ÉTUDES

1. LES SURCAPACITÉS BANCAIRES

Michel Dietsch [4], *Institut d'études politiques de Strasbourg, en collaboration avec le service des Études bancaires du Secrétariat général de la Commission bancaire*

Les restructurations bancaires en cours sont souvent justifiées par la volonté de réaliser des économies d'échelle et de profiter des synergies de coûts entre activités. On reconnaît en même temps la nécessité de réduire les surcapacités. Au plan théorique, en effet, l'un des moyens les plus simples de réaliser des économies d'échelle est de réduire les surcapacités. D'un point de vue économique, les surcapacités sont les capacités que toute entreprise « regrette » d'avoir installées. Des capacités de production jugées « normales » au moment où elles ont été introduites peuvent s'avérer après coup « excédentaires », si elles sont supérieures à celles qui permettraient de minimiser les coûts. Si l'entreprise avait la possibilité d'utiliser sur une plus large échelle ses équipements — en d'autres termes, si elle pouvait produire davantage — son coût moyen de production pourrait baisser. Dans une situation de surcapacité, ou bien chaque offreur ne vend pas les quantités suffisantes qui lui permettraient d'extraire la totalité des économies d'échelle de ses investissements et ses coûts unitaires sont alors trop élevés, ou bien le nombre d'offreurs est trop élevé, ce qui est le cas, en particulier, si le niveau de l'activité ne peut croître parce que la demande elle-même n'augmente plus.

Toutefois, si l'impact des surcapacités sur les coûts est bien établi au plan théorique, au plan pratique, rares sont les études qui ont tenté d'évaluer l'existence des surcapacités bancaires et d'en chiffrer le montant, que ce soit en Europe ou aux États-Unis.

L'objet de cet article est de proposer une mesure des surcapacités bancaires en France depuis la fin des années 1980. On étend à cette fin au secteur bancaire une méthodologie de mesure de la surcapacité initialement développée pour le secteur manufacturier. Cette méthodologie repose précisément sur l'idée que les industries en situation de surcapacité se caractérisent par des coûts excessifs et des rendements d'échelle croissants. Mais avant de présenter ces modalités de mesure, il est utile de s'interroger sur l'apparition et la persistance apparente des surcapacités dans la banque.

1.1. Pourquoi les surcapacités apparaissent-elles ?

Dans l'industrie bancaire, trois grands types de forces sont généralement considérées comme pouvant être à l'origine de surcapacités au cours de la dernière décennie 5 : la déréglementation financière, la réduction de la demande pour les produits bancaires et les innovations technologiques.

Considérons tout d'abord la réglementation. Le mouvement de déréglementation de la fin des années 1980 a fait passer l'industrie bancaire de la situation caractéristique d'un oligopole bénéficiant d'une protection publique à une situation tout aussi caractéristique de très forte rivalité stratégique entre les offreurs. Ce changement du contexte concurrentiel a rendu caduques certains investissements bancaires, en particulier dans la banque de détail. Ainsi, par exemple, on peut considérer comme une conséquence de la réglementation antérieure le nombre jugé souvent excessif de guichets bancaires. Faute de pouvoir mener une concurrence par les prix, les banques se sont livrées partout à une concurrence par la proximité et l'accessibilité des services, entraînant ce que l'on a appelé une « course aux guichets », d'autant plus profitable que les dépôts à vue n'étaient pas rémunérés. L'effet d'une réglementation des prix bancaires sur le nombre de banques a d'ailleurs été clairement mis en évidence au plan théorique 6 ou appliqué 7.

---

4 Les vues exprimées ici n'engagent que l'auteur et non le Secrétariat général de la Commission bancaire.

5 Voir Frydl, « Excess Capacity in the Financial Sector: Causes and Issues » Federal Reserve Bank of New York, June 1993, Dietsch M. « Les surcapacités bancaires en France », Revue d'économie financière, mars 1994 et Davis P. et Salo S. « Excess Capacity in EU and US Banking Sectors – Conceptual, Measurement and Policy Issues » London School of Economics-Financial Markets Group Special Paper Series, n° 105, August 1998.

6 Voir Chiappori P-A., Perez-Castillo D. et T. Verdier, « Spatial Competition in the Banking System: localization, Cross Subsidies and the Regulation of Deposits Rates », European Economic Review n° 39, 1995.

La déréglementation suffit-elle cependant à expliquer les surcapacités bancaires ? Il est permis d'en douter. Tout d'abord, les clients peuvent apprécier les services de proximité offerts par leurs banques. Comme pour nombre de biens de grande consommation, les consommateurs décident sur la base des prix mais aussi de la qualité. Pour satisfaire les préférences de leurs clients, les banques peuvent donc être incitées à réaliser des investissements immatériels coûteux afin de maintenir de bonnes relations de clientèle et une réputation suffisante pour garantir la qualité de leur offre. La libéralisation des prix ne saurait donc, à elle seule, expliquer l'apparition des surcapacités dans les réseaux. On peut remarquer, ensuite, que la création d'un marché unique de capitaux en Europe n'a pas provoqué à ses débuts les rapprochements transfrontaliers attendus, ce qui montre que des barrières non réglementaires subsistent sur les marchés bancaires. Enfin, le mouvement de déréglementation financière a aujourd'hui plus de dix ans en Europe. Ses effets ont donc largement eu le temps de se produire sur cette période. En réalité, d'autres forces que les forces (dé)réglementaires expliquaient déjà hier et expliquent encore aujourd'hui l'apparition de capacités excédentaires et cette volonté de les réduire qui est à l'origine de la course à la taille à laquelle les banques se livrent aujourd'hui.

La saturation de la demande de produits bancaires est souvent considérée comme l'une de ces forces. Le développement de la vente de produits financiers par des institutions non bancaires et par les marchés financiers a sûrement détourné des banques certains de leurs clients traditionnels. Aujourd'hui, en France, le premier fournisseur de financements externes aux grandes entreprises n'est plus le système bancaire mais le marché financier. Par ailleurs, les plus grandes banques françaises ne tirent plus, sur certains clients, l'essentiel de leurs recettes des marges d'intérêt mais des commissions. Mais, pour autant, la banque peut-elle être considérée comme une industrie « en déclin » ?

En menant une analyse attentive des bilans bancaires, Berger, DeYoung, Genay et Udell 8 montrent que les banques restent dans tous les pays industrialisés les principaux pourvoyeurs de services financiers pour deux grands secteurs de l'économie : les particuliers et les petites et moyennes entreprises. On sait aussi qu'elles sont les intermédiaires obligés des entreprises pour l'accès aux marchés. Ainsi, par exemple, selon les mêmes auteurs, en 1998, le montant total des prêts bancaires syndiqués représentait un total de 574 milliards de dollars, supérieur à celui des émissions de dettes sur les marchés de titres (413 milliards) et à celui des émissions d'actions (70 milliards). Le « déclin » de la banque paraît donc en réalité plus apparent que réel.

