunda, em que o acto em si foi descrito pelo facto de à altura dos recenseamentos qualquer individuo não residir na sua região de nascimento.

Uma vez que a estratégia de modelação está descrita adiante, aquando da apresentação dos resultados, aqui ficam somente alguns aspectos teóricos que podem ajudar à sua compreensão.

Deste modo, os *p-values* obtidos para as variáveis predictoras quer individualmente, quer em conjunto com outras, foram obtidos através do teste de *Wald*:

$$W = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$$

No entanto, este teste é um teste conservativo, pois por vezes falha a rejeição da hipótese nula quando o coeficiente é significativo.

Deste modo, uma alternativa ao teste de *Wald*, é a utilização da razão de verosimilhanças para testar a significância da variável predictor. Para isso torna-se necessário ajustar um modelo sem a referida variável e outro com, de forma a se medir o aumento no valor de $-2LL$ e testando a sua significância pelo teste do qui-quadrado $[\chi^2]$:

$$\begin{aligned} G &= Deviance(modelo sem variável) - Deviance(modelo com variável) \\ &= -2 \ln \left[\frac{Verosimilhança sem variável}{Verosimilhança com variável} \right] \end{aligned}$$

Após atingimos o modelo final, com todas as variáveis e interações entre variáveis significativas incluídas, podemos olhar para o valor do teste de *Hosmer e Lemeshow*, que nos indica se o modelo se ajusta aos dados ou não:

$$X_P^2 = \sum_{j=1}^J \left(\frac{y_i - n_j \hat{\pi}_j}{\sqrt{n_j \hat{\pi}_j (1 - \hat{\pi}_j)}} \right)^2 = \sum_{j=1}^J \frac{(O_j - E_j)^2}{E_j}$$

em que os dados estão agrupados em J células, y_i o número de sucessos na célula j , $\hat{\pi}_j$ a probabilidade estimada da célula j , e n_j o número de elementos da célula j . Esta terminologia é o equivalente à utilização da notação tradicional do qui-quadrado, onde O_j corresponde ao número de sucessos observados e E_j ao número de sucessos estimados, ambos na célula j .

Tal como o teste de *Hosmer e Lemeshow*, que serve para avaliar a bondade do ajustamento, podemos ainda utilizar a chamada curva “ROC”, que é um teste que tende

a ser avaliado por duas medidas, a sensibilidade e a especificidade, e serve para avaliar a capacidade discriminante do modelo.

É o valor referente à área abaixo da curva que nos vai indicar a capacidade discriminante do modelo e permitir a sua comparação com curvas de diferentes testes (tabela 8).

Tabela 8 – Capacidade de Discriminação da curva “ROC”

$A \leq 0,5$	Não existe Discriminação
$0,6 \leq A < 0,7$	Discriminação Pobre
$0,7 \leq A < 0,8$	Discriminação Aceitável
$0,8 \leq A < 0,9$	Discriminação Boa
$0,9 \leq A < 1$	Discriminação Excelente

Passando agora à análise de resíduos, e começando pelos resíduos *deviance*:

$$d(y_j, \hat{\pi}_j) = \pm \left\{ 2 \left[y_j \ln \left(\frac{y_j}{m_j \hat{\pi}_j} \right) + (m_j - y_j) \ln \left(\frac{m_j - y_j}{m_j (1 - \hat{\pi}_j)} \right) \right] \right\}^{1/2}$$

Os valores associados aos resíduos *deviance* devem variar entre -2 e 2, de forma a nenhuma observação ser tida como aberrante.

Por outro lado, e ainda servindo como ferramenta de diagnostico, podemos utilizar a representação gráfica dos resíduos *Dfbeta*, que nos permitem registar quais as observações que revelam um mau ajustamento ao modelo, medido o nível de alteração de coeficientes quando se elimina uma dada observação:

$$\Delta\beta_j = \frac{(X'CX)^{-1}X'_je_j}{1 - h_j}$$

Por último, podemos ainda referir a *distância de Cook*, que nos permite observar se existem pontos influentes ou não entre as observações em estudo. A distância de *Cook* é assim, um indicador da influência da i-ésima observação no vector de parâmetros estimado e pode ser calculada através da seguinte fórmula:

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_1)^T (Z^T W Z) (\hat{\beta} - \hat{\beta}_1)}{p \hat{\phi}}$$

em que $\hat{\beta}$ e $\hat{\beta}_1$ representam as estimativas de máxima verosimilhança do vector parâmetro β obtidas da amostra sem a observação (y_i, z_i) e da amostra com todas as observações.

3.4. O teste de independência do χ^2

Face a inúmeras dificuldades com que nos deparamos na aplicação do modelo logístico, e também devido a termos aplicado o mesmo modelo mas com duas diferentes interpretações do acto de migrar (uma baseada na região de nascimento, e outra na região de residência cinco anos antes do recenseamento geral da população), optou-se no final, por testar se estas duas variáveis se encontravam correlacionadas.

Para o efeito, foi aplicado o teste de independência do χ^2 , teste este que permite testar se dois ou mais grupos independentes diferem relativamente a uma dada característica. Neste caso foi considerado o acto de migrar por grupos de idade.

A estatística de teste é dada por:

$$\chi^2 = \sum_{i=1}^L \sum_{j=1}^C \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

em que C e L são o número de colunas e linhas, respectivamente, e as frequências esperadas são dadas por:

$$E_{ij} = \frac{L_i \times C_j}{N}$$

onde:

$$L_i = \sum_{j=1}^c o_{ij}$$

e

$$C_j = \sum_{i=1}^L o_{ij}$$

Referindo-se a primeira aos totais em linha, e a segunda aos totais em coluna.

4. Estimação Indirecta dos Fluxos Migratórios Internos

Tal como referimos inicialmente, o estudo da temática das migrações engloba bastantes dificuldades, pois além de não se poder definir uma característica única, deparamo-nos também com a inexistência de dados, não nos permitindo realizar uma análise fiável dos fluxos migratórios existentes.

Deste modo, uma das formas de podermos contornar estas dificuldades passa pela aplicação de metodologias de estimação indirecta (Rogers *et al.*, 2002), que nos permitem, além de estimar, observar e analisar os padrões de migração inter-regionais em Portugal continental, utilizando para isso, numa primeira etapa, a informação referente à propensão migratória das crianças, numa segunda, a informação relativa às taxas líquidas de migração, e numa terceira, conjugar a propensão migratória de “primeiros” e “segundos” migrantes.

Mais do que uma aproximação à realidade das migrações internas, a aplicação desta metodologia permitir-nos-á identificar possíveis padrões migratórios. Simultaneamente, a aplicação de três métodos distintos possibilita também obter diferenciadas aproximações à realidade das migrações internas portuguesas que apenas se torna possível através da utilização de uma abordagem indirecta.

Os métodos de estimação indirecta são, muitas vezes, compostos por duas partes, uma referente à representação matemática do processo em estudo e sua realização, e outra referente à metodologia que se pretende utilizar para identificar um dos muitos padrões gerados pela referida representação matemática (Rogers e Liu, 2005).

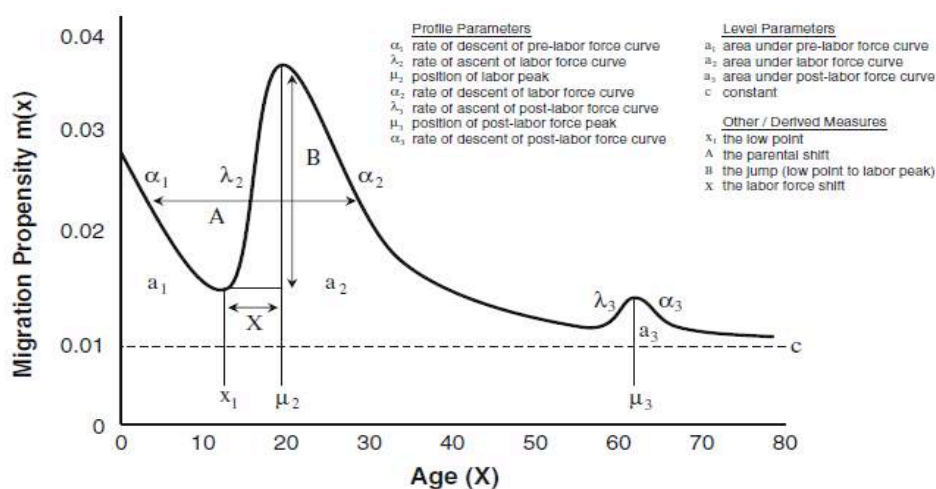
4.1. Especificidades Etárias

A observação atenta dos movimentos migratórios permite-nos detectar a existência de algumas especificidades que lhe são próprias, tais como as diferenças nas probabilidades de migrar associadas à idade (Raymer e Rogers, 2007), o que constitui uma das condições base para o procedimento da aplicação da estimação indirecta à temática das migrações.

Segundo Rogers *et al.* (2007), as probabilidades mais elevadas ocorrem nos primeiros anos de vida adulta do indivíduo, quando, por exemplo, estes deixam a casa dos pais para frequentar um grau de ensino mais elevado, ingressar no meio militar, constituir família, ou simplesmente começar a trabalhar, resultando no que frequentemente se apelida de “pico laboral” (*gráfico 3*). Inversamente, as probabilidades de migrar mais baixas registam-se na fase final da adolescência e, normalmente, após a entrada no mercado de trabalho até ao início da reforma.

No que concerne à probabilidade de migrar das crianças, constata-se que esta é reflexo da migração dos pais, geralmente jovens adultos, e enquanto a migração na infância atinge valores mais elevados do que na adolescência, a entrada na idade de reforma, em especial nos países desenvolvidos, conduz a um aumento da probabilidade migratória que se traduz num “pico de retorno” perto dos 65 anos.

Gráfico 3 – Padrões específicos da Migração por Grupos Etários



Fonte: Rogers *et al.* (2007)

O modelo de migração completo (Rogers *et al.*, 2007) é formado por quatro componentes: (1) força pré-laboral (crianças); (2) força laboral (adultos); (3) força pós-laboral (idosos); e (4) curva constante; o que se pode traduzir através da seguinte expressão:

$$m(x) = N_1(x) + N_2(x) + N_3(x) + c$$

$$m(x) = a_1 \exp(-\alpha_1 x) + a_2 \exp\{-\alpha_2(x - \mu_2) - \exp[-\lambda_2(x - \mu_2)]\}$$

$$a_3 \exp\{-\alpha_3(x - \mu_3) - \exp[-\lambda_3(x - \mu_3)]\}$$

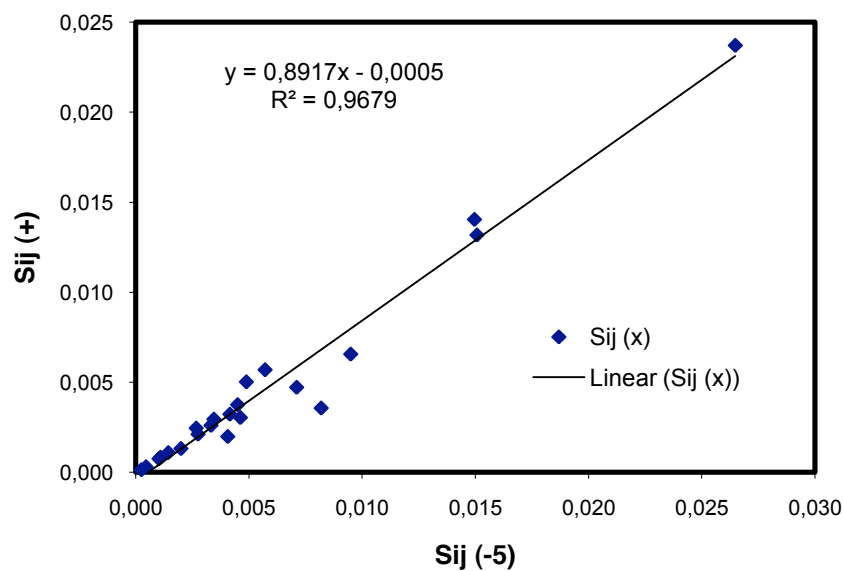
Onde $m(x)$ é probabilidade de migrar na idade x ; N_1 a força pré-laboral (crianças); N_2 a força laboral (adultos); N_3 a força pós-laboral (idosos); c a constante; α e μ , os parâmetros; e x a idade.

4.2. A Estimação Baseada na Propensão Migratória das Crianças

As especificidades etárias observadas nos padrões migratórios parecem, de certa forma, sugerir a existência de uma ligação entre as probabilidades migratórias das crianças e as probabilidades correspondentes aos restantes grupos de idades (*gráfico 4*). Situação que pode ser observada através da consideração de equações de regressão que relacionem a migração dos restantes grupos de idade $S_{ij}(x)$ com a propensão migratória das crianças $S_{ij}(-5)$:

$$S_{ij}(x) = a + b S_{ij}(-5) + \varepsilon$$

Gráfico 4 – Propensão migratória total com base na propensão migratória infantil, 1991
(Portugal Continental)



A utilização da propensão migratória das crianças permite, deste modo, utilizar a técnica da regressão linear simples entre a proporção de crianças dos 0-4 anos que nasceram em i e vivem em j , e a proporção de pessoas que, em cada grupo etário, viviam em i há cinco anos e à data do recenseamento vivem em j . O que permite explicar a probabilidade específica de migrar, por idade, a partir de um parâmetro estimado com base na informação da migração infantil (Rogers *et al.*, 2003).

Esta metodologia assenta na concepção de que as migrações dos adultos dependem, linearmente, das migrações das crianças entre os 0 e os 4 anos (dito de outra forma, as

crianças quando migram, não migram sozinhas, mas sim acompanhadas por familiares). Quando estas crianças não residem no local onde nasceram é porque migraram pelo menos uma vez, e como estes indivíduos nasceram em média cerca de dois anos e meio antes, é pouco provável que tenham migrado efectivamente mais do que uma vez. Por outro lado, como parecem existir padrões de migrações semelhantes nas várias idades, essa relação linear que explica a variação das migrações dos adultos e jovens em função das crianças, é alargada aos restantes grupos etários (Rogers *et al.*, 2007).

À semelhança das estimativas indirectas da mortalidade, que se baseiam numa única taxa de mortalidade infantil para estimar a curva de mortalidade em todas as idades, as probabilidades migratórias resultam da estimativa das “taxas de sobrevivência” dos migrantes numa determinada idade x , representada por $S_{ij}(x)$:

$$S_{ij}(x) = \frac{Migrantes_{ij}(x)}{População\ Total_i(x)}, x = idade$$

Neste sentido, $Migrantes_{ij}(x)$ representa o número de indivíduos que à data do recenseamento residem num determinado local j , mas que 5 anos antes residiam em i e $População\ Total_i(x)$ representa o somatório de todos os indivíduos, com idade x , que residiam em i , 5 anos antes da data do recenseamento. Deste modo, $S_{ij}(x)$ é uma medida que traduz, para um indivíduo com uma idade x e que reside em i , a probabilidade de sobrevivência em j , t anos mais tarde (neste caso $t=5$).

Para estimar as taxas de sobrevivência específicas das migrações, calcula-se em primeiro lugar o rácio de migração infantil $-r_{ij}(x, -5)$, também denominado de ATI (*age-to-infant migration ratio*), que traduz a proporção entre a probabilidade de migrar em qualquer idade e a probabilidade de migração das crianças com idades compreendidas entre os 0 e os 4 anos (ou seja, que nasceram todas até 5 anos antes):

$$r_{ij}(x, -5) = \frac{S_{ij}(x)}{S_{ij}(-5)}, x = 0, 5, 10, \dots, 70 +$$

Este rácio permite-nos obter estimativas para $S_{ij}(x)$ 10 anos mais tarde (intervalo usual entre recenseamentos):

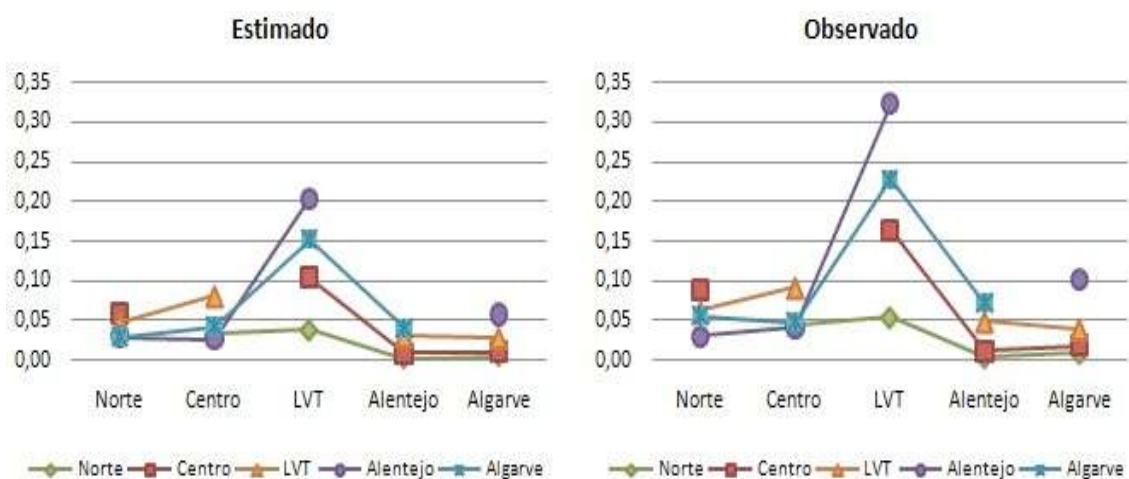
$$\hat{S}_{ij}^t = r_{ij}^{t-10} S_{ij}^t(-5)$$

Iniciando a nossa análise de resultados através de uma representação gráfica global dos resultados obtidos (*gráficos 5 e 6*), constatamos que, apesar da existência de um desfasamento significativo entre os valores estimados pela primeira metodologia e os efectivamente observados, a região de Lisboa e Vale do Tejo se destaca como a mais atractiva.

