Capítulo I

Economias de escala e de gama

e eficiência produtiva na banca

Introdução

Neste capítulo, vamos abordar a teoria e o cômputo de *economias de escala* e *de gama* e da eficiência custo aplicadas à banca.

As instituições de crédito, nomeadamente as portuguesas, baseiam a sua política de crescimento, quer internamente quer através de fusões e de acordos de cooperação, na necessidade de adquirirem dimensão suficiente para conseguirem sobreviver no mercado da União Europeia por via de economias de escala. Se o crescimento da escala de produção puder ser baseada num crescimento menor dos factores de produção, então os consequentes custos mais baixos promoverão, no longo prazo, uma situação tendencialmente de monopólio natural onde surgirão bancos de maior dimensão e mais eficientes.

Scherer (1980) sustenta que há várias fontes de redução de custos com o aumento da escala de produção. Em primeiro lugar, o carácter indivisível de alguns factores de produção em relação ao output pode permitir a redução dos custos médios à medida que o nível de produção aumenta. Em segundo lugar, o aumento da escala pode permitir uma mais eficiente afectação dos recursos por via da especialização. Em terceiro lugar, algumas inovações tecnológicas (como software e hardware) podem ser incorporadas, mais facilmente, quando a escala é maior. Finalmente, os bancos de maior dimensão não necessitam de ter, proporcionalmente ao total do activo, as mesmas disponibilidades que os bancos pequenos; de igual forma, os maiores bancos estão em melhores condições para

captar recursos e oferecer produtos diversificados, diminuindo o risco não sistemático (Kolari e Zardkoohi, 1987).

As economias de gama ou de diversificação ocorrem, como vimos, quando a produção de vários produtos por uma mesma empresa é superior àquela produzida por várias empresas, cada uma produzindo um único produto. O apuramento destes ganhos de diversificação poderá conduzir à aposta em bancos com capacidade de oferecer linhas de produtos alargados em detrimento de bancos especializados.

Vários autores[1] apontam para possíveis origens da diminuição de custos por via da diversificação. Em primeiro lugar, se um banco tem excesso de capacidade em algum sector, a produção de mais outputs leva à maior diluição dos custos fixos. Em segundo lugar, a utilização da informação de mercado[2] para um mix alargado de produtos irá diminuir os custos médios associados. Em terceiro lugar, a diversificação dos activos por diferentes grupos pode promover a diminuição do risco da carteira e das taxas de juro. Em quarto lugar, haverá a possibilidade da redução dos custos de transacção a suportar pelos clientes dos bancos que operem com diferentes produtos de um mesmo banco.

De igual forma, vamos situar o nosso estudo sobre as diferentes aproximações à eficiência na banca.

[1] Baumol, Panzar e Willig (1988), pág. 75-79

[2] Os elevados custos derivados da obtenção de informação imediata do mercado foram relevados já em 1975 por O.R. Willimson em *Markets and Hierarchies: Analysis and Antitrust Implications,* Free Press

O problema da eficiência está relacionado com a escala operativa, mas também com a existência, ou não, de desperdícios na afectação dos recursos; ou seja, os custos mais ou menos elevados de uma instituição de crédito podem ter origem quer na eficiência de escala quer na eficiência produtiva.

As medidas de eficiência de escala estão apenas associadas à dimensão, enquanto a eficiência-X (eficiência produtiva) mede a maior ou menor aproximação à fronteira eficiente para aquela dimensão. Os estudos recentes para a análise da eficiência têm-se baseado quer em aproximações paramétricas quer em não – paramétricas (DEA — data envelope analysis). O primeiro capítulo pretende enquadrar estes desenvolvimentos.

I.1. Alguns conceitos da teoria da empresa aplicáveis a empresas multiproduto

Baumol et al.(1988) constituem uma importante referência no tratamento da extensão dos conceitos da teoria de custos de empresas produzindo um só produto para empresas multiproduto.

I.1.1. Economias de escala

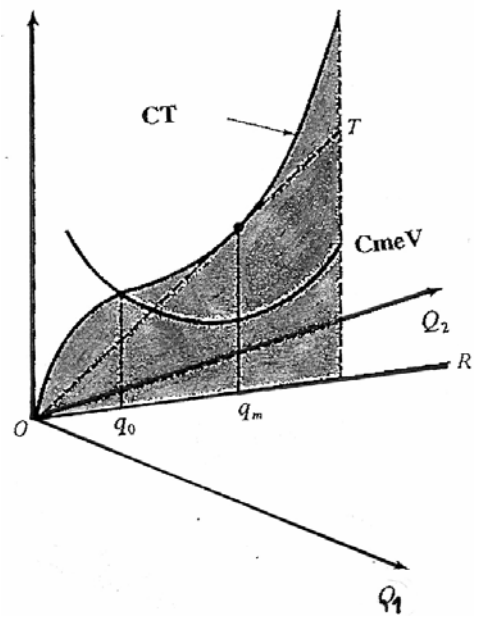
O conceito de custo médio para uma empresa multiproduto (que vamos designar por *Custo Médio Vectorial - CMeV)* tem como pressuposto que o custo total varia proporcionalmente à medida que o nível da produção (de um vector de output) se altera, sendo dado por

$$CMeV(Q) = \frac{CT(Q)}{\sum_{i=1}^{n} Qi} = \frac{CT(tq_0)}{t\sum_{i} Qi}$$

onde q_0 é o vector unidade referente a determinados outputs e t é o número de unidades no vector tq_0=Q.

A figura 1 pretende, através de uma representação tri-dimensional, inscrever o CMeV, os custos totais e as economias de escala de uma empresa multiproduto.

Figura 1 — Economias de escala para uma empresa multiproduto



Fonte: Baumol W. et al.(1988), p.50

O plano Q1OQ2 representa o espaço do output; a proporção em que os outputs representados por Q1 e Q2 são produzidos é dada pelo vector ou raio OR[3]. O vector da produção movimenta-se ao longo do vector OR. Da mesma forma que, no caso de uma empresa produzindo um único produto, o mínimo do Custo Variável Médio (CMeV), no ponto $q_{0,}$ corresponde à escala mais eficiente (para a empresa produzindo Q1 e Q2 na proporção definida pelo vector OR).

O *grau de economias de escala globais* ($EEG_N(Q)$) definido para o conjunto de toda a produção N ($N=\{1,2,..n\}$) é dado por

[3] A expressão *raio* é associada à denominação inglesa de *RAC-Ray Average Cost* avançada por Baumol, Panzar e Willig(1988)

$$EEG_N(Q) = \frac{CMeV(Q)}{CM(Q)} = \frac{CT(Q)}{\sum_{i=1}^{n} Qi * \frac{\partial CT(Qi)}{\partial Qi}}$$

Sendo o custo médio vectorial dado por $CMeV(tQ) = \frac{CT(tq_0)}{t\sum_i Qi}$, a sua primeira derivada em ordem a t será

$$\frac{dCMeV(tQ)}{dt} = \frac{1}{(t\sum Qi)^2}\left[(t\sum Qi)Q * \frac{1}{CT(tq_0)} - CT(tq_0)\sum Qi\right]$$

Se designarmos por ε a elasticidade do CMeV em relação a t, no ponto q_0, teremos

$$\varepsilon = \frac{dCMeV(tQ)}{dt} * \frac{t}{CMeV(tQ)} = \left(\frac{1}{EEG_N(tQ)} - 1\right)$$

ou seja, para t=1,

$$EEG_N(Q) = \frac{1}{\varepsilon + 1} = \frac{1}{\sum_{i=1}^{n} \frac{\partial \ln CT}{\partial \ln Q_i}}$$

O grau de economias de escala pode ser interpretado como a elasticidade do output em relação ao custo necessário para a sua produção (Baumol et al., 1988, p.51). Haverá economias, deseconomias de escala ou constância à escala no ponto q_0 se a derivada nesse ponto do CMeV for negativa ou positiva, sendo $EEG_N>1$, $EEG_N<1$ ou $EEG_N=1$, respectivamente.

Até agora, estivemos interessados em analisar, através da introdução dos conceitos de custo médio vectorial e do grau de economias de escala, de que forma é que os custos variavam em função de alterações proporcionais nas quantidades do conjunto dos produtos (ao longo do vector OR). Um outro aspecto importante é saber de que forma os custos variam em função da alteração do output de um único produto, mantendo-se as quantidades dos outros produtos constantes. É assim que surge o conceito de *custo incremental (CI)* da produção do produto i, definido como o custo total (*CT(Q)*) de uma empresa multiproduto, dado um determinado vector de outputs subtraído do custo decorrente do abandono da produção do produto i (*CT(*Q_{m-i}*)*), mantendo-se todos os restantes outputs constantes. Ou seja,

$$CI(Q)=CT(Q)-CT(Q_{m-i})$$

em que Q_{m-i} é o vector da produção com o elemento zero na posição i. O *custo médio incremental (CMeI)*[4] é definido como o aumento do custo resultante do nível de produção extra de um determinado produto i, quando comparado com o custo da sua não produção, dividido pelo output desse produto i:

$$CMeIi(Q) = \frac{CIi(Q)}{Qi}$$

[4] O termo *incremental* é preferível ao *variável* adoptado por autores como Turner e Areeda (1975) pela possibilidade de incluir, no cálculo do custo médio, alguma parcela correspondente a custos de cariz fixo (cfr. Baumol et al. 1988, p.67)

O *grau de economias de escala específicas* do produto i *($EEE_i(Q)$)*, é calculado, da forma habitual, através do quociente entre os custos médios e os marginais

$$EEE_i(Q) = \frac{CMeI_i}{\frac{\partial CT}{\partial Q_i}}$$

Como habitualmente, quando $EEE_i(Q)>1$, ou seja, quando o custo marginal é inferior ao custo médio, este último tem como primeira derivada um valor negativo, decrescendo à medida que Qi cresce (haverá economias de escala nesse intervalo). De igual forma, quando $EEE_i(Q)<1$, o custo marginal crescerá à medida que Qi aumenta, havendo deseconomias de escala. Quando $EEE_I(Q)=1$ estamos sobre o mínimo do custo médio, havendo economias constantes à escala.

A especificação adoptada para a função custo é um factor relevante para a interpretação do grau de economias de escala. A função custo de um departamento ou secção de um banco pode ser substancialmente diversa da referente a todo o banco. Estes dois níveis de análise são referidos na literatura bancária como *plant (branch) economies* e *firm economies*[5].

Nos primeiros estudos, a distinção entre estes dois níveis *plant* e *firm* era feita através da introdução de variáveis *dummy* (Benston, 1965;Bell e Murphy, 1968[6]).

[5] Optamos pelas designações originais anglo-saxónicas, dada a sua consagração na literatura bancária portuguesa.
[6] Citados por P.Molyneux et al (1996), p.157

Estudos mais recentes incluem na função custo variáveis referentes ao nível *plant*. Por exemplo, se se optar por uma especificação da função custo do tipo

$$CT = C\ (Q,W,B)$$

em que CT representa os custos da empresa bancária, Q o vector dos produtos bancários, W o vector de preços associado aos factores de produção e B o vector correspondente à variável (de estrutura) número de balcões - as economias de escala são calculadas no pressuposto de que o número de balcões permanece constante, durante o processo de estimação.

Para a estimação de economias de escala ao nível *firm*, tanto o número de balcões como as produções podem variar. Neste caso, a especificação da função custo do tipo

$$CT = C(Q,W)$$

não inclui explicitamente o número de balcões como variável autónoma; ou seja, a variação de custos decorrentes da alteração do número de balcões associada a mudanças de escala operativa é traduzida directamente pelo grau de economias de escala.

Na prática, muitos bancos, para aumentarem o seu nível de actividade, são obrigados a aumentar a sua rede de balcões, pelo que seria de desprezar o nível *plant*. Por outro lado, a formulação da função custos incluindo explicitamente o

vector B mostra-se superior, em termos de "implementação econométrica", à da formulação *firm* (Barros e Pinho, 1995, p. 42).

A harmonização das duas perspectivas foi resolvida pela introdução do conceito de *augmented economies of scale* — Clark, (1988, p.30) fala, neste caso, em *firm economies of scale* - que vamos designar por *grau de economias de escala totais (EET)*, medida por

$$EET = \frac{1}{\sum_{i=1}^{n} \frac{\partial \ln CT}{\partial \ln Q_i}} + \frac{1}{\sum_{i=1}^{n} \frac{\partial \ln Ct * \partial \ln B}{\partial \ln B * \partial \ln Q_i}}$$

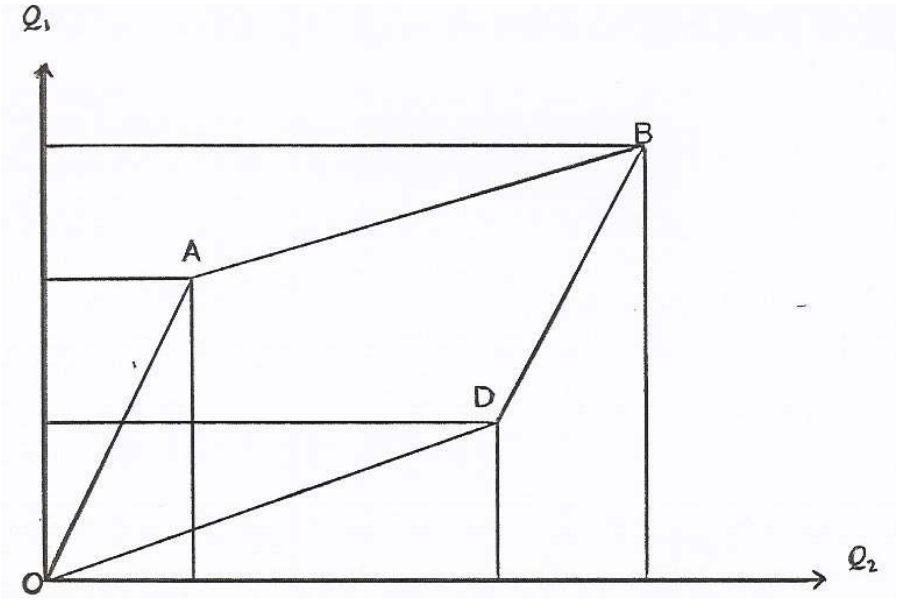
e permitindo autonomizar os efeitos da alteração de B, na segunda parcela de EET.

Estas medidas de mensuração das economias de escala pressupõem que o *mix* de produtos se mantém inalterado com a expansão da actividade, ou seja, inserem-se no conjunto das medidas de economias de escala *vectoriais* (*ray economies of scale)*.