Les vagues successives d'innovations techniques et financières qu'ont connues les banques depuis la fin des années soixante peuvent constituer une troisième force à l'origine de surcapacités. La banque a connu deux révolutions « industrielles » successives en quelques années. La première, qui date de la fin des années 1980, a été caractérisée par l'émergence des nouveaux instruments financiers et le développement de l'ingénierie financière. Elle a conduit à un développement spectaculaire des activités de gestion d'actifs et de banque d'investissement. Elle a conduit également à une convergence des pratiques et des comportements financiers des agents non financiers autant que des institutions financières à travers le monde. Par conséquent, les économies d'échelle peuvent aujourd'hui en principe être exploitées sur une base globale. La seconde, plus récente, est celle des technologies de l'information. Celles-ci sont considérées comme étant la source d'importantes économies d'échelle. Ainsi, Bauer et Hancock 9 ont estimé à 85 % la baisse moyenne des coûts des virements électroniques de dépôts pour les banques américaines entre 1979 et 1994. Là encore, ces progrès incitent à exploiter les économies d'échelle à un niveau d'activité beaucoup plus élevé. Une autre conséquence est que l'obsolescence des matériels devient plus rapide, ce qui est aussi un facteur potentiel de surcapacités.

Il ressort de cette brève analyse des causes d'apparition des surcapacités que celles-ci sont multiples. Si la déréglementation financière a pu être à l'origine de surcapacités au début des années 1990, les profondes mutations technologiques en constituent sans doute la principale cause dans la période récente.

1.2. Pourquoi les surcapacités subsistent-elles ?

La concurrence devrait provoquer la réduction des surcapacités dans le court terme. Cependant, des capacités excédentaires peuvent subsister à moyen terme si les coûts induits par leur réduction sont supérieurs aux gains provenant de cette réduction. Sur les marchés bancaires, les conditions « normales » ne sont pas nécessairement celles de la concurrence parfaite. Ainsi, des barrières technologiques ou d'autres types de barrières endogènes à l'entrée peuvent expliquer pourquoi il est coûteux de réduire les surcapacités.

---

7 Voir Hannan T. « Bank Commercial Loan Markets and the Role of Market Structur : Evidence from Surveys of Commercial Lending », Journal of Banking and Finance, n° 15, 1991.

8 « Globalization of Financial Institutions: Evience from Cross-Border Banking Performance » Brooking-Wharton Papers on Financial Services Third Annual Conference, October 1999.

9 Bauer P. et Hancock D. « Scale Economies and Technological Change in Federal Reserve ACH Payment Processing », Federal Reserve Bank of Cleveland Review, n° 31, 1995.

L'importance des coûts fixes non récupérables (« sunk costs ») constitue sans doute la plus significative de ces barrières. Ces coûts sont ceux qui procurent des gains de long terme à un offreur mais qui ne peuvent être récupérés si celui-ci quitte l'industrie. Cela vient du fait qu'il n'existe pas de marchés liquides sur lesquels l'offreur puisse revendre les actifs fixes correspondants. Si ces coûts sont faibles, l'industrie bancaire peut réduire les surcapacités. En revanche, s'ils sont élevés, des surcapacités peuvent être maintenues au cours du temps. En d'autres termes, des surcapacités peuvent subsister si une large part des coûts bancaires sont des coûts non récupérables.

Sans faire un inventaire complet des coûts non récupérables, on peut avancer que, dans la banque de détail, de tels coûts peuvent tout d'abord provenir de l'existence de relations de long terme. Les banques et les clients peuvent gagner à l'existence de relations de clientèle dans la mesure où l'information que la banque retire de la répétition des relations lui permet à la fois d'être plus compétitive que les banques concurrentes et de lui laisser la possibilité d'extraire un pouvoir de marché de cette information privée 10. Toutefois, en raison précisément de son caractère privé, la banque ne peut négocier aisément cette information auprès d'autres banques. Cela explique le caractère non récupérable des investissements dans la relation de clientèle. L'existence de coûts de changement de banque pour les clients peut aussi être à l'origine de coûts non récupérables. Cette caractéristique du fonctionnement des marchés bancaires procure des avantages de long terme aux banques dans la mesure où elle rend les clients moins sensibles aux prix. Elle peut donc les inciter à livrer une concurrence stratégique dans laquelle elles acceptent une vente « à perte » à court terme, de façon à « capturer » les clients sur le long terme. Les banques réduisent alors leur profitabilité de courte période de façon à extraire un pouvoir de marché des relations de clientèle dans le long terme. Elles supportent ainsi des coûts non récupérables.

En définitive, on peut distinguer deux formes de capacités excédentaires : 1°) celles qui sont « involontaires » parce qu'elles sont provoquées par des changements majeurs des structures et des conditions de fonctionnement des marchés bancaires, 2°) celles qui peuvent être considérées comme « volontaires », soit parce les banques considèrent qu'il est utile de les maintenir dans l'intérêt de leurs clients, ceux-ci préférant des services plus accessibles ou de meilleure qualité, soit parce qu'elles résultent d'une rivalité stratégique entre les banques.

On considère ici avant tout la première forme de surcapacités, les surcapacités involontaires, celles qui n'existent que parce que le niveau des activités bancaires est inférieur au niveau d'équilibre dans des conditions de marché normales. La mesure des surcapacités que nous proposons dans ce qui suit correspond en effet à cette approche. Les surcapacités sont considérées comme excédentaires par comparaison avec les capacités maintenues « volontairement » pour satisfaire la demande des clients. Cependant, notre mesure des surcapacités recouvre les surcapacités « volontaires » résultant des stratégies offensives des banques.

1.3. Comment mesurer les surcapacités bancaires ?

Le taux d'utilisation des capacités peut être défini, d'une manière générale, comme le rapport entre la production potentielle, celle que les entreprises déclarent pouvoir produire en l'état actuel de leurs capacités de production, et la production effective. C'est la définition courante des comptables nationaux. La mesure de cette surcapacité est souvent réalisée à partir des résultats d'enquêtes auprès des entreprises. Dans la banque, de telles données n'existent pas. C'est pourquoi on recourt à une autre mesure, fondée sur la théorie microéconomique des coûts, et déjà utilisée pour mesurer les surcapacités dans les industries manufacturières 11.

Cette mesure (voir annexe 12) consiste à évaluer le surcroît de coûts imputable au fait que le niveau d'activité d'une entreprise est inférieur à celui que permettrait ses capacités courantes. Elle opère donc à un niveau d'activité Y alors qu'elle dispose des capacités qui lui permettrait d'opérer à un niveau Y* supérieur à Y et d'extraire ainsi toutes les économies d'échelle. En d'autres termes, si la banque utilisait tous ses actifs fixes en produisant le niveau Y*, elle serait en mesure d'obtenir le coût unitaire moyen minimum. La surcapacité est en conséquence mesurée par la différence entre le « coût potentiel » de production de l'« output » effectif Y, qui correspond au minimum de son coût moyen de court terme, et le « coût effectif ».

