

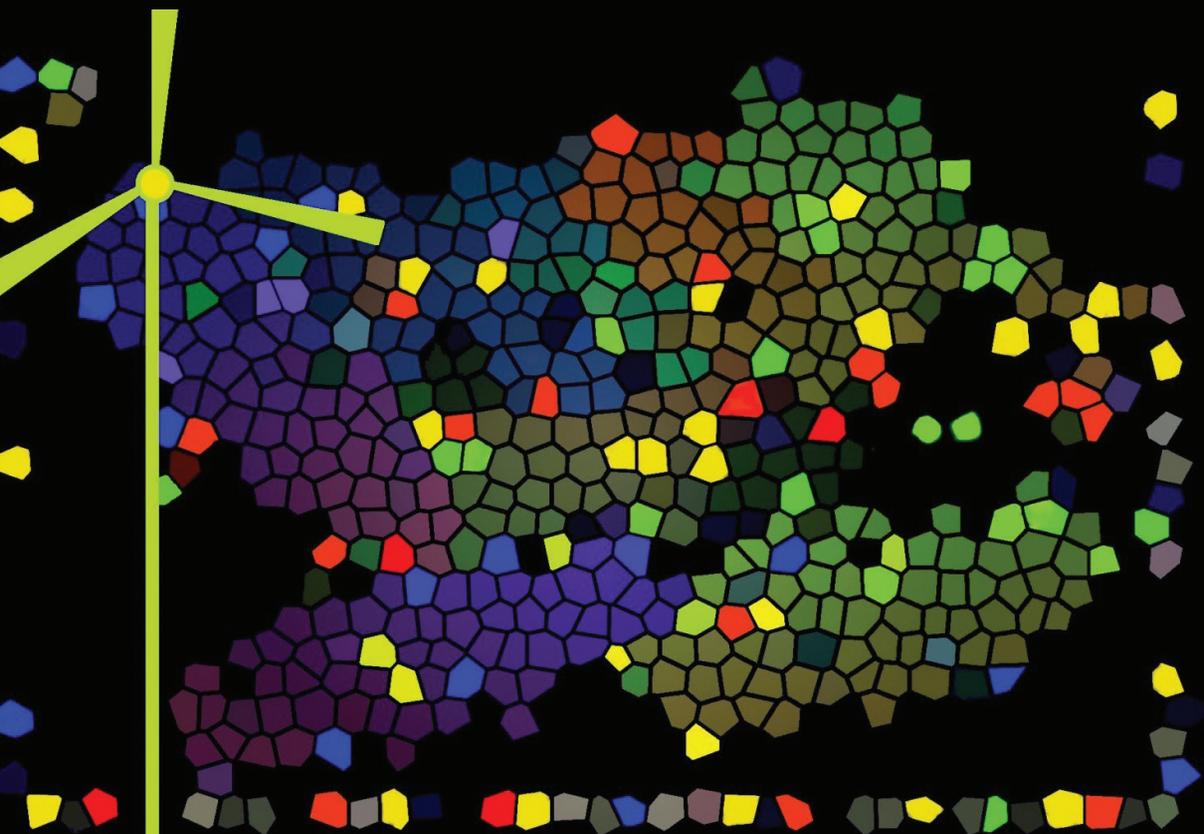
IMPRESA DA UNIVERSIDADE DE COIMBRA
COIMBRA UNIVERSITY PRESS

ASSESSMENT METHODOLOGIES

ENERGY, MOBILITY AND OTHER
REAL WORLD APPLICATION

PEDRO GODINHO
JOANA DIAS

EDITORS



**EFFECTS OF SECTORAL AGGREGATION
ON AN INPUT-OUTPUT TABLE**

Carmen Ramos¹

Abstract

The importance of the consequences of sectoral aggregation, from the perspective of input-output analysis is the foundation of this paper. This issue is addressed from two perspectives: firstly, the effects of aggregation on the amount of information contained in an input-output matrix, in this case the statistical theory of information is used (Theil, 1957).

Moreover, the changes experienced by the output multipliers due to different sectoral aggregations are analyzed. In this sense we have considered various proposals as the hypothetical extraction (Dietzenbacher and Van der Linden, 1997) and Pure Linkage (Sonis and others, 1995).