Por outro lado, podemos também referir o facto de, o factor proximidade deter um papel determinante quando se decide migrar, pois se o destino preferencial dos migrantes não recair sobre Lisboa e Vale do Tejo, a sua opção preferencial é sempre a região geograficamente mais próxima.

Apesar das regiões do Alentejo e Algarve serem as que menos residentes registam em Portugal Continental, destaca-se o facto de serem estas que *contribuem* com a maior proporção de migrantes, direccionando-se essencialmente para a região de Lisboa e Vale do Tejo. Neste sentido, importa referir também, a existência de um maior volume de movimentos efectivamente observados comparativamente com os estimados através da propensão migratória das crianças.

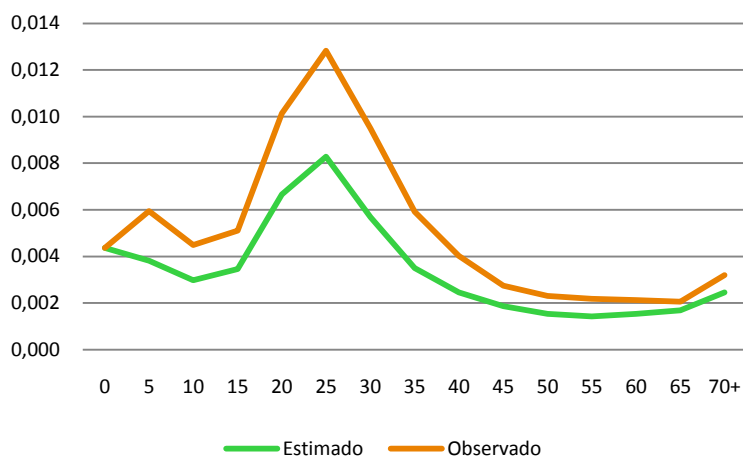
Gráficos 5 e 6 – Fluxos migratórios estimados e observados (Método 1)



Representando graficamente os resultados *médios* obtidos por grupos de idade (*gráfico 7*), apuramos mais objectivamente o que foi referido na análise inter-regional, ou seja, a existência de um desfasamento significativo entre as estimativas e as observações, em que os valores estimados se encontram aquém dos realmente observados. No entanto, este desfasamento é apenas mais significativo entre os 20 e os 34 anos completos, registando-se uma diminuição do seu afastamento nas idades mais avançadas.

Na situação específica do primeiro grupo de idades, ou seja, entre os 0 e os 4 anos, os valores representados não apresentam nenhum distanciamento devido a ser este o ponto de partida para as estimativas dos grupos etários seguintes.

Gráfico 7 – Fluxos migratórios estimados e observados por grupos de idade (Método 1)



4.3. A Estimação Baseada na Migração Líquida Específica por Idade

Analiseemos em seguida as especificidades etárias observadas nos padrões migratórios de um outro ângulo, o da migração líquida (Rogers e Liu, 2005). A migração líquida pode obter-se (1) através da diferença entre a população recenseada e a estimada para um determinado período (t), utilizando apenas os dados referentes à fecundidade e mortalidade, pois a estimativa efectuada deixa de parte as migrações (através da Equação de Concorância, por exemplo), ou (2) simplesmente pela diferença entre os valores da imigração e emigração ($I - E$).

No entanto, se não houver disponibilidade desses dados ou mesmo lacunas de informação ou má qualidade dos dados, torna-se possível efectuar a estimativa dos fluxos migratórios através do método padrão residual, em que se pressupõe que quaisquer mudanças na população se devem à migração líquida e não ao seu crescimento natural (positivo ou negativo).

Deste modo, utiliza-se o que podemos denominar de rácio de migração etária total (age-to-age summed), através do qual podemos estimar os fluxos migratórios entre as diversas regiões através de uma regressão linear entre uma variável dependente e duas variáveis independentes (Rogers *et al.*, 2006), segundo a fórmula seguinte:

$$N_i^t(x) = \beta_1 X_{1i} + \beta_2 X_{2i}$$

Onde:

$N_i^t(x)$ = migração líquida da região i , por idades e tempo t ;

β_1 = primeiro parâmetro, a partir do qual se estima $S_{ji}^t(+)$;

$X_{1i} = \frac{K_j^t(x)}{K_i^t(x)} r_{ji}^t(x, +)$;

β_2 = segundo parâmetro, a partir do qual se estima $S_{ij}^t(+)$;

$X_{2i} = r_{ij}^t(x, +)$.

Neste sentido, as estimativas do total de movimentos são dadas por $S_{ij}^t(+)$ =

$\sum_{x=1}^n S_{ij}^t(x)$, e os perfis etários definidos pela relação proporcional $r_{ij}^t(x, +) = \frac{S_{ij}^t(x)}{S_{ij}^t(+)}$.

Finalmente, as estimativas por grupo etário são dadas por $S_{ij}^t(x) = S_{ij}^t(+)$ $r_{ij}^t(x, +)$, que, no seu conjunto, resultam na seguinte expressão:

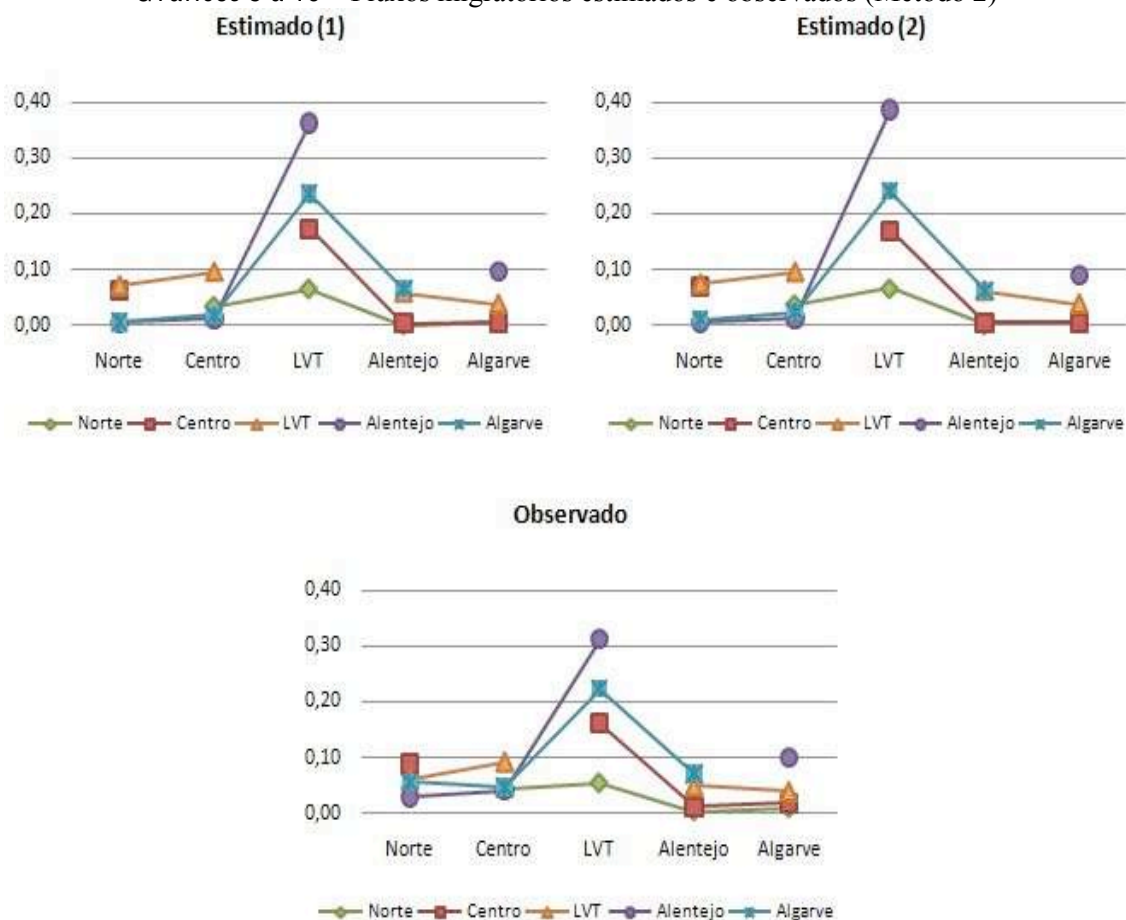
$$N_i^t(x) = \frac{K_j^t(x)}{K_i^t(x)} [S_{ji}^t(+)] r_{ji}^t(x, +) - [S_{ij}^t(+)] r_{ij}^t(x, +)$$

Uma vez que esta metodologia gera dois coeficientes distintos (β_1 e β_2), estes permitem realizar duas estimativas para cada movimento migratório, tal como veremos de seguida.

Analisando agora, a representação gráfica dos resultados obtidos através do segundo método (*gráficos 8 a 10*), constatamos que, tal como no primeiro método, a região mais atractiva para os migrantes é a região de Lisboa e Vale do Tejo e que, os padrões migratórios associados à proximidade geográfica também se mantêm. No entanto, e uma vez que as duas metodologias são diferentes e se baseiam na questão “*onde residia há cinco anos atrás?*”, as estimativas efectuados serão, seguramente diferentes, efectuando estimativas globais mais elevadas comparativamente à primeira metodologia.

Por outro lado, e uma vez que esta segunda metodologia nos permite obter duas estimativas para o mesmo movimento de uma só vez, importa ainda referir também que o desfasamento registado entre as duas não é digno de grande relevância, tal como veremos em seguida.

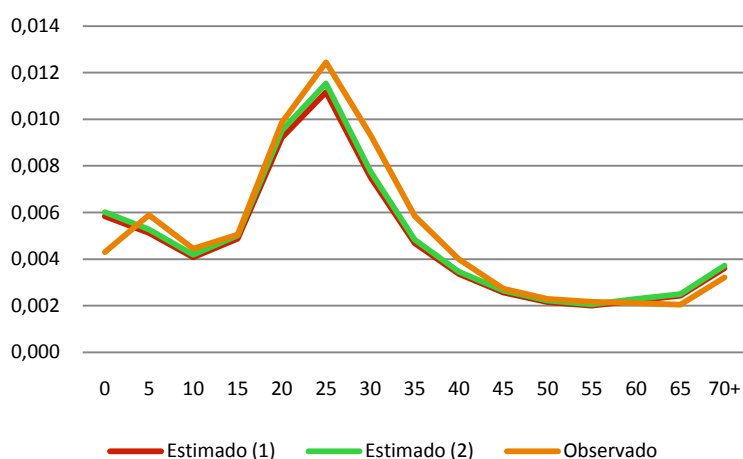
Gráficos 8 a 10 – Fluxos migratórios estimados e observados (Método 2)



Quanto aos resultados *médios* obtidos por grupos de idade (*gráfico 11*), verificamos de imediato a proximidade entre as duas estimações efectuadas. Contudo, contrariamente à primeira metodologia, os valores estimados não se encontram claramente abaixo dos efectivamente verificados, pelo contrário, registando-se situações em que as estimativas ultrapassam as observações.

Uma outra diferença relevante, relativamente à primeira metodologia apresentada, reside no facto de nos ser permitido estimar as proporções referentes às crianças com idades até aos 5 anos exactos (grupo de idades 0-4), o que nos possibilita, neste caso, constatar a existência de um desfasamento mais significativo entre registos efectivos e estimativas.

Gráfico 11 – Fluxos migratórios estimados e observados por grupos de idade (Método 2)



4.4. A Estimação Baseada em duas Distribuições Sucessivas de Stocks Populacionais

Contrariamente aos dois métodos anteriores, que se baseiam no local de residência cinco anos antes do recenseamento, os movimentos migratórios - $S_{ij}(x)$ - calculados através desta metodologia, assentam na região de nascimento. Através do “cruzamento” entre a região de nascimento e aquela onde residia cinco anos antes do recenseamento, este método permite-nos distinguir entre “primeiros” (indivíduos que migram a partir da sua região de nascimento) e “segundos” migrantes (indivíduos que migram a partir de uma região que não é o seu local de nascimento).

Por último, o terceiro método proposto foi desenvolvido com base na seguinte fórmula $S = K^{t+5}(x + 5) [K^t(x)^{-1}]$, onde K^t é a matriz com os elementos ${}_kK_j^t(x)$ que apresenta o número de indivíduos que residem na região j , com a idade x , no ano t , e que nasceram na região k .

Neste sentido, tomando a mesma probabilidade migratória para todos os migrantes, podemos concluir que:

$$K^{t+5} = SK^t$$

Sendo apenas necessário transformar K^t em K^{t+5} , invertendo a matriz original e resolvendo para S :

$$\hat{S} = K^{t+5}[K^t]^{-1}$$

No entanto, e como se verificará mais à frente neste trabalho, as probabilidades migratórias são diferentes entre aqueles que realizam uma primeira migração e os que já vão numa segunda. Assim, e uma vez que a probabilidade de migrar é mais elevada no caso dos migrantes secundários, torna-se necessário relacionar as duas probabilidades com o intuito de se estimar com maior precisão os diversos movimentos migratórios.

Ledent, citado por Rogers *et al.* (2003), propôs então uma alteração a esta metodologia, que se baseia essencialmente na criação de um rácio, tal como nas duas metodologias anteriores, que se obtém através de um ano t , e posteriormente será aplicado ao ano $t+n$:

$$r_{ij}(x_s, x_p) = \frac{{}_kS_{ij}(x)}{{}_iS_{ij}(x)} = \frac{{}_sS_{ij}(x)}{{}_pS_{ij}(x)}$$

Onde ${}^PS_{ij}(x)$ corresponde à propensão migratória dos primeiros migrantes, e ${}^SS_{ij}(x)$ à dos segundos. Importa também referir que, este rácio (STP – secondary to primary migration ratio), fixa a probabilidade de migrar dos primeiros migrantes na unidade (1), enquanto para segundos migrantes essa probabilidade é superior a 1, e para aqueles que ficam na mesma região, ainda que esta seja diferente da sua região de nascimento, valores inferiores a 1.

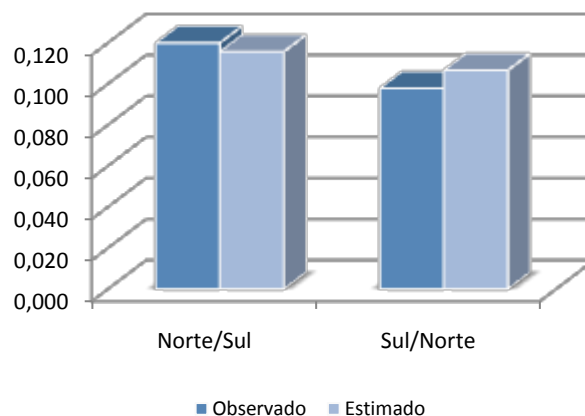
Tal com já afirmamos, este terceiro método difere bastante dos anteriores, na medida em que se torna necessário saber o local de nascimento e, se cinco anos antes, o migrante se moveu directamente da sua região de nascimento (“primeiro migrante”), ou de uma outra (“segundo migrante”). Deste modo, e uma vez que, as crianças com menos de cinco anos não podem ter residido numa outra região há cinco anos atrás, o primeiro grupo de idades, que contempla as crianças com idades até aos 4 anos completos, será deixado de parte.

Por outro lado, o facto de não termos acesso a uma base de dados que nos permita cruzar as diferentes variáveis e obter as tabelas necessárias para esta abordagem, optamos por recorrer à amostra que nos foi facultada. Neste sentido, e uma vez que se trata de uma amostra, por não terem sido contempladas observações para todos os casos, não conseguimos aplicar esta metodologia às cinco NUTS II continentais, tendo sido necessário optar pelo seu agrupamento em apenas duas regiões (Norte e Sul), em que o Norte é composto pelas regiões Norte e Centro, e o Sul, pelas de Lisboa e Vale do Tejo, Alentejo e Algarve.

Começando, por uma abordagem global, ou seja, uma abordagem que apenas contemple a totalidade dos movimento migratórios estimados e observados entre as regiões do Norte e do Sul, podemos constatar que, ainda que a diferença não seja muito significativa, é a região Norte que mais migrantes atrai relativamente ao Sul. Por outro lado, as estimativas efectuadas também não registam um afastamento significativo dos movimentos realmente observados (*gráfico 12*).

No entanto, e tal como nos dois métodos anteriores, esta é uma situação que não é a mais indicada para verificar as diferenças entre valores efectivos e estimativas, pois a diferença real, os seus erros serão sempre calculados tendo em conta a idade, como veremos mais adiante.

Gráfico 12 – Fluxos migratórios estimados e observados (Método 3)



Após uma análise global, e observando os *gráficos 13 e 14*, constatamos que, tal como foi dito anteriormente, o Norte é a região que mais população atrai, sendo na sua maioria jovens adultos entre os 20 e os 34 anos completos. Por outro lado, apesar de ser também sobre estas idades que o Sul exerce mais atractividade, estes são em muito menor número, e tomando o grupo de idades dos 25 aos 29 anos, verificamos que o Norte conseguiu atrair cerca de 2,5% daqueles que residiam no Norte cinco anos antes, e o Sul apenas conseguiu atrair do Norte 1,3%.

Podemos ainda acrescentar que, as estimativas realizadas se mantêm muito perto dos valores efectivamente observados até cerca do grupo de idades 50 – 54 anos, no caso da direcção Norte/Sul, e 35 – 39 no caso Sul/Norte.

Gráficos 13 e 14 – Fluxos migratórios estimados e observados por grupos de idades (Método 3)



4.5. Avaliação Crítica dos Resultados

A aplicação de uma metodologia com o intuito de se efectuar uma qualquer previsão, encontra-se sempre associada a um determinado erro (Rogers *et al.*, 2007), pois apesar de ser possível identificar o padrão migratório dos movimentos inter-regionais, a sua quantificação é uma dificuldade acrescida.