Le problème revient alors à mesurer ce coût potentiel. Une solution cohérente avec notre approche consiste à mesurer la surcapacité en comparant simplement les prix actuels des divers « inputs », ou prix de marché, aux prix

---

10 Par « pouvoir de marché », on entend ici la possibilité de vendre les produits à un prix supérieur au coût marginal, c'est-à-dire de réaliser des profits supérieurs à la normale. Voir aussi Sharpe S. « Asymmetric Information, Bank Lending, and Implicit Contracts: A Stylized Model of Customer Relationships », Journal of Finance, vol. 65, n° 4, September 1990.

11 Voir Berndt E. R. and Hesse D. M. « Measuring and assessing capacity-utilization in the manufacturing sectors of nine OECD countries », European Economic Review, vol. 30, 1986.

12 Une présentation formelle de la méthodologie et du modèle des coûts est fournie dans l'article de Chaffai M. et Dietsch M. « Capacity-Utilization and Cost-Efficiency in the European Banking Industry », mimeo, CEPF-IEP de Strasbourg (septembre 1999).

optimaux, ou « prix de référence ». Ces derniers sont ceux qui correspondent à la situation dans laquelle le coût total de court terme est égal au coût total de long terme. On les calcule en estimant par l'économétrie la fonction de coût de long terme. Si les prix de marché s'avèrent supérieurs aux prix de référence, on en déduit que la banque est incitée à réduire ses capacités de production, c'est donc qu'elle est en situation de surcapacité.

1.4. L'évolution des surcapacités bancaires sur la période 1988-1998

1.4.1. Les données

Les données proviennent des bases Banco et Bafi (Base des agents financiers) qui contiennent les informations comptables et prudentielles remises à la Commission bancaire par les établissements de crédit. Les données Banco couvrent la période 1988-1992, les données Bafi la période 1993-1998. L'échantillon Banco comprend 290 banques à vocation générale en 1988 et, sous l'effet de la démographie bancaire, principalement déterminée par le regroupement des caisses d'épargne, 190 banques en fin de période. L'échantillon Bafi comprend 198 banques en 1993 et se réduit à 169 en 1998. Les échantillons comprennent des banques AFB, des banques mutualistes et des caisses d'épargne. Ils ne contiennent que des banques universelles et écartent tous les établissements spécialisés. Un seuil de taille minimale a également été fixé (le total de bilan doit être supérieur à 2,2 milliards de francs). La raison principale de l'exclusion des petits établissements est leur trop grande spécialisation, même si, le cas échéant, ils ne sont pas juridiquement agréés en tant qu'établissements spécialisés. En revanche, aucune limite supérieure n'a été fixée en matière de taille. Les échantillons comprennent donc les plus grands établissements. Néanmoins, les organes centraux ou têtes de groupe des réseaux mutualistes ou coopératifs — CNCA, Caisse centrale des Banques populaires, CENCEP, etc — ont été exclus de manière à préserver l'homogénéité du champ d'analyse 13. Le choix de ne retenir que des banques à vocation générale a conduit aussi à écarter la quasi-totalité des banques étrangères de l'échantillon, à l'exception de trois d'entre elles qui peuvent être considérées comme des banques universelles 14.

1.4.2. Le modèle des coûts bancaires et le choix des actifs fixes

Les surcapacités sont mesurées à partir de l'estimation économétrique d'une fonction translog des coûts variables (voir annexe méthodologique pour les détails de la modélisation). Le coût variable est ici mesuré par la somme des salaires et charges sociales, des coûts financiers et des autres coûts variables non associés à l'utilisation du capital physique. Les produits bancaires retenus dans la fonction de coût sont : (1) les dépôts d'épargne sur livrets et les comptes à terme, (2) les crédits, (3) les actifs de placement et d'investissement et (4) les commissions. Ces dernières fournissent une mesure approchée du montant des autres services offerts par la banque aux clients. Les prix unitaires des « inputs » variables sont : (1) le prix du travail, mesuré en rapportant les salaires et charges au nombre d'employés, (2) le coût moyen des ressources à court terme empruntées par la banque sur les marchés, (3) le coût moyen des ressources empruntées à long terme sur les marché.

Deux types d'« inputs » quasi fixes 15 sont introduits dans la fonction de coût : les dépôts à vue et les immobilisations réelles. Le choix des dépôts à vue comme actif quasi fixe peut être justifié simplement. Dans la banque, la constitution d'un noyau dur de déposants (« core deposits ») est généralement considérée comme un investissement commercial à long terme. Elle nécessite des investissements immatériels, en publicité, en réputation ou en relations de clientèle, autant que des investissements matériels dans les canaux de distribution et systèmes d'information. Ces investissements ne peuvent être complètement reflétés dans les immobilisations corporelles. C'est pourquoi on intègre aussi les dépôts comme actifs fixes. Comme on l'a dit, les surcapacités sont mesurées en comparant les prix effectifs des actifs quasi fixes à leurs prix de référence. Certains des prix de marché des actifs fixes n'étant pas connus, nous avons utilisé des mesures approchées de ces prix. Pour les dépôts à vue, le choix s'est porté sur le rendement des actifs (ROA). L'hypothèse sur laquelle repose ce choix est que si l'investissement en dépôts de la clientèle est un investissement de long terme, le coût de ce capital doit être égal, à l'équilibre de long terme, au rendement moyen du capital bancaire, mesuré précisément par le ROA. Pour le capital physique, nous avons utilisé comme mesure du prix du capital le coût moyen des ressources empruntées

13 En conséquence, seuls les établissements adhérents — les Caisses régionales du Crédit agricole, les Banques populaires, les Caisses de Crédit mutuel et les Caisses d'épargne — figurent dans le panel. Dans le cas d'une répartition relativement délimitée des activités entre la tête de groupe et les entités de réseaux, nous avons appliqué les prix de dépôts et des ressources de marchés du groupe à chacune de ces entités.

14 Les opérations habituelles de retraitement et de « nettoyage » des données de panels ont été appliquées pour constituer les échantillons définitifs. En outre, toutes les variables monétaires ont été déflatées par l'indice des prix du PIB et les flux ont été corrigés dans le cas où un établissement n'a pas exercé son activité sur l'ensemble d'un exercice comptable annuel.

