

EL PROBLEMA DE LA AGREGACIÓN SECTORIAL EN EL MARCO INPUT-OUTPUT

Carmen Ramos Carvajal; cramos@correo.uniovi.es

Esteban Fernández Vázquez; evazquez@correo.uniovi.es

Rubén Álvarez Herrero; rherrero@correo.uniovi.es

Departamento de Economía Aplicada. Universidad de Oviedo

RESUMEN

Un aspecto de gran relevancia en los estudios input-output es el de la agregación sectorial utilizada, ya que según como se haya realizado ésta se pueden obtener resultados notablemente diferentes.

El objetivo de esta comunicación es cuantificar los efectos que se desprenden de utilizar un mayor o menor nivel de agregación, para ello introduciremos el concepto de *Pérdida de la cantidad de información*, derivado de la teoría estadística de la información. Dicho concepto será establecido a partir de la entropía y de la cantidad de información cuadráticas. Con el fin de comprender la naturaleza de la pérdida experimentada en la cantidad de información, propondremos una descomposición de esta medida a partir de los distintos factores que la constituyen. Se concluirá el desarrollo teórico señalando algunas de sus propiedades más relevantes.

Por último, y con el fin de ilustrar los aspectos anteriormente señalados, cuantificaremos, a partir de las tablas input-output de la región asturiana, la cantidad de información asociada a cada matriz y la pérdida experimentada en la misma según los distintos niveles de agregación empleados.

1-MOTIVACIÓN

Un aspecto relevante en los estudios input-output es el grado de agregación con el que se trabaja. Frecuentemente podemos consultar estudios aplicados con un grado de agregación bastante elevado. Lo cual, por un lado, permite una “simplificación” en el manejo de los datos, pero, por otro,

constituye una fuente de error, ya que al agregar a un número muy reducido de sectores estaremos considerando ramas constituidas por sectores muy dispares¹.

En esta comunicación pretendemos enfatizar en este aspecto y analizarlo con algún detalle. En suma, consideraremos qué repercusiones tiene trabajar con un nivel u otro de agregación. Para efectuar este análisis vamos a utilizar la teoría estadística de la información, la cual proporciona un amplio conjunto de herramientas, no sólo relacionadas con la estimación de coeficientes input-output², sino también con la fiabilidad de estas estimaciones o con la diferenciación del grado de desagregación de una tabla³.

En la presente comunicación expondremos algunos conceptos básicos de teoría de la información, a partir de los cuales desarrollaremos nuestro cometido. En concreto, nos referiremos a la entropía y cantidad de información de Shannon y cuadrática.

Posteriormente, introduciremos el concepto de pérdida de la cantidad de información y analizaremos algunas de sus propiedades. Efectuaremos una descomposición de la misma en tres componentes: la aportación de las filas, de las columnas y de las celdas.

Asimismo se analizará la relación existente entre dicha pérdida de información y la dispersión existente en la tabla.

Concluiremos con un análisis empírico de la tabla input-output de Asturias de 2000, sobre la que se aplicarán los conceptos anteriormente presentados.

2-ALGUNOS ANTECEDENTES DE LA TEORÍA DE LA INFORMACIÓN

Incertidumbre, en la acepción más usual de la palabra, es la falta de certeza ante una determinada situación, la cual puede ser debida a dos razones fundamentalmente: o bien se trata de una experiencia no realizada nunca anteriormente o bien los resultados de la misma son de carácter

¹ Puede verse el análisis efectuado por Szyrmer (1989) en el que se considera la influencia del nivel de agregación en la estimación de coeficientes mediante el método RAS.

² Golan, A., G. Judge y S. Robinson (1994): Recovering information from incomplete or partial multisectoral economic data. *The Review of Economics and Statistics*. Nº 76, pág. 541-549.

³ Ver Ramos, Álvarez, Fernández y García (2003): Algunas aplicaciones de la Teoría Matemática de la información al Análisis Input-Output. XI Jornadas de ASEPUMA, Oviedo.

aleatorio. Nosotros, en términos generales, adoptaremos esta segunda posibilidad.

La incertidumbre imposibilita conocer el resultado de una experiencia antes de llevarla a cabo, en este sentido, disminuir la incertidumbre supone la obtención de alguna información referente al resultado concreto que puede aparecer. Por ello, la información representa una disminución de la incertidumbre.