Neste sentido, e uma vez que podemos aceder aos valores efectivamente observados, torna-se pertinente avaliar os resultados obtidos através dos cálculos referentes ao Erro Percentual Absoluto Médio (MAPE), calculando para cada fluxo migratório:

$$MAPE = \frac{\sum_x \frac{\|\hat{S}_{ij}^{2001}(x) - S_{ij}^{2001}(x)\|}{S_{ij}^{2001}(x)}}{N} \times 100$$

Enquanto que para calcular o erro total, ou seja, o erro para o conjunto de todos os fluxos migratórios, usamos:

$$MAPE_{ij} = \frac{\sum_{i=1}^n \sum_{j \neq i}^n \sum_x \frac{\|\hat{S}_{ij}^{2001}(x) - S_{ij}^{2001}(x)\|}{S_{ij}^{2001}(x)}}{n(n-1)N} \times 100$$

No entanto, e como o MAPE apenas nos dá uma quantificação absoluta do erro associado, não nos permitindo observar o seu real comportamento, uma outra alternativa será o Erro Percentual Algébrico Médio (MALPE), realizando-se para isso uma ligeira alteração nas fórmulas acima transcritas (N = número total de grupos de idade):

$$MALPE = \frac{\sum_x \frac{\hat{S}_{ij}^{2001}(x) - S_{ij}^{2001}(x)}{S_{ij}^{2001}(x)}}{N} \times 100$$

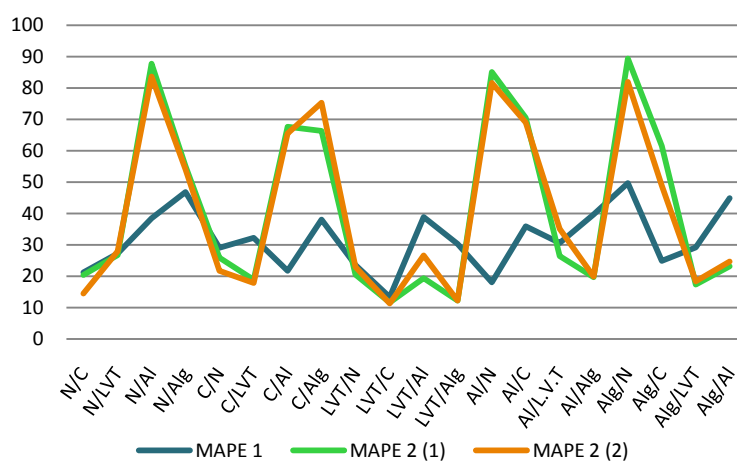
$$MALPE_{ij} = \frac{\sum_{i=1}^n \sum_{j \neq i}^n \sum_x \frac{\hat{S}_{ij}^{2001}(x) - S_{ij}^{2001}(x)}{S_{ij}^{2001}(x)}}{n(n-1)N} \times 100$$

Finalmente, com o intuito de avaliar com maior rigor os resultados obtidos, estes serão também analisados através da significância do modelo ajustado (R^2), ou seja, através da avaliação da capacidade explicativa que os valores estimados têm sobre os valores realmente observados.

Avaliando então os resultados obtidos através das duas metodologias para os fluxos inter-regionais, podemos constatar que os erros obtidos através da utilização do primeiro método são, na sua generalidade, inferiores aos obtidos através da segunda metodologia (*gráfico 15*). No entanto, a segunda metodologia apresenta, erros mais elevados para os movimentos registados entre as duas regiões do norte do país (Norte e Centro) e as duas mais Sul (Alentejo e Algarve), tendo em conta que Lisboa e Vale do Tejo é a região mais a centro e que apesar de ser uma região das mais povoadas, não regista erros dos mais elevados.

Podemos, assim, afirmar que, à excepção de Lisboa e Vale do Tejo, as estimativas efectuadas para fluxos migratórios entre regiões com maior e menor número efectivo de residentes, registam erros totais mais elevados.

Gráfico 15 – Erro Percentual Absoluto Médio (Métodos 1 e 2)

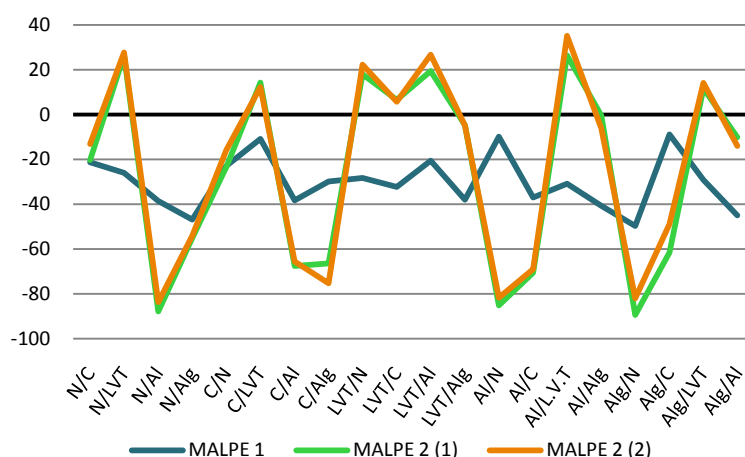


No entanto, os erros apresentados (MAPE), apenas nos transmitem uma ideia do erro global, não permitindo uma avaliação mais profunda, e para isso teremos que recorrer ao MAPLE (*gráfico 16*).

A sua representação gráfica, permite-nos concluir que, apesar do primeiro método continuar a ser aquele que apresenta melhores resultados gerais, as suas estimativas ficam sempre aquém dos valores efectivos. Por outro lado, no respeitante ao segundo método, os seus resultados indicam-nos que os maiores erros registados anteriormente também se referem a *subestimativas* quando os restantes se distribuem em torno de zero, ora assumindo valores positivos, ora negativos.

Como era de esperar, os piores resultados obtidos referem-se novamente às regiões destacadas anteriormente.

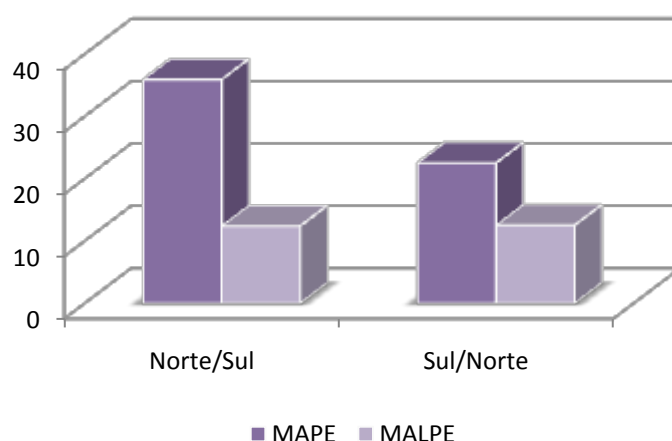
Gráfico 16 – Erro Percentual Algébrico Médio (Métodos 1 e 2)



Uma vez que não nos foi possível aplicar a terceira metodologia ao universo dos dados, e a sua divisão regional não corresponde às NUTS II continentais, a interpretação dos seus erros globais, terá de ser elaborada separadamente (*gráfico 17*).

Apesar de observarmos um erro absoluto (MAPE) associado à migração no sentido Norte/Sul, quando comparado com a sua movimentação inversa, podemos constatar ainda, através do MALPE, que as estimativas efectuadas foram, na sua maioria, elaboradas acima dos valores efectivamente observados.

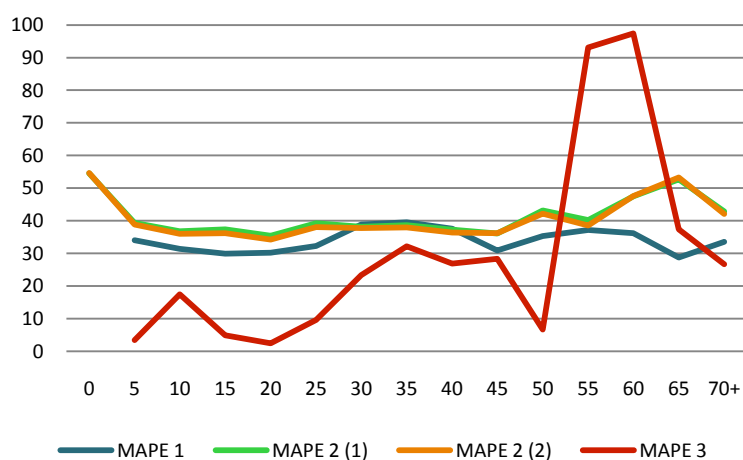
Gráfico 17 – MAPE e MALPE (Método 3)



No respeitante aos erros registados nas estimativas por idades (*gráfico 18*), e aqui já com a representação conjunta da terceira metodologia, verificamos que são, novamente, os resultados obtidos através do primeiro método que registam erros médios absolutos mais baixos comparados com a segunda metodologia, não se denotando, no entanto, o afastamento registado entre as duas metodologias no respeitante aos fluxos migratórios inter-regionais.

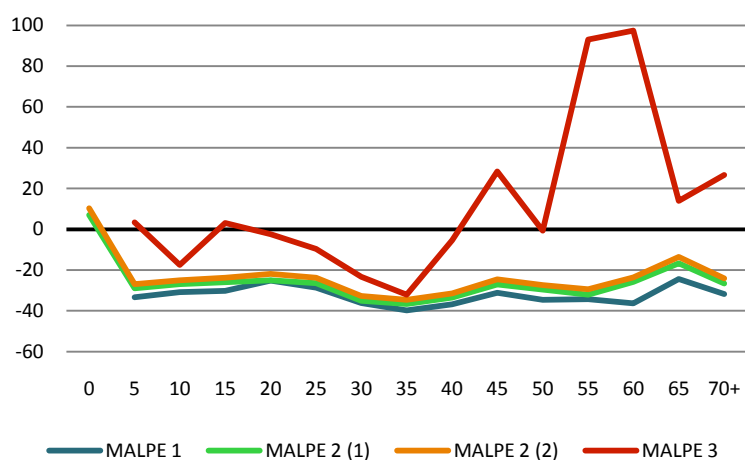
No entanto, e na sua globalidade, são as estimativas efectuadas a partir da aplicação do terceiro método, que registam erros mais baixos, especialmente até aos 54 anos completos, para depois registarem uma subida acentuada e de seguida descer ao nível das duas metodologias anteriores.

Gráfico 18 – Erro Percentual Absoluto Médio por grupos de idade (3 Métodos)



Passando agora à análise do MAPLE (*gráfico 19*), podemos constatar que as estimativas efectuadas pelos dois primeiros métodos ficavam, de uma forma geral, abaixo dos valores efectivamente observados, exceptuando-se no primeiro grupo de idades (0 – 4). O terceiro método, por outro lado, não regista um comportamento constante até ao grupo de idades entre os 45 e os 49 anos, a partir do qual, as suas estimativas foram superiores aos valores efectivos.

Gráfico 19 – Erro Percentual Algébrico Médio (3 Métodos)



Apesar de se terem registado erros bastante elevados, isso não significa que as estimativas obtidas não consigam explicar os valores efectivamente observados.

No que concerne à capacidade explicativa dos fluxos migratórios estimados em relação àqueles que foram efectivamente observados, constatamos que todos eles, independentemente de inter-regionais ou por grupos etários, apresentam uma capacidade explicativa (R^2) bastante significativa (tabelas 4 e 5). Contudo, é na situação referente aos fluxos migratórios inter-regionais que se registam os valores menos significativos.

Enquanto no primeiro método as estimativas efectuadas conseguem explicar entre 71 e 98% dos valores efectivamente registados, a segunda metodologia, e aqui encontramos níveis de significância idênticos para as duas estimativas geradas, varia entre 71 e 96% (tabela 8).

Tabela 8 – Capacidade explicativa das estimativas por fluxo migratório

Reg. 1996	Reg. 2001	R^2		
		Met. 1	Met. 2 (1)	Met. 2 (2)
Norte	Centro	0,96	0,94	0,94
Norte	L.V.T.	0,98	0,96	0,96
Norte	Alentejo	0,94	0,85	0,85
Norte	Algarve	0,99	0,93	0,93
Centro	Norte	0,98	0,95	0,95
Centro	L.V.T.	0,98	0,96	0,96
Centro	Alentejo	0,95	0,94	0,94
Centro	Algarve	0,94	0,90	0,90
L.V.T.	Norte	0,82	0,79	0,79
L.V.T.	Centro	0,88	0,88	0,88
L.V.T.	Alentejo	0,96	0,74	0,74
L.V.T.	Algarve	0,94	0,87	0,87
Alentejo	Norte	0,95	0,95	0,95
Alentejo	Centro	0,96	0,93	0,93
Alentejo	L.V.T.	0,99	0,97	0,97
Alentejo	Algarve	0,91	0,85	0,85
Algarve	Norte	0,78	0,71	0,71
Algarve	Centro	0,71	0,71	0,71
Algarve	L.V.T.	0,98	0,95	0,95
Algarve	Alentejo	0,84	0,71	0,71
Total	Total	0,96	0,94	0,94

No respeitante à situação por idades, podemos constatar o que foi acima referido, pois a percentagem mínima da capacidade explicativa subiu para 86% no primeiro método, e para 88% e 89%, nas duas estimativas geradas pela segunda metodologia (tabela 9). Por outro lado, os máximos subiram também para 98%, no caso da primeira metodologia, e para 99% nas duas estimativas correspondentes ao segundo método.

Tabela 9 – Capacidade explicativa das estimativas por grupos de idade

Grupos de Idade	R ²		
	Met. 1	Met. 2 (1)	Met. 2 (2)
0 - 4	-	0,93	0,94
5 - 9	0,97	0,90	0,91
10 - 14	0,96	0,91	0,91
15 - 19	0,98	0,96	0,97
20 - 24	0,98	0,99	0,99
25 - 29	0,97	0,97	0,97
30 - 34	0,95	0,95	0,95
35 - 39	0,97	0,93	0,93
40 - 44	0,92	0,90	0,89
45 - 49	0,94	0,94	0,93
50 - 54	0,90	0,95	0,95
55 - 59	0,86	0,95	0,96
60 - 64	0,96	0,95	0,95
65 - 69	0,91	0,88	0,87
70 +	0,98	0,99	0,98
Total	0,96	0,94	0,94

Quanto ao terceiro método aplicado, e tal como afirmámos anteriormente, o facto de não termos podido utilizar a mesma base de dados, a capacidade explicativa apenas foi passível de calcular para os movimentos inter-regionais, deixando de parte a sua divisão por grupos de idade, bem como para o seu total (tabela 10).

Assim, constatamos que, apesar de se registar uma capacidade explicativa de 90% dos valores estimados em relação ao observado na direcção Norte/Sul, o sentido inverso apenas consegue uma capacidade explicativa de 39%, o que não invalida que, na sua globalidade, as estimativas efectuadas consigam explicar 82% dos valores efectivamente registados.

Tabela 10 – Capacidade explicativa Método 3

Reg. 1996	Reg. 2001	R ²
Norte	Sul	0,90
Sul	Norte	0,39
Total	Total	0,82

5. Redução do Erro da Estimativa Indirecta

Tal como se verificou no capítulo anterior, registaram-se alguns erros bastante elevados, especialmente quando analisados os movimentos migratórios entre regiões com maior e menor dimensão populacional, exceptuando o caso da região de Lisboa e Vale do Tejo. Neste sentido, e uma vez que apenas os dados usados para as duas primeiras metodologias são idênticos, optamos por deixar de parte a terceira metodologia, e tentar estabelecer algumas linhas gerais que possam ser utilizadas em abordagens futuras.

5.1. “A Estimação da Curva”

Como primeira tentativa, recorreremos à aplicação de um modelo de regressão que nos permita identificar qual o melhor ajustamento para posteriormente voltarmos a realizar uma nova estimativa.

Assim, ajustou-se um modelo de regressão que nos permitisse estimar a curva associada aos erros registados por grupos de idade, sendo para o efeito apenas utilizado o MALPE, pois é o erro que melhor consegue transmitir o comportamento das estimativas efectuadas.

Então, se $MALPE = \frac{\hat{S}_{ij}^{2001}(x) - S_{ij}^{2001}(x)}{S_{ij}^{2001}(x)}$, o valor efectivamente observado será dado por

$S_{ij}^{2001}(x) = \frac{\hat{S}_{ij}^{2001}(x)}{1 + a}$, onde a corresponde à equação da recta de regressão do modelo mais significativo.

No ajustamento dos modelos foram seleccionadas diferentes tipos de ajustamento com o intuito de se escolher posteriormente o modelo que melhor se ajusta aos erros obtidos, ou seja, o modelo com maior capacidade explicativa (R^2).

Para isso, tal como afirmámos anteriormente, foram somente utilizados os MALPE correspondentes aos dois primeiros métodos, no entanto, os erros referentes ao primeiro modelo foram ajustados duas vezes, pois o facto da estimativa correspondente ao primeiro grupo de idades (0 – 4) registar sempre um erro nulo, levou-nos a testar as duas formas.

Dos resultados obtidos (tabelas 11 a 14), seleccionaram-se os movimentos cujos modelos registaram uma capacidade explicativa superior a 50%, destacando-se assim as

seguintes 8 diferentes direcções: Norte/LVT, Norte/Algarve, Centro/LVT, Centro/Algarve, LVT/Alentejo, LVT/Algarve, Alentejo/LVT e, finalmente, Algarve/LVT. No entanto, na situação em que o grupo de idades dos 0 – 4 foi posto de parte, apenas cinco fluxos migratórios se revelaram significativos (Norte/LVT, Norte/Algarve, Centro/LVT, LVT/Algarve, e Algarve/LVT).