Suponhamos[7] que estamos em presença de um banco de pequena dimensão, A, e que se pretendem calcular as economias de escala para o banco B, de maior dimensão:

[7] Cfr Barros e Pinho (1995), p.42-43

Figura 2 - Expansão de actividade e output mix



Fonte: Barros e Pinho, *Estudos sobre o sistema bancário português,* 1995, pág.43

A expansão da actividade do banco B faz-se ao longo de AO e AB e não de OB. Ou seja, se se abandonar a hipótese da permanência do *mix* de produtos à medida que varia a actividade, o cômputo de economias de escala (Berger et al., 1987[8]) poderá ser feito, através da elasticidade do custo relativamente à produção, em termos incrementais, ao longo de uma linha de expansão (*linha de expansão de economias de escala - LEEE-* de A para B*)*:

$$LEEE = \frac{1}{\sum_{i=1}^{n}\left(\frac{\left(Q_i^{\,B} - Q_i^{\,A}\right)/Q_i^{\,B}}{\left(CT(Q_i^{\,B}) - CT(Q_i^{\,A})\right)/CT(Q_i^{\,B})}\right) * \frac{\partial \ln CT(Q_i^{\,B})}{\partial \ln Q_i^{\,B}}}$$

[8] Citado por Barros e Pinho (1995), p..42

onde Q_i^A e Q_i^B são o volume de um determinado produto i produzido pelos bancos A e B.

I.1.2. Economias de gama[9]

Além das economias resultantes da escala ou da dimensão da produção de uma empresa há, também, a possibilidade da redução dos custos, por via da produção de diferentes bens ou serviços por uma única empresa, comparativamente à produção por empresas especializadas; as primeiras economias designam-se por economias de escala e as segundas por *economias de gama*. Da existência de economias de gama, virtualmente em todas as empresas, decorre a importância do estudo das empresas multiproduto (Willig, 1979).

Poderemos referir dois tipos de economias de gama. As *economias de gama internas ou de produção* resultantes da junção de serviços como, por exemplo, produção e marketing; e as *economias de gama externas ou de consumo* derivadas da possibilidade dos consumidores se abastecerem de vários produtos ou serviços, no mesmo local ou na mesma empresa.

Se os custos associados à produção conjunta de dois produtos forem designados por $CT\ (Q_1,Q_2)$ e as funções custo decorrentes da produção separada

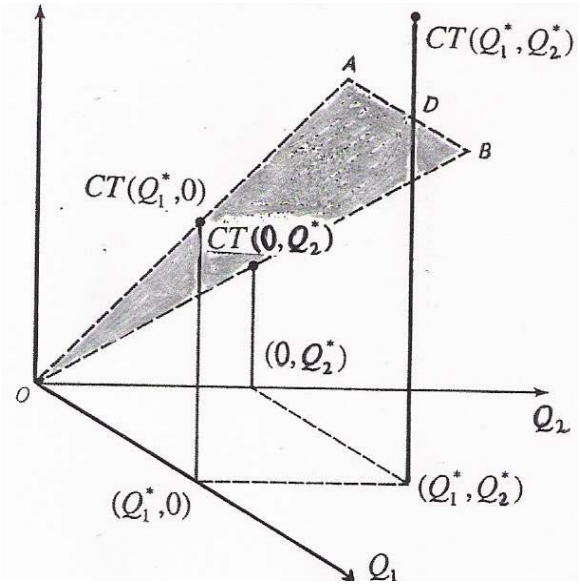
[9] Adoptamos a expressão *economias de gama* como tradução da terminologia anglo-saxónica economies of *scope*. Outros autores preferem traduzir como *economias de diversificação, de fins, de âmbito.* Autores como Pinho (1995) optam pela manutenção do termo original *scope*.

forem representadas por $CT(Q_1)$ e $CT(Q_2)$, então, estamos perante economias de gama se e só se

$$CT(Q_1, Q_2) < CT(Q_1) + CT(Q_2)$$

Haverá deseconomias de gama sempre que o sentido da desigualdade anterior for do tipo "maior do que".

Figura 3 - Economias de gama



Fonte: Adaptado de Baumol, Panzar e Willig, 1988, pág.72

Graficamente (figura 3) o conceito de economias de gama implica a comparação entre a soma dos custos associados à produção especializada de cada um dos produtos, representada por $CT(Q_1^*,0)+CT(0,Q_2^*)$, soma das alturas dos respectivos segmentos de recta inscritos no plano dos custos sobre os eixos, e $CT(Q_1^*,Q_2^*)$, altura no plano dos custos no ponto (Q_1^*,Q_2^*) — vector soma de

$(Q_1^*,0)$ e $(0,Q_2^*)$. Se $CT(Q_1^*,Q_2^*)$ se situar abaixo do hiperplano OAB, então, a condição para a existência de economias de gama é assegurada.

O *grau de economias de gama (EG)* de uma empresa pode ser definido como a medida do crescimento relativo dos custos resultante da separação da produção em dois (ou mais) produtos, Q_1 e Q_2.:

$$EG = \frac{CT(Q_1) + CT(Q_2) - CT(Q_1, Q_2)}{CT(Q_1, Q_2)}$$

Considerando n produtos, teremos:

$$EG = \frac{\sum_{i=1}^{n} CT(Q_i) - CT(Q)}{CT(Q)}$$

Se EG<0 excluir produtos da gama oferecida aumenta os custos médios de produção; ou seja, o valor, em módulo, obtido pela medida EG significa os ganhos de diversificação da produção em percentagem dos custos totais de produção. Se EG>0 ou EG=0 estaremos perante economias de gama decrescentes ou constantes, respectivamente.

Empiricamente, a medida EG é pouco utilizável, comportando alguns problemas.

Em primeiro lugar, a informação disponível não permite, de um modo geral, imputar os custos a cada produto; este problema é ultrapassável pelo recurso à condição suficiente para a existência de economias de gama (complementaridade dos outputs) ─ como veremos a propósito da função Translog.

Em segundo lugar, sendo a Translog a forma funcional actualmente mais utilizada em economia bancária não é definida no caso da não produção de qualquer produto, inviabilizando o cômputo de EG ─ esta questão foi resolvida, nomeadamente, pelo recurso à Translog Híbrida.

Em terceiro lugar, o *mix* de produtos vai-se alterando com a dimensão das empresas bancárias, tornando relevante o estudo da possível *subaditividade* das funções custos. Se o *mix* de outputs pode ser produzido a um custo mais baixo por uma única empresa do que por um conjunto de pequenas empresas, então, a função custo da indústria diz-se *subaditiva* e o monopólio natural tende a prevalecer. No caso de duas empresas A e C e dois produtos Q_1 e Q_2, a subaditividade pode ser expressa por

$$CT(Q) < CT(Q_1^A, Q_2^C) + CT(Q_1^A, Q_2^C)$$

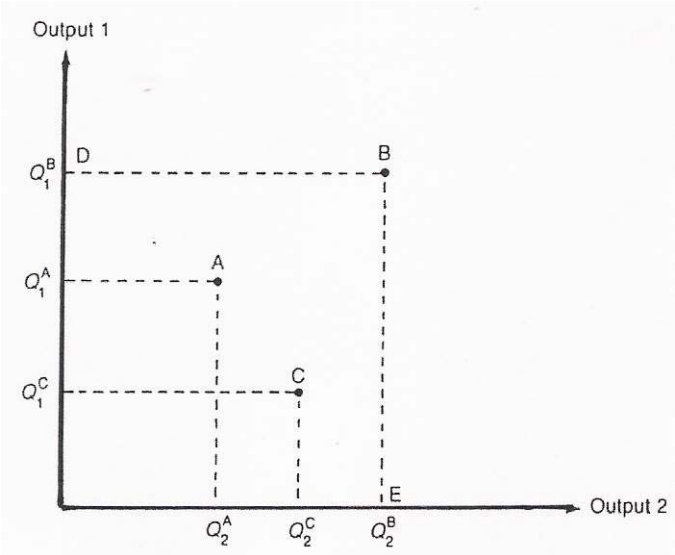
onde $Q = (Q_1, Q_2)$ e $Q = \left(Q_1^A + Q_2^C, Q_1^A + Q_2^C\right)$. Note-se que as empresas A e C têm *mix* de produtos opostos, sendo, portanto, empresas complementares. A subaditividade das funções custo é uma medida da eficiência relativa dos bancos de grande dimensão e de pequena dimensão, levando em linha de conta, simultaneamente, a escala e a diversificação.

Berger et al(1987) propuseram o conceito de l*inha de expansão de subaditividade* (*expansion path subadditivity*) — LES — que traduz em que medida uma empresa mais diversificada (B) será mais eficiente do que duas empresas mais especializadas, complementares e de menor dimensão(A e C).

$$LES(Q^B) = \frac{CT(Q^A) + CT(Q^C) - CT(Q^B)}{CT(Q^B)}$$

LES traduz a alteração do custo total resultante da partição de um banco de grande dimensão B em dois de mais pequena dimensão A e C.

Figura 4 - A linha de expansão de subaditividade



Fonte: P.Molyneux et al.(*1996*), *pág.212*

O cômputo das economias de gama através de LES traduz de que forma o custo total de uma empresa multiproduto deve ser maior ou menor (pontos A, C ou B da figura 4). A forma tradicional da medida das economias de gama é um caso particular de LES, quando os pequenos bancos se especializam nos pontos D e E da figura 4 (considerando $CT(Q_1^B,0) + CT(0,Q_2^B) - CT(Q^B) > 0$, por exemplo, estamos perante economias de gama crescentes).

Refira-se, por fim, que se uma empresa multiproduto tem uma grande diversidade de outputs, mesmo no caso em que o custo médio vectorial é decrescente, a não existência de economias de gama pode impedir o monopólio natural[10]. Uma empresa multiproduto, face à inexistência de economias de gama, tenderá a dividir-se em diversas empresas especializadas.

Da natureza multiproduto da empresa bancária decorre que, para o seu estudo, o conceito de economias de gama assuma uma grande importância.

A partir dos anos setenta, dada a crescente complexidade da empresa bancária, começou a ser ventilado o carácter multiproduto das instituições de crédito. Os desenvolvimentos de Baumol et al. (1988) foram sendo adaptados à actividade bancária. O cômputo das economias de escala específicas, a juntar às tradicionais economias de escala globais, aliadas ao das economias de gama constituem importantes contribuições para o estudo das empresas bancárias.

[10] Baumol et al. (1988), pág.88.

I.2. Os debates teóricos recentes na economia bancária e as suas implicações para o cômputo das economias de escala e de gama

Actualmente, existem dois debates fundamentais, e ainda não resolvidos, na literatura de economia bancária. O primeiro quanto à definição de empresa bancária e o segundo acerca da especificação da função custo (ou produção), comportando implicações quanto à medida das economias de escala e de gama (e da eficiência).

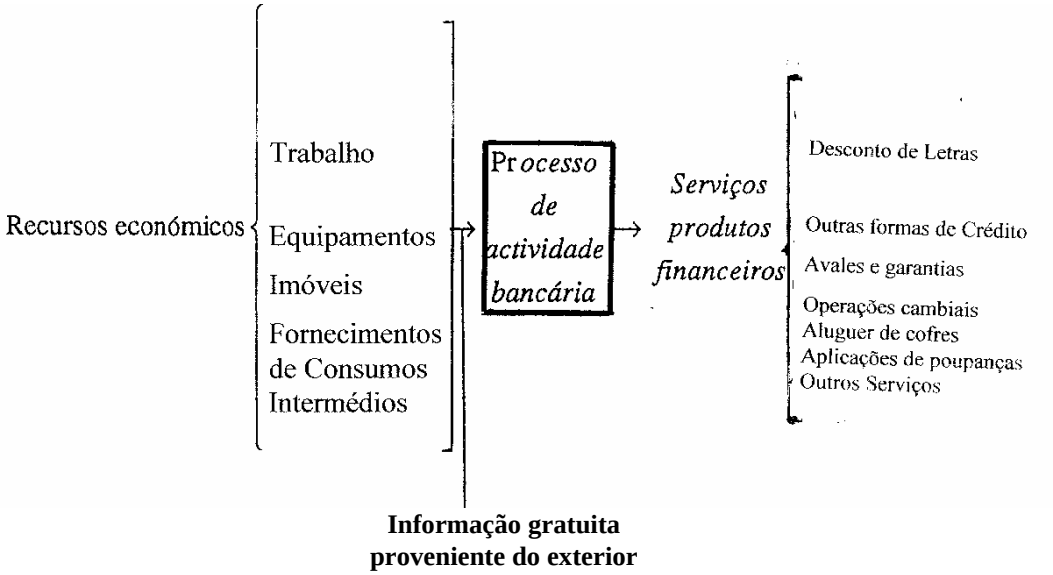
I.2.1. A definição de empresa bancária

A especificação de uma medida adequada do output de uma empresa bancária, ou seja, o que é que um banco produz e como se pode medir essa produção, é um dos debates teóricos não resolvidos na literatura de economia bancária. Ao contrário de uma empresa industrial, o output de um banco não pode ser medido através de quantidades físicas; a acrescer a esta dificuldade, uma empresa bancária caracteriza-se por ser multiproduto.

A estimação de funções de produção e de custos aplicadas à actividade bancária não foi objecto de estudos senão a partir de finais dos anos sessenta. Uma das razões explicativas dessa tardia aplicação é, justamente, a dificuldade de medição das produções bancárias decorrente da impossibilidade da sua mensuração em termos físicos.

O processo de produção bancária caracteriza-se pela captação de uma série de informação, pelo seu tratamento e pela disponibilização de nova informação sob a forma de serviços e de produtos financeiros:

Figura 5 - O processo da actividade bancária



Fonte: José Martins Barata (1984), *Modelo económico bancário: caso português,* ISE, p.13

A medida da produção bancária tem sido objecto de numerosas leituras, divergindo com base em questões teóricas e outras com suporte em problemas de mera operacionalidade — informação disponível e processos de estimação.

Os estudos efectuados em torno da medida do output bancário podem ser divididos em três grupos[11].

[11]D'Oliveira, Eduardo L.(1984), *The theory of banking: a critical survey of the literature,* The Open University, Julho, pp.27-31

Um primeiro grupo de trabalhos assume que o output bancário é constituído apenas por um único produto (por exemplo as *aplicações creditícias*) ou por um conjunto de serviços homogeneizados (por exemplo, sob a denominação de *depósitos*). As críticas fundamentais que se podem fazer a esta interpretação são: por um lado, o considerar que o output é homogéneo, ou seja, é posta de parte a ideia de que diferentes estruturas de depósitos, por exemplo, produzem diferentes custos operacionais; por outro lado, a simplificação excessiva da actividade bancária através de um único indicador (crédito ou depósitos) sabendo que aquela traduz a ligação, simultânea, entre aforradores e investidores. A caracterização da empresa bancária como produzindo um único produto homogéneo, ainda que teoricamente esteja actualmente posta de lado, pode decorrer da assumpção de formas funcionais que não comportam o caso multiproduto (que é o caso da formulação Cobb-Douglas e, de certa forma, da forma funcional tipo CES). Autores como Alhadeff (1954) e Schweiger (1962)[12] foram pioneiros neste tipo de abordagem.

Um segundo grupo de estudos associa a produção bancária a um fluxo, ao contrário dos trabalhos anteriores, fazendo a segmentação do output entre *empréstimos* e *não empréstimos* (incluindo estes últimos serviços de consultadoria financeira, aluguer de cofres, etc.). Esta abordagem constitui um avanço em relação aos primeiros estudos uma vez que comporta o carácter multiproduto da empresa bancária. Greenbaum (1967) e Powers (1969)[13] foram os autores que introduziram esta leitura.