Fisher (1922) propone una definición cuantitativa de la cantidad de información contenida en los datos que se obtienen cuando realizamos un experimento aleatorio. La cantidad de información que una muestra extraída de una población X contiene sobre un parámetro θ que toma valores en un conjunto abierto Θ de $\Re$ y del que depende la función de distribución $F(x)$ es

$$I_F(X, \theta) = E \left[\left(\frac{\partial \ln f(x, \theta)}{\partial \theta} \right)^2 \right]$$

donde $f(x, \theta)$ es la función de densidad de la variable para ese valor del parámetro.

Con el desarrollo de las telecomunicaciones surge la necesidad de cuantificar la información que se transmite a través de un canal de comunicación. En 1924 Nyquist publicó un trabajo en el que se señalaba la relación existente entre la velocidad telegráfica (V) y el número de símbolos enviados (n) o sucesivos valores de la corriente:

$$V = k \log n$$

Donde k es una constante que depende del número de valores sucesivos que se envíen en un segundo.

Basándose en trabajos de Nyquist y Kupfmüller, Hartley (1928) demostró que para transmitir una determinada cantidad de información, un canal de comunicación tiene que someterse a un intercambio entre su duración y el ancho de banda. La información se definió como una selección arbitraria de símbolos procedentes de un conjunto previamente establecido.

La medida de la información fue definida en términos del logaritmo del número de símbolos igualmente disponibles para la comunicación. Hartley en 1928 realizó un trabajo en el que considera al transmisor de un mensaje equipado con un conjunto de símbolos de los que selecciona uno a

continuación de otro, generando así una secuencia de los mismos. Definió la información de un mensaje como el logaritmo del número de posibles secuencias de símbolos que pueden seleccionarse:

$$I = \log N^k = k \log N$$

Donde I es la cantidad de información, k representa el número de símbolos del mensaje y N el número de símbolos del alfabeto. Si se toma $n = N^k$, se puede interpretar que el número posible de secuencias es el número de resultados de un experimento aleatorio, donde todas las secuencias tienen la misma probabilidad de ser elegidas. Es decir, si fuese independiente la selección de los símbolos tendríamos

$$\underbrace{N * N * N \dots * N}_k = N^k$$

Sin embargo, esta medida presenta algunos inconvenientes ya que no tiene en cuenta las probabilidades correspondientes a los resultados y, por lo tanto, concede el mismo valor a una experiencia con dos resultados equiprobables que a otra cuyas probabilidades respectivas son 0.99 y 0.01

A partir de todas estas contribuciones Shannon y Wiener (1948) iniciaron el nacimiento de la teoría de la información. La idea central de la teoría matemática de la comunicación de Shannon-Wiener es que la comunicación es un proceso estadístico, que puede ser descrito únicamente en términos probabilísticos. Un mensaje no llevará información si es posible predecirlo. Cuando un mensaje tenga la misma probabilidad de ser emitido que cualquier otro, el mensaje llevará la máxima información. Es decir, se trata de una medida que viene asociada, en cierta manera, con nuestra libertad de elección.

3-MEDIDAS DE ENTROPÍA

Shannon basándose en el indicador de Hartley, propuso como medida de incertidumbre asociada a un experimento aleatorio, una función que denominó entropía⁴.

⁴ Esta palabra deriva del griego y significa “replegarse hacia el interior”.

3.1.1. La entropía de Shannon

Sea X una variable aleatoria discreta con distribución de probabilidad $P=(p_1, p_2, \dots, p_n)$, se denomina entropía de Shannon de dicha variable o de la distribución P a la expresión siguiente:

$$H(X) = H(p_1, p_2, \dots, p_n) = -\sum_{i=1}^n p_i \log p_i$$

La entropía de Shannon está acotada inferiormente por cero, alcanzándose este valor cuando la variable es degenerada, esto es, cuando son nulas $n-1$ probabilidades y la restante es la unidad. Alcanzará su mayor valor, $\log(n)$, en el caso de que todas las probabilidades coincidan, $p_1=p_2=\dots=p_n=\frac{1}{n}$, esto es, cuando la distribución es uniforme.

La entropía de una distribución puede ser entendida como el desorden existente en la misma, es decir, la incertidumbre asociada a un determinado fenómeno.

Consideremos ahora una variable aleatoria bidimensional discreta (X, Y) que toma valores (x_i, y_j) , $\forall i=1, 2, \dots, n$ y $\forall j=1, 2, \dots, t$, con probabilidades $p(x_i, y_j) \geq 0$, tales que

$$\sum_{i=1}^n \sum_{j=1}^t p_{ij} = 1$$

Se puede definir la entropía asociada a dicha variable del siguiente modo:

$$H(X, Y) = -\sum_i \sum_j p_{ij} \log p_{ij}$$

El comportamiento de esta medida es análogo al visto para una variable.