Tabelas 11 a 14 – R² “Curve Estimation”

Met. 1						Met. 1 (sem 0-4)					
Orig/Dest	Norte	Centro	LVT	Alentejo	Algarve	Orig/Dest	Norte	Centro	LVT	Alentejo	Algarve
Norte	-	0,378 ⁵	0,615 ⁵	0,469 ⁵	0,883 ²	Norte	-	0,340 ⁵	0,502 ⁵	0,103 ⁵	0,910 ⁵
Centro	0,277 ⁵	-	0,699 ⁵	0,242 ²	0,562 ⁵	Centro	0,196 ⁵	-	0,557 ⁵	0,126 ⁵	0,350 ⁵
LVT	0,084 ³	0,263 ⁵	-	0,740 ³	0,862 ⁵	LVT	0,193 ⁵	0,272 ⁵	-	0,134 ⁵	0,899 ⁵
Alentejo	0,468 ⁵	0,330 ³	0,587 ⁵	-	0,481 ⁵	Alentejo	0,470 ⁵	0,330 ³	0,492 ⁵	-	0,149 ⁵
Algarve	0,336 ³	0,023 ⁵	0,626 ⁵	0,485 ³	-	Algarve	0,251 ⁵	0,143 ⁵	0,661 ⁵	0,449 ³	-

Met. 2 (1)						Met. 2 (2)					
Orig/Dest	Norte	Centro	LVT	Alentejo	Algarve	Orig/Dest	Norte	Centro	LVT	Alentejo	Algarve
Norte	-	0,379 ⁵	0,617 ⁵	0,471 ⁵	0,884 ²	Norte	-	0,379 ⁵	0,617 ⁵	0,472 ⁵	0,884 ²
Centro	0,281 ⁵	-	0,702 ⁵	0,249 ²	0,567 ⁵	Centro	0,281 ⁵	-	0,702 ⁵	0,248 ²	0,567 ⁵
LVT	0,087 ³	0,261 ⁵	-	0,737 ³	0,859 ⁵	LVT	0,087 ³	0,261 ⁵	-	0,737 ³	0,859 ⁵
Alentejo	0,470 ⁵	0,347 ³	0,588 ⁵	-	0,493 ³	Alentejo	0,470 ⁵	0,347 ³	0,588 ⁵	-	0,493 ³
Algarve	0,342 ³	0,020 ⁵	0,627 ⁵	0,486 ³	-	Algarve	0,343 ³	0,020 ⁵	0,627 ⁵	0,487 ³	-

¹Linear; ²Logarithmic; ³Inverse; ⁴Quadratic; ⁵Cubic

Uma vez que os modelos ajustados (tabelas 15 a 18), e que se revelaram com maior capacidade explicativa, apenas correspondem a três ajustamentos diferentes (Cúbico, Logaritmico e Inverso), as suas equações são dadas por:

$$\text{Cúbico: } a = \beta_0 + (\beta_1 * X) + (\beta_2 * X^2) + (\beta_3 * X^3)$$

$$\text{Logaritmico: } a = \beta_0 + (\beta_1 * \ln(X))$$

$$\text{Inverso: } a = \beta_0 + (\beta_1 * \ln(X))$$

Tabelas 15 a 18 – Coeficientes dos Ajustamentos

Met. 1						Met. 1 (sem 0 - 4)					
Direcção	β_0	β_1	β_2	β_3	Dist.	Direcção	β_0	β_1	β_2	β_3	Dist.
N/LVT	0,1201	-0,1601	0,0150	-0,0004	Cubic	N/LVT	0,0374	-0,1256	0,0109	-0,0002	Cubic
N/Alg	-0,0671	-0,2157	-	-	Logarit	N/Alg	-0,4118	0,0472	-0,0099	0,0004	Logarit
C/LVT	0,1200	-0,1979	0,0207	-0,0006	Cubic	C/LVT	-0,0686	-0,1192	0,0113	-0,0003	Cubic
C/Alg	0,1549	-0,1895	0,0194	-0,0007	Cubic	C/Alg	-	-	-	-	-
LVT/Al	-0,4765	0,3984	-	-	Inverse	LVT/Al	-0,2373	-0,0001	0,0002	-0,0001	Inverse
LVT/Alg	0,0159	-0,1058	0,0128	-0,0005	Cubic	LVT/Alg	-	-	-	-	-
Al/LVT	0,0976	-0,1931	0,0210	-0,0006	Cubic	Al/LVT	-	-	-	-	-
Alg/LVT	0,0159	-0,1471	0,0143	-0,0003	Cubic	Alg/LVT	-0,3661	0,0123	-0,0047	0,0004	Cubic

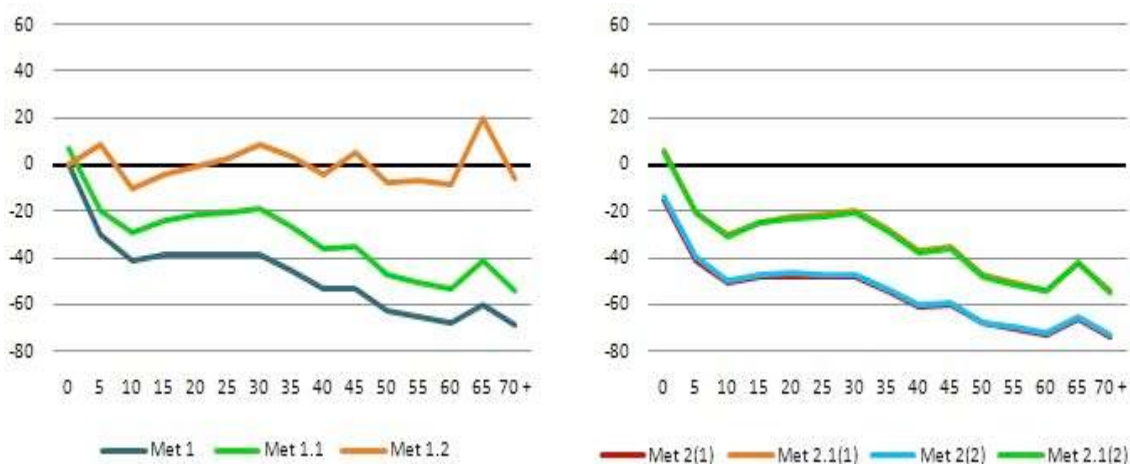
Met. 2 (1)						Met. 2 (2)					
Direcção	β_0	β_1	β_2	β_3	Dist.	Direcção	β_0	β_1	β_2	β_3	Dist.
N/LVT	0,9219	-0,2763	0,0259	-0,0006	Cubic	N/LVT	0,9409	-0,2790	0,0261	-0,0006	Cubic
N/Alg	-0,2031	-0,1845	-	-	Logarit	N/Alg	-0,1785	-0,1902	-	-	Logarit
C/LVT	0,9090	-0,3403	0,0356	-0,0010	Cubic	C/LVT	0,8728	-0,3339	0,0349	-0,0010	Cubic
C/Alg	-0,3406	-0,1093	0,0112	-0,0004	Cubic	C/Alg	-0,5459	-0,0753	0,0077	-0,0003	Cubic
LVT/Al	0,0254	0,7865	-	-	Inverse	LVT/Al	0,0874	0,8339	-	-	Inverse
LVT/Alg	0,3928	-0,1471	0,0178	-0,0008	Cubic	LVT/Alg	0,3953	-0,1474	0,0178	-0,0008	Cubic
Al/LVT	1,0068	-0,3547	0,0385	-0,0011	Cubic	Al/LVT	1,1465	-0,3794	0,0412	-0,0012	Cubic
Alg/LVT	0,6085	-0,2356	0,0229	-0,0005	Cubic	Alg/LVT	0,6498	-0,2417	0,0235	-0,0005	Cubic

Com o intuito de exemplificarmos as melhorias obtidas com a aplicação através da utilização da fórmula atrás mencionada – $S_{ij}^{2001}(x) = \frac{\hat{S}_{ij}^{2001}(x)}{1+a}$ – destacamos o exemplo de dois dos 8 modelos com maior capacidade explicativa.

Neste sentido, e em primeiro lugar, podemos observar a situação do fluxo migratório proveniente do Norte e com direcção ao Algarve (*gráficos 20 e 21*), que registou uma grande melhoria, pois o erro associado à estimativa proveniente da primeira metodologia, mais precisamente o correspondente ao modelo ajustado sem o primeiro grupo de idades (Met 1.2). No entanto, também a estimativa resultante do modelo ajustado com todos os grupos de idade (Met 1.1), melhorou significativamente quando comparada com os erros obtidos através da metodologia correspondente ao primeiro método (Met 1).

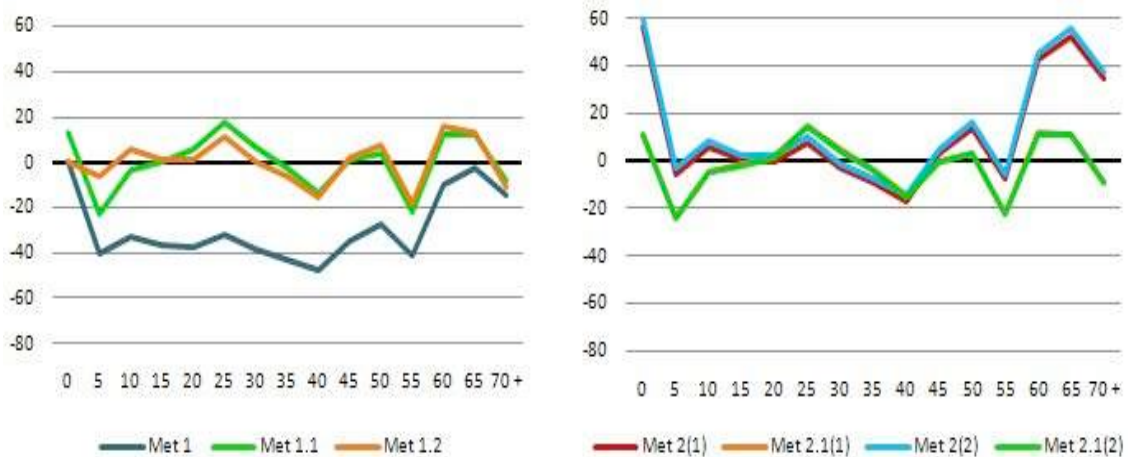
Por outro lado, e apesar de também se terem registado melhorias significativas através dos erros obtidos através da primeira metodologia, o melhor resultado é mesmo aquele cujos coeficientes resultam do modelo ajustado sem o grupo de idades dos 0 – 4 anos.

Gráficos 20 e 21 – Movimento Norte/Algarve



Observando agora um segundo exemplo, correspondente à migração estimada entre as regiões do Algarve e de Lisboa e Vale do Tejo (*gráficos 22 e 23*), podemos constatar que qualquer um registou melhorias bastante significativas nas estimativas efectuadas. No entanto, podemos destacar a estimativa efectuada através dos coeficientes obtidos, novamente, pelo ajustamento do modelo sem o primeiro grupo de idades (Met 1.2).

Gráficos 22 e 23 – Movimento Algarve/L.V.T



A redução dos erros de estimativa por nós obtida, revela-se muito útil, pois o facto das migrações não serem constantes, ou seja, não demonstrarem um comportamento estável ao longo dos anos, tal como no caso específico da fecundidade e mortalidade, resulta numa grande dificuldade na realização de estimativas fidedignas.

Neste caso específico, podemos afirmar que os modelos ajustados através dos erros sem o primeiro grupo de idades, se destacam essencialmente quando as estimativas iniciais ficam muito aquém dos valores realmente observados. Por outro lado, nas restantes ocasiões, qualquer uma das *re-estimativas* efectuadas parece produzir melhorias significativas.

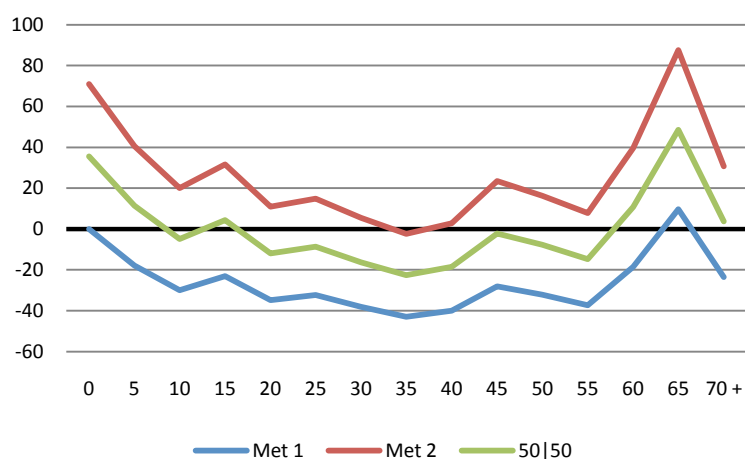
5.2. Conjugação Metodológica

Durante a análise e avaliação dos resultados obtidos, denotamos que as estimativas efectuadas a partir das próprias metodologias registavam, muitas vezes comportamentos complementares entre si, podendo assim ser conjugados, através de uma simples média aritmética ou de uma média ponderada.

Começando assim por “reduzir” as nossas estimativas a apenas duas, através da realização de uma média aritmética entre as duas estimativas reproduzidas pela segunda metodologia, iremos de seguida analisar alguns dos resultados obtidos.

Referindo como exemplo o fluxo migratório com origem na região Norte e com direcção a Lisboa e Vale do Tejo, e começando por uma simples média (0,5|0,5) dos resultados obtidos através das duas metodologias, podemos constatar que se registou uma melhoria bastante significativa, no entanto, apesar de se registar uma melhoria geral em alguns grupos de idades, outros houve em que o erro associado a uma das estimativas aumentou (*gráfico 24*).

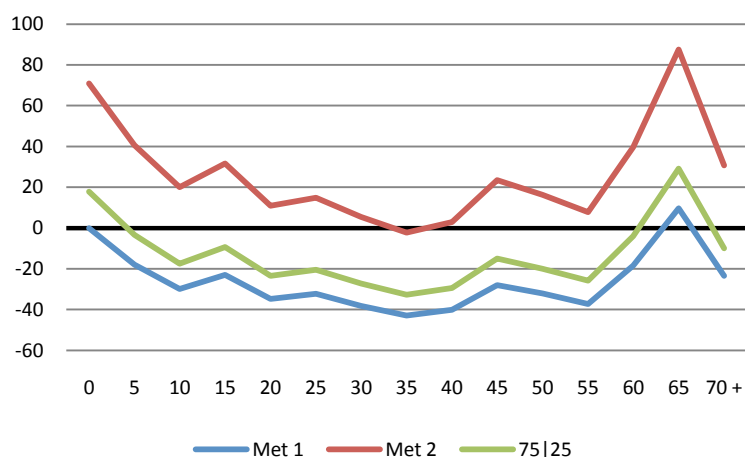
Gráfico 24 – Movimento Norte/LVT, com ponderação 50|50



Neste sentido, e com o intuito de chegarmos ao melhor resultado, optámos por, de seguida, voltar a conjugar as duas metodologias, mas agora, através de uma média ponderada, dando maior peso à primeira (0,75) metodologia (*gráfico 25*).

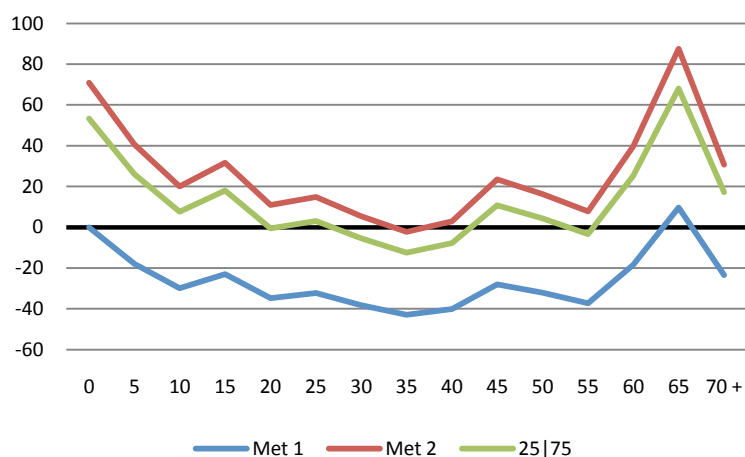
No entanto, e uma vez que, neste caso, a primeira metodologia tinha sido aquela que mais se afastou dos valores efectivos, o resultado obtido não foi o melhor.

Gráfico 25 – Movimento Norte/LVT, com ponderação 75|25



Invertendo, de seguida a ponderação efectuada (*gráfico 26*), ou seja, dando agora um maior peso à segunda (0,75) metodologia em relação à primeira (0,25), observamos uma melhoria global em relação às duas tentativas anteriores, no entanto, a maior diminuição do erro associado ocorreu nos grupos de idades intermédios (10 – 59 anos), para nos seus extremos (0 – 9 e 60 – 70+) voltar a registar erros mais elevados.

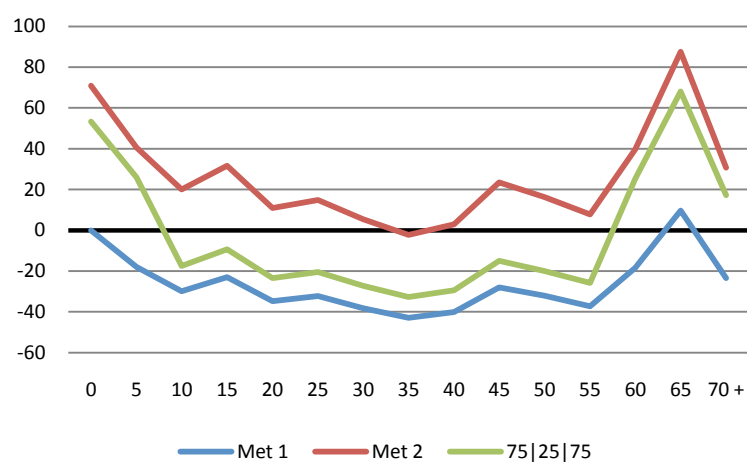
Gráfico 26 – Movimento Norte/LVT, com ponderação 25|75



Esta última situação, alerta-nos para o facto de as estimativas efectuadas não estarem a acompanhar o comportamento dos migrantes ao longo das idades, e deste modo, pareceu-nos correcto dividir a ponderação de acordo com o comportamento observado por idades. Começamos então, por atribuir 0,75 às crianças com idades até aos 9 anos, 0,25 aos adultos com idades entre os 10 e os 59 anos, e por último, novamente 0,75 àqueles que registavam mais de 60 anos de idade, isto começando pela primeira metodologia (*gráfico 27*).