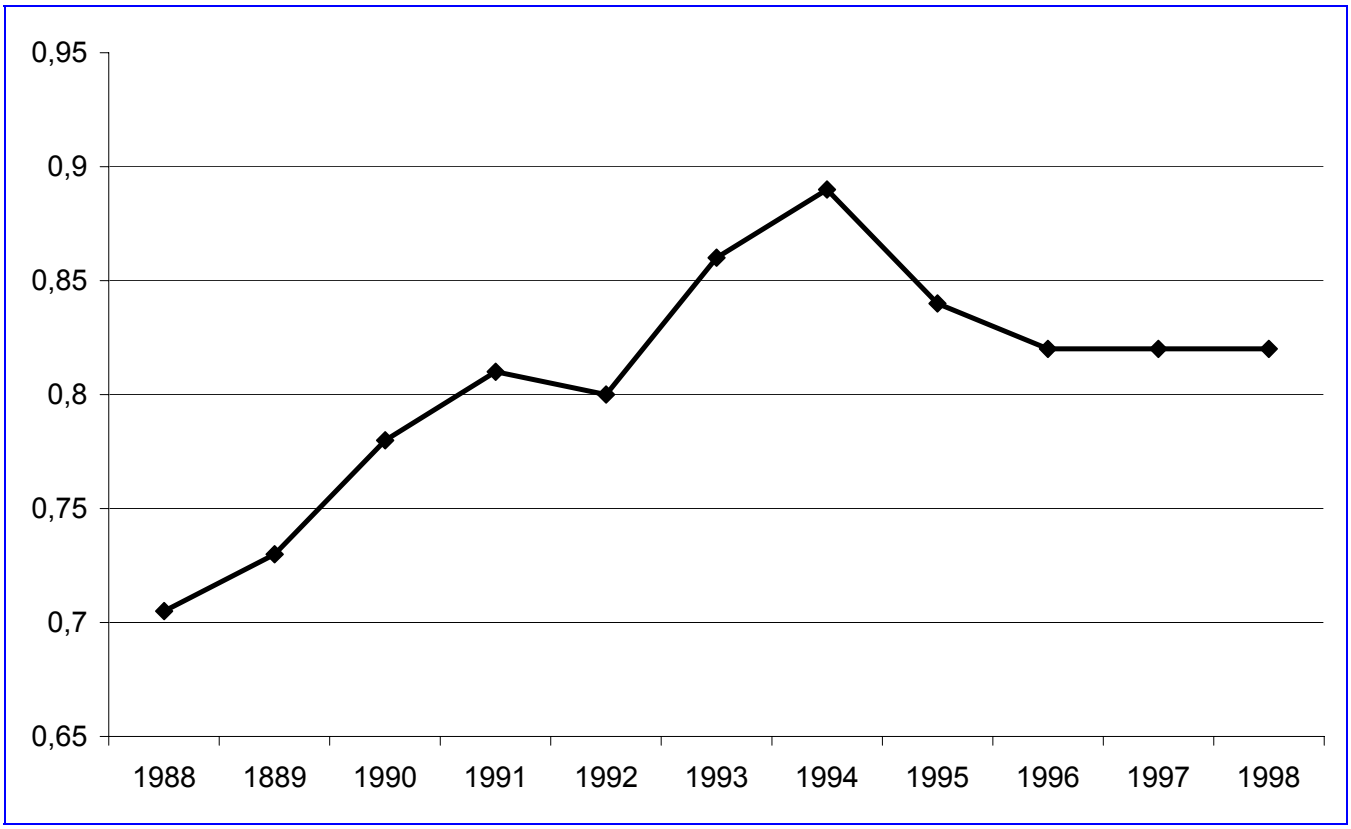
15 Un actif quasi fixe est un actif fixe à court terme, mais variable à moyen et long terme.

par chaque banque sur les marchés de capitaux (exception faite des ressources interbancaires). Ce taux représente le coût d'opportunité des investissements en actifs immobiliers et autres équipements bancaires. Une analyse de la sensibilité des mesures de surcapacité aux variations de ces prix a confirmé ces choix méthodologiques 16.

1.4.3. L'estimation des surcapacités et leur évolution au cours de la période 1988-1998

Le tableau 1 et le graphique 1 présentent l'évolution du taux d'utilisation des capacités bancaires sur la période 1988-1998. Ils sont divisés en deux sous-périodes, qui correspondent aux deux bases de données utilisées (Banco puis Bafi). Les mesures des deux sous-périodes ne sont pas directement comparables, en raison des différences des définitions comptables des « inputs » et des « outputs » dans les deux bases de données. Les taux présentés sur le tableau 1 sont les taux moyens sur l'ensemble des banques de la population étudiée (les écarts-types figurent entre parenthèses sur ce tableau).

Graphique 1 : Évolution du taux d'utilisation des capacités de 1988 à 1998



Source: Commission bancaire

16 De plus, un test statistique du ratio de vraisemblance a été conduit pour justifier le choix des deux actifs fixes dans la fonction des coûts de court terme. L'hypothèse nulle que la fonction du coût variable ne comprend qu'un des deux « inputs » fixes au lieu des deux est rejetée.

Tableau 1 : Valeur moyenne des taux d'utilisation des capacités dans les banques françaises au cours de la période 1988-1998

	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
Taux d'utilisation des capacités*	0,705 (0,10)	0,73 (0,08)	0,78 (0,09)	0,81 (0,08)	0,80 (0,09)	0,86 (0,17)	0,89 (0,13)	0,84 (0,16)	0,82 (0,14)	0,82 (0,15)	0,82 (0,14)
* Un taux égal à 1 correspond à une situation de pleine capacité.											

Source : Commission bancaire

Les résultats montrent :

1°) que les surcapacités dans les banques de dépôts françaises se situent autour de 17 % en moyenne sur la période 1993-1998 (la moyenne du taux d'utilisation des capacités sur la période des six ans est égale à 82,9%) ;

2°) que les surcapacités ont eu tendance à décroître, en moyenne, au cours de la période 1988-1992, puis à croître au cours de la période la plus récente. Cette divergence d'évolution traduit sans doute l'évolution dans les causes d'apparition des surcapacités évoquée plus haut : alors qu'à la fin des années 1980 les principales causes provenaient probablement de la déréglementation et des innovations de produits, au cours des années 1990 il faut sans doute davantage les chercher du côté des innovations de processus et du progrès des technologies de l'information. De plus, on observe que le taux d'utilisation des capacités croît en 1993 et 1994, dans une période où la demande des produits bancaires était en net recul, et qu'il diminue nettement à partir de l'année 1995. Ce résultat peut être mis en rapport avec le renforcement durant cette période de la rivalité stratégique entre banques qui s'est exprimé notamment sous la forme de guerres de prix ;

3°) que la dispersion des surcapacités est relativement limitée, comme en témoigne la valeur plutôt faible de l'écart-type par rapport à la moyenne. Une description plus précise de la dispersion figure sur le tableau 2 pour la période 1993-1998. Elle montre que les valeurs des quartiles supérieur et inférieur sont relativement proches de la médiane ;

4°) que les taux d'utilisation des capacités bancaires sont un peu moins élevés que les taux de capacité mesurés dans l'industrie à partir des enquêtes de conjoncture, notamment au cours de la première période. Les taux d'utilisation cités par ces enquêtes étaient de 82,5 % en 1996 et 85 % en 1999. De plus, les mesures présentées ici, réalisées sur données microéconomiques, visent à faire ressortir l'existence de surcapacités structurelles ou de long terme alors que les enquêtes de conjoncture ont simplement pour objet de mesurer les variations du taux d'utilisation des capacités dans le court terme.

Par ailleurs, on observe que les surcapacités à l'intérieur d'un même réseau diminuent systématiquement lorsque le réseau procède à des fusions de banques.