3.1.2. La entropía cuadrática

Existen diferentes generalizaciones de la medida de entropía, una de las cuales es debida a Havrda y Charvat (1967), los cuales propusieron la entropía de grado β .

Consideremos una población sobre la que se define una variable aleatoria X que toma unos valores $(x_1, \dots, x_n)$ cuyas probabilidades asociadas serán $(p_1, \dots, p_n)$, donde $\sum_{i=1}^n p_i = 1$.

Se puede establecer la siguiente definición de entropía de grado β ($\beta \neq 1$), $\beta > 0$:

$$H^\beta(X) = H^\beta(p_1, \dots, p_n) = \frac{1}{2^{1-\beta} - 1} \sum_i (p_i^\beta - p_i)$$

Esta medida coincide con la entropía de Shannon cuando β se aproxima a la unidad. Para el valor concreto $\beta=2$ surge la denominada entropía cuadrática; así pues, dada una variable aleatoria X que lleva asociado un sistema de probabilidades $\{p_1, p_2, \dots, p_n\}$ se define la entropía cuadrática a partir de la expresión

$$H^2(X) = 2 \sum_i p_i (1 - p_i) = 2 \left[1 - \sum_i p_i^2 \right]$$

Esta medida se halla acotada entre cero y dos, alcanzando su menor valor en el caso de que la distribución sea degenerada y el máximo cuando las probabilidades son coincidentes.

Dado que nuestro interés se centra en la determinación de la incertidumbre asociada a una tabla input-output, consideraremos una variable aleatoria bidimensional, que presenta una probabilidad conjunta p_{ij} .

La entropía cuadrática toma la expresión siguiente:

$$H^2(X, Y) = 2 \sum_i \sum_j p_{ij} (1 - p_{ij}) = 2 \left[1 - \sum_i \sum_j p_{ij}^2 \right]$$

Dicha medida se encuentra acotada entre cero y dos, con interpretación análoga a la señalada para una variable unidimensional.

3.1.3. Cantidad de información

A partir del concepto de incertidumbre podemos definir el de cantidad de información que la variable aleatoria Y contiene sobre X , la cual puede interpretarse como la reducción de la incertidumbre de X entre la situación inicial y la que se produce después de conocer Y .

La expresión de la cantidad de información será:

$$I(X, Y) = H(X) - H(X/Y) = H(X) + H(Y) - H(X, Y)$$

donde $H(X/Y)$ representa la entropía condicionada de X dada la variable Y .

Operando convenientemente llegamos a la expresión siguiente⁵:

$$I(X, Y) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j}$$

La cantidad de información es mayor o igual que cero y no superior a la entropía de X , por lo tanto, podemos escribir

$$0 \leq I(X, Y) \leq H(X) = \log(n)$$

Dada una v.a. bidimensional (X,Y) , las probabilidades asociadas serán p_{ij} $\forall i=1,2,\dots,n$ y $\forall j=1,2,\dots,n$, por lo que la cantidad de información cuadrática se puede definir de la siguiente forma

$$I^2(X,Y)=2\left[1-\sum_i p_i^2-\sum_j p_j^2+\sum_i\sum_j p_{ij}^2\right]$$

Esta medida se encuentra acotada entre cero y dos.

4-CONEXIÓN DE UNA TABLA INPUT-OUTPUT CON LA TEORÍA DE LA INFORMACIÓN

Una tabla input-output recoge tanto las relaciones existentes entre los distintos sectores como la demanda final de los mismos.

Las ecuaciones que recogen estos flujos son:

$$\begin{aligned} X_1 &= x_{11} + x_{12} + \dots + x_{1n} + y_1 \\ X_2 &= x_{21} + x_{22} + \dots + x_{2n} + y_2 \\ &\vdots \\ X_n &= x_{n1} + x_{n2} + \dots + x_{nn} + y_n \end{aligned}$$

Donde X_i representa el output total del sector i -ésimo, x_{ij} el flujo de bienes y servicios del sector i -ésimo al j -ésimo e y_i es la demanda final del i -ésimo sector.

Las anteriores ecuaciones recogen los flujos “en filas”, pero análogamente podrían haber sido presentados “en columnas” del siguiente modo:

⁵Obsérvese que $H(y) = -\sum_i p_j \log p_j$

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{x}_{11} + \mathbf{x}_{12} + \dots + \mathbf{x}_{1n} + \mathbf{g}_1 \\ \mathbf{X}_2 &= \mathbf{x}_{21} + \mathbf{x}_{22} + \dots + \mathbf{x}_{2n} + \mathbf{g}_2 \\ &\vdots \\ \mathbf{X}_n &= \mathbf{x}_{n1} + \mathbf{x}_{n2} + \dots + \mathbf{x}_{nn} + \mathbf{g}_n \end{aligned}$$

Donde g_i se refiere a los inputs primarios del sector i -ésimo.