[12] Cfr.E.L.d'Oliveira (1984), p.10
[13] Cfr. E.L. d'Oliveira (1984), p.11

Uma terceira categoria de abordagem da produção bancária põe a tónica no facto da actividade bancária ser um processo diferenciado. Assume como aproximação do output, nomeadamente, os diferentes tipos de depósitos, o número de contas, recorrendo a medidas físicas (em vez de medidas monetárias como nas duas leituras anteriores). Autores como Benston (1972) e Bell e Murphy(1967)[14] inserem-se nesta corrente. Note-se que, no caso português, esse tipo de informação não está disponível, tendo-se, se outras razões não existissem, de definir o output bancário por outra via.

A definição de produção bancária e a sua mensuração teve, nestas últimas três décadas, numerosas opções como o *total do activo*, os *depósitos totais,* os *depósitos à ordem,* as *aplicações creditícias a particulares e empresas* e os *créditos interbancários,* as *operações em títulos,* a soma da *margem financeira* (definida como a diferença entre *juros e proveitos* equiparados e *juros e custos* equiparados) e *outros resultados correntes* (diferença entre *proveitos por natureza e custos por natureza),* o *número de contas de aplicações creditícias e de depósitos*[15].

[14] Cfr. J.Martins Barata (1981), p.114-115

[15] Para uma revisão da literatura em torno da mensuração do output bancário cfr.P.Molyneux(1996), p.152-6

Toda esta problemática da definição da produção bancária tem como suporte teórico a definição económica de empresa bancária que trouxe à literatura duas aproximações: a da *produção* e a da *intermediação* (Humphrey,1985).

Segundo a *abordagem produção,* a empresa bancária é uma empresa de serviços caracterizada por captar recursos (depósitos à ordem, depósitos a prazo) com vista à sua aplicação (créditos, aplicações em títulos, participações de capital); ou seja, as empresas bancárias são produtoras de serviços associadas a créditos e depósitos, sendo os factores de produção o trabalho e o capital. Para o cômputo dos custos, todos os serviços são considerados como outputs distintos, sendo geralmente o número de contas de depósitos e de empréstimos a unidade de medida adoptada para a mensuração do output. Os custos totais considerados nas análises empíricas excluem, logicamente, os encargos financeiros, comportando somente os custos operatórios ou de produção.

Segundo a *óptica da intermediação*, desenvolvida a partir de finais dos anos setenta (Sealey e Lindley, 1977), a empresa bancária é vista como simples intermediária nos mercados financeiros, ou seja, o processo produtivo exige a recolha ou o empréstimo de fundos que serão, em seguida, objecto de aplicação ou empréstimo. Nesta óptica, além do trabalho e do capital também os depósitos são considerados como inputs. Consistente com esta abordagem, nos custos totais, são incluídos os custos operacionais ou de produção e os custos financeiros.

Estudos mais recentes, no seguimento de Pulley e Humphrey (1993), incluem os depósitos (o valor monetário dos depósitos, sobretudo os depósitos à ordem)

também como uma medida do output, tendo por base o grande peso que eles assumem no consumo dos recursos reais e o seu importante contributo para a geração do valor acrescentado dos bancos. Segundo estes autores, rubricas do passivo como os depósitos absorvem, da mesma forma que os créditos concedidos, recursos como capital e mão-de-obra e, portanto, deverão ser consideradas como produção (e, simultaneamente, como input). Esta abordagem, conhecida como a do *valor acrescentado*, é seguida, entre outros, por Mendes e Rebelo (1999) para o estudo do sector bancário português de 1990-95.

Uma abordagem mais ambiciosa seria incluir nos *outputs* algumas produções intangíveis, como o grau de satisfação dos clientes, a qualidade dos serviços e produtos, o grau de satisfação dos empregados. No entanto, dada a dificuldade de mensurar esses agregados, a generalidade dos estudos bancários não relevam essas variáveis (Resti, 2000, p.566).

Embora conceptualmente as duas abordagens originais ─ *produção e intermediação* ─ sejam muito diferentes, os resultados empíricos quanto a economias de escala e de gama parecem não ser significativamente afectados pelas diversas definições de inputs, outputs e custos (Clark, 1988, p. 24).

Já o mesmo não se verifica em relação ao cômputo da eficiência onde a escolha dos inputs e dos outputs parece trazer grandes alterações nos resultados (Barros e Pinho, 1995, p. 41).

Berger et al (1987, p. 511) sustentam que a escolha apropriada entre as duas aproximações depende da questão avançada:

> *However, for questions related to the economic viability of banks, the intermediation approach is preferred because it is more inclusive of the total cost banking(...) .*

Desta forma, parece-nos a abordagem intermediação mais apropriada para a caracterização da empresa bancária, dado o importante peso da actividade interbancária e dos custos financeiros na totalidade dos custos. A justificar esta escolha temos, ainda, uma razão de ordem empírica: no caso português, a informação quanto ao número de contas de crédito e de depósitos, nas diferentes empresas bancárias, não se encontra disponível, inviabilizando, à partida, a aplicação da abordagem produção (Mendes, 1994).

Finalmente, avancemos uma questão relacionada com a própria definição de custos totais. O conjunto dos custos operacionais e financeiros, que temos designado por custos totais, corresponderá, de facto, à totalidade dos custos da empresa bancária? Por outras palavras, além dos custos explícitos de produção (operacionais e financeiros) não serão de considerar os custos implícitos ou custos de oportunidade associados à aplicação alternativa dos recursos bancários? O que corresponde à introdução da problemática do risco. Estas questões vão ser tratadas, no final deste capítulo, após o estudo das funções custo.

I.2.2. A especificação da função custo

Antes de nos debruçarmos sobre o segundo grande debate teórico da literatura de economia bancária (que diz respeito, relembremos, à escolha da especificação da função custo mais adequada ao estudo da banca), vamos deter-nos um pouco em torno da dualidade entre as funções de custo e de produção.

I.2.2.1.Dualidade e natureza das funções custo

A função custo pode ser estabelecida a partir da função de produção sob certas condições, permitindo uma abordagem alternativa, via custos, das condições tecnológicas do processo produtivo.

A aplicação da teoria da dualidade, desenvolvida por R.W.Shephard, e, em particular, o *lema de Shephard*[16] permite concluir que, sob determinadas condições de regularidade, as funções custo e de produção são duais uma em relação à outra. Esta conclusão é particularmente importante na execução de estudos empíricos:

> *Applied researchers need no more longer begin their study of the firm with detailed knowledge of the technology and with access to relatively obscure data. Instead, they can concentrate on devising and estimating flexible functions of observable market prices and output and be assured that they are carrying along all economically relevant aspects of the underlying technology.*[17]

[16]Para mais desenvolvimentos ver, por exemplo, Fare e Primont (1997).
[17]Jehle, Geoffrey A. (1991), *Advanced Microeconomic Theory*, Prentice-Hall International Editions, p.238

Diewert (1992) demonstra que o recurso a funções custo em detrimento das funções produção tem a vantagem de simplificar o processo de estimação; além de que os parâmetros da função custo podem ser estimados de uma forma precisa, através do recurso a variadas técnicas da função custo. Binswanger (1974)[18] mostrou que as funções custo são homogéneas em relação aos preços, não sendo necessária a existência da homogeneidade de grau um na função de produção para o processo de estimação; por outro lado, o problema da existência de elevada multicoliniaridade entre as variáveis inputs que ocorre no processo de estimação das funções de produção não ocorre na estimação das funções custo (não existe, regra geral, uma alta multicolinearidade entre os preços dos inputs).

A função custo total de uma empresa multiproduto, como um banco, é função dos inputs totais, dos preços desses inputs e dos outputs. Considerando apenas dois factores de produção, trabalho e capital, e recorrendo à simbologia habitual, teremos

$$CT = p_L L + p_K K$$

Dividindo o custo total pelo vector dos outputs, obtemos o custo médio

$$\text{CMe} = \frac{CT}{Q}$$

Ou seja, o custo total pode ser dado por

[18] Referido por P.Molyneux et al.(1996), p.150

$$CT = CMe * Q = \frac{p_L L + p_K K}{Q} F(L,K)$$

Se se considerar o vector dos preços dos inputs constantes, as propriedades da função de produção serão satisfeitas pela correspondente função custo.

Se a função custo for diferenciável, as suas elasticidades em relação aos preços dos inputs medem o peso relativo ou quotas dos custos de cada um dos inputs (*shares*)[19]

$$S_i = \frac{p_i x_i}{CT} = \frac{\partial \ln CT}{\partial \ln P_i}, \text{ com } 1 \leq i \leq n$$

Se a função custo for homogénea de grau um em relação aos preços dos inputs, o somatório de S_i terá de ser igual a um.

Recorrendo à dualidade entre produção e custos, as economias de escala podem ser definidas como o somatório das elasticidades da função custo em relação ao nível de cada output

$$EE_i = \sum_i \frac{\partial \ln CT}{\partial \ln Q_i}, \text{ com } 1 \leq i \leq n$$

Se EE for igual a um, os custos de produção aumentam na mesma proporção que a produção da empresa (bancária), havendo economias constantes à escala; se EE for superior a um, os custos aumentam mais que proporcionalmente que os

[19] Utilizamos a simbologia *S* para seguir a terminologia anglo-saxónica de *cost shares*

outputs, pelo que haverá deseconomias de escala; se EE for inferior a um, teremos economias de escala.

Finalmente, da teoria da produção, decorrendo da dualidade entre produção e custos, a função custo apresenta as seguintes propriedades[20]: positividade (a função custo é positiva para valores positivos dos preços dos inputs e para níveis positivos de output); homogeneidade (a função custo é homogénea de grau um em relação aos preços dos inputs); monotonicidade (a função custo é crescente com os preços dos inputs e com o nível da produção); e concavidade (a função custo é côncava).

I.2.2.2. Formas funcionais para as empresas multiproduto

Duas especificações dominam a literatura bancária: a função Cobb-Douglas e a função Translog. A partir dos anos setenta, tem vindo a preterir-se a primeira (pelos problemas de tratamento dos rendimentos de escala variáveis em relação ao output e da complementaridade dos outputs, ou seja, do cômputo dos rendimentos de gama) em favor da especificação Translog (embora, também esta, com o problema da dificuldade da análise da hipótese da sub-aditividade).

Mas quais são as razões, *a priori*, que poderão fazer com que se escolha uma forma para uma função custo multiproduto (ou uma função de produção) em detrimento de outra? Por outras palavras, quais as propriedades desejáveis para uma função custo multiproduto?

Seguindo J. Lau (1986) podemos agrupar essas propriedades em cinco grupos: consistência teórica, domínio de aplicação, flexibilidade, facilidade de tratamento informático e conformidade com os dados empíricos.

A consistência teórica requer que a forma funcional para a função custo cumpra as propriedades apontadas por Jorgenson (1986), pelo menos localmente.

O critério do domínio de aplicação refere-se ao conjunto de valores das variáveis independentes para o qual são cumpridas as condições decorrentes da consistência teórica; da mesma forma que anteriormente, o domínio de aplicação pode não corresponder a todas as escolhas dos parâmetros, mas apenas para um pequeno conjunto de preços dos inputs estar assegurada a consistência teórica.

A flexibilidade da forma funcional da função custo refere-se à não exigência de quaisquer restrições nos valores na primeira e segunda derivadas parciais, ou seja, permite que os dados empíricos forneçam informação acerca do comportamento dos parâmetros.

A facilidade do recurso a tratamento informático das formas funcionais da função custo implica que: os parâmetros possam ser inferidos claramente dos dados (*linearidade nos parâmetros* seguindo Diewert, 1992); o número de parâmetros a estimar deve ser o menor possível (em muitos casos o número de observações é relativamente reduzido e da necessidade da estimação de um número elevado de parâmetros pode decorrer uma perda de graus de liberdade excessiva).

---

[20] Seguindo D.W. Jorgenson (1986), *Econometric Methods for Modelling Producer Beahaviour*, Handbook of Econometrics, pág.1841-915

A conformidade com os dados empíricos conhecidos tem de ser assegurada pela forma funcional da função custo escolhida.

Lau (1986) mostrou que, quer a forma generalizada de Leontief quer a Translog da função custo perdem a característica da flexibilidade se assegurarem a consistência teórica: apenas mantêm esta última característica sob fortes restrições dos parâmetros. Para uma empresa multiproduto, como um banco, uma forma funcional da função custo deve ser capaz de produzir estimativas aceitáveis para os vectores dos outputs, quando não é produzido um ou mais produtos; formas funcionais do tipo da Cobb-Douglas, da CES e da Translog não cumprem este critério. A forma da Translog híbrida, já testada empiricamente, parece cumprir razoavelmente os cinco grupos de critérios apontados. No entanto, estudos recentes (Altunbas et al., 2000; Humphrey e Vale, 2003) defendem a pouca flexibilidade da forma Translog, optando por formas mais flexíveis (no caso a Fourier ─ Altunbas et al.,2000─ ou a Fourier e a Spline ─ Humphrey e Vale, 2003).

Algumas das formas funcionais de funções produção/custo de que a literatura bancária mais se tem socorrido são: a Cobb-Douglas, a CES, a Translog, a Translog Híbrida e a Fourier.

A função Cobb-Douglas

A primeira formulação de uma função de produção foi a avançada em 1928 por Cobb-Douglas (1928) em que os autores faziam a sua aplicação para a indústria transformadora norte-americana, apresentando-se na seguinte formulação

$$Q = AL^{\alpha_1} K^{\alpha_2}$$

ou, logaritmizando

$$\ln Q = \ln A + \alpha_1 \ln L + \alpha_2 \ln K$$

em que Q representa a produção do período, A os inputs fixos, incluindo o efeito do progresso tecnológico, L e K os habituais factores de produção trabalho e capital. As elasticidades do output em relação aos factores de produção trabalho e capital são representadas, respectivamente, por α_1 e α_2. Com efeito,

$$\varepsilon_L = \frac{\partial Q}{\partial L} * \frac{L}{Q} = \frac{\partial \ln Q}{\partial \ln L} = \alpha_1 \qquad \text{e} \qquad \varepsilon_K = \frac{\partial Q}{\partial K} * \frac{K}{Q} = \frac{\partial \ln Q}{\partial \ln K} = \alpha_2$$

Os parâmetros A, α_1 e α_2 são positivos.

O grau de intensidade de utilização dos factores pode ser avaliada por

$$\frac{\varepsilon_K}{\varepsilon_L} = \frac{\alpha_2}{\alpha_1}$$

ou seja, se numa função Cobb-Douglas o rácio $\frac{\alpha_2}{\alpha_1}$ for maior do que noutra, dizemos que a primeira é mais capital intensiva do que a segunda.

A existência de economias de escala pode ser detectada pelo grau de homogeneidade, uma vez que a função Cobb-Douglas é uma função homogénea de grau $\alpha_1 + \alpha_2$ $\left[F(\lambda L, \lambda K) = \lambda^{\alpha_1+\alpha_2} F(L,K)\right]$; se $\alpha_1 + \alpha_2 > 1$ existem economias de escala, se $\alpha_1 + \alpha_2 < 1$ existem deseconomias de escala e se $\alpha_1 + \alpha_2 = 1$ estamos perante economias constantes à escala.