Tableau 2 : La dispersion des taux d'utilisation des capacités* par année (1993-1998)

	Minimum	Quartile inférieur	Médiane	Quartile supérieur	Maximum
1993	0,405	0,786	0,896	0,962	1,637
1994	0,556	0,789	0,891	0,961	1,627
1995	0,525	0,754	0,864	0,934	1,694
1996	0,537	0,734	0,857	0,928	1,182
1997	0,334	0,707	0,850	0,931	1,404
1998	0,478	0,719	0,844	0,928	1,227
* Un taux égal à 1 correspond à une situation de pleine capacité.					

Source : Commission bancaire

1.4.4. Les surcapacités selon la taille des banques

Le tableau 3 est relatif à la période 1993-1998. Il montre comment les surcapacités varient en fonction de la taille des établissements. Trois classes de taille ont été construites : les petites banques, dont le total de bilan est inférieur à 10 milliards de francs, les banques moyennes dont le total de bilan est compris entre 10 et 20 milliards de francs et les grandes banques dont le total de bilan dépasse ce dernier seuil. Les taux du tableau 3 sont les taux moyens par classe au cours de la période.

Tableau 3 : Les taux d'utilisation des capacités selon la taille (1993-1998)

	Minimum	Quartile inférieur	Médiane	Quartile supérieur	Maximum
Petites	0,334	0,824	0,888	0,943	1,695
Moyennes	0,538	0,770	0,864	0,944	1,092
Grandes	0,405	0,631	0,825	0,928	1,404
* Un taux égal à 1 correspond à une situation de pleine capacité.					

Source : Commission bancaire

On observe que la valeur médiane du taux d'utilisation des capacités tend à diminuer à mesure que la taille augmente. Les surcapacités tendent donc, en moyenne, à être plus importantes dans les grandes banques. Toutefois, ce résultat est à prendre avec prudence. La dispersion du taux d'utilisation à l'intérieur de chaque classe de taille est, en effet, plus forte que la dispersion moyenne entre classes de taille.

1.4.5. La robustesse des résultats

Pour tester la robustesse des résultats précédents, nous avons tout d'abord utilisé d'autres spécifications des fonctions de coûts de court terme en faisant varier le choix des « outputs » bancaires et nous avons comparé les résultats des nouvelles estimations avec les précédentes. On observe en particulier que le niveau des taux d'utilisation des capacités diminue si l'on ne considère pas les dépôts sur livrets comme un « output » mais, au contraire, comme un « input » (financier) variable. On peut inférer de ce résultat qu'une partie des capacités sont mobilisées par le fonctionnement des comptes sur livrets. Les taux d'utilisation des capacités diminuent (les surcapacités augmentent) également si l'on traite les dépôts sur livrets comme des « inputs » quasi fixes et qu'on les ajoute aux dépôts à vue.

Ensuite, la mesure de la surcapacité utilisée dans cette étude étant, par construction, sensible au choix des prix de marché des actifs fixes, d'autres tests ont été menés pour vérifier la stabilité des résultats quand on change la définition du prix de ces actifs. Ces tests montrent que les surcapacités tendent à être un peu plus importantes si l'on prend comme mesure approchée du coût des dépôts à vue (l'un des deux actifs fixes) d'autres mesures du prix, en particulier le taux moyen de rémunération des comptes à terme et des livrets ou la marge bancaire sur les opérations avec la clientèle, plutôt que le ROA. On vérifie ainsi, conformément à la logique de la modélisation choisie, que si l'on retient un prix de l'actif fixe plus faible, les surcapacités mesurées sont plus importantes. Toutefois, un résultat intéressant des comparaisons est que le profil temporel caractéristique du taux de surcapacité est confirmé par tous les modèles, c'est-à-dire quel que soit le choix des indicateurs de prix. De plus, les écarts de taux d'utilisation des capacités bancaires selon la taille des établissements ou selon le type de réseaux subsiste.

1.5. La relation entre les surcapacités et l'efficience coût de long terme

1.5.1. Les scores d'efficience

L'efficience coût mesure la capacité d'une entreprise à utiliser ses « inputs » de façon à éviter tout gaspillage (efficience dite technique) et à choisir correctement les combinaisons d'« inputs », compte tenu des prix de marché des « inputs » (efficience dite allocative) 17. Dans la courte période, une situation de surcapacité n'implique pas nécessairement l'inefficience. Elle signifie, en effet, que les actifs quasi fixes sont excédentaires et ne peuvent être utilisés plus intensément, alors que l'inefficience coût de court terme signifie que les « inputs » variables sont mal utilisés ou mal alloués, ce qui accroît les coûts. En d'autres termes, dans le court terme, une banque pourrait être efficiente en termes de coût même si elle maintenait des capacités excédentaires. Néanmoins, dans le long terme, la surcapacité implique l'inefficience coût. L'excès d'« inputs » fixes implique des inefficiences à la fois techniques et allocatives.

L'efficience de long terme a été mesurée à partir d'une fonction de coût total de long terme 18. L'analyse est ici restreinte à la seconde période 1993-1998. Les scores d'efficience moyens par banque ont été calculés sur la période des six ans en utilisant la méthode DFA (Distribution Free Approach). Les résultats sont conformes à ceux de l'étude précédente, à savoir un niveau d'efficience médian de l'ordre de 80 % (une inefficience moyenne de 20 %) et une relative indépendance de l'efficience par rapport à la taille et au type de réseaux.

Pour analyser les relations entre l'efficience et la surcapacité, nous avons calculé les corrélations entre les scores d'efficience (l'efficience croît quand le score tend vers 1) et le taux d'utilisation des capacités (les surcapacités diminuent quand le taux tend vers 1 ou est supérieur à 1). Une corrélation positive signifie donc qu'une réduction des surcapacités coïncide avec une plus forte efficience, ce qui est le résultat théorique de long terme attendu.

Le coefficient de corrélation a d'abord été calculé sur l'ensemble des banques de l'échantillon, tous réseaux confondus pour la période 1993-1998. Pour cette période, ce coefficient est égal à - 0,043. Il est donc très faible et peu différent de zéro. Il n'existe donc pas, dans l'ensemble de la population étudiée, de relation significative entre les deux variables, contrairement aux attentes.

L'inefficience augmente lorsque les surcapacités bancaires sont plus importantes.

Pour affiner l'analyse de cette relation, nous avons rangé les banques en trois classes selon l'importance des surcapacités. On constate alors que la relation attendue entre l'inefficience et l'importance des surcapacités apparaît dans la classe des plus fortes surcapacités (tableau 4).

Tableau 4 : Les coefficients de corrélation entre les scores d'efficience coût et les taux d'utilisation des capacités par classe de surcapacités

	Faibles surcapacités	Surcapacités moyennes	Fortes surcapacités	Ensemble
Coefficient de corrélation	- 0,283[(a)]	- 0,245[(b)]	0,464[(c)]	- 0,043[(a)]
[(a)] Non significatif ; [(b)] Significatif au seuil de 5 % ; [(c)] Significatif au seuil de 1 %.				