En concreto, si nos centramos en la submatriz de consumos intermedios que representa las transacciones en una economía, se puede establecer un sistema de probabilidades⁶, al dividir el contenido de cada celda entre el total de las mismas, esto es,

$$p_{ij} = \frac{x_{ij}}{X}$$

donde $X = \sum_i \sum_j x_{ij}$.

Por lo tanto, y si aplicamos cualquiera de las fórmulas anteriormente presentadas, podremos determinar la incertidumbre y la cantidad de información asociada a una tabla input-output.

5-AGREGACIÓN Y PÉRDIDA DE INFORMACIÓN

A continuación pasaremos a hacer una breve referencia a la información estadística empleada y posteriormente definiremos la pérdida de información asociada a la agregación sectorial.

5.1. INFORMACIÓN ESTADÍSTICA UTILIZADA: LA TIOA DE 2000

Las tablas input-output (TIO) se presentan frecuentemente en distintos niveles de desagregación; así, por ejemplo, la TIO simétrica correspondiente a Asturias referida al año 2000 aparece desglosada en 65 y 4 sectores. Esta matriz es la primera, en nuestra región, que sigue las directrices contables del SEC-95. Posiblemente, si empleamos la matriz con 65 ramas el gran volumen de información dificultaría la comprensión del objeto de nuestro estudio, mientras que si, por el contrario, nos decantamos por trabajar con 4 ramas habremos perdido mucha información relevante. Por todo ello, consideramos interesante disponer de una herramienta que permita diferenciar entre los

⁶ Este aspecto está recogido en Tilanus y Theil (1965) o en Doblado (1988), entre otros.

distintos niveles de agregación; nuestra propuesta es utilizar la cantidad de información como elemento diferenciador.

5.2. PÉRDIDA DE LA CANTIDAD DE INFORMACIÓN: DESCOMPOSICIÓN

Tomemos como punto de partida una matriz que esté constituida por n sectores. A partir de los flujos que aparecen en la tabla se determinarán las probabilidades p_{ij} como se ha señalado anteriormente.

En lo que sigue de esta comunicación nos referiremos a la cantidad de información cuadrática, ya que Doblado (1988) ha realizado un ejercicio similar a éste, aplicando la cantidad de información de Shannon.

La cantidad de información asociada a dicha tabla se obtendrá a partir de la incertidumbre cuadrática, es decir,

$$I^2(X, Y) = 2 \left[1 - \sum_i p_i^2 - \sum_j p_j^2 + \sum_i \sum_j p_{ij}^2 \right]$$

Supongamos, ahora, que se desea realizar una agregación de los sectores que constituyen la tabla, pasaremos a disponer de m grupos de filas y columnas; dichos grupos serán denotados, respectivamente, por S_g y S_h , $\forall g=1,2,\dots,m$ y $\forall h=1,2,\dots,m$. Por lo tanto, las probabilidades resultantes de la agregación (p_{gh}) variarán, en relación a las iniciales (p_{ij}), es decir, $p_{gh} \geq p_{ij} \forall i \in S_g$ y $\forall j \in S_h$. La expresión de la cantidad de información cuadrática después de la agregación será:

$$I_A^2(X, Y) = 2 \left[1 - \sum_g p_g^2 - \sum_h p_h^2 + \sum_g \sum_h p_{gh}^2 \right]$$

Obsérvese que $p_g \geq p_i, \forall i \in S_g$ y $p_h \geq p_j, \forall j \in S_h$, por lo tanto, $\forall i \in S_g$ y $\forall j \in S_h$, se verificará $p_g^2 \geq p_i^2, p_h^2 \geq p_j^2$ y $p_{gh}^2 \geq p_{ij}^2$.

La pérdida que experimenta la cantidad de información ante esta agregación puede expresarse de la siguiente manera

$$\begin{aligned} \Delta I^2(XY) &= I^2(X, Y) - I_A^2(X, Y) = \\ &= 2 \left[\left(\sum_g p_g^2 - \sum_i p_i^2 \right) + \left(\sum_h p_h^2 - \sum_j p_j^2 \right) + \left(\sum_i \sum_j p_{ij}^2 - \sum_g \sum_h p_{gh}^2 \right) \right] \end{aligned}$$

Esto es, hemos descompuesto el cambio en la cantidad de información debido a la agregación en tres términos: el cambio que se experimenta en las

filas, $\left(\sum_g p_g^2 - \sum_i p_i^2 \right)$, el que se experimenta en columnas, $\left(\sum_h p_h^2 - \sum_j p_j^2 \right)$, y el debido a las celdas, $\left(\sum_i \sum_j p_{ij}^2 - \sum_g \sum_h p_{gh}^2 \right)$.

