Análisis sobre concentración y economías de escala en la industria bancaria dentro de la literatura económica (el caso de la banca mexicana)*

Rafael Bouchain Gallcia•

Introducción

A partir de los setenta el sistema bancario mexicano experimentó una fuerte tendencia a la concentración vía fusiones entre los distintos bancos. Para dar una idea de la celeridad de dicho proceso, presentamos los siguientes datos:

Años	*Número de Bancos*	*Reducción porcentual*
1970	240	—
1975	139	42%
1979	100	28%
1982	60	40%
1988	18	80%

* El autor agradece los valiosos comentarios del Dr. Antonio Gutiérrez, así como la colaboración de Susana Carreón.

• Investigador Asociado del Instituto de Investigaciones Económicas, UNAM.

Este proceso de concentración bancaria que caracterizó al comportamiento del sector, ha sido promovido además por las autoridades centrales en dos etapas: de 1974-1976 a 1982 fue inducida por la autorización a la banca especializada para integrarse como banca múltiple; y de 1982 a 1988 cuando la nacionalización bancaria propició una rápida concentración.

A través de la concentración, el Gobierno Federal ha buscado la racionalización de la estructura de la industria bancaria (particularmente durante la nacionalización), tanto en el número como en la escala de las instituciones, por lo que se ha permitido la fusión de bancos pequeños, relativamente atrasados, con altos costos de operación y sin acceso a economías de escala, con otros de mayor tamaño.

¿Por qué la búsqueda de economías de escala? Para lograr, por un lado, una minimización de costos y un mejoramiento en la rentabilidad del negocio bancario con vistas a enfrentar la competencia en mejores condiciones, sobre todo ante la virtual apertura del sistema financiero mexicano; asimismo consolidar la presencia regional de los bancos. Por otro lado, la sociedad debiera esperar una reducción en los márgenes de intermediación (precio de los servicios), y una mejor calidad y variedad de los servicios bancarios, lo cual sería posible en la medida en que se pudiera consolidar un tamaño óptimo de dichos bancos.

En la economía industrial, el análisis de las economías de escala resulta relevante para la determinación del tamaño mínimo eficiente de la planta y permite verificar en qué dirección, a determinados niveles altos o bajos de concentración, las empresas sufren deseconomías de escala, las cuales redundan en mayores costos de operación y menores márgenes de competencia.

Trasladado ese tipo de análisis al sector bancario, el objetivo del presente trabajo es realizar una revisión de las investigaciones realizadas sobre concentración y economías de escala aplicadas al análisis de la industria bancaria, ya que sus resultados son útiles en el estudio de las estructuras de los mercados financieros, y pueden ser retomados en el diseño de la política financiera. Dichos trabajos destacan, entre otros aspectos, la definición de las funciones de producción y de costos así como las variables que pueden ser incorporadas, sin dejar de subrayar algunas de las dificultades que presenta la información estadística en su diseño. Se hará referencia a trabajos realizados principalmente en Estados Unidos, ya

que sólo existen tres estudios para el caso mexicano, los cuales analizan el periodo previo a la nacionalización de la banca. Esperamos que una continuación del presente pueda realizarse sobre el periodo que cubre la nacionalización de la banca, retomando las experiencias expuestas aquí.

Para los fines de este trabajo se hace, primero, un repaso de los principales elementos involucrados en la definición del concepto de economías de escala; luego se exponen algunas ideas derivadas de la literatura pionera en este tipo de análisis, para abordar, más tarde, las líneas de investigación de los escasos trabajos realizados para el caso de México (los cuales se inspiraron en aquella bibliografía), exponiendo sus resultados. En la última parte presentamos la metodología más recientemente utilizada, que incluye la definición de funciones translogarítmicas que, como veremos, permiten una mejor especificación de la variedad de productos bancarios (resultado de la innovación de servicios financieros) y de insumos, más acorde con la liberalización financiera que se viene experimentando.

Definición de conceptos

El concepto de economías de escala

Las economías de escala se encuentran asociadas con dos cuestiones: la eficiencia en la asignación de los recursos al elevar el nivel del producto de una empresa, y el ahorro (economía) del gasto en dichos recursos. De esta manera las economías de escala comprenden, tanto las economías técnicas o rendimientos a escala, que se deducen de la forma de la función de producción, y las economías monetarias que resultan de incluir la función de costos respectiva.

Los rendimientos a escala comprenden las leyes de producción que describen las posibilidades técnicamente viables de combinar y variar las proporciones de todos los factores para la expansión de la producción en el largo plazo; si hacemos referencia al corto plazo estamos aludiendo a la ley de los rendimientos decrecientes del factor variable, o la ley de proporciones variables, en la que se supone que por lo menos el factor fijo permanece constante mientras se aumenta el o los factores variables.

De esta manera, al analizar las causas del incremento de la producción en el largo plazo (en el cual presuponemos la variación de las cantidades utilizadas de todos los factores en las mismas proporciones), hemos de ubicarnos en las leyes de los rendimientos a escala, las cuales se presentan en tres casos generales: si la producción se incrementa en la misma proporción que el aumento de los insumos, se dice que existen rendimientos a escala constantes; si el aumento del producto está dado en una menor proporción que el de los factores, resultan los rendimientos decrecientes a escala, y, si se produce un incremento más que proporcional en el producto respecto de los insumos, existen rendimientos crecientes a escala.

Las economías de escala también se pueden dividir en pecuniarias y reales, las primeras se derivan de pagar precios inferiores por los factores empleados en la producción y distribución, hecho que se encuentra relacionado con las empresas mayores que adquieren grandes cantidades de factores, (o que resultan meramente monetarias, por ejemplo la reducción del interés); por su lado, las economías reales se encuentran ligadas a la reducción del volumen de insumos utilizados, esto es, materias primas, capital y mano de obra.