Source : Commission bancaire

Il est intéressant de constater que la relation positive attendue entre l'inefficience et l'importance des surcapacités est assez nette (coefficient = 0,464) dans la classe de fortes surcapacités. Les coûts fixes élevés associés à l'existence de surcapacités y sont donc à l'origine d'une insuffisance relative des performances productives dans le long terme. En d'autres termes, la production est insuffisante pour minimiser les coûts moyens de long terme.

Par contre, la relation est négative dans la classe intermédiaire et elle est non significative dans la classe de faibles surcapacités. On peut donc se demander pourquoi certaines banques relativement efficientes — qui parviennent à abaisser les coûts moyens de long terme — détiennent des surcapacités. Trois hypothèses peuvent être avancées ici. La première est que les surcapacités résultent d'une baisse — temporaire — de la demande des produits

17 Voir Burkart O., Dietsch M. et Gonsard H. « L'efficience coût et l'efficience profit des banques françaises dans les années 1990 », Bulletin de la Commission bancaire, avril 1999.

18 Le modèle économétrique des coûts estimé est identique à celui qui sert au calcul des surcapacités, à ceci près que la variable expliquée est le coût total qui comprend le coût d'utilisation des « inputs » fixes et que tous les « inputs » sont supposés variables.

bancaires. Cette hypothèse doit être rejetée dans la mesure où la demande de ces produits a augmenté à partir de 1995, alors même que les surcapacités commençaient à croître.

Une seconde hypothèse est que les surcapacités bancaires sont maintenues parce qu'elles procurent des avantages aux banques et à leurs clients. Ainsi, par exemple, le maintien de réseaux de distribution serait utile pour satisfaire la préférence des clients pour des produits plus accessibles. Si cette hypothèse est vérifiée, les banques qui possèdent les plus fortes surcapacités doivent aussi posséder les parts de marché les plus fortes, notamment dans la banque de détail. Or on observe que ce sont, au contraire, les banques qui possèdent de fortes positions de marché qui ont les surcapacités en moyenne les plus faibles : le tableau 5 montre que les surcapacités sont d'autant plus importantes que la présence des banques sur les marchés locaux est faible, en moyenne (tableau 5). Ce résultat est conforme à notre approche de la mesure des surcapacités. Des parts de marché plus importantes permettent logiquement d'amortir des coûts fixes plus élevés. En conséquence, des guerres stratégiques de capacités ne semblent pas en mesure d'améliorer les performances de long terme des banques. Il existe sans doute, au contraire, un potentiel de réduction des coûts dans une stratégie de conquête de parts de marché par regroupement ou restructuration de réseaux. Cette deuxième hypothèse n'est donc pas suffisante.

Tableau 5 : Relation entre le taux de présence des banques, l'efficience coût de long terme et le taux d'utilisation des capacités (1993-1998)

Taux de présence territorial (*)	Taux moyen d'utilisation des capacités	Score moyen d'efficience coût
Faible	0,78	0,78
Moyen	0,80	0,85
Bon	0,87	0,79
Élevé	0,87	0,83

* Ce taux rapporte le nombre de guichets d'une banque dans les départements où elle est présente au nombre total de guichets des banques présentes dans les mêmes périmètres. C'est une mesure approchée de la part de marché des banques sur les marchés bancaires locaux.

Source : Commission bancaire

Enfin, une troisième hypothèse est que les banques les plus efficientes à long terme, celles qui ont par conséquent mis en place les combinaisons techniques qui leur permettent d'être à terme les plus performantes, se heurtent dans le court terme au problème de l'élimination des capacités associées aux anciennes technologies. Comme on l'a vu plus haut, à court terme, les coûts de réduction des surcapacités peuvent être supérieurs aux gains associés à leur élimination, ce qui explique le maintien de surcapacités.

1.6. Conclusion

Cette étude présente des mesures du taux d'utilisation des capacités de production dans les banques françaises de 1988 à 1998. Deux phases ont marqué l'évolution des surcapacités sur cette période. La première est caractérisée par une réduction des surcapacités. Elle s'achève au milieu des années 1990 et intervient après les chocs créés par la déréglementation et les innovations de produits de la fin des années 1980. La deuxième phase débute vers 1995. Elle est marquée par une augmentation des surcapacités. Celle-ci intervient dans une période où de fortes guerres stratégiques apparaissent entre banques et alors même que la demande pour les produits bancaires est de nouveau en hausse. En théorie, des surcapacités peuvent être maintenues « volontairement » dans le court terme pour satisfaire des objectifs de parts de marché à moyen terme. Cependant, on montre que les surcapacités sont en moyenne d'autant plus faibles que l'implantation territoriale des banques est forte. De plus, la période est aussi caractérisée par une nouvelle vague d'innovations qui concernent cette fois davantage les processus de production que les produits. Les surcapacités peuvent donc aussi résulter de la difficulté à se défaire des capacités associées aux anciennes technologies.

On vérifie aussi que les performances productives des banques, mesurées par leur efficience de long terme, tendent à être plus faibles lorsque les surcapacités sont relativement importantes.