Esta ponderação de acordo com as idades dos migrantes, apesar de registar algumas melhorias, ainda não é a mais indicada.

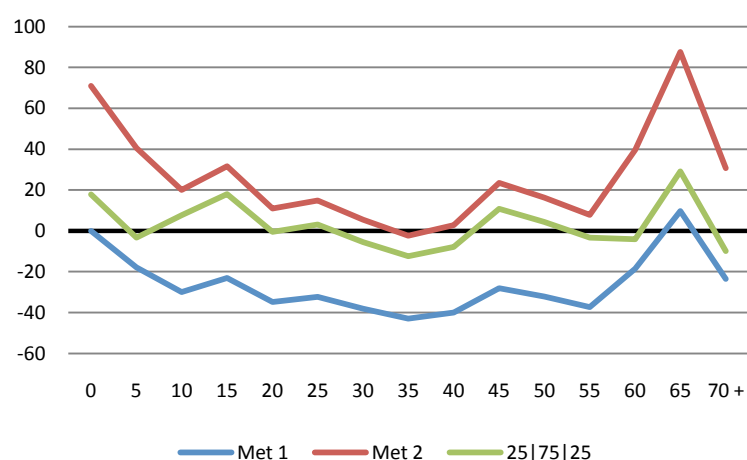
Gráfico 27 – Movimento Norte/LVT, com ponderação 75|25|75



Por último, realizamos uma outra abordagem, que foi inverter a ponderação anterior, e onde se tinha dado maior peso ao primeiro método, agora passou para o segundo, e vice-versa (*gráfico 28*).

Observando então, os resultados obtidos, podemos verificar que esta foi a ponderação que deu, até agora, melhores resultados, encontrando um ponto de equilíbrio entre as duas metodologias e reduzindo o erro associado.

Gráfico 28 – Movimento Norte/LVT, com ponderação 25|75|25



A conjugação metodológica que observamos acima, demonstrou que o facto das duas metodologias produzirem estimativas diferenciadas, estas têm potencialidade para efectuarmos uma redução do erro.

Podemos então afirmar que, de uma forma geral a ponderação deve ser efectuada tendo em conta os grupos etários, e que a primeira metodologia efectua estimativas mais correctas para os grupos de idades situados nos extremos, e a segunda metodologia, para aqueles que se encontram a “meio” da estrutura etária.

Assim, e tal como referimos, em aplicações futuras, e perante o quadro de envelhecimento constante que caracteriza o comportamento demográfico actual, a conjugação destas duas metodologias é uma solução viável para obtermos estimativas mais correctas.

5.3. A estrutura Etária

Tal como a análise preliminar dos dados já tinha indicado, a estrutura populacional do continente português sofreu muitas alterações entre os recenseamentos de 1991 e 2001, este poderá ser também um ponto de partida para que consigamos elaborar estimativas com menor percentagem de erro associada.

Neste sentido, e baseando-nos na primeira metodologia, a estimativa de fluxos migratórios calculada através de um rácio que relaciona os diferentes grupos etários com o dos 0 a 4 anos – $r_{ij}(x, -5) = \frac{S_{ij}(x)}{S_{ij}(-5)}$ – supõe que a estrutura da população se mantém inalterada durante os dez anos seguintes.

Assim, e uma vez que a propensão migratória das crianças até aos 4 anos de idade ($S_{ij}(-5)$) é a base do cálculo deste rácio, pensamos em “corrigir” o valor de chegada (2001) através do de partida (1991) em conjunto com algumas variáveis que pudessem explicar a diferença registada. Para isso, tentamos aplicar uma regressão linear múltipla, em que $S_{ij}^{2001}(-5) = S_{ij}^{1991}(-5) + a + b + c$, mas devido à repetição, tanto das regiões de origem com as de destino, as variáveis explicativas iriam repetir-se tantas vezes quanto as próprias regiões, o que não permite um bom ajustamento do modelo.

Por outro lado, e tal como foi possível observar através da conjugação das metodologias, além de termos uma estrutura populacional em constante evolução, os comportamentos migratórios associados aos diferentes grupos de idade não é partilhado entre si, pelo que poderíamos obter resultados através dos dois grupos de idades que se situam nos extremos:

$$\hat{S}_{ij}^{2001} = \frac{1}{2} * \frac{S_{ij}^{2001}(-5)}{S_{ij}^{1991}(-5)} * S_{ij}^{1991} + \frac{1}{2} * \frac{S_{ij}^{2001}(70+)}{S_{ij}^{1991}(70+)} * S_{ij}^{1991}$$

No entanto, estas abordagens não registaram os resultados desejados, pelo que nos parece mais viável uma conjugação metodológica que uma alteração ao próprio método.

6. Análise Estatística das Migrações Internas em Portugal Continental

Além de ser extremamente importante realizar uma estimativa bastante fidedigna dos fluxos migratórios, sejam eles a nível interno ou internacional, importa também verificar e analisar os comportamentos registados de acordo com algumas variáveis que possam caracterizar esses movimentos.

Por outro lado, o desenvolvimento de políticas demográficas mais eficientes só faz sentido se os diversos comportamentos caracterizadores da população forem conhecidos, pois só assim é que as políticas desenvolvidas conseguem responder às necessidades dos diferentes indivíduos que compõem as diferentes sociedades. Na situação específica das migrações, uma estimativa precisa permitirá regular e analisar melhor quer os movimentos de entrada, quer de saída das diferentes regiões ou países.

Deste modo, e mais uma vez, recorrendo às bases de dados amostrais provenientes dos recenseamentos gerais da população elaborados em 1991 e 2001, aplicamos uma regressão logística onde a variável resposta é o acto de migrar ou permanecer na região de residência. Para que pudéssemos realizar esta abordagem estatística, a partir dos dados referidos anteriormente, foi introduzida a uma nova variável foi formulada através do cruzamento entre a região de residência à altura do recenseamento e a sua NUT II de residência cinco anos antes, onde o facto de a residência há cinco anos atrás ser numa região diferente daquela em que reside à data do recenseamento, significa que existiu migração. Com o intuito de realizarmos uma análise completa e que nos permita caracterizar da melhor forma as migrações internas registadas no continente português, elaborámos ainda uma outra base, à qual foi aplicada a mesma metodologia, mas que a sua variável resposta foi construída baseando-nos no local de nascimento e onde reside na altura do recenseamento.

Assim, o facto de, à data do recenseamento, qualquer indivíduo residir numa região que não seja aquela onde nasceu ou residia cinco anos antes, será considerado um migrante interno. A variável resposta será assim, composta por duas categorias, onde o código “0” corresponde a não ter migrado, e “1” a ter migrado.

No respeitante às restantes covariáveis (explicativas) incluídas na base, estas correspondem ao ano do recenseamento (1991 e 2001), ao sexo dos migrantes, às NUTS II de residência (Norte, Centro, Lisboa e Vale do Tejo, Alentejo e Algarve) à data do recenseamento, do nascimento, e há cinco anos atrás, separadamente, e por último, à

idade dos migrantes. Contabilizamos assim, um total de 6 covariáveis, onde apenas a idade não é uma covariável categorizada, e tal como na estimação indirecta as Regiões Autónomas dos Açores da Madeira foram deixadas de parte.

A construção de duas bases distintas, permitir-nos-á analisar as diferenças entre os resultados obtidos, e elaborar duas análises que se complementam.

6.1. A Residência cinco anos antes

Deste modo, e começando pela base em que a variável resposta foi calculada de acordo com a sua residência cinco anos atrás (tabela 19), começamos por testar as variáveis uma a uma, separadamente, de forma a identificarmos quais as significativas para o modelo. Assim, e fixando um nível de significância a 5% ($\alpha = 0,05$), constatamos que apenas existe evidência estatística se se rejeitar a hipótese nula ($p\text{-value} < \alpha$), e apesar de se registarem algumas variáveis cujos coeficientes de algumas categorias não sejam significativas com $\alpha = 0,05$ ($p\text{-value} < \alpha$), como no caso das covariáveis “ano” e “sexo”, onde:

$$H_0: \beta_i = 0 | \beta_0, \beta_1, \beta_{i-1}; \beta_{i+1}; \beta_p \text{ VS } H_1: \beta_i \neq 0 | \beta_0, \beta_1, \beta_{i-1}; \beta_{i+1}; \beta_p \quad (i = 1, \dots, p)$$

Esta situação, indica assim, que não existe evidência estatística para podermos afirmar que existem diferenças significativas entre as duas categorias que compõem cada uma das covariáveis, tal como se pode através da sua análise individual, pois tal como podemos observar na tabela 19, o facto de um individuo pertencer ao sexo masculino, sobe a sua possibilidade de migrar em 5,5% (O.R. = $e^\beta = 1,055$) em relação ao sexo feminino. Tal como foi afirmado anteriormente, através da análise dos $p\text{-values}$ obtidos, as percentagens registadas não são de grande realce.

Por outro lado, o facto de um migrante residir, à data do recenseamento na região do Algarve, significa que a sua possibilidade de migrar diminui cerca de 55,1% comparativamente aos residentes na região Norte, enquanto aqueles que residem no Alentejo apenas registam uma possibilidade de migrar inferior cerca de 25,7%. Por

outro lado, aqueles que residem nas regiões do Centro e de Lisboa e Vale do Tejo, registam possibilidades de migrar praticamente idênticas à da região Norte.

Passando agora à região de residência cinco anos antes do recenseamento, e aqui não se regista a ocorrência de eventos para a região de Lisboa e Vale do Tejo, verificamos que o facto de um indivíduo residir na região Centro cinco anos antes do recenseamento, aumenta a sua possibilidade de migrar em cerca de 16,7% quando comparada com os residentes da região Norte. Num outro sentido, e sempre tomando como ponto de comparação a região Norte, aqueles que residiam na região do Algarve viam a sua possibilidade de migrar diminuir 52,6%, enquanto na região do Alentejo diminui apenas cerca de 25%.

No que concerne à região de nascimento, o facto de um qualquer indivíduo ter nascido na região de Lisboa e Vale do Tejo, faz com que a sua possibilidade de migrar seja superior cerca de 1,5 vezes em relação aos que nasceram na região Norte, enquanto para aqueles que nasceram na região Centro, esse valor desce para 26,6%.

Por outro lado, enquanto para aqueles que nasceram na região do Alentejo, a sua possibilidade de migrar é inferior cerca de 21,5% quando comparada com a região Norte, os indivíduos que nasceram no Algarve apenas vêem a sua possibilidade diminuída cerca de 1,2%.

Por último, no que concerne à idade, podemos afirmar que, por cada ano a mais na idade de um indivíduo, a sua possibilidade de migrar diminui cerca de 1,8%.

Neste caso referente à idade, podemos ainda estimar a probabilidade de migrar de um indivíduo daqui a dez anos:

$$\pi(idade = 10) = \frac{\exp(0,4794 + (-0,0182) * 10)}{1 + \exp(0,4794 + (-0,0182) * 10)} = \frac{1,3464}{2,3464} = 0,57$$

Significa assim, que a probabilidade de um indivíduo migrar aos 10 anos é de 43%.

Tabela 19 – Regressão Logística Univariada

Variáveis	β	Std. β	<i>p-value</i>	R	Intervalo Confiança 95%	
Ano	0,0354	0,0350	<u>0,3120</u>	1,0360	[0,9673; 1,1096]	
Sexo	0,0530	0,0350	0,1300	1,0545	[0,9845; 1,1294]	
NUT II	Centro	-0,0109	0,0521	0,8350	0,9892	[0,8931; 1,0956]
	LVT	-0,0003	0,0516	0,9960	0,9997	[0,9035; 1,1060]
	Alentejo	-0,2965	0,0547	0,0000	0,7434	[0,6678; 0,8276]
	Algarve	-0,7996	0,0652	0,0000	0,4495	[0,3956; 0,5109]
NUT II 5 Ant.	Centro	0,1542	0,0468	0,0010	1,1667	[1,0645; 1,2787]
	Alentejo	-0,1738	0,0579	0,0027	0,8405	[0,7504; 0,9415]
	Algarve	-0,7471	0,0635	0,0000	0,4738	[0,4183; 0,5366]
NUT II Nasc.	Centro	0,2358	0,0567	0,0000	1,2659	[1,1327; 1,4147]
	LVT	0,4047	0,0532	0,0000	1,4989	[1,3505; 1,6636]
	Alentejo	-0,2423	0,0619	0,0001	0,7848	[0,6952; 0,8861]
	Algarve	-0,0117	0,0587	0,8420	0,9884	[0,8810; 1,1089]
Idade	-0,0182	0,0007	0,0000	0,9819	[0,9805; 0,9834]	

Após uma análise detalhada das covariáveis individualmente, e de identificarmos quais as que têm coeficientes significativos ao nível de 20%, ajustamos um modelo deixando de parte a covariável ano (tabela 20).

Uma vez que todas elas, e com $\alpha = 0,20$, voltaram a registar coeficientes significativos no seu conjunto, o próximo passo será incluir as covariáveis que foram, inicialmente deixadas de parte, e verificar se, apesar de não se terem revelado significativas individualmente, no seu todo, acrescentam algo ao modelo.

Tabela 20 – Modelo com variáveis significativas 1

Variáveis		β	Std. β	p -value	OR	Intervalo Confiança 95%
Sexo		0,0888	0,0367	0,0157	1,0929	[1,0169; 1,1745]
NUT II	Centro	-0,0036	0,0553	0,9479	0,9964	[0,8940; 1,1104]
	LVT	-0,0700	0,0550	0,2030	0,9324	[0,8371; 1,0385]
	Alentejo	-0,2328	0,0597	0,0001	0,7923	[0,7048; 0,8907]
	Algarve	-0,6767	0,0701	0,0000	0,5083	[0,4431; 0,5831]
NUT II 5 Ant.	Centro	0,1094	0,0553	0,0479	1,1156	[1,0010; 1,2434]
	Alentejo	0,0160	0,0748	0,8303	1,0162	[0,8775; 1,1767]
	Algarve	-0,9492	0,0812	0,0000	0,3871	[0,3301; 0,4539]
NUT II Nasc.	Centro	0,2142	0,0633	0,0007	1,2388	[1,0942; 1,4025]
	LVT	0,5199	0,0604	0,0000	1,6819	[1,4942; 1,8931]
	Alentejo	-0,1540	0,0764	0,0438	0,8572	[0,7380; 0,9957]
	Algarve	0,5909	0,0750	0,0000	1,8056	[1,5588; 2,0916]
Idade		-0,0189	0,0008	0,0000	0,9812	[0,9798; 0,9827]

Uma vez mais, e agora ajustando um outro modelo com todas as covariáveis inicialmente testadas, podemos afirmar que estas passaram a ser significativas para o modelo ($\alpha = 0,05$), tal como podemos constatar na tabela 21.

No entanto, e apesar de termos encontrado quais as covariáveis cujos coeficientes são significativos para o modelo, teremos de incluir também no modelo as interacções possíveis, de modo a no final chegarmos a um modelo com apenas aquelas que revelam ter coeficientes significativos e acrescentam alguma capacidade explicativa ao modelo.

Tabela 21 – Modelo com variáveis significativas 2

Variáveis	β	Std. β	p -value	OR	Intervalo Confiança 95%	
Ano	0,0088	0,00 7	0,0168	1,0088	[1,0016; 1,0161]	
Sexo	0,0888	0,0368	0,0157	1,0929	[1,0169; 1,1745]	
NUT II	Centro	-0,0034	0,0553	0,9512	0,9966	[0,8942; 1,1107]
	LVT	-0,0708	0,0550	0,1982	0,9317	[0,8364; 1,0377]
	Alentejo	-0,2311	0,0597	0,0001	0,7937	[0,7060; 0,8923]
	Algarve	-0,6763	0,0701	0,0000	0,5085	[0,4433; 0,5834]
NUT II 5 Ant.	Centro	0,1090	0,0553	0,0489	1,1151	[1,0005; 1,2428]
	Alentejo	0,0164	0,0749	0,8264	1,0166	[0,8778; 1,1772]
	Algarve	-0,9546	0,0813	0,0000	0,3850	[0,3283; 0,4514]
NUT II Nasc.	Centro	0,2132	0,0633	0,0008	1,2377	[1,0932; 1,4012]
	LVT	0,5209	0,0604	0,0000	1,6835	[1,4956; 1,8950]
	Alentejo	-0,1544	0,0764	0,0433	0,8569	[0,7377; 0,9954]
	Algarve	0,5932	0,0750	0,0000	1,8098	[1,5624; 2,0965]
Idade	-0,0190	0,0008	0,0000	0,9812	[0,9797; 0,9827]	

A tabela 22 apresenta assim, o modelo e todas as variáveis e interacções e covariáveis que se mostraram significativas para o modelo. No entanto, e apesar de estas terem sido validadas com $\alpha = 0,05$, resta verificar se acrescentam, ou não, alguma capacidade explicativa ao modelo ajustado anteriormente, ou seja, ao modelo que fechamos sem interacções.

A validação destas interacções foi realizada através da sua inclusão individualmente no modelo, e após se terem definido aquelas que se mostraram significativas, é que se enquadraram todas no modelo, para posteriormente retirarmos as que não se mostraram significativa no seu conjunto.

Apesar de durante o processo de modelação estatística termos utilizado o teste de razão de verosimilhanças entre dois modelos ajustados com o intuito de dissiparmos algumas

dúvidas relativamente à significância de determinada covariável para o modelo, para testarmos a interações utilizou-se ainda o critério de AKAIKE que nos indicará qual o modelo com melhor capacidade explicativa. Assim, o AIC associado ao modelo sem interações (tabela 21) é de 16991, enquanto o correspondente ao modelo com as interações significativas (tabela 22) é de 19646, sendo ligeiramente inferior, o que significa que é o melhor modelo e que as interações acrescentam capacidade explicativa ao modelo ajustado.