Hay y Morris,[1] con base a un estudio de Haldi y Whitcomb, mencionan tres fuentes de economías de escala: primero, las que se relacionan con la reducción de costos de unidades individuales resultado del uso de indivisibilidades cuando se eleva la capacidad de operación del equipo, así como del buen conocimiento de las propiedades geométricas y de las capacidades; segundo, las economías en los costos de plantas y áreas de proceso, así cuando cada unidad de equipo se construye a un tamaño óptimo, una alta escala de utilización puede permitir la especialización y la división del trabajo; y, tercero, relacionado con los costos de operación, se tiene la especialización al asignar a los trabajadores en actividades especiales, de manera que el incremento en el tamaño de la empresa va a permitir economías en el equipo de mantenimiento, donde los costos por máquina se reducen a medida que aumenta el número de máquinas operando, de la misma manera se pueden esperar economías en el uso de materiales.

[1] Hay, Donald y Derek J. Morris. *Industrial Economics and Organization,* Oxford University Press, 1991, cap. Costs and supply conditions.

Adicionalmente podemos incorporar otros elementos relacionados con las economías de escala. Una corrida larga de producción puede permitir la reducción de los costos, particularmente los fijos; y las economías de alcance son una forma particular si queremos analizar una empresa que elabora varios productos a la vez, de manera que el costo de producir dos productos al mismo tiempo resulta menor que el de producirlos por separado.

Las operaciones multiplanta también representan economías de escala no atribuíbles a la operación de una simple planta (Bain), Scherer presenta algunas razones que explican y justifican la existencia de este tipo de operaciones multiplanta: la reducción de los costos de transporte que implica la existencia de mercados dispersos; la necesidad de incrementar la capacidad durante el tiempo, así una empresa puede elegir entre producir a una gran o pequeña escala, ya que si se eleva la demanda y la capacidad resulta insuficiente, se origina un costo penalizador que será transferido a los consumidores; y por último tenemos la posibilidad de especialización en diferentes plantas (diversificación) y la flexibilidad en la producción.

Ahora bien, la elevación en la escala implica gastos o deseconomías en los costos de administración, pero éstos pueden influir en una mejor organización que tiene lugar en la gran empresa de la actualidad, razón por la cual tales costos pueden estar relacionados con la eficiencia.

Según Koutsoyiannis[2] las posibles fuentes de economías de escala serían:

A. De producción. a) En la mano de obra: especialización y perfeccionamiento, ahorro de tiempo, automatización y por valor acumulado, b) en la técnica y el capital: especialización e indivisibilidades, acondicionamiento de maquinaria a fines múltiples, costos iniciales, relación técnica entre "volumen de maquinaria" e insumos que la producen, capacidad instalada de reserva en márgenes y obras encargadas de reparaciones, c) de inventarios; repuestos, materias primas y productos terminados.

[2] Kautsoyiannis, A. *Microeconomía moderna,* Buenos Aires, Argentina, Amorrortu editores, 1985, caps. 3 y 4.

B. De venta o de comercialización: publicidad, propaganda en gran escala, distribución, departamentos de servicio y cambios de modelos.

C. Gerenciales: especialización y experiencia del trabajo en equipo, descentralización, mecanización y técnicas gerenciales que ahorran tiempo.

D. Transporte y almacenamiento.

La función de producción

Una función de producción (que incluye solamente los métodos técnicamente eficientes), describe la relación técnica entre insumos factoriales y volúmenes de producción por unidad de producto, la isocuanta es la línea que presenta dichos métodos técnicamente eficientes a un determinado nivel de producto. Esta isocuanta muestra la sustituibilidad de dichos factores, en diversas formas: lineal, que supone una perfecta sustituibilidad de factores; quebrada, que supone sustituibilidad limitada entre los factores capital y trabajo (isocuanta de programación lineal) y convexa o lisa, que supone sustituibilidad continua en un cierto intervalo. La función de producción puede ser descrita así:

$$q = F(X_1, \ldots, X_n)$$

$$q = f(\textit{de producción})$$

donde:

$$X_1, \ldots, X_n = \textit{insumos}$$

La función de costos

Tradicionalmente la teoría enfatiza la escala de operaciones como el mejor determinante de los costos y debido a que los empresarios pagan por el uso de los recursos, una empresa producirá un nivel de producto seleccionando el método o técnica que minimice su gasto (costo) en recursos, si el precio de nuestros insumos está dado por w entonces nuestra función de costo queda:

$$c = w_1X_1 + \ldots + w_nX_n$$

$$c = f(\textit{de costo})$$

donde

$$w_1, \ldots, w_n = \textit{precio-de-insumos}$$

la función de costos puede expresarse como:

$$c = f(q)$$

y la curva de costo promedio:

$$\frac{c}{q} = \frac{f(q)}{q}$$

La forma de la curva de costo va a depender tanto de la forma de la función de producción (eficiencia tecnológica) como de los precios relativos de los insumos. En la teoría tradicional, las curvas de costo promedio de corto y de largo plazos tienen la forma de U, pero no se puede afirmar que en el corto plazo la empresa esté minimizando sus costos, pues dicha minimización solo se puede observar en la curva de costo promedio de largo plazo, esta última nos va a relatar los puntos de la curva de costo promedio de corto plazo en su tendencia en el largo plazo, así suponemos que únicamente en el largo plazo la firma va a maximizar utilidades minimizando costos.

Nuestro interés se vuelca en observar la función de costo de largo plazo, la cual nos mostrará si es que existe incremento o decremento en los costos al elevar la escala de operaciones. Esta curva de costo de largo plazo también se denomina curva de planificación, que suponemos en forma de U pero aplanada.

Por último, debemos señalar que las economías de escala van a determinar la forma de la curva de costo promedio de largo plazo, es decir en su forma se van a reflejar las leyes de rendimientos a escala antes descritos (constantes, crecientes o decrecientes).