Annexe

1.7. LA MESURE DES SURCAPACITES A PARTIR DE LA FONCTION DE COUTS

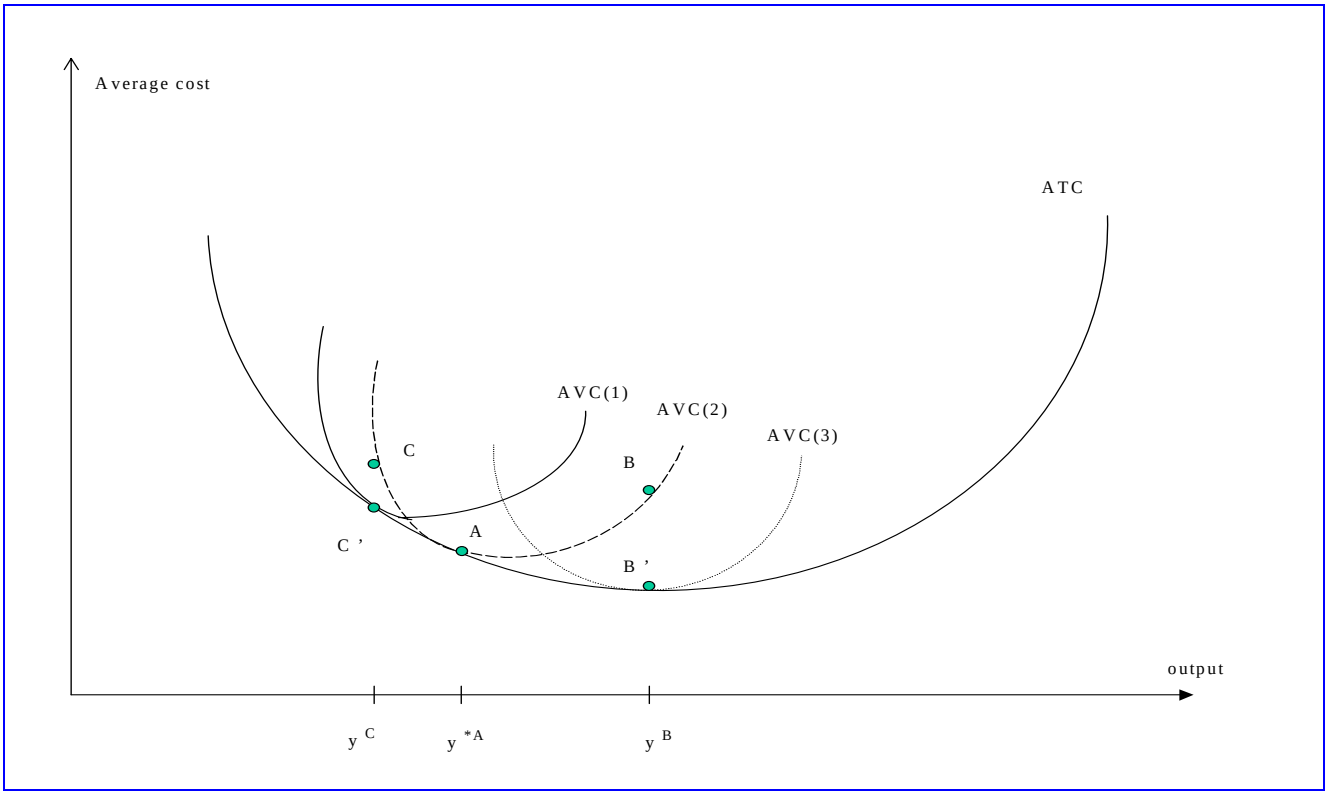
1.7.1. La mesure

Une mesure dite « primale » de la capacité d'utilisation d'une entreprise (CU_p) consiste à mesurer l'écart entre son niveau de production effectif actuel Y et le niveau de production potentiel Y* qu'elle peut atteindre en l'état actuel de ses capacités de production, c'est-à-dire en l'état actuel de ses actifs fixes, en supposant qu'elle minimise le coût moyen de court terme 19. Ainsi, $CU_p = Y/Y^*$. Si $CU_p > 1$, cela signifie que $Y > Y^*$, auquel cas l'entreprise est incitée à accroître ses investissements pour réduire ses coûts. À l'inverse, si $CU_p < 1$, alors $Y < Y^*$, et l'entreprise est incitée à réduire ses investissements. Si $CU_p = 1$, le stock d'actifs fixes dont elle dispose est celui qui minimise le coût moyen de production de la quantité courante Y et elle n'est donc pas incitée à le modifier.

Cette mesure est illustrée sur la figure 1 dans le cas à un « input » quasi fixe (c'est-à-dire fixe à court terme, mais variable à long terme) et un « output ». Sur la figure, les courbes AVC(i) représentent trois courbes de coût moyen de court terme différentes, correspondant à trois niveaux différents de l'« input » quasi fixe. La courbe ALT représente la courbe du coût moyen de long terme, construite en supposant que tous les « inputs » sont variables. On note que l'entreprise C ne minimise pas son coût moyen de court terme. Ses surcapacités en actif quasi fixe sont mesurées par : $y^C/y'^A < 1$. L'entreprise B est dans la situation opposée de sous-capacité. Elle pourrait atteindre une situation de pleine capacité en augmentant la quantité d'actif quasi fixe. Ce faisant, elle réduirait son coût moyen de court terme et viendrait se positionner au point B'. Les entreprises représentées par les points A, C' et B' opèrent à pleine capacité.

Cette mesure de la surcapacité en termes des niveaux de production est celle des comptables nationaux. Elle repose en ce cas sur des enquêtes régulières sur les capacités utilisées par les entreprises. Mais elle n'est pas facile à mettre en œuvre au niveau microéconomique, car elle suppose de disposer de nombreuses données, ce qui entraîne des coûts importants. C'est pourquoi on utilise une mesure « duale ».

Figure 1 : Illustration dans le cas d'une technologie à un « input » fixe et un « output »



19 Graphiquement, cela correspond au point de tangence entre la courbe de coût moyen de court terme et la courbe de coût moyen de long terme.

Cette mesure duale consiste à évaluer le surcroît de coûts imputable au fait que l'on produit le niveau de production Y alors que l'on dispose des capacités permettant de produire le niveau Y*. C'est donc la différence entre, d'un côté, le « coût potentiel » de production de l'« output » effectif Y, qui correspond à une situation dans laquelle l'entreprise minimiserait le coût moyen de court terme 20, et de l'autre, le « coût effectif ».

Cette approche présente l'intérêt de fournir une mesure du taux d'utilisation des capacités en termes des prix des « inputs » et qui est cohérente avec l'hypothèse de minimisation des coûts. Cette mesure consiste à comparer les prix actuels des divers « inputs », ou prix de marché, aux prix optimaux, ou « prix de référence » (« shadow prices »). Ces derniers sont ceux qui correspondent à l'équilibre de long terme dans lequel le coût total de court terme est égal au coût total de long terme. Si les prix de marché sont supérieurs aux prix fictifs, on en déduit que l'entreprise est incitée à réduire ses capacités de production, c'est-à-dire à réduire ses actifs quasi fixes. En d'autres termes, elle est en situation de surcapacité 21.

On trouve le vecteur des prix de référence des « inputs » quasi fixes en dérivant le coût total de court terme par rapport aux « inputs » quasi fixes et en supposant que l'entreprise est à l'équilibre de court terme. La mesure duale de la capacité CU_d est alors définie par le ratio du coût de référence total par rapport au coût effectif total :

$$CU_d = \frac{\mathrm{TC}^*}{\mathrm{TC}} = \frac{\mathrm{CV}(\mathrm{Y}, \omega_x, \mathrm{Z}) + \omega^{*\prime}_z \mathrm{Z}}{\mathrm{CV}(\mathrm{Y}, \omega_x, \mathrm{Z}) + \omega^{\prime}_z \mathrm{Z}} = 1 + \frac{(\omega^*_z - \omega_z)' \mathrm{Z}}{\omega'_x \mathrm{X} + \omega'_z \mathrm{Z}} \quad (5)$$

où Y = (Y_1, Y_2,…, Y_p) est le vecteur des « outputs » produits par une entreprise, X = (x_1, x_2,…, x_p), le vecteur de ses « inputs » variables, Z = (z_1, z_2 ,…, z_k), celui des « inputs » quasi fixes, ω= (ω_{x1},ω_{x2}, …,ω_{xp} ; ω_{z1},..,ω_{zk}) = (ω_x,ω_z) représente le vecteur des prix de marché des divers « inputs » et ω_z^* est le vecteur des prix de référence des « inputs » quasi fixes.

Si CU_d <1, cela signifie que, pour réaliser un niveau de production Y, le coût total de long terme est inférieur au coût total de court terme. L'entreprise est incitée à réduire ses actifs quasi fixes. Elle est en situation de surcapacité. De façon équivalente, cela signifie que les prix actuels des actifs fixes sont supérieurs aux prix de référence. Une situation de surcapacité est ainsi définie comme une situation dans laquelle les prix actuels des actifs quasi fixes sont trop élevés par rapport aux prix de référence. Dans le cas d'une technologie à un « input », si le prix de référence est supérieur (inférieur) au prix actuel, il n'y a pas d'ambiguïté quant au résultat : CU_d est supérieur (inférieur) à 1. Dans le cas où il existe plusieurs « inputs » quasi fixes, en revanche, la relation entre les prix de référence et les prix courants est plus complexe. Par exemple, si CU_d < 1, il est seulement possible de dire qu'au moins l'un des facteurs quasi fixes est sur-utilisé, ce qui explique l'existence de surcapacités.