Deste modo, e em conformidade com o teste de razão de verossimilhanças, os resultados obtidos e consistentes com o melhor modelo podem ser visualizados na tabela 22.

Tabela 22 – Modelo com variáveis significativas 3

Variáveis		β	Std. β	p-value	OR	Intervalo Confiança 95%
Ano		0,0090	0,0037	0,0146	1,0090	[1,0018; 1,0163]
Sexo		-0,0521	0,0738	0,4801	0,9492	[0,8215; 1,0969]
NUT II	Centro	-0,133	0,1149	0,2450	0,8749	[0,6985; 1,0900]
	LVT	0,2746	0,1133	0,0154	1,3160	[1,0539; 1,6434]
	Alentejo	-0,3574	0,1282	0,0053	0,6995	[0,5441; 0,8992]
	Algarve	-0,4429	0,1449	0,0022	0,6422	[0,4834; 0,8531]
NUT II 5 Ant.	Centro	-0,0248	0,1047	0,8127	0,9755	[0,7946; 1,1976]
	Alentejo	0,0062	0,1362	0,9638	1,0062	[0,7704; 1,3142]
	Algarve	-0,7289	0,1440	0,0000	0,4825	[0,3638; 0,6399]
NUT II Nasc.	Centro	0,2030	0,0637	0,0014	1,2251	[1,0812; 1,3881]
	LVT	0,5038	0,0609	0,0000	1,6550	[1,4689; 1,8648]
	Alentejo	-0,1683	0,0771	0,0291	0,8451	[0,7265; 0,9830]
	Algarve	0,6221	0,0760	0,0000	1,8628	[1,6050; 2,1619]
Idade		-0,0239	0,0031	0,0000	0,9764	[0,9705; 0,9823]
Sexo:Idade		0,0033	0,0015	0,0296	1,0033	[1,0003; 1,0063]
NUT II Centro:Idade		0,0030	0,0023	0,1948	1,0030	[0,9985; 1,0075]
NUT II LVT:Idade		-0,0086	0,0024	0,0004	0,9914	[0,9868; 0,9961]
NUT II Alentejo:Idade		0,0029	0,0026	0,2615	1,0029	[0,9978; 1,0080]
NUT II Algarve:Idade		-0,0060	0,0031	0,0476	0,9940	[0,9880; 0,9999]
NUT II 5 Ant Centro:Idade		0,0031	0,0022	0,1484	1,0031	[0,9989; 1,0074]
NUT II 5 Ant Algarve:Idade		0,0006	0,0028	0,8401	1,0006	[0,9950; 1,0062]
NUT II 5 Ant Alentejo:Idade		-0,0070	0,0032	0,0297	0,9931	[0,9868; 0,9993]

Apesar de termos chegado a um modelo final, com todas as variáveis e interações significativas, isto não indica que o modelo se ajuste bem aos dados, e para isso recorreremos ao teste de *Hosmer e Lemeshow* (tabela 23) de forma a analisarmos a bondade do ajustamento. Assim, podemos verificar que a bondade do ajustamento é bastante considerável, pois regista um *p-value* de 0,9567, o que significa que este modelo se ajusta aos dados ($\alpha = 0,05$), ou seja, os valores estimados pelo modelo são muito próximos dos valores observados.

H_0 : O modelo ajusta-se aos dados *vs* H_1 O modelo não se ajusta aos dados

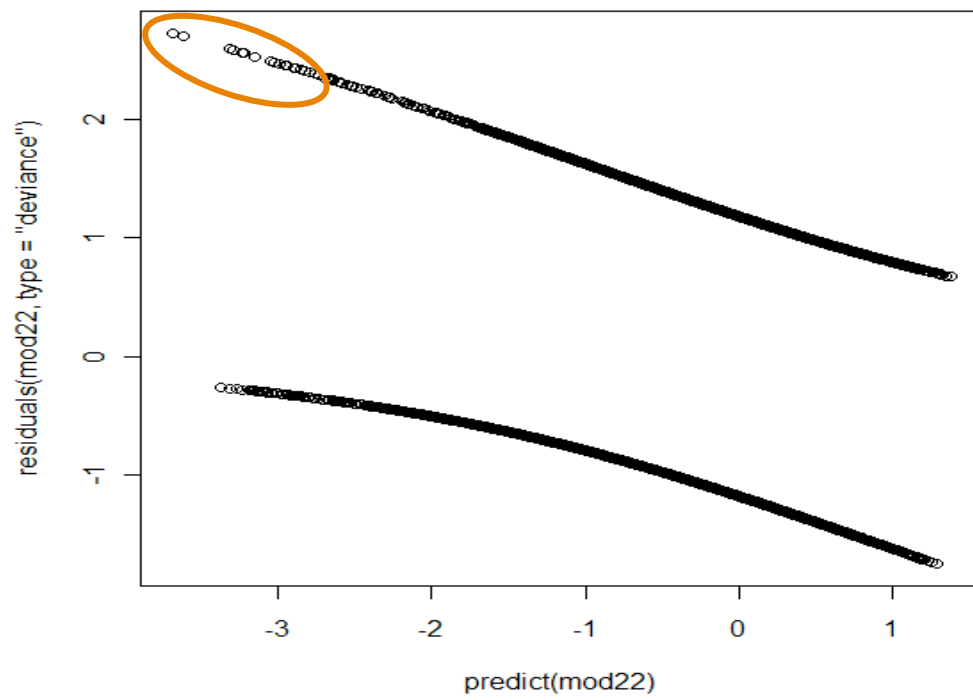
No entanto, e apesar deste modelo se ajustar ao dados, a sua capacidade explicativa é muito reduzida, conseguindo explicar apenas 7,1% da variação da variável resposta, ou seja, do acto de migrar ou não migrar, onde $\rho^2 = 1 - \frac{RD/gl_{RD}}{ND/gl_{ND}}$.

Tabela 23 – *Hosmer e Lemeshow*(1)

χ^2	f	<i>p-value</i>
2,6049	8	0,9567

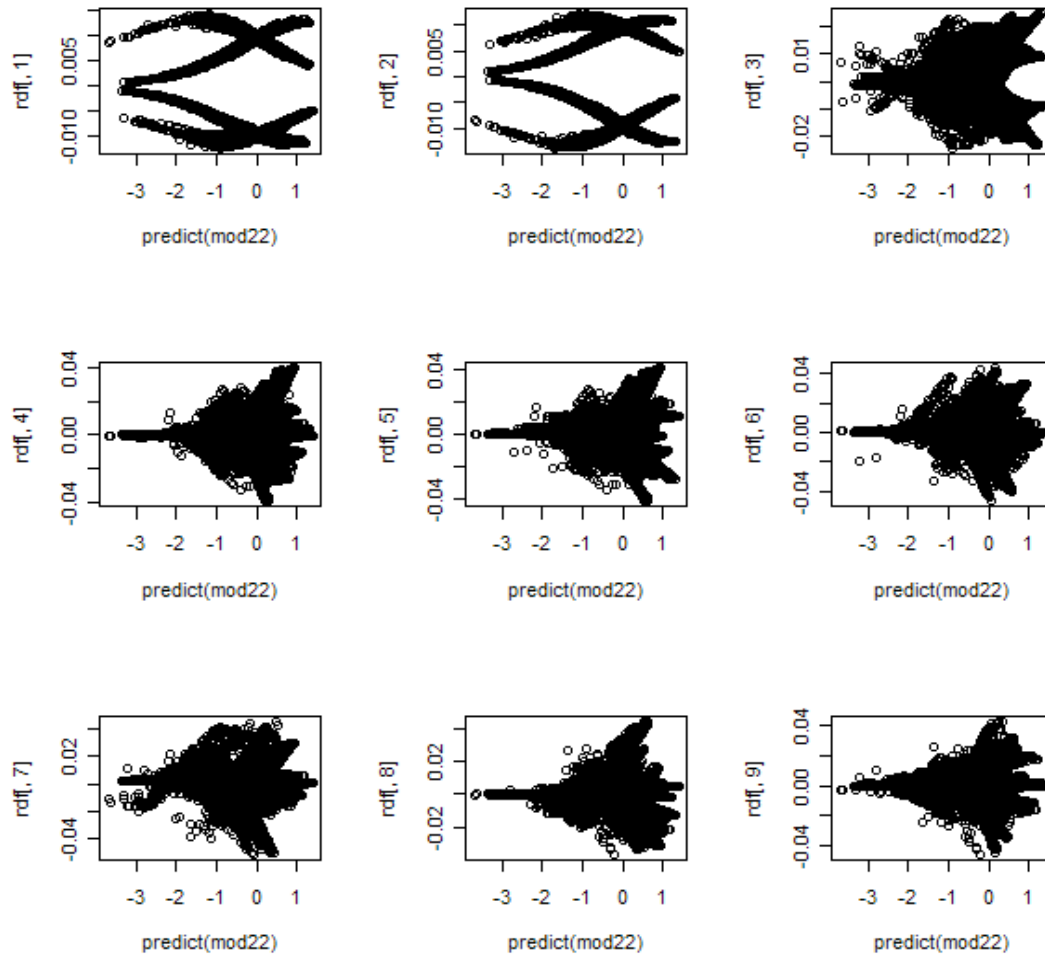
Passando agora à análise de resíduos, e começando pela representação gráfica dos resíduos *deviance*, podemos constatar que, uma vez que estes deveriam variar entre -2 e 2, que se registam algumas observações que poderão ser consideradas aberrantes. Esta situação pode ser identificada através da observação do *gráfico 29*.

Gráfico 29 – *Resíduos Deviance* (1)



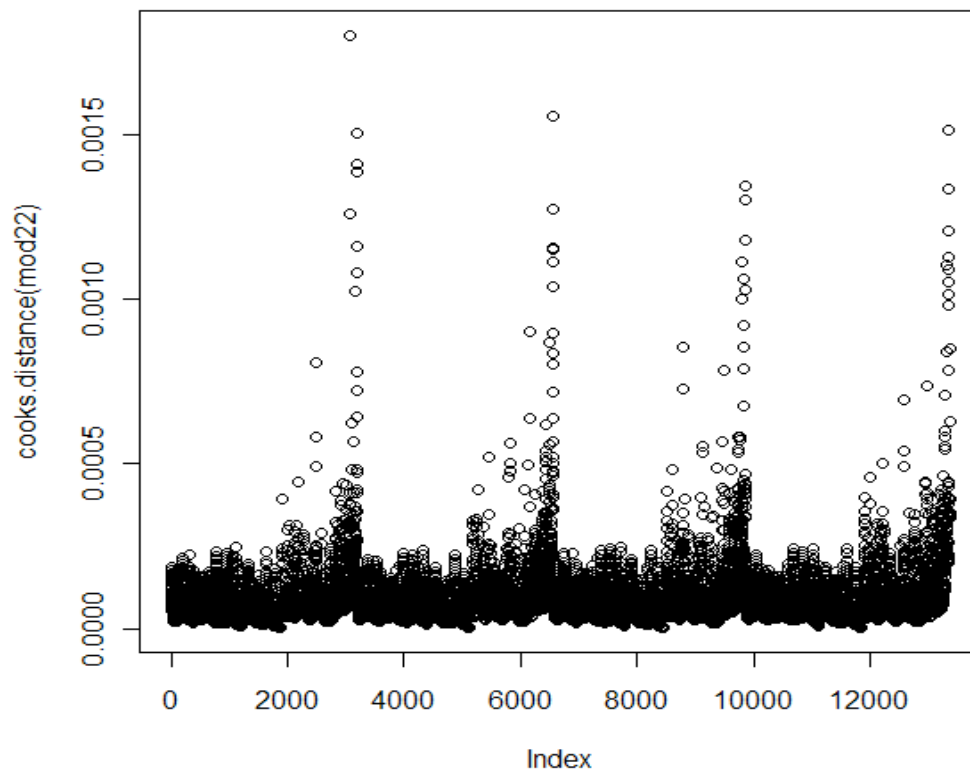
Representando graficamente agora os resíduos *dfbetas*, e tal como podemos observar em alguns gráficos obtidos (*gráfico 30*), constatamos que não se registam observações que se destacam excessivamente.

Gráfico 30 – Resíduos *Dfbetas* (1)



Representando graficamente os *valores da distância de Cook* (gráfico 31), podemos constatar que apesar de se observarem alguns pontos mais distantes, em termos de escala, esta é muito reduzida, pelo que não podemos afirmar a existência de pontos influentes.

Gráfico 31 – Distância de *Cook* (1)



No respeitante ao efeito que as observações têm nos valores preditos, ou seja, a sua repercussão (*leverage*), o facto de estarmos perante muitas observações não nos permite medir de forma precisa o efeito que a observação tem nos valores preditos e, por outro lado, estaríamos a avaliar demasiadas observações com base num só limite.

Recorrendo agora ao SPSS, com o intuito de avaliarmos a capacidade de discriminação do modelo, optou-se por representar a curva ROC (*gráfico 32*) de modo a analisarmos a área existente abaixo (tabela 24). Desta forma, o registo de um valor de 0,685 indica que o modelo ajustado tem um fraco poder discriminativo (discriminação pobre), no entanto, e tomando um $\alpha = 0,05$, podemos ainda afirmar que a própria área sob a curva é significativamente superior a 0,5 ($p\text{-value} = 9,163\text{E-}279$).

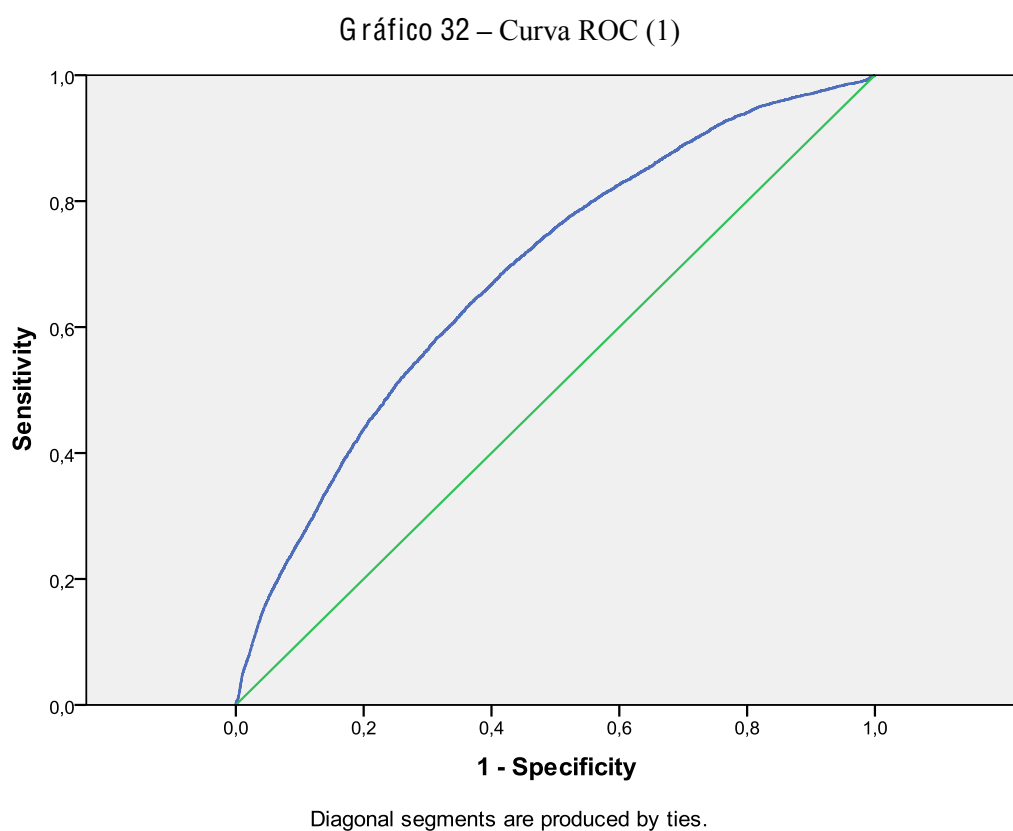


Tabela 24 – Área abaixo da Curva

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0,685	0,005	0,000	0,676	0,694

Através dos resultados obtidos, recorreremos às *coordenadas da curva* com o intuito de definirmos um ponto de corte equilibrado, que corresponde a 0,44, onde a *sensibilidade* é de 0,635 e “1 – *especificidade*” de 0,365.

Constatamos assim, que o modelo tem uma capacidade de discriminação (*sensibilidade*) de 63,1%, ou seja, classifica correctamente 63,1% dos indivíduos que cinco anos atrás residiam numa região diferente daquela em que residem à data do recenseamento. Por outro lado, a sua *especificidade* é de 63,9%, o que indica que o modelo classifica correctamente 63,9% dos indivíduos que não alteraram a sua residência no referido espaço de 5 anos.

6.2. A região de nascimento

Após uma análise cuidada dos resultados referentes à base de dados em que o acto de migrar foi definido com base na residência onde residia cinco anos antes do recenseamento, passamos agora a observar os resultados obtidos através da alteração do acto de migrar para ser definido através do seu local de nascimento (tabela 25).

Deste modo, registou-se logo uma alteração, pois apesar de quando testados os coeficientes da covariáveis de forma individual, as covariáveis “ano” e “sexo” continuarem, com $\alpha = 0,20$, a não serem significativas para o modelo, juntou-se também a covariável correspondente à NUT II de residência cinco anos (NUT II 5 Ant.) antes do recenseamento geral da população. Não existe assim, evidência estatística para podermos afirmar que as covariáveis não são significativas.

$$H_0: \beta_i = 0 | \beta_0, \beta_1, \beta_{i-1}; \beta_{i+1}; \beta_p \text{ VS } H_1: \beta_i \neq 0 | \beta_0, \beta_1, \beta_{i-1}; \beta_{i+1}; \beta_p \quad (i = 1, \dots, p)$$

Deste modo, e observando os resultados obtidos para as covariáveis significativas, podemos afirmar que o facto de um individuo residir na região do Centro à data do recenseamento, a sua possibilidade de migrar aumenta cerca de 15,8% relativamente à região Norte. Enquanto que para o Alentejo, essa possibilidade desce para 5,8%, e no caso Algarvio, a possibilidade de migrar desce cerca de 27,4% em relação ao Norte.