Análisis de economías de escala en la industria bancaria

En este campo encontramos una amplia literatura en inglés que analiza los casos de Estados Unidos, Gran Bretaña y Canadá. El avanzado desarrollo de los sistemas financieros en los países industrializados hace que este tipo de investigaciones resulten útiles tanto en la observación del desempeño de las firmas financieras al interior del mercado, como para la evaluación de la política financiera por parte de las autoridades, particularmente la que se relaciona con las fusiones, transferencias y subsidios.

En nuestra opinión, existen por lo menos dos problemas centrales en la definición del análisis de economías de escala en la industria bancaria. Por un lado la dificultad en la definición del concepto de "producto bancario", donde se puede proceder de dos maneras: a través de la homogeneización del producto bancario en un solo producto, cuya naturaleza es "heterogénea", o bien realizando una especificación de tipo multiproducto acorde con las economías de alcance por medio de una función translogarítmica. El segundo problema consiste en la existencia de una gran variedad de procedimientos para la especificación de la función de costos, los cuales también dependen, tanto de la calidad y cantidad de las variables a incluir en dicha función, como de las restricciones que existen por el lado de la información disponible.

Los primeros trabajos se centraron en la determinación del costo unitario (promedio) de operación como variable dependiente. Alhadeff (1954)[3] seguido por Horvitz (1963),[4] analizaron las economías de escala mediante una relación entre el costo total de operación (por dólar en préstamo y valores) con el tamaño de los bancos (clasificados por tipos y montos de depósitos). Estos análisis arrojaron evidencias de economías de escala crecientes para el estrato inferior de los bancos estadounidenses en 1963, costos constantes para el estrato intermedio y deseconomías para el estrato superior.

[3] Alhadeff, D.A. *Monopoly and Competition in Banking*, Berkeley: University of California Press, 1954.

[4] Horvitz, P.M. "Economies of Scale in Banking", en *Private Financial Institutions*. Englewood Cliff, N.J., Prentice-Hall, 1963.

Schweiger y McGee (1961)[5] y Gramley (1962)[6] aplicaron el análisis de regresión múltiple donde la variable dependiente fue el costo total de operación como porcentaje de los activos totales, esta última representó la variable independiente más importante, los resultados para 1961 evidenciaron la declinación de los costos hasta que los bancos alcanzaron activos que definieron el estrato superior, y a partir de ahí aunque los costos siguieron declinando lo hicieron de una manera más lenta.

Brigham y Grebler (1963),[7] en un estudio similar al de Gramley, corroboraron la existencia de economías de escala. Una mención especial requiere el extenso trabajo de George Benston (1965, 1972 y 1982)[8] quien en opinión de González Méndez fue quien formalizó de mejor manera y por primera vez la relación entre economías de escala y costo marginal.

En 1968, Bell y Murphy[9] utilizaron una función cobb-douglas para la especificación del producto bancario, determinado exógenamente bajo el supuesto de que los bancos minimizan su costo de operación. El uso de esta metodología dio pie a los trabajos de Daniel, Longbrake y Murphy (1973)[10] que incluyeran funciones cobb-douglas para tres tipos de tecnología bancaria mostrando deseconomías en los bancos que carecieron de computadoras, Murray y White (1980)[11] de manera similar encontraron economías de escala para las instituciones de depósito del Canadá durante

[5] Schweiger, I. y J.S. McGee. "Chicago Banking: The Structure and performance at Banks and Related Financial Institutions in Chicago and other Areas". *The Journal of Bussiness*, 34, july; 201-306, 1961.

[6] Gramley, L.E. *A Study of Scale Economies in Banking*, Kansas City, Mo. Federal Reserve Bank of Kansas City, 1962.

[7] Brigham, E.F. y L. Grebler. *Savings and Mortgage Markets in California*, Pasadena, California, Savings and Loan League, 1963.

[8] Benston, George J. "Economies of Scale and Marginal Cost in Banking Operations". *National Banking Review* 2(4), june; 507-49, 1965. "Economies of Scale of Financial Institutions". *Journal of Money, Credit and Banking*, (4), may; 312-41, 1972. Benston, G., G. Hanweck and D. Humphrey. "Scale Economies in Banking: A Restructuring and Reassessment". *Journal of Money*, Credit and Banking (14), november; 435-56, 1982.

[9] Bell, D. W., y N.B. Murphy. "Economies of scale and Division of Labor in Commercial Banking," *Southern Economic Journal*, 35:131-39, october, 1968.

[10] Daniel, D.L., W.A. Longbrake and N.B. Murphy. "The Effect on Technology on Bank Economies of Scale for Demand Deposits". *Journal of Finance*, 28, march; 1931-46, 1973.

[11] Murray, J.D. y R.W. White. "Economies of Scale and Deposit-Taking Financial Institutions in Canada". *Journal of Money, Credit and Banking*, 12(1), febrero, 1980.

1972-1975, y por último también un análisis de Longbrake y Haslem (1975)[12] que se caracteriza por el uso de este tipo de funciones.

Investigaciones sobre la determinación de economías de escala en la industria bancaria mexicana

En contraste con la extensa literatura sobre el tema consagrada a los países desarrollados, solo encontramos tres trabajos de investigación para el caso mexicano. El texto pionero en el análisis fue el de Héctor González Méndez titulado "Economías de Escala y Concentración Bancaria: el caso de México" (1981)[13] que utilizó estimaciones de regresión. A este estudio le sigue "Comportamiento de la Función de Costos de la Banca Múltiple y Alternativas de su Evolución", del mismo autor, realizado en 1981[14] en el que se incluye una función cobb-douglas. Por último se encuentra el trabajo de Enrique González Sánchez titulado "Economías de escala en la industria bancaria mexicana", publicado en 1988.[15]

Abordaremos dichos trabajos por orden cronológico con la finalidad de lograr una mejor coherencia y secuencia en su presentación.

Los modelos

En su primer trabajo, González Méndez define un concepto de planta (unidad básica o complejo de producción) no restringido al establecimiento bancario, ya que la escala de producción del mercado financiero se encuentra ligada a la cobertura geográfica, y

[12] Longbrake, W.A. y J.A. Haslem. "Productive Efficiency in Commercial Banking". *Journal of Money, Credit and Banking*, 7, august; 317-30, 1975.