Como señala Doblado (1988) la pérdida de información será debida a la heterogeneidad de los inputs, a la heterogeneidad de los outputs y a la de las celdas.

Con la finalidad de ilustrar este punto vamos a considerar distintas agregaciones de los sectores de la tabla de Asturias de 2000 y cuantificaremos la pérdida de información global y su descomposición entre el peso que tienen los inputs, los outputs y celdas en la variación global. Como hemos señalado anteriormente la tabla de Asturias aparece presentada a 65 y 4 sectores. Nosotros hemos procedido a efectuar una agregación intermedia a ambas, en concreto, a 31 sectores, siguiendo el modelo de la TIO de Andalucía de 1995. Dicha agregación aparece recogida en el anexo.

Cuadro N° 1. Cantidad de información asociada a diferentes niveles de agregación

$I^2(X,Y)$	TIOA-65	TIOA-31	TIOA-4
Valor	1.8207	1.7379	0.8342
% respecto a su cota	91.03%	86.90%	41.71%

Como podemos apreciar la cantidad de información contenida en las dos primeras tablas es bastante elevada en relación a su cota superior; esta cantidad va disminuyendo a medida que se agrega a un menor número de sectores, es decir, mayor agregación lleva asociada una menor cantidad de información.

A continuación vamos a cuantificar la pérdida global de la información asociada a la agregación tanto global, como desglosada en los outputs, los inputs y las celdas.

Cuadro N° 2. Pérdida de información: descomposición

	Pérdida global	Pérdida outputs	Pérdida inputs	Pérdida celdas
TIO 65-31	0.0828	12.74%	41.34%	45.92%
TIO 65-4	0.9865	22.10%	41.05%	36.85%
TIO 31- 4	0.9037	22.73%	41.02%	36.24%

Como puede apreciarse en el cuadro anterior la pérdida global tiene un pequeño peso relativo cuando se pasa a agregar de 65 a 31, y toma valores bastante superiores en las otras dos situaciones. En los tres casos considerados los componentes que tienen más peso son los de columnas y celdas.

A partir de la información anterior podríamos determinar qué agrupación de sectores tiene más importancia en la pérdida de información, para ello deberíamos determinar las contribuciones de los distintos conjuntos de ramas en el total. Así pues, cuando pasamos a agregar a cuatro sectores, las agrupaciones Industria y Servicios son las que proporcionan una mayor contribución a la pérdida tanto en filas como columnas. Si la agregación es a 31 ramas, los grupos con mayor peso serían Metalurgia, Servicios bancarios y Otros Servicios. En general, estas ramas tienen como rasgo común que están constituidas por un número relativamente elevado de sectores, posiblemente con unas características no demasiado homogéneas.

5.3. LA NO NEGATIVIDAD DE LA PÉRDIDA DE INFORMACIÓN

Si nos fijamos en los sumandos que constituyen la pérdida de información, se puede observar que los dos primeros son positivos (los correspondientes a inputs y outputs) y el último negativo (el referente a las celdas). A continuación, pasamos a analizar el signo de la pérdida de cantidad de información.

Sabemos que $H^2(X)+H^2(Y)> H^2(X,Y)$, por lo tanto,

$$2 \left[1 - \sum_i p_i^2 - \sum_j p_j^2 + \sum_i \sum_j p_{ij}^2 \right] > 0$$

Análogamente,

$$2 \left[1 - \sum_g p_g^2 - \sum_h p_h^2 + \sum_g \sum_h p_{gh}^2 \right] > 0$$

entonces, podemos escribir

$$1 + \sum_i \sum_j p_{ij}^2 > \sum_i p_i^2 + \sum_j p_j^2$$

$$1 + \sum_g \sum_h p_{gh}^2 > \sum_g p_g^2 + \sum_h p_h^2$$

Si sumamos ambas inecuaciones se obtiene

$$2 + \sum_g \sum_h p_{gh}^2 + \sum_i \sum_j p_{ij}^2 > \sum_g p_g^2 + \sum_h p_h^2 + \sum_i p_i^2 + \sum_j p_j^2$$