Por outro lado, e uma vez que a região de nascimento também mostrou ser significativa para o modelo, podemos referir que o facto de se ter nascido na região de Lisboa e Vale

do Tejo, aumenta a possibilidade de migrar em cerca de 50,9% em relação à região Norte, e no caso daqueles que nasceram na região do Algarve, esse valor desce ligeiramente para 49,1%.

Por último, no respeitante à covariável idade, podemos afirmar que por cada ano a mais na idade de um indivíduo a sua possibilidade de migrar diminui cerca de 0,32%.

Ainda referente à idade, podemos ainda estimar a probabilidade de migrar de um indivíduo daqui a dez anos:

$$\pi(idade = 10) = \frac{\exp(1,1252 + (-0,0032) * 10)}{1 + \exp(1,1252 + (-0,0032) * 10)} = \frac{2,9838}{3,9838} = 0,75$$

Significa assim, que a probabilidade de um indivíduo migrar aos 10 anos é de 25%.

Tabela 25 – Regressão Logística Univariada

Variáveis	β	Std. β	p -value	OR	Intervalo Confiança 95%	
Ano	0,0081	0,0407	<u>0,8430</u>	1,0081	[0,9308; 1,0918]	
Sexo	0,0056	0,0407	<u>0,8900</u>	1,0057	[0,9286; 1,0892]	
NUT II	Centro	0,1466	0,0636	0,0211	1,1579	[1,0222; 1,3117]
	LVT	0,0263	0,0606	0,6639	1,0267	[0,9117; 1,1562]
	Alentejo	0,0568	0,0620	0,3601	1,0584	[0,9372; 1,1953]
	Algarve	-0,3205	0,0664	0,0000	0,7258	[0,6372; 0,8267]
NUT II 5 Ant.	Centro	-0,0128	0,0543	<u>0,8140</u>	0,9873	[0,8875; 1,0983]
	Alentejo	-0,0592	0,0642	<u>0,3562</u>	0,9425	[0,8311; 1,0688]
	Algarve	-0,1255	0,0660	<u>0,0571</u>	0,8820	[0,7750; 1,0038]
NUT II Nasc.	Centro	0,1690	0,0654	0,0097	1,1842	[1,0417; 1,3461]
	LVT	0,4115	0,0609	0,0000	1,5091	[1,3392; 1,7005]
	Alentejo	0,0245	0,0642	0,7027	1,0248	[0,9037; 1,1621]
	Algarve	0,3993	0,0646	0,0000	1,4908	[1,3134; 1,6921]
Idade	-0,0032	0,0008	0,0001	0,9968	[0,9952; 0,9984]	

Tal como no ponto anterior, e mantendo sempre $\alpha = 0,05$, ajustamos um modelo com todas as covariáveis que se mostraram anteriormente significativas (tabela 26), e neste caso, também todas elas continuaram a ser significativas para o modelo.

Mais uma vez, e após as covariáveis que mostraram ser significativas individualmente não terem “perdido” a sua significância, ajustamos um novo modelo com as que deixamos inicialmente de parte.

Tabela 26 – Modelo com variáveis significativas 1

Variáveis		β	Std. β	<i>p-value</i>	OR	Intervalo Confiança 95%
NUT II	Centro	0,1 15	0,0639	0,0115	1,1753	[1,0369; 1,3322]
	LVT	-0,0053	0,0613	0,9314	0,9947	[0,8820; 1,1218]
	Alentejo	0,0294	0,0635	0,6436	1,0298	[0,9093; 1,1663]
	Algarve	-0,4120	0,0683	0,0000	0,6623	[0,5790; 0,7572]
NUT II Nasc.	Centro	0,1719	0,0656	0,0088	1,1875	[1,0442; 1,3506]
	LVT	0,4703	0,0619	0,0000	1,6004	[1,4175; 1,8070]
	Alentejo	0,0605	0,0655	0,3561	1,0624	[0,9343; 1,2080]
	Algarve	0,4801	0,0663	0,0000	1,6162	[1,4193; 1,8405]
Idade		-0,0031	0,0008	0,0002	0,9969	[0,9953; 0,9985]

Ajustando então, um outro modelo com todas as covariáveis inicialmente testadas, podemos afirmar que, apenas a NUT II de residência cinco anos antes do recenseamento se tornou significativa para o modelo ($\alpha = 0,05$), tal como podemos constatar na tabela 27.

Resta, de seguida, incluir também no modelo as interações possíveis, de modo a no final chegarmos a um modelo com apenas aquelas que revelam ter coeficientes significativos e acrescentam alguma capacidade explicativa ao modelo.

Tabela 27 – Modelo com variáveis significativas 2

Variáveis		β	Std. β	<i>p-value</i>	OR	Intervalo Confiança 95%
NUT II	Centro	0,1937	0,0647	0,0027	1,2137	[1,0692; 1,3777]
	LVT	0,0271	0,0623	0,6635	1,0275	[0,9094; 1,1608]
	Alentejo	0,0628	0,0650	0,3342	1,0648	[0,9374; 1,2096]
	Algarve	-0,3313	0,0699	0,0000	0,7180	[0,6260; 0,8234]
NUT II 5 Ant.	Centro	-0,1710	0,0618	0,0057	0,8428	[0,7466; 0,9513]
	Alentejo	-0,1202	0,0823	0,1441	0,8868	[0,7547; 1,0420]
	Algarve	-0,4861	0,0854	0,0000	0,6150	[0,5202; 0,7271]
NUT II Nasc.	Centro	0,2438	0,0705	0,0005	1,2761	[1,1114; 1,4653]
	LVT	0,5560	0,0674	0,0000	1,7437	[1,5280; 1,9898]
	Alentejo	0,1206	0,0801	0,1320	1,1282	[0,9643; 1,3198]
	Algarve	0,7434	0,0817	0,0000	2,1031	[1,7919; 2,4683]
Idade		-0,0031	0,0008	0,0001	0,9969	[0,9953; 0,9985]

Neste sentido, e após termos testado cada interação de forma individual, incluído todas as que se revelaram individualmente significativas no modelo, e retirando as que deixaram de ser significativas, chegamos a um modelo em que apenas as interações entre a NUT II de residência cinco anos antes do recenseamento, bem como a NUT II de nascimento, e a idade, se mostraram significativas para o modelo (tabela 27).

A tabela 28 apresenta assim, o modelo e todas as variáveis e interações e covariáveis que se mostraram significativas para o modelo. No entanto, e apesar de estas terem sido validadas com $\alpha = 0,05$, resta verificar se acrescentam, ou não, alguma capacidade explicativa ao modelo ajustado anteriormente, ou seja, ao modelo que fechamos sem interações.

Apesar de durante o processo de modelação estatística, tal como no ponto anterior, termos utilizado a razão de verosimilhanças entre dois modelos ajustados com o intuito de dissiparmos algumas dúvidas relativamente à sua significância para o modelo, utilizou-se ainda critério de AKAIKE que nos indicará qual o modelo com melhor capacidade explicativa. Assim, o AIC associado ao modelo sem interações (tabela 27) é de 14107, enquanto o correspondente ao modelo com as interações significativas (tabela 28) é de 14004, sendo ligeiramente inferior, o que significa que é o melhor modelo e que as interações acrescentam capacidade explicativa ao modelo ajustado.

Tabela 28 – Modelo com variáveis significativas 3

Variáveis		β	Std, b	<i>p-value</i>	OR	Intervalo Confiança 95 %
NUT II	Centro	0,1802	0,0651	0,006	1,1974	[1,0540; 1,3603]
	LVT	0,0065	0,0626	0,9171	1,0065	[0,8903; 1,1380]
	Alentejo	0,0746	0,0657	0,2565	1,0774	[0,9472; 1,2256]
	Algarve	-0,3325	0,0705	0,0000	0,7171	[0,6246; 0,8233]
NUT II 5 Ant,	Centro	0,3562	0,1213	0,0033	1,4278	[1,1258; 1,8110]
	Alentejo	-0,3313	0,1573	0,0352	0,7180	[0,5275; 0,9773]
	Algarve	-0,3940	0,1630	0,0156	0,6744	[0,4900; 0,9282]
NUT II Nasc,	Centro	-0,2490	0,1414	0,0782	0,7795	[0,5909; 1,0285]
	LVT	-0,1933	0,1315	0,1414	0,8242	[0,6370; 1,0664]
	Alentejo	0,3103	0,1610	0,0539	1,3639	[0,9949; 1,8698]
	Algarve	0,8243	0,1631	0,0000	2,2803	[1,6565; 3,1391]
Idade		-0,0037	0,0020	0,0647	0,9963	[0,9924; 1,0002]
NUT II Centro:Idade		-0,0139	0,0026	0,0000	0,9862	[0,9811; 0,9913]
NUT II LVT:Idade		0,0057	0,0036	0,1194	1,0057	[0,9985; 1,0129]
NUT II Alentejo:Idade		-0,0016	0,0038	0,6790	0,9984	[0,9909; 1,0060]
NUT II Algarve:Idade		0,0132	0,0030	0,0000	1,0133	[1,0074; 1,0192]
NUT II 5 Ant Centro:Idade		0,0188	0,0028	0,0000	1,0189	[1,0133; 1,0246]
NUT II 5 Ant Algarve:Idade		-0,0052	0,0036	0,1543	0,9949	[0,9878; 1,0019]
NUT II 5 Ant Alentejo:Idade		-0,0022	0,0038	0,5624	0,9978	[0,9904; 1,0052]

Após termos “encerrado” o modelo, importa ainda verificar se ajusta aos dados, e para isso recorreremos, novamente, ao teste de bondade do ajustamento *Hosmer e Lemeshow* (tabela 29).

Deste modo e definindo um $\alpha = 0,05$, podemos afirmar que existe evidência estatística para rejeitarmos a hipótese nula, onde H_0 : O modelo ajusta-se aos dados *vs* H_1 : O modelo não se ajusta aos dados, o que significa que os valores estimados pelo modelo não são próximos dos valores observados.

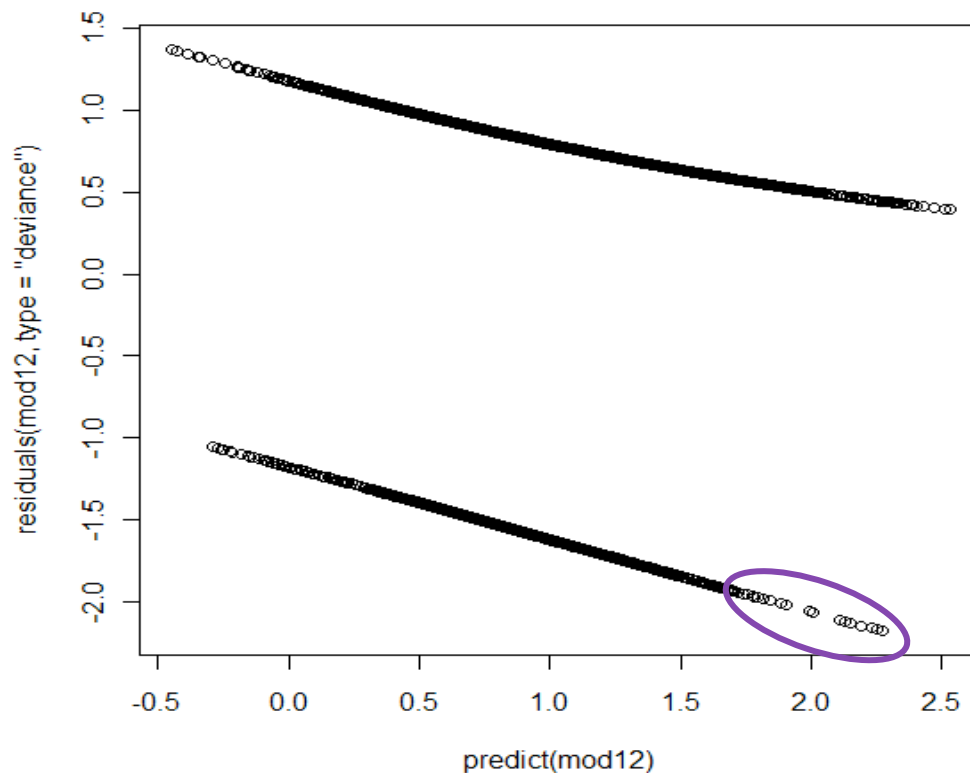
Por outro lado, a sua capacidade explicativa também não é muito elevada, explicando apenas 2,1% da variação da variável resposta, ou seja, do acto de migrar ou não migrar, onde $\rho^2 = 1 - \frac{RD/gl_{RD}}{ND/gl_{ND}}$.

Tabela 29 – Hosmer e Lemeshow

χ^2	f	p-value
16,8036	8	0,0322

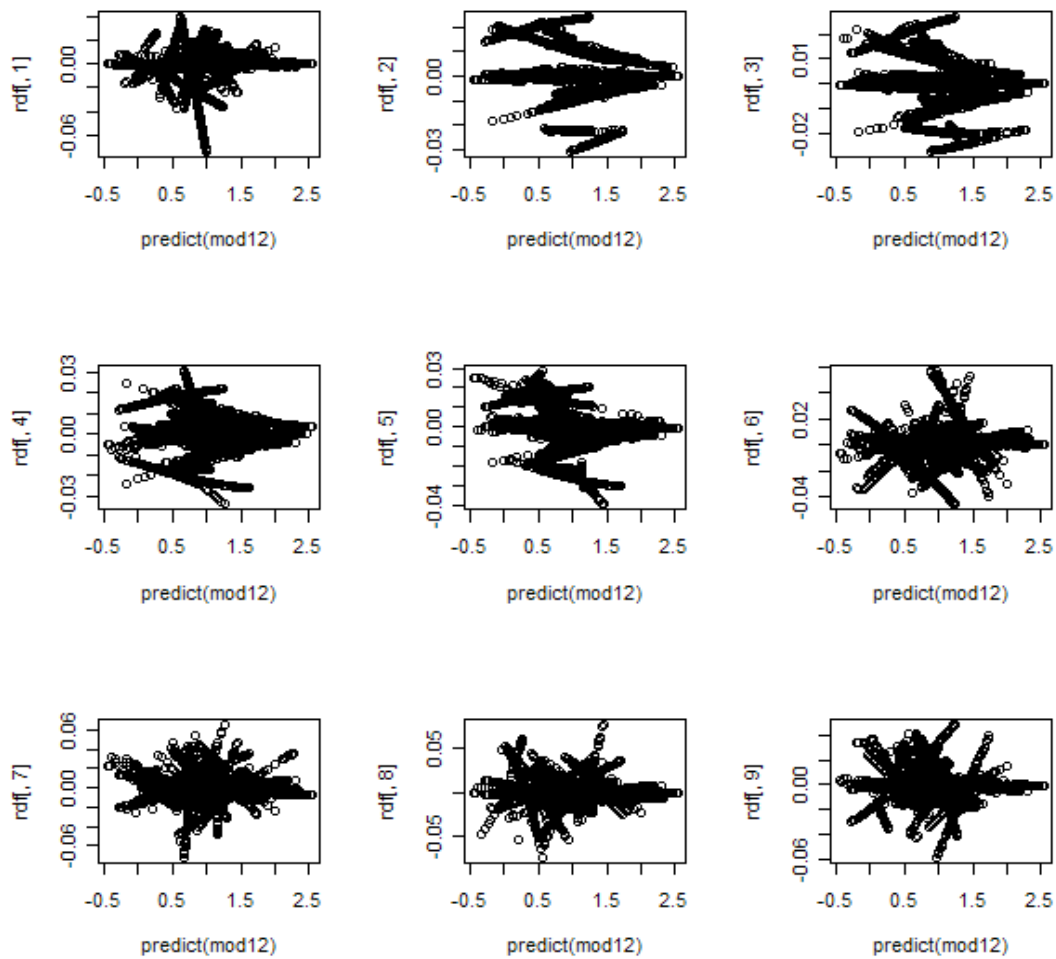
Quanto à análise de resíduos, e começando novamente pela interpretação dos resíduos *deviance*, podemos constatar que, uma vez que estes deveriam variar entre -2 e 2, que se registam algumas observações que poderão ser consideradas aberrantes. Esta situação pode ser identificada através da observação do *gráfico 33*.

Gráfico 33 – Resíduos *Deviance* (2)



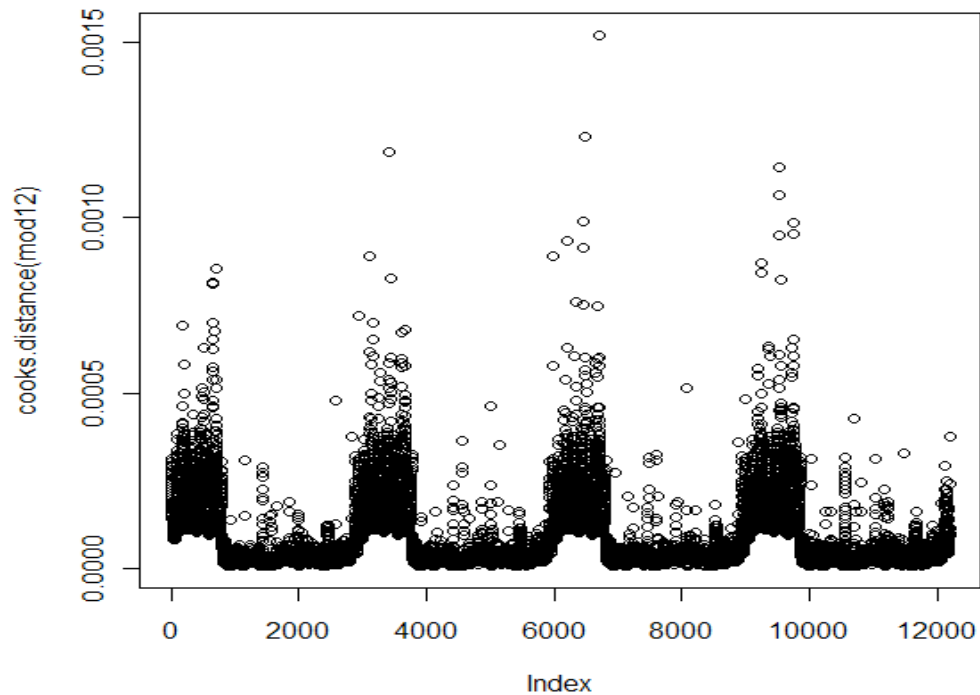
Representando graficamente agora os resíduos *dfbetas*, e tal como podemos observar em alguns gráficos obtidos (*gráfico 34*), constatamos que não se registam observações que se destacam excessivamente.