[13] González Méndez, Héctor. "Economías de Escala de Concentración Bancaria: El caso de México". *Monetaria*, CEMLA, 4(1) enero-marzo; 73-79, 1981.

[14] González Méndez, Héctor. "Comportamiento de la Función de costos de la Banca Múltiple y Alternativas de su Evolución". *Banco de México*, Subdirección de Investigación Económica, Documento No. 36, septiembre de 1981, 33 p.

[15] González Sánchez, Enrique. "Economías de Escala en la Industria Bancaria Mexicana". *Monetaria*, CEMLA, 13(2) abril-julio; 235-41, 1990.

el tamaño depende del número de oficinas. Así identifica la planta con la firma. Cabe señalar que Benston y Longbrake han considerado los costos de operación entre sucursales, ya que resulta más adecuado con las características regionales de la legislación bancaria estadounidense.

El análisis se realizó para los años de 1974 y 1978, debido a que representan puntos de equilibrio, antes y después del decreto de constitución de la banca múltiple. El autor clasificó los costos totales en costos no financieros (CNF), más cercanos a los costos de operación (gastos en factores: trabajo y administración), y los costos financieros (CF), que incluyen los pagos por intereses a los depositantes regulados por las autoridades, por lo que no se encuentran sujetos a cambios con la escala de operaciones; a éstos últimos agregó los gastos en publicidad y propaganda previamente descontados de los costos de operación. Ésta última manipulación resultó de considerar que tales gastos corresponden a la política de penetración y no se encuentran sujetos necesariamente a minimización. De esta manera, los costos no financieros se acercan al costo de operación.

El modelo puede especificarse mediante la siguiente ecuación:

$$CTU = CF + CNF$$

Después, con el fin de eliminar problemas en la medición del producto bancario debidos a la existencia de un producto diferenciado, el autor consideró pertinente incluir las siguientes variables:

1. El volumen de recursos de operación independientemente del origen de los depósitos, lo que define la variable de escala de operación.
2. El número de cuenta-habientes y/o el saldo promedio de los depósitos, pues a un mismo volumen de recursos existe un costo mayor o menor respecto del número de cuentas.
3. Las diferencias entre los precios de los factores de producción entre bancos, aunque supuso que los mismos se contrataron en mercados competitivos y por lo tanto no se presentaron diferencias entre sus precios.
4. La influencia de las diferencias de la estructura de plazos de los pasivos sobre el costo de captación.

González Méndez propone la medición de la escala de operaciones con base al tamaño del producto de cada banco respecto del total, y tomando en cuenta su participación en el número total de cuenta-habientes, ya que una mayor atomización de los depósitos tiende a elevar el costo.

Como resultado, la variable de escala o medida relativa del producto quedó definida como:

$$E_j = \left[\frac{RT_j}{\sum_{j=1}^{m} RT_j}\right]\left[\frac{TCH_j}{\sum_{j=1}^{m} TCH_j}\right]$$

RT_j representa los recursos del banco j,
TCH el número total de cuenta-habientes del banco j,
m el número total de bancos y $O<Ej<1$.

Para especificar la heterogeneidad de los servicios bancarios, el autor definió una variable de liquidez, o de plazo de los pasivos, la cual permite una mejor definición del comportamiento de los costos, lo que se logra al derivar la importancia relativa de los depósitos en cada institución respecto del total de la cartera pasiva. Tenemos:

$$P_j = \sum_{i=1}^{n}\left[\frac{IC_{ij}}{\sum_{i=1}^{n} IC_{ij}}\left(\frac{PC_i}{PM}\right)\right]$$

IC_i representa al instrumento de captación i del banco j,
PC_i el plazo de captación del instrumento i,
PM el plazo máximo de captación y $O<Pj<1$ (representa liquidez total y 1 total iliquidez).

El modelo a estimar quedó especificado de la siguiente manera:

$$CTU = CF + CNF$$

$$CF = F(E,P)$$

$$CNF = f(E,P)$$

En su segundo trabajo González Méndez se propuso la determinación de la minimización de costos con base a una función de producción cobb-douglas como la siguiente:

$$MinC = rK + wL + mT + \lambda(Q - \psi k^{\alpha} L^{\beta} T^{\gamma})$$

C representa el costo total de operación,
K el capital,
L el trabajo,
T las materias primas;
r el costo del capital;
w la tasa de salarios,
m el costo de materias primas y
Q el producto.

Para estimar esta función se definieron dos cuestiones: los supuestos respecto de los factores de la función y la definición del producto bancario (Nerlove, 1965[16] y Walters, 1963[17]). El primer problema fue resuelto suponiendo que tanto el capital como las materias primas (administrativas) presentaron costos uniformes entre los establecimientos bancarios (Bell y Murphy, 1968).[18]

La definición del producto bancario representó un problema más complejo. Powers (1969)[19] y Benston (1972)[20] han propuesto un método para homogeneizar el producto bancario entre establecimientos; esta idea consiste en especificar el producto bancario en términos de unidades monetarias (pesos intermediarios) independientemente de la naturaleza de su origen o destino (moneda nacional o extranjera). Así resulta un producto homogéneo (peso intermediado), que va a depender de costos heterogéneos, que a su vez dependen de "factores estructurales".

La ecuación de costos resultante es la siguiente:

[16] Nerlove, M. *Estimation and Identification of Cobb-Douglas Production Functions*, Chicago Rand McNally Co., 1965.

[17] Walters, A.A. "Productions and Cost Functions: An Econometric Survey". *Econométrica*, enero-abril, 1-66, 1963.

[18] Bell, F.W. y N.B. Murphy. *Op. cit.*

[19] Powers, J.A. "Branch Versus Unit Banking: Bank Output And Cost Economies". *Southern Economic Journal*, October; 153-64, 1969.