Operando convenientemente

$$2 + \sum_g \sum_h p_{gh}^2 + \sum_i \sum_j p_{ij}^2 - \sum_i p_i^2 - \sum_j p_j^2 > \sum_g \sum_h p_{gh}^2 - \sum_i \sum_j p_{ij}^2 + \sum_i p_i^2 + \sum_j p_j^2$$

A partir de donde, se deriva la desigualdad siguiente:

$$\left(\sum_g p_g^2 - \sum_i p_i^2 \right) + \left(\sum_h p_h^2 - \sum_j p_j^2 \right) > \left(\sum_i \sum_j p_{gh}^2 - \sum_g \sum_h p_{ij}^2 \right)$$

Con lo cual, podemos concluir que a medida que se agrega más, la cantidad de información no crece, sino que en general disminuirá. Por tanto, denominaremos a esta medida de cambio en la cantidad de información debido a la agregación como pérdida de la cantidad de información.

5.4. PÉRDIDA DE INFORMACIÓN Y DISPERSIÓN

Para intentar ahondar algo más en el concepto de la pérdida de información se relacionará con el de dispersión. En concreto, expresaremos dicha función en términos de la dispersión de los coeficientes y, por lo tanto, de las relaciones intersectoriales.

Tradicionalmente, la dispersión asociada a una distribución se cuantifica en relación a la media aritmética mediante la varianza, que al tratarse de una medida de dispersión cuadrática se adapta bastante bien a la medida de entropía utilizada, que también lo es.

Si consideramos una variable discreta (X), la expresión de la varianza de las participaciones marginales, será

$$S_X^2 = \left[1 - \frac{H^2(X)}{2} \right] \frac{1}{n} - \frac{1}{n^2} = \frac{1}{2n} \left[2 \left(1 - \frac{1}{n} \right) - H^2(X) \right]$$

La cual se obtiene a partir a partir de la expresión

$$S_x^2 = \sum_i \frac{p_i^2}{n} - \left(\sum_i \frac{p_i}{n} \right)^2$$

Es decir, se ha establecido la varianza como la diferencia entre la cota superior de la entropía cuadrática⁷ y su valor observado, multiplicado por una constante.

Si ahora, expresamos la entropía cuadrática, en términos de la varianza se tiene:

$$H^2(X) = 2 \left[1 - nS_x^2 - \frac{1}{n} \right] = 2 \left[\left(1 - \frac{1}{n} \right) - nS_x^2 \right]$$

Esta expresión representa, la diferencia entre la cota superior de la entropía, esto es, la máxima entropía y la varianza multiplicada por una constante. A partir de lo anteriormente expuesto, se aprecia que la varianza puede ser entendida como una medida opuesta de la entropía, cuando la dispersión es mínima la entropía es máxima y viceversa.

A continuación vamos a calcular la varianza asociada, a los inputs, los outputs y las celdas, a cada una de las tablas anteriormente señaladas.

Cuadro N° 3. Cálculo de la varianza

	S^2 celdas	S^2 outputs	S^2 inputs
TIOA 65	2.9501E-06	0.00041	0.00069
TIOA 31	2.0334E-05	0.00106	0.00173
TIOA 4	0.0091	0.03876	0.03399

A partir del cuadro anterior se puede apreciar que existe una mayor dispersión en los inputs que en los outputs o en las celdas.

Si calculamos ahora la entropía se puede apreciar el comportamiento opuesto, que ya habíamos señalado, con respecto a la varianza:

⁷ El término $2 \left(1 - \frac{1}{n} \right)$ es la cota máxima de la entropía cuadrática que se aproxima a 2 cuando n toma un valor muy elevado.

Cuadro N° 4. Cuantificación de la entropía

	H ² celdas	H ² outputs	H ² inputs
TIOA 65	1.9746	1.9156	1.8797
TIOA 31	1.9604	1.8697	1.8287
TIOA 4	1.5838	1.1899	1.2281

A partir de lo anteriormente señalado podemos expresar la pérdida de la cantidad de información en términos de la varianza. Así pues,

$$\Delta I^2(XY) = \Delta H^2(X) + \Delta H^2(Y) - \Delta H^2(XY)$$

entonces podemos escribir:

$$\Delta I^2(XY) = 2 \left[\left(\frac{1}{m} - \frac{1}{n} \right) - (nS_o^2 - mS_{OA}^2) + \left(\frac{1}{m} - \frac{1}{n} \right) - (nS_l^2 - mS_{IA}^2) + \left(\frac{1}{m^2} - \frac{1}{n^2} \right) - (n^2S_C^2 - m^2S_{CA}^2) \right]$$

donde $S_o^2(S_{OA}^2)$ representan la varianza de los outputs antes (después) de la agregación, $S_l^2(S_{IA}^2)$ es la varianza de los inputs antes (después) de la agregación y $S_C^2(S_{CA}^2)$ la varianza de las celdas antes (después) de la agregación.