Gráfico 34 – Resíduos *Dfbetas* (2)



Quanto à distância *de Cook*, representados no *gráfico 35*, podemos constatar que apesar de se observarem alguns pontos mais distantes, em termos de escala, esta é muito reduzida, pelo que não podemos afirmar a existência de pontos influentes.

Gráfico 35 – Resíduos de *Cook* (2)



Neste situação, e tal como no caso do evento ter sido calculado com base na região de residência cinco antes do recenseamento, a análise que o efeito que as observações têm nos valores preditos (*leverage*) foi deixada de parte, poiso facto de estarmos perante muitas observações não nos permite medir de forma precisa o efeito que a observação tem nos valores preditos e, por outro lado, estaríamos a avaliar demasiadas observações com base num só limite.

Recorrendo novamente a outra ferramenta estatística, o SPSS, com o intuito de avaliarmos a capacidade de discriminação do modelo, representamos a curva ROC de modo a analisarmos a área existente abaixo (*gráfico 36*). Desta forma, o registo de um valor de 0,596 indica que o modelo ajustado tem um fraco poder discriminativo (discriminação pobre), no entanto, e tomando um $\alpha = 0,05$, podemos ainda afirmar que a própria área sob a curva é significativamente superior a 0,5 ($p\text{-value} = 2,329\text{E-}60$).

Gráfico 36 – Curva ROC (2)

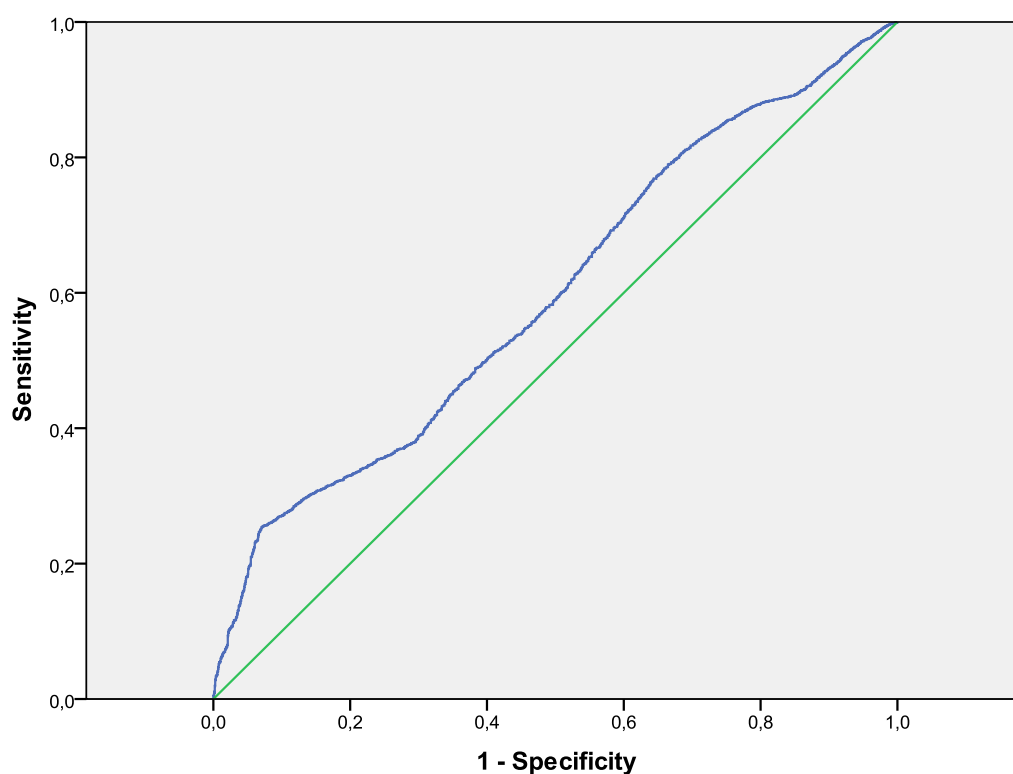


Tabela 30 – Área abaixo da Curva

Area	Std. Error ^a	Asymptotic Sig. ^b	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0,596	0,006	0,000	0,585	0,607

Através dos resultados obtidos, recorreremos às *coordenadas da curva* com o intuito de definirmos um ponto de corte equilibrado, que corresponde a 0,72, onde a *sensibilidade* é de 0,546 e “1 – *especificidade*” de 0,454.

Constatamos assim, que o modelo classifica correctamente 57,0% dos indivíduos que cinco anos atrás residiam numa região diferente daquela em que residem à data do recenseamento. Por outro lado, a sua *especificidade* indica que o modelo classifica correctamente 52,0% dos indivíduos que não alteraram a sua residência no referido espaço de 5 anos.

6.3. Interacções regionais

Durante a construção dos dois modelos apresentados anteriormente, deparamo-nos com a dificuldade em que alguns dos modelos ajustados convergissem, especialmente quando eram utilizadas as covariáveis referentes às NUTS II, quer de residência, na altura do recenseamento e há cinco anos atrás, quer de nascimento.

Assim, e com o intuito de contornarmos estas dificuldades, resolvemos recorrer ao teste de independência do qui-quadrado, de forma a podermos testar se as referidas variáveis se encontravam associadas.

Neste sentido, e apesar de termos estratificado a análise efectuada pelos diferentes grupos de idades, constatamos que, ao nível de significância de 5%, existe evidência estatística para afirmarmos que a região de residência cinco anos antes do recenseamento, a região de nascimento, e a região de residência, em qualquer dos grupos de idades, não são independentes entre si (obtenção de p-values muito baixos para todas as idades independentemente da região de nascimento ou de residência cinco anos atrás). Existe assim, um determinado grau de associação entre as covariáveis em análise, independentemente da forma como foi “interpretado” o acto de migrar.

H_0 : A distribuição das contagens pelos grupos é independente das variáveis V H_1 : A distribuição das contagens pelos grupos não é independente das variáveis

7. Conclusão/Discussão de Resultados

A principal conclusão que podemos extrair da aplicação das duas metodologias em questão prende-se com o facto de, e tendo em conta a atractividade económica, a região de Lisboa e Vale do Tejo, onde se situa a capital do país, ser inquestionavelmente a mais atractiva para os habitantes de todo o restante espaço nacional. Além desta primeira grande conclusão, uma outra se destaca: o padrão migratório inter-regional passível de ser identificado baseia-se na proximidade geográfica, na medida em que, se em primeiro lugar, as migrações registadas tiveram em comum o destino preferencial da região de Lisboa e Vale do Tejo, a segunda opção recaiu sempre sobre as regiões geograficamente mais próximas da região de origem.

Noutro sentido, foi também possível identificar, um outro padrão agora ao nível da estrutura etária. O estudo permitiu reconhecer três fases migratórias distintas, em que a primeira corresponde às crianças que registam idades até aos 9 anos completos; a segunda se identifica com os indivíduos com idades compreendidas entre os 20 e os 34 anos de idade; e por último, a terceira, constituída pelos migrantes com idades superiores a 70 anos.

A análise das diferenças verificadas entre as estimativas efectuadas e os valores observados, permite também concluir que, apesar da existência de algum desfasamento, se observa uma boa aproximação aos padrões de comportamento efectivamente verificados, e que, em qualquer uma das situações apresentadas, o desfasamento entre as estimativas e os valores efectivos, se traduziu, para 2001, numa *subestimação* da proporção migratória efectiva.

Esta situação reforça assim a dificuldade existente em prever ou até estudar o fenómeno migratório, como foi afirmado inicialmente, pois não existe um padrão que possamos identificar e que se mantenha inalterado ao longo dos anos, tal como a inércia dos comportamentos demográficos nos permite nos casos da previsão da fecundidade e da mortalidade.

Por outro lado, constatamos também que a tendência para a deslocação populacional das regiões menos populosas para as com mais população saiu reforçada, pois denotou-se uma clara atractividade das regiões mais populosas e simultaneamente mais urbanas. No entanto, a especificidade da análise, elaborada em função das NUTS II não nos permite identificar se a deslocação populacional direccionada para as regiões litorais se mantém

ou não. Esta avaliação apenas seria possível através da utilização de um nível de desagregação mais elevado (NUTS III, por exemplo), mas que nos traria outros problemas, pois o baixo número de residentes em algumas regiões tornaria impossível a aplicação destas metodologias.

Por outro lado ainda, podemos afirmar que, apesar da estimação indirecta nos permitir identificar um padrão comportamental caracterizador dos fluxos migratórios internos em Portugal continental, é possível e desejável melhorar quantitativamente as estimativas efectuadas.

Por último, importa ainda referir que, devido às mais diversas adversidades, alguns dos objectivos específicos não foram cumpridos, tais como a referida comparação com resultados espanhóis, pois o facto da questão essencial (onde residia há cinco anos atrás?) não estar incluída no recenseamento geral da população de 2001, não nos permitiria realizar uma comparação entre os resultados obtidos nos dois países.

Bibliografia

- Abreu, A. (2004). Migrações e desenvolvimento: alguns comentários críticos, *SOCIUS Working Papers* n.º6/2004, Lisboa.
- Bijak J. e Kupiszewska D. (2008). Methodology for the Estimation of Annual Population Stocks by Citizenship Group, Age and Sex in the EU and EFTA Countries. *Informatica (Ljubljana)*, 32(2), 133–145. <http://www.soton.ac.uk/socsci/socstats/staff/profile.php?name=JakubBijak>
- Carneiro, R. et. al. (2005). *A Mediação Sócio-Cultural: um puzzle em construção*, ACIME, Lisboa.
- Carrilho, M.; Patrício, L. (2002). “A situação demográfica recente em Portugal”, *Revista de estudos Demográficos* n.º38, INE – Lisboa.
- Carrilho, M. and Patrício, L. (2007). A situação demográfica recente em Portugal. *Revista de Estudos Demográficos* nº40, INE – Lisboa.
- Faraway, J.J. (2006), *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Taylor & Francis.
- Genereux, A. (2007). A review of migration and fertility theory through the lens of African immigrant fertility in France. *MPIDR Working Paper* WP 2007–008, Germany, <http://www.demogr.mpg.de/papers/working/wp-2007-008.pdf>
- Hosmer D.W., Lemeshow S. (1989). *Applied Logistic Regression*. New York: Wiley.
- van Imhoff E e Henkens K. (1998). The budgetary dilemmas of an ageing workforce: a scenario study of the public sector in the Netherlands. *European Journal of Population* 14: 39–59. <http://www.nidi.knaw.nl/en/output/1998/ejp-14-01-vanimhoff.pdf/ejp-14-01-vanimhoff.pdf>
- Jenninsen, R. (2003). Economic Determinants of Net International Migration on Western Europe, *European Journal of Population*, Netherlands.
- Jennissen, R. (2004). *Macro-Economic Determinants of International Migration in Europe*, University of Groningen, Netherlands.

- Kupiszewski, M. et. al. (2001). Internal Migration and Regional Dynamics in Europe: Sweden Case Study, in *Working Paper* 01/01, Leeds.
- Ledent J. 1980. Constructing multiregional life tables using place-of-birth-specific migration data. Working Paper WP-80-096, International Institute for Applied Systems Analysis. <http://www.iiasa.ac.at/cgi-bin/pubsrch?WP80096.pdf>
- Lutz W, Goujon A e Doblhammer-Reiter G. 1999. Demographic dimensions in forecasting: Adding education to age and sex. In *Frontiers of Population Forecasting*, Lutz W, Vaupel J, Ahlburg D (eds). The Population Council: New York; 42–58. Supplement to *Population and Development Review* 24(4).
- Lutz W. e Goujon A. 2001. The world's changing human capital stock: Multi-state population projections by educational attainment. *Population and Development Review* 27: 323–339. DOI: 10.1111/j.1728-4457.2001.00323.
- Machado, F. (2003). Imigração e Imigrantes em Portugal – Parâmetros de regulação e cenários de exclusão, *Sociologia, Problemas e Práticas* nº41, ISSN.
- Machado, C. (2004). Estabelecimento dos Fluxos Migratórios para o Brasil: Análise exploratória dos dados censitários de 1991 e 2000” *Colorado Conference on the Estimation of Migration*, Institute of Behavioral Science, University of Colorado at Boulder.
- Magalhães, M. (2003). Migrações Inter NUTS II e projecções regionais de População Residente. *Revista de Estudos Demográficos*, INE – Lisboa 34(5), 61–71
- Maroco, J. (2007). Análise estatística com utilização do *SPSS*. Edições Sílabo, Lisboa.
- Maroco, J. & Bispo, R. (2005). Estatística aplicada às ciências sociais e humanas. Climepsi Editores, Lisboa.
- Marques, J. (2008). Migrantes Europeus: Os portugueses na Suíça. 1ª Ed., ICS, Imprensa de Ciências Sociais, Lisboa
- Nyberg-Sorensen, N. et al. (2002). The Migration-Development Nexus: Evidence and Policy Options, *International Organization for Migration Series* nº8.

- Patrício, L.; Gonçalves, C. (2007). Envelhecimento crescente mas espacialmente desigual, *Revista de Estudos Demográficos* n.º40, INE, Lisboa.
- Peixoto, J. (1998). Selectividade Migratória e Dinâmicas Regionais: as migrações inter-regionais em Portugal nos anos 80”, in *Revista de Estatística* n.º9 3º quadrimestre, INE, Lisboa.
- Peixoto, J. and Oliveira, C. (2001). Migrações Inter-Regionais em Portugal Continental, 1992-1999. *Revista Portuguesa de Estudos Regionais – Região de Lisboa e Vale do Tejo*, n.º2 1º Semestre, INE/DRLVT – Lisboa
- Peixoto, J. (2004). As Teorias Explicativas das Migrações: Teorias Micro e Macro – Sociológicas, *SOCIUS Working Papers* n.º11/2004, Lisboa.
- Pereira, A., (2006). *Guia prático de utilização do SPSS: Análise de Dados para Ciências Sociais e Psicologia*, 6ª Edição, Edições Sílabo, Lisboa.
- Pestana, M.; Gageiro, J. (2005). *Análise de Dados para Ciências Sociais: a Complementaridade do SPSS*, 4ª edição, Edições Sílabo, Lisboa.
- Pires, R. (2002). Mudanças na Imigração: uma análise das estatísticas sobre a população estrangeira em Portugal, 1998-2001, *Sociologia, Problemas e Práticas* n.º39, ISSN.
- Raymer, J. and Rogers, A. (2007). Using Age and Spatial Flow Structures in the Indirect Estimation of Migration Streams. *Demography* 44(2), 199–223
- Rees, P. et al. (1998). Internal Migration and regional Population Dynamics in Europe: Portugal Case Study, *Working Paper* 98/13, Leeds.
- Rees, P.; Kupiszewski, M. (1999). Internal Migration: What Data are available in Europe?, *Journal of Official Statistics* vol.15 n.º4, Statistics Sweden, Sweden.
- Rogers A. (1975). Introduction to Multiregional Mathematical Demography. John Wiley: New York.
- Rogers, A. (2002). The Indirect Estimation of Migration: A Proposal for a Multinacional Study, *IBS Working Paper Pop* 2002-0004, University of Colorado, USA.

- Rogers, A. and Liu, J. (2005). Estimating Directional Migration Flows from Age-Specific Net Migration Data. *Review of Urban and Regional Development* 17(3), 177–196
- Rogers A, Little J e Raymer J. (1999). Disaggregating the historical sources of regional foreign-born and native-born population growth in the United States: a new method with applications. *International Journal of Population Geography* 5: 449–475.
- Rogers, A., Raymer, J. and Jordan, L. (2003). Inferring Migration Flows from Birthplace-Specific Population Stocks. *Population Program, Institute of Behavioral Science, University of Colorado, Working Paper POP 2003-0002*, 55p, Colorado, <http://www.colorado.edu/ibs/pubs/pop/pop2003-0002.pdf>
- Rogers, A., Liu, J., Partida, V. and Tilton, M. (2006). Inferring Directional Migration Flows from Net Migration Data: Mexico. *Population Program, Institute of Behavioral Science, University of Colorado, Working Paper POP 2006-01*, 37p, Colorado, <http://www.colorado.edu/ibs/pubs/pop/pop2006-0001.pdf>
- Rogers, A., Jones, B., Partida, V. and Muhidin, S. (2007). Inferring Migration Flows from the Migration Propensities of Infants: Mexico and Indonesia. *The Annals of Regional Science* 41, 443–465
- Rogers, A., Willekens, F., Little, J. and Raymer, J. (2002). Describing migration spatial structure. *Papers in Regional Science* 81, 29–48
- Siegel JS. (2002). The demography of the labor force and the workforce of organizations. In *Applied Demography: Applications to Business, Government, Law and Public Policy*. Academic Press: San Diego; 329–395
- Smith KD, Perez-Johnson I. e Wooldridge J. (2000). Uncertainty and forecasting local health professional shortages. *Population Research and Policy Review* 19: 477–503. DOI: 10.1023/A:1010700930344
- Turkman, M.A.M, & Silva, G.L., (2000) Modelos Lineares Generalizados da teoria à prática,, Edições SPE, Lisboa.