[20] Benston, G. *Op. cit.*

$$C = AQ^{\alpha}{}_{i}w^{\beta}_{v}\, p^{\mu 1}F^{\mu 2}(ICC)^{\mu 3}R^{\mu 4}G^{\mu 5}e^{\mu}$$

donde:

$$V = \alpha + \beta + \gamma$$

representan los parámetros de economías de escala. Además se tuvieron que definir las siguientes variables:

a) P = plazo promedio de la cartera pasiva, determinado por:

$$P = \sum_{i=1}^{n}\left[\frac{D_i}{\sum_{i=1}^{n} D_i}\left(\frac{PC_i}{PM}\right)\right]$$

b) F = plazo promedio de la cartera activa, dado por:

$$F = \sum_{i=1}^{n}\left[\frac{TP_j}{\sum_{j=1}^{n} TP_j}\,(TA_j)\right]$$

c) ICC = índice de concentración del crédito que mide la distribución de la cartera activa por tipo de clientes:

$$ICC = \sum_{k=1}^{m}\left[\frac{\left(\frac{IC_k}{CT}\right)^2}{\sum_{k=1}^{m}\left(\frac{IC_k}{CT}\right)\left(\frac{IU_k}{IU}\right)}\right]$$

d) R = riesgo de operaciones activas:

$$R = \left(\frac{CU}{CC}\right)^2$$

e) G = velocidad de crecimiento anual de los recursos totales:

$$G = \frac{Q_t - Q_{t-1}}{Q_{t-1}}$$

D_i es el iésimo instrumento de captación,
PC_i el plazo de captación del iésimo instrumento,
PM plazo máximo de captación (104 semanas),
TP_j el tipo de préstamo,
TA_j plazo promedio del tipo de préstamo,
IC_k intervalo de crédito,
IU_k intervalo de acreditados,
CU cartera vencida y
CC cartera de crédito.

Adicionalmente, González Méndez realiza la siguiente consideración: un agregado amplio del producto bancario (a partir de los recursos totales) no permite estipular el sentido en el que se puede producir el agotamiento de las economías de escala, así como de las variables que contribuyen al mismo. Por lo tanto, retomando el planteamiento de Longbrake y Haslem (1975),[21] propone un método para la medición del producto bancario, y captar las formas e influencias de su incremento, resultando la siguiente ecuación:

$$Q = NSH$$

Donde N es el número de cuenta-habientes por oficina bancaria,
S el número de oficinas bancarias y
H es el saldo promedio por cuenta-habiente.

El modelo que resulta permite identificar las economías o deseconomías de escala como derivadas de las alternativas de ampliación de los recursos totales, de manera que:

$$C = AN^{\delta 1}S^{\delta 2}H^{\delta 3}w^{\beta}_{v}P^{\mu 1}F^{\mu 2}(ICC)^{\mu 3}G^{\mu 4}e^{\mu}$$

donde:

$$A = (v)\,(\Psi^{1}_{v}\alpha^{\alpha}_{v}\beta^{\beta}_{v}\gamma^{\gamma}_{v})^{-1}$$

Los parámetros $\delta 1$, $\delta 2$ y $\delta 3$ se pueden interpretar de la misma manera que 1/v; si los valores de los mismos son mayores que 1 indican deseconomías y viceversa.

[21] Longbrake, W.A. y J.A. Haslem. *Op. cit.*

El tercer texto de Enrique González Sánchez, analiza las economías de escala en la banca mexicana previo a la nacionalización bancaria.

El autor utiliza en su investigación tres métodos: el primero corresponde al enfoque de elasticidades de Humphrey, el segundo se denomina técnica de la "empresa sobreviviente", y por medio del último aplica el índice de concentración Herfindahl-Hirshman.

Enrique González Sánchez efectúa un análisis de corte transversal para una muestra de 24 bancos en 1980, año en el que se supone se encontraban realizadas las principales transformaciones producidas por la constitución de la banca múltiple, además de representar un año de relativa estabilidad económica. En el modelo los costos representaron la variable dependiente, de la única variable independiente: los activos totales.

El propósito fue mostrar el cambio de la elasticidad costo a medida que aumentan los activos bancarios, utilizando una ecuación cuadrática del logaritmo de los costos (C) en función del logaritmo de los activos totales; la ecuación de regresión es la siguiente:

$$lnC = a + b(lnA) + \frac{C1}{2}(lnA)^2$$

la elasticidad costo se deriva de la siguiente manera:

$$\frac{dlnC}{dlnA} = b + c(lnA)$$

Un segundo método fue utilizar la técnica de la "empresa sobreviviente", el cual supone que las empresas más eficientes se seleccionan ellas mismas en el proceso de competencia, y a través del tiempo. El procedimiento parte de clasificar las empresas por tamaño, calculando su participación en el producto total de la industria en el tiempo, si la participación declina significa la existencia de un tamaño insuficiente lo que revela un mayor costo de operación.

El autor utilizó la información de los activos bancarios presentada por Eckaus[22] para los cálculos correspondientes a los años de

[22] Eckaus, Richard S. "The structure of the commercial banking system in Mexico, 1940-70", en Organización de Estados Americanos, *Papers on the Capital Markets Development Program*, Washington, junio de 1974, pp. 54-66.

1949, 1960 y 1970 y del anuario de la Asociación Mexicana de Bancos para el año de 1980.

El tercer método corresponde a un análisis de concentración a través del índice Herfindahl-Hirshman (índice H), que representa simplemente la suma de los cuadrados de la participación de cada empresa en el mercado, dicha participación fue medida por el monto de activos de cada banco. La fórmula es la siguiente:

$$H = \sum_{i=1}^{n}\left(\frac{A_i}{\sum A_i}\right)^2$$

A_i es el monto de los activos del banco i, el índice varía de 0 a 1, y a medida que el índice se acerca a la unidad ($H = 1$) la industria se encuentra dominada por una empresa, o aumenta el grado de concentración industrial.