The theoretical results derived from this study will be applied to the Input-Output Table for Asturias (2010). We have also considered different standard aggregations as NACE or CIIU, among others.

Keywords: Input-Output Analysis, Sectoral Aggregation, Loss information, Pure linkage

¹ University of Oviedo (Spain)
E-mail: cramos@uniovi.es

1. Introduction

The subject of aggregation in the input-output analysis may seem obsolete and overcome, since computers and specific computer programs enable users to work with almost any level of disaggregation, hence, why worry about the aggregation when there is the possibility to carry out studies with a very large number of sectors? Nonetheless, this approach is a bit simplistic due to the fact that the performance of economic analyses often requires the use of different statistical sources, and they are usually presented with different aggregations, which makes it necessary to homogenize (aggregate) them. Input-output analysis is similar in this respect: it is very common to have input-output tables with a certain number of sectors whose information needs to be “combined” with other data sources.

On a separate issue, the construction of symmetric matrices requires the joint use of supply and use tables, which previously need to be “made” square, aggregating in an analogous way the products and the industries.

The problems of aggregation have been studied by different authors such as Leontief (1951) himself, Hatanaka (1952), Theil (1957), Ara (1959), Tilanus y Theil (1965), Morimoto (1970), Blin and Cohen (1977), Doblado (1988), Lauritzen (1989) and Russo (2001), among others.

In this study we are going to address the subject of sectoral aggregation from the perspective of the effects it causes, in a double area: focusing on the information itself contained in an input-output table, considering how this one varies when it has more or less aggregated matrices and, on the other hand, focusing on the results of the structural analysis derived from the calculation of the multipliers.

The theoretical results derived from this study will be applied to the Input-Output Table for Asturias (2010) published by SADEI².

² Asturian Society of Economic and Industrial Studies.

2. Statistical theory of information: some concepts

It seems reasonable to assume that the more disaggregated an input-output table is, the more information it provides, as a higher number of data about the economic reality of a country or a region is available. Therefore, using matrices with a smaller number of sectors implies a loss of information that can be more or less considerable depending on how this aggregation is being performed and on the initial distribution of the matrix. In order to analyse this loss of information we are going to use the so-called statistical theory of information.

The statistical theory of information was initiated, among others, by Shannon, who in 1948 proposed entropy as a measure of the amount of information, that is to say, of the disorder or the uncertainty in a certain system. The entropy associated with a random variable X is related to the probability distribution of the variable and it shows how predictable the process subject to uncertainty is.

Let X be a discrete random variable whose associated probability distribution is $P = (p_1, p_2, \dots, p_n)$. We call entropy of the random variable X or of the distribution $P = (p_1, p_2, \dots, p_n)$ the expression

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

This measure can be interpreted as the average information provided by each element ($\log p_i$). The entropy is non-negative, it will take the highest value in the event that the distribution is uniform, in this case, $p_i = 1/n$ and $H(X) = \log n$. In other words, if all the sectors have a very similar (the same) volume of sales or of purchases, the input-output table will be more informative, since the probability of a purchase (sale) taking place is similar for all the sectors, hence it is “difficult” to predict from which ones it comes from. Nevertheless, if the distribution happened to be degenerate³ the entropy would reach its lowest possible value, 0, as in that case a single sector would be the one to make all the purchases

³ A distribution is degenerate when all the values of a variable have null probability, except for one with unitary probability.

(sales), so the origin of the transactions would be easily identifiable (predictable) and the information contained in the table would be null.

Since we are going to work with input-output tables, and consequently with double-entry tables, it will be necessary to consider not a single random variable, but a two-dimensional magnitude (X,Y) , where X represents the sectors that make purchases and Y , the sectors that sell. In that case, Shannon entropy will be formulated as follows

$$H(X,Y) = -\sum_i \sum_j p_{ij} \log p_{ij} \quad (2)$$

where p_{ij} represents the joint probabilities associated with the two-dimensional random variable (X,Y) . The interpretation and bounds of this measure are analogous to those indicated for a one-dimensional variable.