O grau de substituibilidade dos factores produtivos é constante e igual a um (restrição implícita da função Cobb-Douglas). Esta restrição empobrece a aplicabilidade empírica desta forma funcional.

A eficiência do processo produtivo é exprimida pelo valor estimado para o parâmetro A; se as elasticidades do output em relação aos factores produtivos, representadas, como vimos, por α_1 e α_2, forem iguais de uma função de produção para outra assim como as quantidades incorporadas dos inputs, então, um maior nível de produção só poderá ter origem na variação do parâmetro A.

Recorrendo à teoria da dualidade, a função custo total associada a uma função de produção incorporando dois factores de produção, trabalho e capital, será dada por

$$CT = wL + rK$$

onde w e r são os preços unitários dos factores de produção.

Para que sejam asseguradas as condições de minimização do custo para um dado output Q_0 ,estabelecemos o Lagrangeano

$$\Phi = wL + rK - \lambda(AL^{\alpha_1}K^{\alpha_2} - Q_0)$$

Derivando o Lagrangeano em ordem a L, K e a λ e igualando a zero, obtemos o seguinte sistema de equações

$$\begin{cases} \partial\Phi/\partial L = w - \lambda(\alpha_1 AL^{\alpha_1 - 1}K^{\alpha_2} = 0 \\ \partial\Phi/\partial K = r - \lambda(\alpha_2 AL^{\alpha_1}K^{\alpha_2 - 1}) = 0 \\ \partial\Phi/\partial\lambda = AL^{\alpha_1}K^{\alpha_2} - Q_0 = 0 \end{cases}$$

Resolvendo em ordem a L e a K, obtemos a quantidade dos inputs que deveremos incorporar no processo produtivo, para obter a produção Q_0 de forma a minimizar o custo total

$$L = \left[(\alpha_1 r/\alpha_2 w)^{\alpha_2/(\alpha_1+\alpha_2)}\right](Q_0/A)^{1/(\alpha_1+\alpha_2)}$$

e

$$K = \left[(\alpha_2 w/\alpha_1 r)^{\alpha_1/(\alpha_1+\alpha_2)}\right](Q_0/A)^{1/(\alpha_1+\alpha_2)}$$

Substituindo na equação dos custos totais obtém-se

$$CT = w^{\alpha_1/(\alpha_1+\alpha_2)} r^{\alpha_2/(\alpha_1+\alpha_2)} \left[\left(\frac{\alpha_2}{\alpha_1}\right)^{\alpha_1/(\alpha_1+\alpha_2)} + \left(\frac{\alpha_2}{\alpha_1}\right)^{-\alpha_2/(\alpha_1+\alpha_2)}\right]\left(\frac{Q}{A}\right)^{1/(\alpha_1+\alpha_2)}$$

ou, fazendo $\mu = \alpha_1 + \alpha_2$ e $\gamma = \mu\left[A\left(\alpha_1^{\alpha_1}\alpha_2^{\alpha_2}\right)\right]^{-1/\mu}$, teremos

$$CT = \gamma Q^{1/\mu} w^{\alpha_1/\mu} r^{\alpha_2/\mu}$$

ou, recorrendo à logaritmização, obtém-se a seguinte forma linear para a função custo Cobb-Douglas

$$\ln CT = \ln \gamma + (1/\mu)\ln Q + (\alpha_1/\mu)\ln w + (\alpha_2/\mu)\ln r$$

O parâmetro μ representa o grau de economias de escala. A soma dos expoentes dos preços unitários dos factores de produção, ou seja, $(\alpha_1/\mu + \alpha_2/\mu)$ deve ser igual a um (dada a exigência de homogeneidade de grau um dos preços dos factores de produção numa função Cobb-Douglas).

A formulação da função custo Cobb-Douglas pode ser generalizada para um processo produtivo incluindo n factores produtivos (Chajai, 1986), obtendo-se uma expressão para a função custo do tipo

$$CT = AQ^{1/\mu}\prod^{n} w_i^{\alpha_i/\mu}$$

ou, na forma logarítmica

$$\ln CT = \ln A + \frac{1}{\mu}\ln Q + \sum_{i=1}^{n}\frac{\alpha_i}{\mu}\ln w_i$$

onde w_i representa os preços do input i, α_i a elasticidade da produção em relação ao input i, e as outras variáveis com os significados habituais. Ou seja, a função custo logaritmizada é o desenvolvimento da série de ordem um, na vizinhança de zero, em relação às variáveis $\ln Q$ e $\ln w_i$; nesta forma funcional, não são considerados os termos da função de produção de grau igual ou superior a dois.

A formulação Cobb-Douglas da função custo apresenta algumas limitações quanto à sua adequabilidade empírica, nomeadamente: apenas comporta custos sempre crescentes, sempre decrescentes ou sempre constantes (deixa de parte todas as curvas de custo do tipo U); a elasticidade de substituição dos inputs é constante e igual a um; não permite avaliar a existência de economias de gama (a função é somente aplicável a empresas que produzem um único bem homogéneo, o que não traduz o cariz multiproduto da empresa bancária).

Apesar destas limitações, as funções Cobb-Douglas têm sido amplamente aplicadas na literatura bancária. O carácter simples da sua modelização não implica problemas no processo de estimação, embora a sua aplicação traga dificuldades no concernente à definição das variáveis.

A função CES

Uma segunda forma funcional da função de produção é a CES (constant elasticity of substitution) proposta por Arrow et al. (1961), no início dos anos sessenta. Esta formulação é menos restritiva nas suas hipóteses do que a de Cobb-Douglas uma vez que admite qualquer grau de substituição entre os inputs (constante, mas podendo ser diferente de um).

A função de produção CES apresenta a seguinte forma habitual

$$Q = A\left[\alpha L^{-\beta} + (1-\alpha)K^{-\beta}\right]^{-\frac{\nu}{\beta}}$$

sendo L, K dois inputs, A, α e β parâmetros positivos e ν representa o grau de homogeneidade. Se $\beta = 1$ a formulação anterior traduz uma função de produção do tipo Cobb-Douglas.

A correspondente função custo total é dada por:

$$CT = Q^{1/\nu} A^{-1/\nu} \left[\alpha^{1/(1+\beta)} p_1^{\beta/(1+\beta)} + (1-\alpha)^{1/(1+\beta)} p_2^{\beta/(1+\beta)}\right]^{(1+\beta)/\beta}$$

O grau de intensidade de utilização dos factores de produção pode ser medido por $\frac{\alpha}{1-\alpha}$ (quanto menor for este rácio maior a intensidade capitalista).

A existência de economias de escala analisa-se através do valor do parâmetro ν: se $\nu > 1$, existem economias de escala, se $\nu < 1$ estamos perante deseconomias de escala, se $\nu = 1$ não existem economias de escala.

A eficiência do processo produtivo é dada pelo valor assumido pelo parâmetro α.

O progresso tecnológico pode ser analisado da mesma forma que no caso da função Cobb-Douglas (Seijas, 1975, pp. 25-26).

O grau de substituibilidade dos factores de produção, medido pela elasticidade de substituição na função CES, é dado por $\sigma = 1/(1+\beta)$ [21], constante (e daí a designação desta forma funcional: *constant elasticity of substitution*) mas podendo diferir de um.

A função CES exige que a elasticidade de substituição entre cada par dos inputs seja igual, o que se afigura como uma hipótese demasiado restritiva (Varian, segunda edição, p.180). Esta forma funcional apresenta melhor aderência ao estudo de processos tecnológicos com um único output e dois inputs do que ao caso multiproduto[22], o que não é o caso da empresa bancária. Daí o desenvolvimento das chamadas formas funcionais *flexíveis*.

[21] Geoffrey A. Jehle (1991), *Advanced Microeconomic Theory*, Prentice-Hall International Editions, pág.223-4

[22] H.Uzawa (1962), *Production Functions With Constant Elasticities of Substitution,* Review of Economic Studies, pág. 291-99

A função Translog

Uma das formas funcionais flexíveis mais utilizadas na literatura bancária é a Translog (*transcendental logaritmic*) que representa a aproximação em série de Taylor de segunda ordem da função custo, habitualmente em torno do vector unitário ou média ─ estamos perante uma função genérica em que todas as variáveis aparecem em logaritmos. A Translog foi desenvolvida para o caso multiproduto por Diewert (1974).

Ao contrário da especificação Cobb-Douglas, o desenvolvimento da função custo logaritmizada considerando os termos de grau dois permite melhorar a qualidade do ajustamento, quer para uma indústria uniproduto quer para o caso multiproduto, acolhendo a consideração da forma U da função custo (e o cômputo das economias de escala em diferentes pontos) e o tratamento, ainda que parcial, das economias de gama.

A função de produção Translog apresenta a forma geral

$$\ln Q = \alpha_0 + \sum_{i=1}^{n} \alpha_i \ln X_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij} \ln X_i \ln X_j$$

onde $\alpha_{ij} = \alpha_{ji}$ para todos os i,j e X_i e n representa as quantidades dos inputs.

Decorrendo da teoria da dualidade, as propriedades da tecnologia acima descritas podem ser completamente caracterizadas por uma função custo que, na sua forma logaritmizada, pode ser apresentada como[23]

$$\ln CT = \alpha_0 + \sum_{i=1}^{m} \alpha_i \ln Q_i + \sum_{j=1}^{n} \beta_j \ln W_j + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \delta_{ij} \ln Q_i \ln Q_j$$

$$+ \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij} \ln W_i \ln W_j + \sum_{i=1}^{m} \sum_{j=1}^{n} \rho_{ij} \ln Q_i \ln W_j$$

em que CT representa o custo total, Q_i o nível do output i e W_i o preço do input i. Se os parâmetros δ_{ij}, γ_{ij} e ρ_{ij} forem nulos, a expressão reduz-se à forma funcional da Cobb-Douglas. Quando se pretende trabalhar com amostras em que coexistem empresas (bancos) de diferente dimensão e para tentar ultrapassar o problema da consideração da mesma tecnologia sem necessidade da partição em sub amostras, é usual a introdução de uma variável de estrutura que, no caso das empresas bancárias, se associa ao número de balcões (vector B), apresentando-se a função custo do tipo $CT = C(Q, W, B)$ [24].

A forma funcional Translog da função custo apresenta $m+1$ parâmetros independentes α_i , n parâmetros independentes β_j , $m(m+1)/2$ parâmetros dependentes δ_{ij} (uma vez que se assume que $\delta_{ij} = \delta_{ji}$ para i>1, m>j), $n(n+1)/2$

[23]W.E.Diewert (1974), para uma demonstração da passagem da forma funcional Translog da produção para a função custo.

[24]A correspondente forma funcional Translog pode ser consultada em Barros e Pinho (1995), pp.36-37

parâmetros independentes γ_{ij} (dado que se assume que $\gamma_{ij} = \gamma_{ji}$ para i>1, n>j) — constituindo estas duas últimas condições a propriedade de simetria — e nm parâmetros independentes ρ_{ij}.

Para que a Translog possa ser considerada uma função custo, além da propriedade de simetria terá de verificar a propriedade da homogeneidade de grau um dos preços dos factores produtivos, dada pelas seguintes restrições:

$$\begin{cases} \sum_{j=1}^{n} \beta_j = 1 \\ \sum_{j=1}^{n} \gamma_{ij} = 0\,, \quad \forall\, i \\ \sum_{i=1}^{n} \rho_{ij} = 0, \quad \forall\, j \end{cases}$$

A acrescer às propriedades da simetria (os parâmetros de segunda ordem dos inputs e dos outputs deverão ser simétricos) e da homogeneidade de grau um da função custo em relação aos preços dos factores de produção, Jorgerson (1986) adianta mais três condições para que a teoria da produção e a dos custos possam ser integradas: a de que os custos têm de ser exaustivos, ou seja, o valor dos i inputs tem de ser igual ao custo total; a da monotonicidade segundo a qual a função é crescente com o nível da produção e com os preços dos factores de produção; e a da não-negatividade que exige que as elasticidades da função custo em relação aos preços dos inputs sejam não negativas.

A diferenciação parcial da função custo em relação aos preços dos factores de produção dá-nos as equações "shares"[25] S_i, correspondentes às elasticidades da função custo em relação aos preços dos inputs:

$$S_i = \frac{\partial \ln CT}{\partial \ln W_i} = \frac{\partial CT}{\partial W_i} * \frac{W_i}{CT}, 1 < i < n$$

S_i é o "share" (quota) do i-ésimo factor de produção em termos dos custos totais.

No caso da forma funcional Translog as equações dos "shares" podem ser reescritas, atendendo ao *Lema de Shephard* como:

$$S_i = \beta_i + \sum_{j=1}^{n} \gamma_{ij} \ln W_j + \sum_{i=1}^{m} \rho_{ij} \ln Qi$$

Sendo n o número de factores de produção e dado a soma dos "shares" ter de ser um, apenas n-1 equações podem ser estatisticamente independentes.

A especificação Translog para empresas multiproduto permite estimar os parâmetros dos rendimentos de escala global e por produto e testar a hipótese da existência ou não de economias de gama associadas à produção conjunta

O *grau de economias de escala globais* é definido como

$$EEG = \frac{CT}{\sum_{i=1}^{n} Q_i CM_i} = \frac{1}{\sum_{i=1}^{n} \frac{\partial \ln CT}{\partial \ln Q_i}}$$

[25]Optámos por manter a designação anglo-saxónica no seguimento de trabalhos já publicados (ver, por exemplo, Barros e Pinho (1995), p.37)

Em que CT designa o custo total dos factores de produção, Q_i o output i e CM_i o custo marginal relativo ao output i.

No caso da Translog, o grau de economias de escala é dado por

$$EEG = \left[\sum_{i=1}^{m} (\alpha_i + \sum_{j=1}^{m} \delta_{ij} \ln Q_i + \sum_{l=1}^{n} \rho_{il} \ln W_{il}) \right]^{-1}$$

Consoante o valor assumido por EEG pode estar-se perante economias de escala (EEG>1), deseconomias de escala (EEG<1) e inexistência de economias à escala (EEG=1). Os parâmetros a estimar resultam da estimação da função Translog assegurando, simultaneamente, as restrições de homogeneidade e de simetria e o *Lema de Shephard* —- o método de estimação normalmente utilizado é o de Zellner[26].

Na prática, muitos autores[27] calculam EEG na vizinhança do ponto médio $q_i = 1$ e $w_i = 1$, sendo as variáveis em logaritmo centradas em relação à sua média, o que permite simplificar o cálculo de EEG:

$$EEG = \left(\sum_{i=1}^{m} \alpha_i \right)^{-1}$$

O *grau de economias de gama* (EG) é definido como

[26] Mendes, Victor (1991), *Scale and scope Economies in Portuguese Commercial Banking: the years 1965-88,* Economia, 15(3), pp. 457-8

[27] Cfr. Mohamed Sassenou (1992), *Economies des coûts dans les banques et les caisses d'épargne, impact de la taille et de la variété de produits,* Revue Économique, Mars

$$EG = \frac{\sum_i CT(Q_i, W)}{CT(Q, W)}$$

em que $CT(Q_i, W)$ representa o custo de produção relativa ao produto i, quando tomado separadamente e $CT(Q, W)$ designa o custo total dos factores relativos ao conjunto dos produtos.