Las características generales de los resultados de la aplicación de esta medición expresan: a) H aumenta a medida que las desviaciones respecto del tamaño promedio aumentan y a medida que el número de bancos disminuye, esto es:

$$H = \sum\left(\frac{A_i}{\sum A_i}\right)^2$$

$$H = \sum\left(\frac{A_i}{\sum A_i}\right)^2 - \left(\frac{2\sum A_i}{N\sum A_i}\right)\frac{1}{n} + \frac{1}{n}$$

$$H = \sum\left[\frac{A_i^2}{(\sum A_i)^2} - \left(\frac{2A_i}{n\sum A_i}\right) + \frac{1}{n^2}\right] + \frac{1}{n}$$

$$H = \sum\left(\frac{A_i}{\sum A_i} - \frac{1}{n}\right)^2 + \frac{1}{n}$$

$$H = \sum \left[\frac{A_i}{\sum A_i} - \left(\frac{\sum A_i}{\sum A_i} \right) / n \right]^2 + \frac{1}{n}$$

$$H = \sum \left[\frac{A_i}{\sum A_i} - \left(\frac{\sum A_i}{\sum A_i} \right) / n \right]^2 + \frac{1}{n}$$

donde el primer término es la suma al cuadrado de las desviaciones respecto a la participación promedio y 1/n es el inverso del número de bancos (el índice aumenta cuando el primer término aumenta y disminuye con *n*); *b*) el inverso de *H* muestra el número de bancos de igual tamaño que proporcionarían el valor del índice; *c*) *H* no revela la participación de los bancos en el mercado; *d*) *H* es una buena estimación de la concentración con poca información; *e*) *H* puede modificarse para encontrar mejores ajustes estadísticos.

Resultados de los modelos

El primer modelo de González Méndez mostró los siguientes resultados:

- Se utilizaron los logaritmos de las variables exógenas para la especificación de modelos de regresión lineales y cuadráticos en dos versiones, mínimos cuadrados ordinarios y a través del modelo de covarianza con una variable dicotómica.
- Se presentó la evidencia de constantes positivas lo que mostró un costo fijo unitario de operación pero muy cercano al origen, y el signo positivo de la variable dicotómica del modelo de covarianza evidenció que el costo fijo unitario se elevó entre 1974 y 1978.
- Las variables de liquidez y escala explican más del 80% del comportamiento de los *CNF*.
- El *CNF* (costo de operación) es función inversa de la escala y estable entre ambos periodos, así mismo el *CNF* decrece a medida que el plazo de los pasivos se incrementa, y dicho cambio

porcentual se ha hecho más acentuado en el tiempo, lo que puede denotar un encarecimiento de los factores de producción.
- El costo financiero se eleva con la escala de operación de los bancos, esto podría indicar que los bancos más grandes encarecen el costo social debido a la desviación de sus objetivos de maximización hacia el incremento y mantenimiento de su participación en el mercado (gastos en publicidad y propaganda).
- Las pruebas *t* muestran diferencia en los costos fijos que crecen, mientras que en los variables no hay mayor diferencia.
- Las ecuaciones cuadráticas definen una función de costos en forma de U, decrecientes y crecientes a escala, lo que puede conducir a la definición de un tamaño óptimo. Así mismo la curva de costo de largo plazo se desplazó hacia la derecha y hacia arriba.
- Los resultados estadísticos permiten afirmar al autor que existieron economías de escala por lo menos hasta que los bancos alcanzaron el tamaño óptimo de alrededor del 3% de los recursos totales del sistema, y que dicho tamaño se elevó entre 1974 y 1978 (desplazamiento de la curva de escala). El tamaño óptimo se desplazó del 2.75 al 3.63% de los recursos de 1974 a 1978.
- La estructura de plazos afectó las funciones de costos financieros y no financieros, la mayor carga financiera que significaron estos movimientos pudo afectar a los bancos pequeños, de manera que en conjunto, los movimientos de escala y liquidez pudieron presionar hacia una mayor concentración.
- La concentración bancaria presentó un índice muy elevado (46.6% de los recursos en dos instituciones), y se redujeron las condiciones competitivas del sistema bancario, los bancos se redujeron a la mitad entre 1976 y 1978 con la integración de la banca múltiple (de 240 a 100 de 1970 a 1979), así se comprobó la existencia de un mercado oligopólico en el que un poco más de 15 instituciones controlaron el 90% de los activos bancarios.
- Por último, la existencia de economías de escala por encima del tamaño económico viable y el desplazamiento de éste hacia arriba permiten al autor afirmar que existieron razones para pensar en un ahorro financiero proveniente de la fusión de las pequeñas instituciones, ya que operaban con costos elevados.

El segundo modelo de González Méndez arrojó los siguientes resultados:

- La estimación logarítmica de la primera ecuación del modelo se realizó para tres paquetes de bancos, superior, inferior y el total de bancos. Se observó la existencia de economías de escala en el estrato inferior de bancos y deseconomías en los bancos del estrato superior, pero aunque las economías parecen agotarse, en los bancos más grandes lo hacen en una proporción muy reducida, así un incremento en 10% en los recursos de los 18 grandes bancos sólo elevó el costo en 10.9%, dejando poco espacio para las deseconomías.

— Como se mencionó anteriormente, un agregado amplio del producto no permite verificar en qué sentido se agotan las economías, por lo que se incluyó la especificación de Q como el producto de tres variables NSH que representan en su orden, el número de cuenta-habientes por oficina, el número de oficinas bancarias por establecimiento y el saldo promedio por cuenta-habiente; un valor de los parámetros mayor que uno evidencia deseconomías y viceversa.

— Los coeficientes de N y H resultaron estadísticamente significativos, es decir existieron economías de escala en relación al número de cuenta-habientes y al saldo promedio de los depósitos, en contraste, el número de sucursales repercutió en forma proporcional sobre el costo (coeficiente igual a la unidad).

— Los bancos encontraron mayores economías al ampliar el tamaño promedio de las cuentas pasivas destacando los bancos más grandes del sistema; así, los clientes con saldos voluminosos encontraron más atractivo trabajar con bancos grandes que con pequeños.