3. Aggregation in input-output tables

As it has been indicated before, it is often necessary to aggregate the input-output tables. In this regard, consider situations in which it is essential to combine different sources of information where the sector classifications are different or when we use supply and use tables, which are rectangular since they have different classification in rows and columns.

3.1. Aggregation and amount of information

Consider the matrix \mathbf{Z} of intermediate demand

$$\mathbf{Z} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \quad (3)$$

where x_{ij} represents the purchases that the sector j makes with regard to the sector i .

In order to apply the Information Theory in the input-output context, we will consider X and Y as two qualitative variables: X would represent the buying sectors, while Y would represent the selling sectors. The values that appear in the intermediate demand matrix contain the purchases/sales between the different sectors or the number of times that sectors i and j interact (through economic exchanges). From these values we can obtain a relative frequency table, by only dividing each cell between the total of the table, that is⁴,

$$f_{ij} = \frac{x_{ij}}{\sum_i \sum_j x_{ij}} \quad (4)$$

Therefore, from the matrix \mathbf{Z} presented before, we can obtain the following one

$$\begin{bmatrix} f_{11} & f_{12} & \dots & f_{1n} \\ f_{21} & f_{22} & \dots & f_{2n} \\ \dots & \dots & \dots & \dots \\ f_{n1} & f_{n2} & \dots & f_{nn} \end{bmatrix} \quad (5)$$

where f_{ij} are the relative frequencies associated with the table \mathbf{Z} ⁵.

The most frequent notion of probability states that the probability is an ideal number with which the relative frequency converges, when n tends to infinity. On the basis of this concept we will use relative frequencies as probabilities. If we calculated the entropy, we would obtain:

$$H_0(X, Y) = -\sum_{i=1}^n \sum_{j=1}^n p_{ij} \log p_{ij} \quad (6)$$

It is denoted by $H_0(X, Y)$ to indicate that it is the initial moment (before the aggregation).

Let us imagine that it is necessary to perform an aggregation in the sectors that are included in the table due to any of the previously indicated

⁴ See Doblado (1988).

⁵ Observe that f_{ij} is not exactly a technical coefficient, as its denominator is the sum of the intermediate consumption and not the total output. The use of this type of coefficients is of interest in the input-output applications since they allow to get to know the amount of intermediate purchases (sales) of the total of intermediate consumption, eliminating the influence of the primary inputs (final demand).

reasons. More specifically, let us suppose⁶ that the sectors from 1 to m are aggregated, both in rows and in columns. The resulting aggregated matrix would be the following:

$$\begin{bmatrix} \sum_{i=1}^m \sum_{j=1}^m p_{ij} & \sum_{i=1}^m p_{i,m+1} & \cdots & \sum_{i=1}^m p_{in} \\ \sum_{j=1}^m p_{m+1,j} & p_{m+1,m+1} & \cdots & p_{m+1,n} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{j=1}^m p_{nj} & p_{n,m+1} & \cdots & p_{nn} \end{bmatrix} \quad (7)$$

In general terms, it seems logical to think that the aggregation is going to lead to a reduction in the amount of information contained in an input-output table; going forward in this point, we are going to quantify and compare the information before and after the aggregation. As it has been indicated before, the expression that represents the information contained in the table can be found in (6).

After performing referred operation, the entropy will take the value $H_1(X,Y)$ with regard to the final moment. That is, $H_1(X)$ takes the following value

$$\begin{aligned} & - \left[\left(\sum_{i=1}^m \sum_{j=1}^m p_{ij} \right) \log \left(\sum_{i=1}^m \sum_{j=1}^m p_{ij} \right) + \left(\sum_{i=1}^m p_{i,m+1} \right) \log \left(\sum_{i=1}^m p_{i,m+1} \right) + \cdots + \left(\sum_{i=1}^m p_{in} \right) \log \left(\sum_{i=1}^m p_{in} \right) + \left(\sum_{j=1}^m p_{m+1,j} \right) \log \left(\sum_{j=1}^m p_{m+1,j} \right) + \right. \\ & \left. \cdots + \left(\sum_{j=1}^m p_{nj} \right) \log \left(\sum_{j=1}^m p_{nj} \right) + \sum_{i=m+1}^n \sum_{j=m+1}^n p_{ij} \log(p_{ij}) \right] \quad (8) \end{aligned}$$