Como já vimos, consoante EG<0, EG>0 ou EG=0, estaremos perante economias de gama crescentes, decrescentes ou constantes.

Na prática, é extremamente difícil fazer o cômputo do grau de economias de gama uma vez que a informação que permitiria imputar os custos de produção a cada produto não se encontra disponível. No entanto, *uma condição suficiente para a existência de economias de gama é a da complementaridade dos outputs*[28].

Dois outputs q_i e q_j são complementares se e só se $\partial CT / \partial q_i$ for decrescente com q_i, ou seja, se e só se, $\frac{\partial^2 CT}{\partial q_i \partial q_j} < 0$

O que equivale a

$$\frac{\partial^2 \ln CT}{\partial \ln q_i \partial \ln q_j} + \left(\frac{\partial \ln CT}{\partial \ln q_i} * \frac{\partial \ln CT}{\partial \ln q_j} \right) < 0$$

Aplicando à forma funcional Translog obtém-se

[28]W.Baumol et al(1988), sublinhado de M. Sassenou (1992).

$$\delta_{ij} + \left(\alpha_i + \sum_{j=1}^{m} \delta_{ij} \ln q_j + \sum_{j=1}^{n} \rho_{ij} \ln w_j \right) * \left(\alpha_j + \sum_{j=1}^{m} \delta_{ij} \ln q_j + \sum_{j=1}^{n} \rho_{ij} \ln w_i \right) < o$$

Tal como acontece com as economias de escala considera-se habitualmente como ponto de referência o ponto médio $q_i = 1$ e $w_i = 1$, donde resulta a condição simplificada de complementaridade

$$\delta_{ij} + \alpha_i \alpha_j < 0$$

Ou seja, a existirem economias de gama crescentes, a condição anterior verificar-se-á; se $\delta_{ij} + \alpha_i \alpha_j > 0$, então, estamos perante economias de gama decrescentes; se $\delta_{ij} + \alpha_i \alpha_j = 0$, ocorrerão economias de gama constantes.

A função Translog Híbrida

Uma crítica fundamental que se coloca a propósito da formulação Translog tradicional é a de que se um dos produtos não é produzido (o que numa empresa multiproduto é uma hipótese não negligenciável) — e uma vez que todos os outputs estão numa forma logarítmica - a função custo não terá uma representação finita. Neste caso, não poderá ser feito, por exemplo, o cômputo das economias de gama. Para ultrapassar este problema da produção nula, alguns autores sugeriram, no início dos anos oitenta, a Translog Híbrida que corresponde a uma generalização da

função Translog em que os diferentes logaritmos dos produtos $\left(\ln Q_i\right)$ são substituídos pela transformação Box-Cox[29]:

$$Q_i^* = \begin{cases} \dfrac{\left(Q_i^{\lambda} - 1\right)}{\lambda}, se \lambda \neq 0 \\ \ln Q_i, se \lambda = 0 \end{cases}$$

Esta formulação exige a estimação de um parâmetro de transformação, λ, para cada um dos produtos considerados.

Quando λ se aproxima de zero a função custo Translog Híbrida aproxima-se da especificação Translog. Empiricamente, esta parece ser a situação mais corrente[30].

A função custo Translog Híbrida apresenta-se da forma seguinte

$$\ln CT = \alpha_0 + \sum_{i=1}^{m} \alpha_i Q_i^* + \sum_{j=1}^{n} \beta_j \ln W_j + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \delta_{ij} Q_i^* Q_j^*$$
$$+ \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij} \ln W_i \ln W_j + \sum_{i=1}^{m} \sum_{j=1}^{n} \rho_{ij} Q_i^* \ln W_j$$

As condições de simetria (dadas por $\delta_{ij} = \delta_{ji}$ e $\gamma_{ij} = \gamma_{ji}$) bem como as restrições de homogeneidade da função custo em relação aos preços dos inputs ($\sum_{j=1}^{n} \beta_j = 1$, $\sum_{j=1}^{n} \gamma_{ij} = 0, \quad com\ 1 < i < n$ e $\sum_{i=1}^{n} \rho_{ij} = 0, \quad com\ 1 < j < m$) têm de ser

[29] G.Box, D.Cox (1964), *An analysis of transformations*, Journal of Royal Statistical Society, Series B, pp. 211-43

[30] Berger et al (1993), p.225

asseguradas, de forma idêntica à da função Translog tradicional. Na prática, estas restrições são, regra geral, impostas explicitamente no processo de estimação.

Esta forma funcional é bastante flexível[31]. Por exemplo, se $\gamma > 0$, o custo médio vectorial toma a forma de $\cup$ e se $\gamma < 0$ a curva do custo médio vectorial assume a forma de $\cap$.Os "shares" do custo total são explicitados da seguinte forma:

$$S_i = \frac{W_i X_i}{CT} = \beta_i + \sum_{j=1}^{n} \gamma_{ij} \ln W_j + \sum_{i=1}^{m} \rho_{ij} Q_i^*$$

onde X_i representa a quantidade do i-ésimo factor de produção.

A forma funcional Translog Híbrida tem sido sujeita a algumas críticas, nomeadamente, no referente à existência de um grande número de variáveis explicativas, de que pode resultar a existência de multicolinearidade entre elas. Apesar destas limitações, esta forma funcional continua a ser preferida por muitos autores.

A função Fourier

Alguns autores têm relevado as vantagens de se optar por uma especificação da função custo mais flexível do que as anteriores.

A função Fourier é uma forma funcional mista ou semi-não paramétrica, construída a partir de uma Translog completa e incluindo os termos trigonométricos de primeira, segunda e terceira ordem (Berger et al., 1997).

$$\ln CT = \alpha_0 + \sum_{i=1}^{m} \alpha_i \ln Q_i + \sum_{j=1}^{n} \beta_j \ln W_j + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \delta_{ij} \ln Q_i \ln Q_j +$$

$$+ \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij} \ln W_i \ln W_j + \sum_{i=1}^{m} \sum_{j=1}^{n} \rho_{ij} \ln Q_i \ln W_j + \sum_{i=1}^{m} (\sigma_i \cos z_i + \theta_i \sin z_i) +$$

$$+ \sum_{i=1}^{m} \sum_{j=1}^{m} \left[\sigma_{ij} \cos(z_i + z_j) + \theta_{ij} \sin(z_i + z_j) \right] +$$

$$+ \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=j}^{m} \left[\sigma_{ijk} \cos(z_i + z_j + z_k) + \theta_{ijk} \sin(z_i + z_j + z_k) \right]$$

em que CT representa o custo total, Q_i o nível do output i, W_i o preço do input i e z_i os valores "ajustados" de $\ln Q_i$ de forma a variarem no intervalo $[0;2\pi]$[32]. O cálculo de z_i é dado por $z_i = 0{,}2\pi - \mu \times a + \mu \times \ln Q_i$, onde $[a;b]$ é o intervalo de variação de $\ln Q_i$ e $\mu = \frac{0{,}9 \times 2\pi - 0{,}1 \times 2\pi}{b - a}$.

Carbot et al.(2003) Altunbas et al.(2000) e Altunbas et al.(2001) — seguindo Mitchell e Onvural (1996) — adoptaram, para a especificação da Fourier, os termos trigonométricos de primeira e de segunda ordem.

[31] Baumol et al.(1988) citado por Molyneux et al(1996), p.165

[32] Berger et al.(1997) fazem variar os z_i apenas no intervalo $[0{,}1.2\pi;\ 0{,}9.2\pi]$ para evitar problemas de aproximação junto dos extremos – vamos seguir esta sugestão na formulação do cálculo de z_i. De igual forma os autores somam uma unidade aos outputs para evitar problemas no cálculo dos logaritmos.

Os autores baseiam-se nos trabalhos de Gallant (1981, 1982) segundo o qual se consegue obter uma boa aproximação da verdadeira função custo, recorrendo a um número limitado de termos trigonométricos[33]. De facto, foi mostrado por Tolstov (1962) que uma combinação linear das funções seno e coseno (séries Fourier) pode ajustar exactamente o comportamento de uma função multivariável (uma vez que as funções seno e coseno são mutuamente ortogonais no intervalo $[0; 2\pi]$).

No caso da forma funcional Fourier, como resultante da Translog, as equações dos "shares" podem ser explicitadas, atendendo ao L*ema de Shephard* , de forma idêntica[34]:

$$S_i = \beta_i + \sum_{j=1}^{n} \gamma_{ij} \ln W_j + \sum_{i=1}^{m} \rho_{ij} \ln Qi$$

De igual forma, sendo n o número de factores de produção e dado que a soma dos "shares" tem de ser um, apenas n-1 equações podem ser estatisticamente independentes.

Uma vez que o teorema da dualidade exige que a função custo seja linearmente homogénea em relação aos preços dos inputs, os parâmetros têm de obedecer às seguintes restrições:

[33] Altunbas et al.(2001), p.1938.

[34] Alguns estudos estimam a função Fourier não considerando as equações de "shares"(cfr. Altunbas et al., 2000) apenas por uma questão de simplicidade, não havendo nenhum suporte teórico para este procedimento. Ver, no entanto, a aplicação de Dietsche e Lozano-Vivas (2000) — p. 991, nota de rodapé 2.

$$\begin{cases} \sum_{j=1}^{n} \beta_j = 1 \\ \sum_{j=1}^{n} \gamma_{ij} = 0, \quad \forall\, i \\ \sum_{i=1}^{n} \rho_{ij} = 0, \quad \forall\, j \end{cases}$$

Igualmente, como na função Translog, as condições de simetria dos parâmetros têm de estar asseguradas na Fourier ($\delta_{ij} = \delta_{ji}$ para i>1, m>j e $\gamma_{ij} = \gamma_{ji}$ para i>1, n>j).

A especificação Translog, embora sendo a forma funcional mais utilizada nos estudos empíricos, apresenta limitações que se prendem com a tipificação dos custos médios de longo prazo, através de uma curva quadrática com a forma de U, sugerindo que apenas os pequenos bancos ou os de dimensão média poderão apresentar economias de escala[35]. A função Fourier permite a existência de mais de um ponto de inflexão para a curva de custo médio (Humphrey e Vale, 2003, chegaram a representações mais próximas de M do que de U para os custos médios bancários, com base na análise de dados em painel de 130 bancos noruegueses, para o período de 1987-1998).

Berger et al.(1997) defendem que a função tipo Fourier é uma aproximação preferível à função custo Translog, quando se estuda uma indústria multiproduto, como é o caso da bancária, uma vez que apresenta uma maior adaptação aos dados:

[35] Humphrey, David B. e Bent Vale (2003), p.3

> *The Fourier-flexible for is a global approximation because the terms such* $\cos z_i$, $\sin z_i$, $\cos 2z_i$, $\sin 2z_i$, *are mutually orthogonal over the* $[0; 2\pi]$ *interval, so that each additional term can make the approximating function closer to the true path of the data wherever it is most needed.*[36]

De igual forma, Altunbas et al.(2000) sustentam que a função Fourier tem propriedades matemáticas e estatísticas desejáveis, uma vez que é capaz de representar qualquer função com exactidão (mesmo truncada, a forma Fourier pode representar razoavelmente qualquer função). Ao contrário da forma funcional Translog, em que se assume que a função custo da indústria bancária tem, obrigatoriamente, a forma em U (do que, a não ser verdade, decorrerão erros na especificação), a formulação Fourier permite que os dados determinem a forma da função custo.

No entanto, num estudo posterior, Altunbas e Chakravarty (2001) apresentam algumas reservas à aplicabilidade crescente da Fourier nos estudos de economia bancária:

> *The underlying critique (…) rest on the proposition that the mechanical approaches above are unlikely to provide economic insight into the banking system.*[37]

A acrescer a esta suspeição, os autores chamam a atenção para o facto de o melhor ajustamento aos dados não ser sinónimo de melhor capacidade de previsão (afirmação que é aplicável a qualquer processo de estimação).

[36] Berger et al.(1997), p.147

Um outro problema que se pode colocar com a aplicação da forma funcional tipo Fourier é a dificuldade (previsivelmente acrescida em relação aos resultados obtidos pela Translog) na interpretação dos resultados, uma vez que a existência de um número tão elevado de variáveis implica um elevado grau de multicolinearidade.

[37] Altunbas e Chakravarty (2001), p.239, nota de rodapé.

I.3.A estimação da eficiência no sector bancário

I.3.1. Os conceitos de eficiência

Muitos dos estudos de economia bancária têm-se debruçado sobre a problemática dos custos decorrentes de economias de escala e de gama.

Há, no entanto, outros aspectos em torno das condições de produção de empresas multiproduto, como são as instituições de crédito, que necessitam de ser estudados: estamos a referir-nos, em concreto, à (in)eficiência-X. O conceito de eficiência-X foi introduzido por H.Leibenstein (1966) com o objectivo de analisar a eficácia do funcionamento no seio das empresas. A teoria da empresa tradicional (de origem neo-clássica) pressupõe que cada unidade produtiva optimize o seu comportamento em relação quer aos factores de produção quer aos produtos; se o não fizer, o processo concorrencial fará com que as empresas menos eficientes sejam banidas do mercado. Leinbenstein afirma que a existência de ineficiência-X é a regra, ou seja, as empresas que operam segundo a minimização dos custos (sobre a curva fronteira) são excepções, constituindo as barreiras naturais à entrada ou processos regulamentares, as condições permissivas da manutenção de empresas ineficientes no mercado. A consideração da existência de ineficiência custo constitui, portanto, um ponto de clivagem com a teoria neo-clássica da empresa, em que se assume a eficiência dos produtores (Canhoto, 1999, p.29).

Diferenças de comportamento de gestão no controlo dos custos (ou de maximização dos rendimentos), imperfeições do mercado ou imposições

regulamentares parecem, no seu conjunto, ser relativamente mais importantes para o estudo da eficiência do que a escolha da escala ou da gama. Berger et al. (1993)[38], na sua revisão dos estudos sobre eficiência, afirmam que a ineficiência-X é responsável por mais de 20% da totalidade dos custos bancários, enquanto que as ineficiências ligadas à escala e à gama não ultrapassavam os 5% dos custos.

Alguns estudos[39] segmentam a eficiência-X ou custo em eficiência técnica (resultante de desperdício puro ou da subutilização dos recursos) e eficiência preço ou de afectação (uso dos factores de produção nas proporções incorrectas, dados os seus preços relativos). A eficiência técnica é uma noção física resultante da comparação entre a combinação das produções e dos factores produtivos incorporados em relação à melhor combinação tecnológica possível. Já o conceito de eficiência na afectação implica o conhecimento das condições de mercado (preços dos inputs e preços dos outputs) e dos objectivos económicos da própria empresa.