— Los bancos con escala de operación intermedia, que emergieron del amplio proceso de fusiones son los que presentaron costos de operación más bajos y por lo mismo menor congestionamiento; el análisis de los 23 bancos más pequeños reflejó la existencia de economías de escala, aunque mínimas; y el segmento de los 18 bancos más grandes reflejó costos crecientes. Este último resultado sugirió la necesidad de impedir que los bancos más grandes siguieran creciendo, ya que esto tiende a repercutir en mayores costos.

A su vez, los resultados de los tres modelos de Enrique González Sánchez fueron los siguientes:

A. En el modelo de elasticidades, el autor calculó 24 elasticidades para el año de 1980 y encontró resultados similares a los del primer modelo de González Méndez, es decir que tanto los bancos medianos como los grandes presentaron economías de escala ($e<1$). El autor derivó de Galbraith y Schumpeter los beneficios que se pueden obtener de una estructura oligopólica: una mayor investigación y avance tecnológico, pues una empresa necesita tener suficiente control sobre el mercado para aprovechar los beneficios de la innovación, ya que si los competidores pudieran imitarla, rápidamente se eliminarían los incentivos para el cambio.*

B. En cuanto a la aplicación de la técnica de la empresa sobreviviente, se presentó una declinación de la participación de los bancos con menos del 5% del total de los activos, mientras los bancos con más del 5% correspondiente, la aumentaron. Esto confirma la existencia de deseconomías de escala en los bancos pequeños y de economías de escala en los bancos medianos y grandes.

En el decenio 1970-1980, los bancos más pequeños perdieron participación; los poseedores del dos al 5% de los activos mantuvieron la suya; los medianos la aumentaron y los dos más grandes la redujeron. Esto se encuentra relacionado con la introducción de la banca múltiple y permite al autor deducir que favoreció a los bancos de tamaño mediano a expensas de los más pequeños y los más grandes.

Una conclusión importante deriva del hecho de que no existieron bancos con activos entre el 10 y el 20% del total, lo cual indicó que éste nunca fue de un tamaño factible, y que permite deducir que quizá existieron barreras que impidieron pasar del nivel del 10% de activos, ya que tan solo dos bancos se encontraron sobre el 20 por ciento.

C. Respecto de la aplicación del índice Herfindahl-Hirshman se extrajeron las siguientes conclusiones:

— A pesar de que se observó una disminución en la concentración en 1980, la industria estuvo lejos de considerarse compe-

* Estudios similares de Hunter y Timme revelaron que los bancos más grandes no tuvieron una mayor tasa de progreso tecnológico aunque dicha tasa se maximiza con el tamaño de banco más grande.

titiva. El sistema bancario se caracterizó por tener una estructura oligopólica en la cual los dos bancos más grandes tuvieron una influencia considerable.
— El inverso del índice H muestra el número de bancos al igual tamaño que proporcionarían el valor respectivo del índice; éste fluctuó entre 6 y 7 confirmando la estructura oligopólica.
— Posiblemente la reglamentación, en especial la relacionada con la fijación de las tasas de interés por las autoridades monetarias, haya influido para evitar el aumento de la participación de los bancos más grandes.

Con base a las conclusiones a que arribó, el autor afirma que la creación de bancos gubernamentales de mediano tamaño se realizó con la idea de disminuir la concentración.

Aplicación de funciones translogarítmicas para la determinación de economías de escala

Al parecer, el primer trabajo que incluye una función de tipo translogarítmica, mediante la cual es posible incluir una variedad de productos o servicios bancarios, corresponde a Mullineaux (1978),[23] (sobre el uso de dicha función este autor cita a Diewert 1973).[24] En él Mullineaux formalizó la estimación de una función de ganancias, como forma de medir las economías de escala y la eficiencia organizacional en 951 bancos de Estados Unidos en 1971-1972. Este estudio se basó en análisis previos de McFadden y Lau, los cuales partieron de la interrelación entre funciones de ganancias y funciones de producción.

La función de ganancia expresa la maximización de beneficio de una firma en situación competitiva como función de los precios del producto y de los insumos factoriales (variables y fijos), de manera que la eficiencia económica se puede descomponer en eficiencia técnica y de precios, y en ciertos casos es posible identificar la fuente de dicha eficiencia.

[23] Mullineaux, Donald J. "Economies of Scale an Organizational efficiency in banking: a profit-function approach". *The Journal of Finance*, 33(1) march; 259-80, 1978.

[24] Diewert, W.E. "Functional Forms for Profit and Transformation Functions". *Journal of Economic Theory*, (6) june; 284-316, 1973.

La función de transformación es $t(Y,X,F) = 0$ donde Y es un vector m dimensional de una variedad de productos bancarios, X un n vector de insumos variables y F un vector de insumos fijos. Las ganancias (McFadden) se definieron como la diferencia entre los ingresos totales y los costos de los factores variables, de manera que $\pi = R'Y - P'X$, donde π son las ganancias, R es un vector de precios exógenos y tasas de interés de los servicios bancarios y P vector de precios de insumos y tasas de interés.

La función translogarítmica es la siguiente:

$$ln\pi = a_0 + \sum_{i=1}^{n} a_i lnP_i + \sum_{j=i}^{n} b_j lnq_j + \sum_{m=1}^{i} s_m lnv_m$$

$$+ \frac{1}{2}\sum_{m=1}^{i}\sum_{j=1}^{i} h_{mj} lnv_m lnv_j + \sum_{k=1}^{w} C_k Z_k$$

Otro trabajo que incluye la estimación de las economías de escala y las economías de alcance mediante una función translogarítmica es el de Murray y White (1983).[25] En él se destaca la ventaja de esta técnica para la inclusión de un caso multiproducto como es el bancario.