Por lo tanto,

$$\Delta I^2(XY) = 2 \left[2 \left(\frac{1}{m} - \frac{1}{n} \right) + \left(\frac{1}{n^2} - \frac{1}{m^2} \right) - [(nS_o^2 - mS_{OA}^2) + (nS_l^2 - mS_{IA}^2) + (m^2S_{CA}^2 - n^2S_C^2)] \right]$$

Es decir, se puede expresar como la diferencia entre las cotas superiores de las entropía respectivas menos la diferencia de las varianzas⁸. Por lo tanto, la pérdida de información depende de dos factores: el primero de ellos, nos indica que la pérdida será menor si la agregación no es muy drástica, es decir, si no hay una fuerte disminución del número de sectores y el segundo es que cuanto más parecida sea la dispersión de las cuotas, antes y después de la agregación, menos pérdida de información se producirá.

5.5-PÉRDIDA RELATIVA DE INFORMACIÓN

Puede resultar interesante para efectuar comparaciones, acotar superiormente la pérdida de información. Para ello vamos a partir de dos

⁸ Hemos calculado la varianza a partir de las cuotas p_{ij} a partir de la fórmula siguiente

$$\sum_i \sum_j \frac{p_{ij}^2}{n} - (\bar{p}_{ij})^2$$

situaciones que difieran lo más posible en su estructura, esto es, supongamos que en la situación antes de efectuar la agregación nos hallamos en una situación donde los coeficientes son coincidentes, por lo tanto, la dispersión con respecto a la media será nula, en este caso $S_O^2 = S_I^2 = S_C^2 = 0$. Supongamos ahora que después de haber efectuado la agregación la dispersión de los coeficientes alcanza su mayor valor, que según lo anteriormente expuesto se alcanzará si la distribución es degenerada, esto es,

$$S_{OA}^2 = \frac{1}{2m} \left[2 \left(1 - \frac{1}{m} \right) \right] = \frac{1}{m} \left(1 - \frac{1}{m} \right) = \frac{1}{m} - \frac{1}{m^2}$$

Análogamente, para S_{IA}^2 y para S_{CA}^2 tendremos

$$S_{IA}^2 = \frac{1}{m} - \frac{1}{m^2} \text{ y } S_{CA}^2 = \frac{1}{m^2} - \frac{1}{m^4}$$

Por lo tanto, sustituyendo estos valores en la expresión de la pérdida de información, obtendremos

$$\Delta I^2(XY) = 2 \left[2 \left(\frac{1}{m} - \frac{1}{n} \right) + \left(\frac{1}{n^2} - \frac{1}{m^2} \right) - \left[2m \left(\frac{1}{m} - \frac{1}{m^2} \right) + m^2 \left(\frac{1}{m^2} - \frac{1}{m^4} \right) \right] \right] = 2 \left[\frac{1}{n^2} - \frac{1}{n} + 1 \right]$$

Esto es, la cota superior depende del número de sectores que hubiese antes de efectuar la agregación. Por lo tanto, podemos construir una medida relativa de la pérdida de información de la siguiente manera

$$\Delta I^2(XY) = \frac{\left[2 \left(\frac{1}{m} - \frac{1}{n} \right) + \left(\frac{1}{n^2} - \frac{1}{m^2} \right) - \left[(nS_O^2 - mS_{OA}^2) + (nS_I^2 - mS_{IA}^2) + (m^2S_{CA}^2 - n^2S_C^2) \right] \right]}{\left[\frac{1}{n^2} - \frac{1}{n} + 1 \right]}$$

dicha expresión está acotada, entre cero y uno. Si se considera en términos porcentuales, podremos determinar el porcentaje de pérdida de información debido a la agregación.

Procederemos a cuantificar la pérdida de información anteriormente calculada en términos de su cota superior, los valores obtenidos se recogen en el siguiente cuadro:

Cuadro N° 5. Pérdida de información global relativa

	Pérdida global	Pérdida global relativa
TIO 65-31	0.0828	16.81%
TIO 65-4	0.9865	48.57 %
TIO 31- 4	0.9037	45.88 %

Como puede apreciarse a partir del cuadro anterior la pérdida relativa de información que supone de pasar de agregar de 65 o 31 a 4 ramas es de más del 40%. Mientras que dicha pérdida será del 16% cuando 31 sectores en lugar de los 65 iniciales.