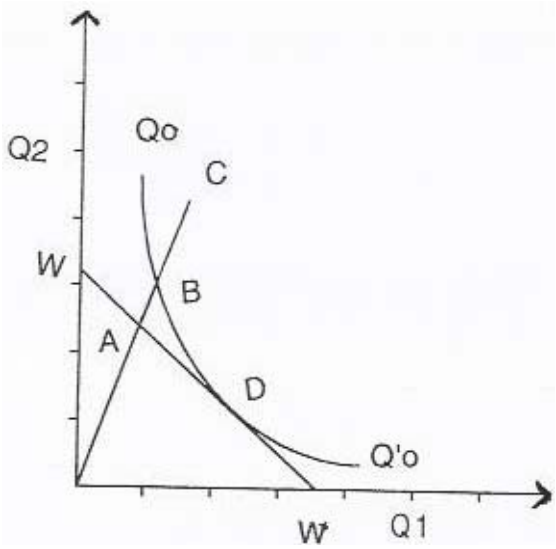
Vamos admitir uma empresa que produz um único produto, Q, recorrendo a dois factores de produção X_1 e X_2[40]. Vamos supor rendimentos constantes à escala (figura 6).

[38]A.N.Berger, W.C.Hunter e S.G.Timme (1993), *The efficiency of financial institutions: a review of preview of research past, present and future,* Journal of Banking and Finance, vol. 17, #2/.3, pág.222

[39]Referido em Mendes (1994), *Eficiência produtiva no sector bancário: uma aplicação do método DEA aos anos 1990-92,* Investigação, FEP, 42

[40] Neste exemplo seguimos de perto Mendes (1994), pp.4-5

Figura 6 - Eficiência-X, técnica e de afectação



Fonte: Adaptado de V.Mendes, *Eficiência produtiva no sector bancário: uma aplicação do método DEA aos anos 1990-92,* Investigação, FEP, 42, 1994, p.4

A isoquanta $Q_0Q_0^{'}$ e a área a sombreado representam as combinações possíveis dos factores de produção que permitem a produção de, pelo menos, Q_0.Dada a tecnologia disponível e os preços dos factores de produção ($\overline{w_1w_2}$), uma empresa para produzir ao mínimo custo deverá situar-se sobre D. Se se localizar em C, para produzir o nível de produto Q_0, não é eficiente. A eficiência-X (EX), ou eficiência produtiva global pode ser dada pelo rácio sugerido por Farrel (1957)

$$EX = \frac{OA}{OC}$$

Este indicador pode ser decomposto em eficiência técnica (ET), dada por

$$ET = \frac{OB}{OC}$$

e eficiência de afectação (EA), dada por

$$EA = \frac{OA}{OB}.$$

O decréscimo potencial dos custos, resultante da eliminação da ineficiência-X, pode ser mensurado por $(1 - EX)$.

Há autores (Evanoff e Israilevich, 1991; Berger et al., 1997; Rhoades ,1998; Altubas et al., 2000) que mencionam, ainda, a ineficiência de escala que resulta da escolha não óptima da escala de produção, ou seja, da existência de rendimentos não constantes à escala. O conceito de ineficiência à escala é diverso do de economias de escala. As economias de escala globais (EEG) são mensuradas como o recíproco da elasticidade custo, ou seja, calculáveis — assumindo-se, como muitas vezes na prática acontece e como já referimos, o cálculo das economias de escala globais na vizinhança do ponto médio q_i e w_i, sendo as variáveis em logaritmo centradas em relação à sua média — como $EEG = \left(\sum_{i=1}^{m} \alpha_i \right)^{-1}$; ou seja, as economias de escala estão associadas a alterações incrementais da produção. Já o conceito de ineficiência à escala mensura a alteração na produção que é necessária para produzir na escala mínima eficiente. Evanoff e Israilevich (1991) sugerem, como medida aproximada

para o cômputo da (in)eficiência à escala, a alteração percentual nos custos ponderados pelos activos[41], dada por:

$$IE = \left[\frac{C\hat{T}_j}{ACTT_j} - \frac{C\hat{T}_{ef}}{ACTT_{ef}} \right] / \left[\frac{C\hat{T}_j}{ACTT_j} \right]$$

em que IE representa a ineficiência à escala; $C\hat{T}_j$ representa a curva custo fronteira estimada para a média dos bancos de dimensão j; $ACTT_j$ activos totais para a média dos bancos de dimensão j; $C\hat{T}_{ef}$ representa a curva custo fronteira estimada para a média dos bancos na escala mínima eficiente e $ACTT_{ef}$ representa os activos totais para a média dos bancos na escala mínima eficiente. O indicador IE assumirá o valor nulo para os bancos que se situam na escala ef e será positivo para os outros casos (quanto maior IE maior a ineficiência à escala).

Num estudo recente de Altunbas et al (2000) sobre a banca japonesa, tendo por base os dados dos anos de 1993 a 1996, foi analisado o impacto do risco (medido pelo capital financeiro e por um rácio de liquidez) e da qualidade (mensurado pelas provisões para créditos vencidos) sobre as economias de escala e a ineficiência-X. Os autores chegaram à conclusão de que a dimensão óptima dos bancos era consideravelmente menor se se incluíssem na curva estocástica[42] de custo os factores de risco e de qualidade mencionados.

[41] Altunbas et al (2000) recorreram a esta formulação, tendo segmentado a amostra dos bancos por classes de dimensão j, segundo o montante dos activos totais.
[42] Os autores recorreram à forma flexível da função custo do tipo Fourier.

Contrariando as conclusões de Berger et al.(1993), Altunbas et al (2000), para o caso específico do Japão, obtiveram como resultado que as ineficiências ligadas à escala superavam a ineficiência - X.

Rhodes (1998, p. 282) define a eficiência total como a mais completa medida da eficiência e resulta do produto da eficiência-X e da eficiência à escala.

I.3.2. Aproximações paramétricas e não paramétricas

O processo de estimação baseado nas formas funcionais mencionadas em *I.2.2.2.* não permite uma separação directa entre a ineficiência decorrente da escala e a ineficiência-X. Por outro lado, o maior problema para o cômputo da ineficiência é conseguir separar a efectiva ineficiência de outros factores aleatórios que afectem o comportamento dos custos.[43]

Para tentar resolver esta dificuldade no cômputo da eficiência-X, foram utilizados na literatura bancária diferentes métodos.

O método SFA (*stochastic frontier approach*) avança que os custos observados num banco podem desviar-se da curva fronteira eficiente devido quer à ineficiência quer a flutuações aleatórias; para separar estas duas parcelas assume-se que o termo da ineficiência segue, habitualmente, uma distribuição assimétrica semi-

[43] No nosso trabalho vamos trabalhar exclusivamente com funções custo e não com funções lucro. Apesar das vantagens que o recurso a funções lucro pode trazer, em particular para a análise das fusões e aquisições (Berger, 1995), a especificação da função lucro exige que se conheçam os preços dos inputs e dos outputs.

normal e os erros aleatórios seguem uma distribuição simétrica normal (Aigner et al.,1977). Pode ser aplicada a dados *cross section* ou a dados em painel.

O método TFA (*thick frontier approach)* defende que os desvios dos custos previstos e o mais baixo quartil de custos, numa determinada classe, representam o erro aleatório, enquanto que a diferença nos custos previstos entre o quartil mais baixo e o mais elevado está associado à ineficiência-X (Berger e Humphrey, 1991, 1992). É uma abordagem menos estruturada que as abordagens convencionais, gerando menor informação, podendo ser aplicada a dados *cross section* ou a dados em painel.

O método DFA (*distribution - free approach)* baseia-se na hipótese de que a eficiência é estável ao longo do tempo, enquanto que os erros aleatórios tendem a compensar-se mutuamente, ao longo do tempo (Berger, 1993; Dietsch e Lozano-Vivas, 2000). É baseado num sistema translog de custos e equações "share"dos inputs, e estima a ineficiência custo para cada produtor por um período de tempo (exige a disponibilidade de dados em painel).

O método DEA (*data envelopment analysis)*, sendo um método determinístico, assume que todos os desvios entre os custos observados e a curva fronteira, de custos mínimos, são explicados pela ineficiência (Farrel, 1957).

Outros métodos decorrentes da aproximação estocástica são analisados por Berger, Hunter e Timme (1993). O que é necessário reter, segundo estes autores, é que a opção por cada um destes métodos de estimação parece ter implicações importantes ao nível dos resultados (Berger, Hunter e Timme, 1993, pág.228).

Schure et al.(2004) recorrem a uma função Cobb-Douglas para estimar a eficiência na banca europeia, após 1992, baseando-se no método RTFA (*recursive thick frontier approach)* desenvolvido por Wagenvoort et al.(1999), assentando na estimação recursiva dos estimadores dos mínimos quadrados (OLS) de subclasses da base de dados.[44]

Também Resti (2000) faz uma comparação entre os métodos "clássicos" da estimação da eficiência bancária (recurso a uma função custo Translog para a a aproximação da fronteira estocástica ─ SFA[45]─, e recurso ao Data Envelopment Analysis ─ DEA, método não paramétrico) e os métodos "novos" (DEA estocástica e DEA estocástica multiplicativa), tendo concluído que os métodos estocásticos "tradicionais" têm dificuldade em segmentar a ineficiência –X (em técnica e de afectação) e os novos métodos DEA estocásticos parecem conseguir alguma superioridade na capacidade de análise dos dados.[46]

[44] Schure et al.(2004) apresentam , pormenorizadamente o algoritmo de RTFA. Este algoritmo RTFA já tinha sido aplicado por Schure e Wagenvoort (1999) ao conjunto dos sistemas bancários da União Europeia entre 1993-97, tendo os autores chegado à conclusão da existência de ineficiência-X em praticamente todos os países ─ apenas a banca holandesa e a da Grã-Bretanha estavam perto da fronteira eficiente. Portugal, por exemplo, apresentava níveis de ineficiência superiores a 50% na banca comercial, percentagem só ultrapassada pelo sector bancário grego.

[45] Para uma análise das origens e desenvolvimentos da análise da eficiência recorrendo à fronteira estocástica (SFA) ver Kumbhakar e Lovell (2003), pp.8-11.

[46] Para mais desenvolvimentos ver Resti, Andrea (2000), "Efficiency measuremente for multi-product industries: A comparison of classic and recent techniques based on simulated data", European Journal of Operational Research, 121, pp. 559-578.

Vão ser referidos em pormenor os dois métodos que são usados, mais recorrentemente, nos estudos de economia bancária para a estimação das funções custo fronteira: os métodos estocásticos SFA (*stochastic frontier approach*) e o método DEA (*Data Envelopment Analysis*) que recorre a técnicas de programação linear.

I.3.2.1. Os modelos estocásticos SFA

A análise da eficiência das empresas bancárias por via da estimação paramétrica tem sido, cada vez mais, objecto de numerosos estudos. A grande maioria dos trabalhos debruça-se sobre a mensuração da ineficiência-X sem a segmentar em técnica e de afectação.

A estimação da ineficiência-X agregada

Nesta aproximação, a função custo tem três parcelas: a primeira correspondente aos custos dos produtores mais eficientes, a segunda derivada da ineficiência-X (montante dos custos associados ao desperdício de recursos) e a terceira decorrente dos efeitos aleatórios não controlados pela empresa.

$$CT = CT(Q_i, W_i, B_i) + u_i + v_i$$

A $CT(Q_i, W_i, B_i)$ designa-se por curva custo fronteira ou fronteira estocástica ─ ou seja, são estimados os custos necessários para atingir um

determinado nível de produção sem desperdícios de recursos. A curva fronteira representa a relação entre custos, níveis de produções e preços dos factores produtivos dos bancos relativamente mais eficientes, ou seja, os de "melhor prática". Se $v_i = 0$, o modelo da fronteira estocástica (SFA) coincide com o modelo da fronteira determinística.

A u_i impõe-se que seja estritamente positivo, uma vez que a ineficiência aumenta os custos; habitualmente assume-se que u_i segue uma distribuição semi-normal. A v_i associa-se uma distribuição normal de média zero e com desvio-padrão de σ_v. Assume-se, ainda, que o erro da função custo, ε, é dado por $\varepsilon = u + v$ e u e v são variáveis aleatórias independentes.

A sua função densidade de probabilidade será dada por

$$g(\varepsilon) = \frac{2}{\sigma} f(\frac{\varepsilon}{\sigma})[1 - F(\varepsilon\lambda / \sigma)]$$

sendo $\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}$, $\lambda = \frac{\sigma_u}{\sigma_v}$, f(.) é a função densidade da normal reduzida e F(.) é a função distribuição da normal reduzida.

Jondrow et al. (1982)[47] mostraram que o rácio da variação das variáveis ν e υ, seja λ, pode ser tomada como uma medida da ineficiência relativa de um banco:

$$\lambda = \frac{\sigma_u}{\sigma_v}$$

A estimação da ineficiência de um banco, dada pelo parâmetro λ, pode ser feita de uma forma directa, através da maximização da função de log-verosimilhança ($\ln\varphi$), decorrente de $g(\varepsilon)$ - estimação pelo método da máxima verosimilhança:

$$\ln\varphi = \frac{N}{2}\ln\frac{2}{\pi} - N\ln\sigma + \sum_{i=1}^{N}\ln\left[F\left(\varepsilon_i\lambda/\sigma\right)\right] - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\varepsilon_i^2$$

em que N corresponde ao número de observações, $\lambda = \sigma_u/\sigma_v$, $\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}$, $\varepsilon_i = u_i + v_i$ e F(.) é a função distribuição da normal reduzida.

Os resíduos estimados ε_i podem ser decompostos nas suas parcelas, através do método apresentado por Jondrow et al. (1982), segundo o qual a esperança matemática da distribuição condicionada para o modelo semi-normal é dada por

$$E(u_i\left|\varepsilon_i\right.) = \frac{\sigma_u\sigma_v}{\sigma}\left[\frac{f(\varepsilon_i\lambda/\sigma)}{F\left(\varepsilon_i\lambda/\sigma\right)} + \frac{\varepsilon_i\lambda}{\sigma}\right]$$

$E(u_i\left|\varepsilon_i\right.)$ é um estimador cêntrico, mas não consistente do parâmetro u_i, porque, independentemente de N, a variância do estimador é positiva[48].

A eficiência produtiva global (eficiência-X), definida como o rácio entre o mínimo e o actual custo, é dada por:

$$EP - X = \frac{CT}{CT + \left(U + V\right)} = \frac{CT(\ln q_i, \ln w_i)}{CT_i} = e^{-u_i - v_i}$$

[47]J.Jondrow, C.A.Lovell, I.S.Materov e P.Schmidt (1982), *On estimation of technical inefficiency in the stochastic frontier production model,* Journal of Econometrics, 19, pág.233-38

onde q_i e w_i são vectores incluindo os logaritmos dos outputs e os preços dos inputs do i-ésimo produtor. Ignorando o termo de perturbação estocástico v_i , o indicador da eficiência produtiva global obter-se-á por:

$$\text{EP-X} = e^{-u}$$

A estimação da ineficiência-X através do modelo semi-normal e considerando a ineficiência na sua totalidade (ou seja, não fazendo a distinção entre ineficiência técnica e de afectação) tem sido a opção de vários trabalhos (Allen e Rai, 1993, Yuenger, 1993, Mester, 1993, Noulas, Miller e Ray, 1994, Kaparakis, Miller e Noulas,1994 [49]).