La minimización de costos en el modelo queda especificada de la siguiente manera:

$$MinC = \sum_{j=1}^{m} p_j x_j$$

Sujeto a la función de producción:

$$F(y_0 \ldots y_n; x_1, \ldots, x_m) = 0$$

donde C es el costo total del capital, trabajo, depósitos y acciones, p_j el precio unitario del insumo j, x_j la cantidad del insumo j, y_i la cantidad del producto i.

[25] Murray, John D., and Robert White "Economies of Scale and Economies of Scope in Multiproduct Financial Institutions: A Study of British Columbia Credit Unions". *The Journal of Finance*, 38, june 887-902, 1983.

La función translogarítmica queda:

$$lnC = \alpha_0 + \sum_{i=1}^{n} \alpha_i lny_i + \sum_{j=1}^{m} \beta_j lnp_j + \frac{1}{2}\sum_{j}^{n}\sum_{k}^{n} \sigma_{ik} lny_i lny_k$$

$$+ \frac{1}{2}\sum_{j}^{m}\sum_{k}^{m} \gamma_{jk} lnp_j lnp_k + \sum_{i}^{n}\sum_{j}^{m} \delta_{ij} lny_i lnp_j$$

que es una función de costo que debe ser linealmente homogénea en todos los precios de los insumos, cóncava en p_j, y creciente en y_i y p_j. Las condiciones de homogeneidad se satisfacen cuando:

$$\sum_j^m \beta_j = 1, \ \sum_j^m \delta_{ij} = 0, \ \sum_j^m \gamma_{jk} = 0$$

Las economías de escala que resultan están dadas por un factor común @ que se obtiene diferenciando la ecuación con respecto a y^i:

$$\eta = \frac{dlnC}{@} = \sum_i^n \frac{\partial lnC}{\partial lny_i}$$

$$\eta = \sum_i^n \alpha_i + \sum_k^n \sigma_{ik} lny_k + \sum_i^n + \sum_j^m \delta_{ij} lnp_j$$

si

$$\eta$$

es mayor que *1* existen deseconomías de escala, igual que *1* rendimientos constantes y menor que 1 economías de escala.

Los trabajos de Zardkoohi, Rangan y Kolari (1986)[26] y Loreta Mester (1987)[27] también han utilizado y destacan la utilidad de

[26] Zardkoohi, Asghar, Nanda Rangan and James Kolari. "Homogeneity Restrictions on the translog cost model: a note". *The Journal of Finance*, 41(5) december; 1153-55, 1986.

[27] Mester, Loreta J. "A Multiproduct Cost Study of Savings and Loans". *The Journal of Finance*, 42(2) june; 423-45, 1987.

las funciones translogarítmicas en el análisis del sistema bancario, que a su vez se caracteriza por la existencia de múltiples productos.

Por último tenemos el trabajo de Daniel Gropper (1991)[28] que aplica el análisis translogarítmico para la deducción de los cambios de economías de escala en los bancos comerciales para el periodo 1979-1986 su especificación es muy similar a las anteriores pero destaca la utilización de datos recientes, con lo que trata de analizar las modificaciones a partir de las fuertes innovaciones financieras que se produjeron en los ochenta. Una de las conclusiones que se resaltan en este análisis se deriva precisamente de los efectos del cambio tecnológico sobre el reciente proceso de innovación financiera.

Consideraciones finales

Las diferentes metodologías utilizadas en investigaciones sobre concentración y economías de escala se venían aplicando al análisis del sector industrial propiamente dicho, pero como podemos constatar existe un interés creciente por utilizar estas herramientas en estudios sobre la estructura de mercado del sector financiero, particularmente del bancario.

Si bien los trabajos pioneros datan de la segunda mitad de los cincuenta y la primera de los setenta, en la actualidad se encuentran numerosos estudios referentes a los países industrializados los cuales van aplicando las innovaciones metodológicas y las más recientes técnicas matemáticas. Por otra parte, aunque los trabajos correspondientes a países subdesarrollados son escasos, podemos afirmar que existe un creciente interés por este tipo de investigaciones, fundamentalmente a raíz del crecimiento del sector bancario iniciado en los setenta y en relación a los efectos de la crisis de la deuda externa de principios de los ochenta.

Las líneas de investigación seguidas para el estudio de la industria bancaria mexicana han arrojado mediciones precisas sobre la estructura de dicho mercado. Destacan, entre otros, los siguientes resultados: la constatación de un ágil proceso de con-

[28] Gropper, Daniel M. "An Empirical Investigation of Changes in Scale Economies for the Commercial Banking Firm, 1979-1986". *Journal of Money, Credit and Banking*, 23(4) november; 718-27, 1991.

centración bancaria, la existencia de una estructura oligopólica dominada fundamentalmente por los dos bancos más grandes; el establecimiento de barreras a la entrada de los bancos para pasar al estrato entre el 10 y 20% de los activos totales, el desplazamiento de los costos fijos, y, por último, la existencia de economías de escala para los bancos medianos, y menores márgenes de economías en los grandes y chicos, de esta manera fue posible determinar un tamaño óptimo de banco.

Podemos afirmar que los resultados arrojados por los modelos aplicados a México son de suma utilidad para el análisis del proceso que va de los setenta a principios de los ochenta, cuando se produjo la nacionalización, pero hace falta abrir una línea de investigación que permita actualizar dicho análisis. Es posible que en la actualidad la estructura de la industria bancaria mexicana conserve ciertos elementos característicos del periodo previo, pero consideramos que existen algunos elementos novedosos, tales como los efectos de la ágil concentración durante la nacionalización bancaria, el repunte de Serfín con activos de alrededor del 17%, mientras Banamex y Bancomer conservan una participación superior al 20% en ese rubro, la competencia de las casas de bolsa en el mercado de dinero y, por último, la constitución de los grupos financieros.

Consideramos que una continuación de la investigación sobre esta línea debiera incorporar la especificación de funciones translogarítmicas, pues estas permiten analizar la existencia de productos múltiples que es una de las características de los procesos de innovación y modernización financiera actuales.