6- CONCLUSIONES

La teoría de la información es una herramienta de importante aplicación en el ámbito del análisis input-output. Un aspecto de interés en los estudios aplicados lo constituye la consideración de la influencia que tiene trabajar con distintos niveles de agregación.

La pérdida de la cantidad de información asociada a la agregación sectorial constituye un buen indicador que permite diferenciar entre la oportunidad de efectuar diferentes agrupaciones.

Se ha utilizado como caso de estudio la TIO de Asturias de 2000, analizando la pérdida de información en diferentes situaciones: agregación a 65, a 31 y a 4 ramas.

7-REFERENCIAS BIBLIOGRÁFICAS

- GIL, P. (1981): *Teoría matemática de la información*. Ed. ICE.
- GOLAN, A., G. JUDGE y S. ROBINSON (1994): Recovering information from incomplete or partial multisectoral economic data. *The Review of Economics and Statistics*. N° 76, pág. 541-549.
- IEA (1998): Sistema de Cuentas Económicas de Andalucía. Marco Input-Output, 1995. Publicación electrónica: <http://www.juntadeandalucia.es/institutodeestadistica/mapa/prodWeb/areapweb002.htm#101>.

- INFANTE, R. (2000): Teoría de la información. Aplicaciones, en *Perspectivas en estadística e investigación operativa*. Pascual, A. y Parras, L. Editores. Servicio de publicaciones de la Universidad de Jaén.
- PARDO, L. (1997): *Teoría de la información estadística*. Ed. Hespérides.
- PÉREZ, R. (1985): *Estimación de la incertidumbre, la incertidumbre útil y la inquietud en poblaciones finitas. Una aplicación a las medidas de desigualdad*. Tesis doctoral. Universidad de Oviedo.
- PULIDO, A. y A. FONTELA (1993): *Análisis input-output. Modelos, datos y aplicaciones*. Ed. Pirámide.
- RAMOS, C. (1997): La distancia como medida de similitud en el análisis input-output. XI Reunión ASEPELT, Bilbao.
- RAMOS, C., ALVAREZ, R., FERNÁNDEZ, E. y GARCÍA, A. S. (2003): Algunas aplicaciones de la teoría matemática de la información al análisis input-output. XI Jornadas de ASEPUMA, Oviedo.
- RÍO, M^a J. y R. PÉREZ (1987): Sobre la medición de la concentración industrial. III Jornadas de Economía Industrial, Madrid.
- RÍO, M^a J. y C. RAMOS (1987): El concepto de distancia y algunas aplicaciones económicas. *Encuentro de Matemática Aplicada a la Empresa*, Zaragoza.
- SADEI (2003): *Marco Input-Output de Asturias*. Publicación electrónica: <http://www.sadei.es>
- SZYRMER, J. (1989): Trade-off between error and information in the RAS procedure. En Miller, Polenske y Rose *Frontiers of Input-Output Analysis*, Oxford University Press, págs. 258-278.
- TILANUS, C.B. y H. THEIL (1965): The Information approach to the evaluation of input-output forecast. *Econometrica*, Vol. 32, N° 4, pág. 847-862.

Anexo 1

Ramas	TIOAS-31
1- Agricultura, ganadería y silvicultura	1+2
2-Pesca	3
3-Extracción de productos energéticos	4+5+6
4-Extracción de otros minerales	7+8
5-Alimentación, bebidas y tabaco	9+10+11+12+13
6-Industria textil y de confección	14+15
7-Industria del cuero y del calzado	16
8-Industria de la madera y corcho	17
9-Industria de la edición, papel y artes gráficas	18+19
10-Coquerías y refino de petróleo	20
11-Industria química	21
12-Industria del caucho y materias plásticas	22
13-Otros productos minerales no metálicos	23
14-Metalurgia y productos metálicos	24+25
15- Maquinaria y equipo mecánico	26
16- Material eléctrico, electrónico óptico	27+28+29+30
17- Fabricación de material de transporte	31+32
18-Industrias manufactureras diversas	33+34
19-Energía eléctrica, gas y agua	35+36
20-Construcción	37
21-Comercio y reparación	38+39+40
22-Hostelería	41
23-Transportes y comunicaciones	42+43+44+45+46
24-Intermediación financiera	47+47 bis+48+49
25-Administración pública	55
26-Educación de mercado	56
27-Educación de no mercado	58
28-Actividades sanitarias; servicios sociales de mercado	57
29- Actividades sanitarias; servicios sociales de no mercado	59
30-Otros servicios	50+51+52+53+54+60+61+62+63
31-Hogares que emplean personal doméstico	64