Dans la perspective définie par cette approche, la mesure des surcapacités bancaires peut être réalisée à partir de l'estimation d'un modèle économétrique des coûts bancaires.

1.7.2. Le modèle économétrique des coûts bancaires

La fonction des coûts bancaires estimée est une fonction translog à deux facteurs quasi fixes. La fonction se présente comme suit :

$$\begin{aligned} Ln\,\mathrm{VC}_{\mathrm{it}} &= \alpha 0 + \sum_{\mathrm{j}} \alpha_j \,\mathrm{Ln}\,\omega_{\mathrm{jit}} + \sum_{\mathrm{j}} \sum_{j'} \alpha_{j\,\mathrm{j}'} \,\mathrm{Ln}\,\omega_{\mathrm{jit}}\,\mathrm{Ln}\,\omega_{\mathrm{j'it}} + \\ &\sum_{\mathrm{h}} \beta_h \,\mathrm{Ln}\,\mathrm{Y}_{\mathrm{hit}} + \sum_{\mathrm{h}} \sum_{h'} \beta_{hh'} \,\mathrm{Ln}\,\mathrm{Y}_{\mathrm{hit}}\,\mathrm{Ln}\,\mathrm{Y}_{\mathrm{h'it}} + \sum_{j} \sum_{h} \eta_{\mathrm{jh}} Ln\omega_{jit} LnY_{hit} + \\ &\sum_{\mathrm{l}} \theta_{\mathrm{l}} Ln\,\mathrm{Z}_{\mathrm{lit}} + \sum_{\mathrm{l}} \sum_{\mathrm{l'}} \theta_{ll'} \,Ln\,\mathrm{Z}_{\mathrm{l\,it}} Ln\,\mathrm{Z}_{\mathrm{l'it}} + \sum_{\mathrm{j}} \sum_{\mathrm{l}} \delta_{jh} Ln\omega_{jit} LnZ_{lit} + \\ &\sum_{\mathrm{h}} \sum_{\mathrm{l}} \gamma_{\mathrm{hl}} \mathrm{Ln}\,\mathrm{Y}_{\mathrm{hit}} Ln\,\mathrm{Z}_{\mathrm{lit}} + \sum_{\mathrm{k}} \psi_{\mathrm{k}} D_{kit} + \mathrm{v}_{\mathrm{it}} + \mathrm{u}_{\mathrm{it}} \qquad (6) \end{aligned}$$

20 Et qui correspond, à nouveau, au point de tangence entre la courbe de coût moyen de court terme et la courbe de coût moyen de long terme.

21 Une présentation formelle du modèle des coûts est fournie dans l'article de Chaffai M. et Dietsch M. « Capacity-Utilization and Cost-Efficiency in the European Banking Industry », CEPF-IEP de Strasbourg, (septembre 1999).

où VC est le coût variable, qui est mesuré par la somme des salaires et charges sociales, des coûts financiers et des autres coûts variables non associés à l'utilisation du capital physique. Les Y_i sont les produits bancaires : (1) les dépôts d'épargne sur livrets et les comptes à terme, (2) les crédits, (3) les actifs de placement et d'investissement et (4) les commissions. Ces dernières apportent une mesure approchée du montant des autres services rendus par la banque à l'ensemble de ses clients. Les ω_j sont les prix unitaires des « inputs » variables : (1) le prix du travail, mesuré en rapportant les salaires et charges au nombre d'employés, (2) le coût moyen des ressources à court terme empruntées par la banque sur les marchés, (3) le coût moyen des ressources empruntées à long terme sur les marché. Z_1 et Z_2 sont les « inputs » quasi fixes. Z_1 représente les dépôts à vue. Le choix des dépôts à vue comme actif fixe peut être justifié de manière simple. Dans la banque, le développement des dépôts à vue (dans le sens des « core deposits ») peut être considéré comme un investissement commercial à long terme. Le second actif quasi fixe Z_2 représente le stock de capital physique. Faute de pouvoir mesurer directement ce stock à partir des données comptables, nous l'avons estimé en rapportant le flux des dépenses que l'on peut identifier dans les bases Banco et Bafi comme étant liées à l'usage d'actifs physiques quasi fixes (c'est-à-dire à l'utilisation des immeubles, de l'informatique et des autres équipements bancaires) à un coût d'opportunité du capital physique, qui est ici le coût des ressources longues empruntées sur le marché. Le terme v est un terme d'erreur symétrique et u > 0 est un terme d'erreur asymétrique qui représente les inefficiences de coût.

Cette fonction de coûts doit vérifier certaines conditions de régularité : homogénéité et symétrie (ces contraintes sont imposées dans le modèle), concavité par rapport aux prix des « inputs » variables et convexité par rapport aux prix des actifs quasi fixes (ces conditions ne peuvent être imposées sans perte de flexibilité). Des degrés de liberté supplémentaires peuvent être obtenus si on suppose que les banques minimisent les coûts et maximisent aussi les profits. Cela conduit à introduire les équations de parts suivantes dans le modèle :

$$M_{jit} = \alpha_{\mathrm{j}} + \sum_{\mathrm{j'}} \alpha_{j'} \,\mathrm{Ln}\,\omega_{\mathrm{j'it}} + \sum_{\mathrm{h}} \eta_{\mathrm{jh}} \;\mathrm{Ln}Y_{hit} + \sum_{\mathrm{l}} \delta_{j\mathrm{l}}\, Ln\,\mathrm{Z}_{\mathrm{lit}} \qquad (7)$$

$$S_{hit} = \beta_{\mathrm{h}} + \sum_{\mathrm{h'}} \beta_{\mathrm{hh'}}\,\mathrm{Ln}\,Y_{\mathrm{h'it}} + \sum_{\mathrm{j}} \eta_{jh} \;\mathrm{Ln}\omega_{jit} + \sum_{\mathrm{l}} \gamma_{lh} Ln\,\mathrm{Z}_{\mathrm{lit}} \qquad (8)$$

où les M_j sont les parts des « inputs » variables dans les coûts variables et les S_h les parts des revenus issus des activités bancaires dans les coûts variables. Dans l'estimation, nous avons retenu la part des revenus tirés des crédits dans les coûts variables.

Le modèle complet comprenant la fonction des coûts variables et les équations de parts a été estimé en utilisant la méthode itérative SUR (« seemingly unrelated regression »). Nous avons aussi estimé une version de long terme du même modèle, dans laquelle tous les « inputs » sont considérés comme variables, afin de mesurer les inefficiences de coût. Nous utilisons aussi ce modèle pour établir la nature des relations entre les surcapacités et l'inefficience coût.