Por exemplo, se se considerar a especificação da função custo Translog, dois outputs (Q_1 e Q_2), três inputs (cujos preços são dados por W_1, W_2 e W_3), uma variável de estrutura (B) ─ habitualmente associada ao número de balcões, que pretende captar alterações tecnológicas ─ teremos:

$$\ln CT = \alpha_0 + \sum_{i=1}^{2} \alpha_i \ln Q_i + \sum_{j=1}^{3} \beta_j \ln W_j + \frac{1}{2} \sum_{i=1}^{2} \sum_{j=1}^{2} \delta_{ij} \ln Q_i \ln Q_j$$

$$+ \frac{1}{2} \sum_{i=1}^{3} \sum_{j=1}^{3} \gamma_{ij} \ln W_i \ln W_j + \frac{1}{2} \lambda_{bb} \ln B \ln B + \sum_{i=1}^{3} \sum_{j=}^{2} \rho_{ij} \ln W_i \ln Q_j$$

$$+ \sum_{i=1}^{2} \lambda_{bi} \ln B \ln Q_i + \sum_{i=1}^{3} \tau_{bi} \ln B \ln W_i + \varepsilon$$

[48] Greene, W.M. (1993), *The econometric approach to efficiency analysis*, Oxford University Press, pp.80-2

onde $\varepsilon = u + v$ é o erro estocástico, definido da forma habitual, e $\alpha, \beta, \lambda, \delta, \gamma, \rho, \tau$ são os coeficientes a estimar. O cômputo da eficiência global (EP-X) será obtido, decompondo os resíduos estimados ε_i nas suas componentes através do método desenvolvido por Jondrow et al.(1982), obtendo-se, como já vimos, EP-X = e^{-u_i}.

Para além da especificação semi-normal para o parâmetro u ($u_i = N^+(0, \sigma_u^2)$ que é, como referimos, a habitualmente aceite), outras abordagens foram objecto de estudo. Podemos referir o estudo de Stevenson (1980)[50]. Este autor considerou que u seguia uma distribuição normal truncada com média μ_i e variância σ_u^2 ($u_i = N^+(\mu, \sigma_u^2)$ — exigindo-se, portanto, a estimação de mais um parâmetro); o modelo normal truncada pode ser obtido substituindo $\varepsilon_i \lambda / \sigma$, na expressão de $E(u_i | \varepsilon_i)$ por

$$\mu^* = \frac{\varepsilon_i \lambda}{\sigma} + \frac{\mu}{\sigma \lambda}$$

onde μ representa a média da distribuição normal não truncada.

Outras especificações foram sugeridas como o modelo normal-exponencial — em que u segue uma distribuição exponencial — ou o modelo normal-gama — em que u segue uma distribuição gama (Greene, 1990; Stevenson, 1980).

---

[49] Cfr P.Molyneux, Y.Altunbas e E.Gardener (1996), p.254
[50] Cfr P.Molyneux, Y.Altunbas e E.Gardener (1996), p.255

A sensibilidade dos resultados estimados para a ineficiência aos diferentes modelos é comummente aceite[51].

No entanto, a ordenação dos produtores pela sua ineficiência individual (dos produtores ou por classes) não parece ser sensivelmente afectada pelo modelo adoptado. De facto, parece haver uma correlação muito elevada entre os resultados obtidos, assumindo diferentes distribuições para u[52], o que permite, nestas circunstâncias, optar por uma distribuição simples para u (como a semi-normal ou a exponencial) em vez de distribuições mais complexas.

Neste sentido, Altunbas et al. (2000) e Altunbas et al. (2001) recorrem a uma função Fourier, com dados em painel[53] (efeitos aleatórios), para estimar a eficiência nos mercados bancários japonês (estudo de 2000) e europeu (estudo de 2001). Do ponto de vista das estimações empíricas, a adopção de diferentes funções densidade de probabilidade para o termo u_i, como semi-normal, normal truncada, exponencial ou gama interfere com os resultados individuais das estimações, como vimos. O recurso a dados em painel diminui estes problemas, dado que permite a estimação das componentes *cross-section* e temporais de ε (Schmidt e Sickles, 1984[54]).

[51] Kumbhakar, Subal C. e C.A. Knox Lovell (2003), *Stochastic Frontier Analysis*, Cambridge University Press, p.90

[52] Kumbhakar e Lovell (2003), a propósito da estimação de uma curva fronteira custo proposta por Greene(1990) recorrendo a diferentes distribuições para u , chegaram a coeficientes de correlação dos resultados estimados de 0,7467 (entre a exponencial e a gama) a 0,9803 (entre a normal truncada e a semi-normal).

[53] Os dados *cross-section* permitem a estimação da eficiência do conjunto dos produtores (conjunto dos bancos), enquanto que os dados em painel proporcionam a análise da eficiência de cada banco ao longo de uma sequência temporal.

[54] Citados por Mendes, Victor e João Rebelo (1999), *Productivity efficiency, technological change and productivity in Portuguese banking,* Applied Financial Economics, 9, pp.515

Altunbas et al. (2000) defendem a diferenciação entre economias de escala globais ─ EEG ─ e eficiência à escala ─ IE ─(como deixámos explicitado em *I.3.1. Os conceitos de eficiência)*. Por outro lado, o facto de se considerar a eficiência como constante ao longo do tempo entra em ruptura com a incorporação acrescida, nas últimas décadas, do progresso tecnológico na banca. Daí que os autores considerem como variável independente o tempo (variável T), incorporando os efeitos da alteração nos factores tecnológicos. Em Altunbas et al.(2000) T capta a alteração tão só do que se pode designar por *progresso tecnológico puro*. Já em Altunbas et al.(2001) e Carbot et al.(2003) o estudo das alterações tecnológicas aparecem segmentadas em três parcelas : a que mede o *progresso tecnológico puro*, a que mede a alteração do nível tecnológico decorrente da *alteração da escala da produção* e a que mede a alteração tecnológica *não neutral* (que é uma medida da alteração tecnológica associada às alterações nas quantidades dos factores produtivos decorrente de variações dos seus preços). Assim, recorrendo à especificação Fourier teremos:

$$\ln CT = \alpha_0 + \sum_{i=1}^{m} \alpha_i \ln Q_i + \sum_{j=1}^{n} \beta_j \ln W_j + \tau T + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \delta_{ij} \ln Q_i \ln Q_j +$$

$$+ \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij} \ln W_i \ln W_j + \frac{1}{2} \tau_l T^2 + \sum_{i=1}^{m} \sum_{j=1}^{n} \rho_{ij} \ln Q_i \ln W_j + \sum_{i=1}^{m} \psi_i T \ln Q_i + \sum_{j=1}^{n} \theta T \ln W_j +$$

$$+ \sum_{i=1}^{m} (a_i \cos z_i + b_i \sin z_i) + + \sum_{i=1}^{m} \sum_{j=1}^{m} \left[a_{ij} \cos(z_i + z_j) + b_{ij} \sin(z_i + z_j) \right] + \varepsilon$$

em que CT representa o custo total, Q_i o nível do output i, W_i o preço do input i, T a variável tempo, z_i os valores "ajustados" de $\ln Q_i$ e ε o termo de perturbação

compósito. Então, o progresso tecnológico ─ PT ─ pode ser medido como a diminuição dos custos totais ao longo do tempo, produzindo um dado nível de output, supondo os preços dos inputs constantes, ou seja, como

$$PT = \frac{\partial \ln CT}{\partial T} = (\tau + \tau_l T) + \sum_{j=1}^{n} \theta \ln W_j + \sum_{i=1}^{m} \psi \ \ln Q_i$$

Se PT >0 estamos perante a existência de regressão tecnológica; se PT<0 haverá progresso tecnológico (uma vez que estamos a derivar uma função , logaritmizada, de custo). [55]. Se θ for negativo, para um determinado factor, o peso do factor, no total dos custos, decresce ao longo do tempo. Se ψ for negativo, para todo o i, a escala da produção que minimiza o custo médio, para uma dada combinação de produções, cresce ao longo do tempo.

Altunbas et al.(2001) concluíram (para uma amostra de bancos europeus, entre 1989 e 1997) pela existência de economias de escala entre 5% e 7% e de ineficiência –X entre 20% e 25%, variável consoante os países, a dimensão dos bancos e ao longo do período temporal. O progresso tecnológico foi responsável pela diminuição dos custos em cerca de 3% ao ano, aumentando com a dimensão do banco ─ o que sugere que os bancos de maior dimensão, embora sem vantagens comparativas quanto à existência de economias de escala, conseguem captar comparativamente mais vantagens com o progresso tecnológico do que os bancos de menor dimensão.

[55] Também Mendes e Rebelo (1999) recorreram a este indicador para estudar o sector bancário português de 1990-95.

Mendes e Rebelo (1999) concluíram *a contrario* para o caso português (1990-95) período em que os dados apontam para uma regressão tecnológica: o aumento anual de 6% nos custos parece decorrer dos elevados investimentos em novas tecnologias (aliada ao esforço em termos de diversificação de produtos e serviços), parecendo não haver relação entre a dimensão dos bancos e regressão/progresso tecnológico.

A estimação das parcelas da ineficiência-X : a ineficiência técnica e a ineficiência de afectação

Alguns autores avançaram no sentido de tentarem estimar as duas componentes da ineficiência custo: a parcela imputada à ineficiência técnica e a decorrente da ineficiência de afectação.

Schmidt e Lovell (1979) apresentaram a decomposição da ineficiência custo para a formulação Cobb-Douglas. Kopp e Diewert (1982) obtiveram a decomposição da ineficiência-X para a formulação translog, embora sejam grandes as dificuldades de estimação das parcelas[56].

Ferrier e Lovell(1990) pretendem estimar a eficiência-X, justificando, desta forma, a opção pela aproximação produção

[56] Cfr. Resti (2000), p. 562 onde apresenta, sumariamente, o método Kopp-Diewert. Mas a aplicação empírica do método não se apresenta "muito satisfatória", conclui o autor.

> *Berger, Hanweck, and Humphrey(1987) were concerned with competitive viability, and preferred the intermediation approach. We are concerned with cost efficiency, and so we adopt the alternative production approach*[57]

Baseando-se na especificação Translog convencional da função custo com as equações dos "shares", os autores modificaram a formulação inicialmente proposta por Cristensen e Greene (1976), no sentido de permitir a estimação da ineficiência técnica e a de afectação.

Definem a forma funcional Translog recorrendo à seguinte simbologia e pressupostos. O banco incorpora uma série de factores de produção, vector $x = (x_1, x_2, ..., x_n) \in R_+^n$, cujos preços são fixos $w = (w_1, w_2, ..., w_n) \in R_+^n$, para a produção de m produtos $Q = (Q_1, Q_2, ..., Q_m) \in R_+^m$, numa envolvente caracterizada pelas variáveis $z = (z_1, z_{2,} ..., z_k) \in R_+^k$. O produtor pretende produzir os produtos Q, no ambiente z ao mínimo custo. Se a curva de custo mínimo, ou custo fronteira, for Translog então ter-se-á:

[57] G.D.Ferrier, C.A.Knox Lovell (1990), *Measuring cost efficiency in banking - econometric and linear programming evidence,* Journal of Econometrics, 46, p. 231

$$\ln CT = \alpha_0 + \sum_{i=1}^{m} \alpha_i \ln Q_{is} + \sum_{j=1}^{n} \beta_j \ln W_{js} + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_{ij} \ln Q_{is} \ln Q_{js}$$

$$+ \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_{ij} \ln W_{is} \ln W_{js} + \sum_{i=1}^{m} \sum_{j=}^{n} \delta_{ij} \ln Q_{is} \ln W_{js} + T_s + A + u_{0s}$$ [58]

E as condições de "shares"

$$S_{js} = \beta_j + \sum_{k=1}^{n} \beta_{ik} \ln W_{ks} + \sum_{i=1}^{m} \delta_{ij} \ln Q_{is} + b_j + u_{js}$$

para j=1,2,...,n e $B_{js} = b_j + u_{js}$.

Para o processo de estimação do sistema das condições de custos e de "shares" pode recorrer-se ao método da máxima verosimilhança, assumindo que os termos de perturbação seguem as seguintes distribuições:

$$u_{0s} \sim N(0, \sigma_{u_0}^2)$$

$$T_s \sim \left| N(0, \sigma_T^2 \right|$$

$$B_{js} \sim N(b_j, \sigma_{Bj_j}^2)$$ [59]

[58] Ferrier e Lovell (1990), consideraram, adicionalmente, uma variável explicativa Z (ambiente macroeconómico).

[59] As economias de escala podem ser medidas pelo recíproco da elasticidade custo $\left(\sum_{i=1}^{m} \alpha_i \right)^{-1}$; as economias de gama são mensuradas por via da complementaridade $\alpha_{ij} + \alpha_i \alpha_j$, como vimos em I.2.2.2..

Alguns autores optam por não considerar as condições "shares". Berger et al. (1997) excluem as condições dos "shares", não considerando o *Lema de Shephard* (e, consequentemente, diminuindo a precisão dos parâmetros estimados), para permitir o cômputo da ineficiência de afectação. Os autores defendem que obrigar os "shares" a serem consistentes com a equação de custo parece ser uma condição muito forte, uma vez que implica que o "mix" de inputs reaja conforme as alterações relativas dos preços, ou seja, a ineficiência-X será tão somente de cariz técnico.

Segundo Ferrier e Lovell (1990), os custos observados podem divergir da fronteira custo, devido a três razões: ineficiência técnica, $T \geq 0$, ineficiência de afectação, $A \geq 0$ e desvio aleatório, u. Os "shares" observados divergem dos "shares" eficientes por duas razões: ineficiência de afectação dos factores de produção, b_j (sendo $\sum_{j=1}^{n} b_j = 0$ e n o número de factores de produção) e o termo de perturbação, u_j (sendo $\sum_{j=1}^{n} u_j = 0$).

No sentido de associar, para efeitos de estimação, os dois parâmetros que mensuram a ineficiência de afectação, os autores avançam com a relação

$$A = b' F b$$

em que $b' = (b_1, b_2, ..., b_n)$, F é uma matriz (n x n) diagonal, com os elementos da diagonal principal $F_{jj} \geq 0$, j=1,2,...n. A relação acima satisfaz, simultaneamente, as três condições:

* $A = 0 \Leftrightarrow b_j = 0, \forall j$, de forma que os custos serão acrescidos se e só se foram tomadas decisões erradas quanto à afectação dos factores de produção

** $corr(A|b_j|) > 0$, ou seja, os custos aumentarão, quando erros, em ambas as direcções, ocorrerem

*** $\text{corr}(A|\text{var}(b_j)) > 0$, ou seja, os custos serão mais penalizados por grandes erros que por pequenos.

Ou seja, o custo derivado da ineficiência de afectação é a soma quadrática ponderada (cujos ponderadores são os parâmetros a estimar) dos erros. Para o cálculo da ineficiência de afectação recorre-se à estimação de $\hat{b}_j$ e de $\hat{F}_{jj}$. Os autores, trabalhando com dados *cross-section*, assumem que os custos com ineficiência de afectação não variam ao longo da amostra (ao contrário dos custos derivados da ineficiência técnica).

O custo decorrente da ineficiência técnica pode ser estimado como

$$\hat{T}_s = (\ln CT_s - [\hat{\cdot}] - \hat{A}) * \left(\frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2 + \hat{\sigma}_u^2} \right)$$

em que $[\hat{\cdot}]$ representa a função custo fronteira estimada.[60]

[60] Resti (2000), p.556, apresenta um método alternativo de segmentação da ineficiência-X. Ver, também, Kumbhakar e Lovell (2003) para o caso da função custo translog e dados em painel (p.174-75).

A aproximação econométrica da eficiência obriga à consideração de uma forma funcional para a função custo fronteira, o que parece trazer, consoante a especificação adoptada, resultados diversos. Daí, ter surgido a aproximação não paramétrica para o cômputo da eficiência.

I.3.2.2. A aproximação DEA

O método DEA ("data envelopment analysis") foi introduzido por Farrel (1957) e desenvolvido por Charnes, Cooper e Rhodes (1978) e aplicado ao sector bancário por Ferrier e Lovell (1990) e por Colwell e Davis (1992), entre outros autores. Para o caso português, tomando os anos de 1990-92, V.Mendes (1994) faz o primeiro estudo da aplicação do método DEA ao caso português, tendo chegado à conclusão de que, em média, os bancos operando no nosso país poderão, se adoptarem "condições de produção em eficiência técnica e de afectação, reduzir custos em cerca de 22%". Esta conclusão terá de ser vista, como o autor refere e como veremos, com algumas reservas dados os pressupostos do método.

Com esta metodologia pretende medir-se a eficiência relativa de um conjunto de empresas multiproduto, recorrendo a vários factores de produção, em que a função de produção eficiente apresenta problemas de especificação (e, daí, a não aplicabilidade dos modelos econométricos).

O objectivo da metodologia DEA é o de conseguir uma linha fronteira correspondente ao conjunto de empresas de "melhor prática" (ou seja, as de maior

eficiência relativa) que "envolvam", em termos de posicionamento, todas as outras empresas[61]. O grau de afastamento em relação a essa linha fronteira é uma medida da ineficiência de cada empresa. Colwell e Davis (1992) sugerem a maximização de

$$E_p = \frac{\sum_{j=1}^{m} \nu_j Q_{jp}}{\sum_{i=1}^{n} \upsilon_i X_{ip}}$$

sujeita a $E_p \leq 1$ para todos os p (empresas) e ponderações $\nu_i, \upsilon_j > 0$. O procedimento consiste na repetição, para cada empresa, da maximização de E. Se $E = 1$, a empresa em questão é relativamente eficiente face ao conjunto da amostra, ou seja, é uma empresa de "melhor prática"; se $E < 1$, a empresa é relativamente ineficiente. Repare-se que o método DEA fornece medidas de (in)eficiência relativa e não absoluta.

Ferrier e Lovell (1990) apresentam o seguinte programa linear de minimização:

$$\text{Min}_{x_{js}} \sum_{j=1}^{n} w_{js} x_{js}$$

[61]D.D.Evanoff, P.R.Israilevich (1991), *Productive efficiency in banking*, Federal Reserve Bank of Chicago, 15, pág. 11-32

$$\text{sujeito a}\begin{cases} Q_{is} \le \sum_{s=1}^{N} \mu_s Q_{is}, i = 1,2,\ldots m \\ x_{ij} \ge \sum_{s=1}^{N} \mu_s x_{js}, j = 1,2,\ldots,n \\ z_h \le \sum \mu_s z_{hs}, h = 1,2,\ldots,k_r \\ z_h \ge \sum \mu_s z_{hs}, h = k_{r+1},\ldots,k \\ \mu_s \ge 0, s = 1,2,\ldots,N \\ \sum_{s=1}^{N} \mu_s = 1 \end{cases}$$

onde w_j é o preço do factor de produção j; x_j representa o factor de produção j; Q_i é a produção do produto i; $\mu = (\mu_1, \mu_2, \ldots, \mu_N)$ é o vector intensidade que permite formar combinações convexas das quantidades observadas dos factores de produção e dos produtos; os elementos de z são exógenos (nas primeiras k_r restrições, os elementos de z pretendem ser o maior possível, dados os inputs, enquanto nas restantes $(k - k_r)$ restrições, os elementos de z deverão ser o mais possível limitados, dados os produtos); s representa o índice da empresa e N é o tamanho da amostra.

Este programa linear será resolvido tantas vezes quanto o número de observações da amostra.

Enquanto que é possível o cômputo de economias de escala pelo método DEA (não nos vamos deter neste problema uma vez que ele se encontra suficiente claro

em Ferrier e Lovell (1990) e não constitui nosso objecto de estudo neste ponto), já o mesmo não acontece com o de economias de gama[62].

A solução vectorial x^* minimiza a função objectivo, dados os preços dos factores de produção da empresa s, w_s, e o vector dos produtos, Q_s. O índice de eficiência produtiva global da empresa s é dado pelo quociente entre o custo mínimo e o custo efectivo:

$$EX = \frac{w_s' x_s^*}{w_s' x_s}$$

O montante em que os custos são aumentados devido à ineficiência-X, ou seja, à soma da ineficiência técnica, T_s, e ineficiência de afectação, A_s, é dado por

$$(T + A)_s = \left[\left((w_s' x_s^*)/(w_s' x_s)\right)^{-1} - 1\right]$$

Repare-se que se $x_s = x_s^*$ então, a ineficiência-X é nula; no caso contrário, a ineficiência-X será positiva.

A ineficiência-X pode ser decomposta em ineficiência de afectação, A_s, dada por

$$A_s = \left[(w_s' x_s^*)/(w_s'(\sum_{j=1}^{n} w_{js} x_{js}^*) x_s)\right]^{-1} - 1$$

e a ineficiência técnica será calculada por

[62] G.D.Ferrier, C.A.K.Lovell (1990), p. 235-6

$$T_s = (T + A)_s - A_s$$

O método DEA tem algumas limitações.

Uma primeira limitação, já referida, tem a ver com o cariz meramente relativo do cômputo da eficiência: as empresas de "melhor prática" são as mais eficientes em relação às outras empresas da amostra, mas poderão não ser eficientes.

Uma segunda limitação diz respeito ao pressuposto da não existência de factores aleatórios e, consequentemente, da atribuição de toda a diferença entre os valores da fronteira estimada e os valores observados à ineficiência (daí que os resultados da ineficiência tendam a ser empolados relativamente a outros métodos de estimação)[63].

Uma terceira limitação dos métodos não paramétricos é que são muito susceptíveis a observações extremas ("outliers"), uma vez que estabelecem linhas custo fronteira baseadas nas instituições com custos comparativamente mais baixos[64].

Finalmente, alguns autores salientam que, nos métodos não paramétricos, não é considerada a estrutura de mercado, quando a eficiência está ligada àquela[65].

[63] Mester, L.J. (1993), *Efficiency in the savings and loan industry,* Journal of Banking and Finance, 17, pág.267-87

[64]Colwell, R.J. e E.P:Davis (1992), *Output, Productivity and Externalities - the case of banking,* Bank of England, Discussion Paper nº3

[65]Berg, S.A. e M.Kim (1991), *Oligopolystic interdependence and banking efficiency: an empirical evaluation,* Norges Bank Research Paper, 5

O método escolhido para a estimação de economias de escala e economias de gama parece não apresentar grandes diferenças em termos de estimação, enquanto que a opção de um método econométrico ou um método não paramétrico parece influenciar os resultados estimados (Ferrier e Lovell, 1990).

I.4. Custos totais: custos explícitos de produção versus custos económicos

A inclusão do custo de oportunidade do capital investido no custo total foi sugerida, recentemente, na literatura de economia bancária (Clark, 1996). Vejamos em pormenor o procedimento proposto.

A afectação óptima de recursos, como elemento central da eficiência, significa que nenhuma realocação alternativa de recursos é capaz de produzir um maior acréscimo de riqueza para os accionistas. Assim, uma vez que a maximização da riqueza dos accionistas é sinónimo da maximização do valor actual líquido total do banco, este deve continuar a realizar projectos de investimento até que o valor actual líquido (VAL) do projecto de investimento marginal atinja zero.

O VAL de qualquer projecto de investimento i é equivalente ao valor actualizado do rendimento económico produzido pelo projecto, sendo o rendimento económico definido como o lucro que excede o custo de oportunidade do capital. Portanto,

$$VAL_i = \sum_t ((R_{it} - CP_{it}) - COP_{it}) / (1+k)^t$$

em que, R_{it}, CP_{it} COP_{it} representam a receita, o custo de produção explícito e o custo de oportunidade do capital do projecto i, no momento t. Sendo o numerador o equivalente do cash-flow certo correspondente a projectos sem risco gerado pelo projecto i, a taxa de actualização apropriada para o denominador é a taxa de juro de

risco nulo, k. O VAL total ou a riqueza dos accionistas RA podem ser expressos tendo em consideração os n projectos assumidos (i=1,2,…,n). Quer dizer :

$$RA = \sum_i VAL_i = \sum_i \sum_t (R_{it} - CP_{it} - COP_{it}) / (1+k)^t$$

O custo económico deve ser minimizado, para que a riqueza do accionista seja máxima.

O custo económico (CE) é definido como a soma dos custos explícitos de produção e do custo de oportunidade de capital, ou seja,

$$CE = \sum_t \sum_i (CP_{it} - COP_{it}) / (1+k)^t = \sum_t (CP + COP)_t / (1+k)^t$$

em que $(CP + COP)_t$ representa o total dos custos de produção e de oportunidade para o conjunto dos projectos de investimento (i=1,2,…,n), no momento t, e k é a taxa de juro livre de risco.

Em qualquer ponto da curva de custos de longo prazo, para se garantir a eficiência económica deve minimizar-se, em cada momento de tempo, a soma dos custos explícitos de produção e dos custos de oportunidade do capital associados à reafectação de recursos operada.

A determinação dos custos de produção explícitos é linear. Obtêm-se, somando as despesas dos inputs adquiridos para produzir os níveis óptimos dos outputs seleccionados pelos bancos. Assim,

$$CP_{bt} = \sum_{j} w_{jbt} \times q_{jbt}$$

em que w_{jbt} e q_{jbt} representam respectivamente os preços e quantidades de cada input j utilizado pelo banco b, no momento t. Os custos explícitos de produção são definidos, de acordo com a literatura, como a soma das despesas operacionais e das despesas financeiras.

Para a determinação do custo de oportunidade devemos considerar as aplicações alternativas de fundos, o retorno esperado e o risco. O custo de oportunidade de um projecto i é o retorno que os investidores esperam obter se investirem os fundos, em alternativa, em títulos financeiros de nível de risco comparável. Projectos mais arriscados exigem retorno mais elevado, pelo que têm um custo de oportunidade mais alto. A contribuição de um título para o risco de uma carteira diversificada depende da sensibilidade do retorno do título aos movimentos do mercado como um todo. Esta sensibilidade pode ser medida pelo coeficiente beta (β), numa regressão relativa a um banco b, do retorno dos accionistas de um título b, no momento t (r_{bt}), contra o retorno de uma carteira de mercado no momento t (r_{mt}), isto é:

$$r_{bt} = \alpha_b + \beta_b r_{mt} + u_{bt}$$

em que u_{bt} representa o termo de perturbação aleatório do título b, no momento t.

Conhecido (por estimação) o coeficiente beta de cada banco, o custo de oportunidade do capital relativo a um dado banco, num certo período, em termos de

taxa, é aproximado, usando o modelo designado como *capital asset pricing model* (CAP). Ou seja,

$$r'_{bt} = r_{ft} + \beta_b (r_{mt} - r_{ft})$$

em que r'_{bt} representa o custo de oportunidade do capital do banco b, no momento t, r_{ft} a taxa de juro livre de risco em t e r_{mt} o retorno da carteira de mercado para o mesmo momento.

Uma vez que todos os investidores devem ser remunerados à taxa do custo de oportunidade, o custo de oportunidade do banco calcula-se, através do produto da referida taxa pelo valor de mercado do capital (VMC) do banco investido, no início do período,

$$COP_{bt} = r'_{bt} \times VMC_{bt-1} = \left[r_{ft} + \beta_b (r_{mt} - r_{ft}) \right] \times VMC_{bt-1}$$

O custo total do banco pode, agora, ser calculado pela soma dos custos explícitos de produção e pelo custo de oportunidade do capital investido no momento t,

$$CE_{bt} = CP_{bt} + COP_{bt} = \left(\sum_j w_{jbt} \times q_{jbt} \right) + \left[r_{ft} + \beta_b (r_{mt} - r_{ft}) \right] \times VMC_{bt-1}$$

Esta medida dos custos introduz os rendimentos ajustados pelo risco que podiam ser ganhos em investimentos com escalas e mix de produtos alternativos, pois os rendimentos destas alternativas estariam incluídos no rendimento da carteira de mercado.

Conclusão

Neste capítulo, percorreu-se a literatura bancária recente em torno da problemática das economias de escala e de gama e da eficiência produtiva. Dos debates teóricos fundamentais acerca da definição económica de empresa bancária elegemos a abordagem *intermediação*, quer por razões de natureza económica (peso que os custos financeiros assumem na totalidade dos custos bancários) quer de ordem empírica (inexistência de informação para a adopção da abordagem *produção* para o caso português).

Foram apresentadas formas funcionais das funções custo, concluindo-se pela necessidade, *a priori,* do recurso a formas flexíveis do tipo Translog por permitirem o tratamento do carácter multiproduto dos bancos. A especificação Fourier surge como uma das formas funcionais eleitas, em estudos mais recentes, dada a capacidade de adaptabilidade da função aos dados pela introdução de senos e cossenos.

A estimação da eficiência-X é elaborada por via de aproximações paramétricas e não paramétricas, não apresentando os resultados empíricos obtidos por estas duas vias grande coerência: da imputação dos desfasamentos, entre a fronteira estimada e os valores observados, à ineficiência de um método não paramétrico, como o DEA, decorrem valores de ineficiência superiores, quando se recorre a este método, do que no caso da aproximação econométrica, como a SFA. Abordou-se, ainda, a existência de ineficiência à escala (conceito diverso do de

economias à escala) e relevou-se a necessidade da consideração de uma variável explicativa associada ao progresso tecnológico, como parte fundamental para o estudo da eficiência.

Estudos recentes chamam a atenção para a imprecisão que decorre da não inclusão do *custo de oportunidade*, ou seja, da problemática do risco, aquando da estimação de funções custo. Fazendo apelo ao artigo de J.Clark (1996) pretendeu-se apenas introduzir o tema para posteriores desenvolvimentos.

As operações de concentração bancária têm-se multiplicado, ao longo das últimas duas décadas, implicando processos de reestruturação importantes. No próximo capítulo, a concentração bancária e os efeitos sobre a eficiência serão objecto de estudo. Mencionam-se as tendências e as razões explicativas da concentração bancária, seriando-se as suas consequências. É de sublinhar, desde já, que os estudos empíricos que se têm debruçado sobre a estimação das consequências da concentração bancária sobre a eficiência têm extraído resultados não concordantes. Finalmente, vão ser apresentadas algumas abordagens empíricas que têm sido elaboradas para permitir o estudo da concentração.