3.2. Empirical results

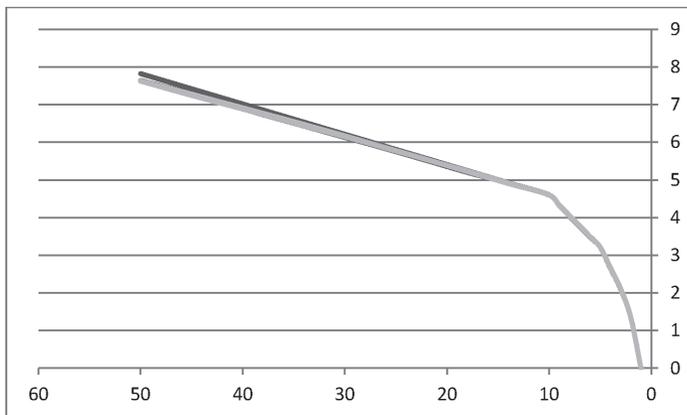
With the aim of illustrating this analysis, we have performed simulations on the basis of random samples of size 2500 (since economies formed by 50 sectors⁷ have been assumed). The samples have been randomly generated and the probability distributions originated correspond to intermediate situations, that is to say, there are neither degenerate samples

⁶ This supposition doesn't diminish the generality of the reasoning that has been used.

⁷ Different analyses with different number of sectors have been carried out and the results have been analogous.

nor uniformly distributed samples. On the basis of each of these samples we have performed different aggregations and determined the amount of information in each level of grouping.

The graph which represents the different levels of information according to the performed aggregation is the following:



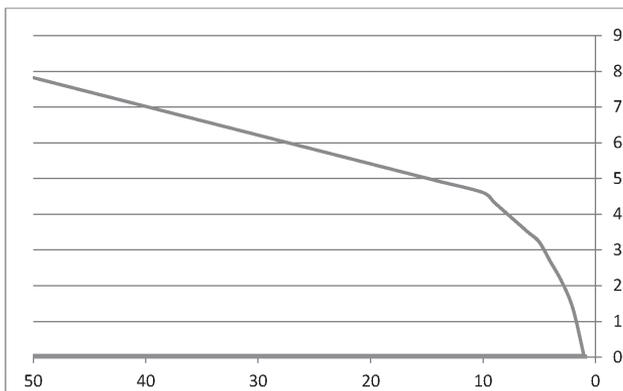
Graph 1: Information associated with an input-output table according to the performed aggregation

Since the aggregation is carried out starting from the highest number of sectors to the lowest, the ordinate axis is arranged in the reverse direction (from highest to lowest). In the vertical axis, the amount of information associated with each level of aggregation is presented.

As can be observed, the entropy progressively decreases and finally becomes null as tables are more aggregated. Furthermore, in the case of very aggregated tables the reduction in the amount of information is greater than when these tables are less aggregated.

It is necessary to point out that the results are very similar, despite the fact that the samples and the aggregations that have been performed differ. It is also true that only intermediate situations have been considered, in other words, the extreme cases haven't been taken into account: neither a uniform distribution of the flows between sectors, nor an economy formed by a single sector that carries out the totality of the transactions.

Below, and in order to complement the graphical analysis, we have considered the extreme situations and we have made a graph where we have included the entropy associated with a uniform distribution and with a degenerate distribution.



Graph 2: Amount of information associated with an input-output table. Extreme situations

The line which is situated over the ordinate axis (coinciding with it) represents the amount of information associated with a degenerate distribution, where, regardless of the level of aggregation, the information is zero. The other curve that appears in the graph represents the amount of information associated with a uniform distribution, for each of the levels of aggregation. Thus, all the intermediate cases, which are the usual ones, will be found between both situations.

Given the form of measurement of this amount of information, it can be observed that only in very extreme situations, close to the degenerate distribution, do the values move away from the upper bound to a great extent. On the basis of the previous results, we can indicate that the measure of entropy proposed by Theil is very sensitive to degenerate distributions, but less sensitive to other kinds of situations.

3.3. Loss of information

A concept that may be of interest is that of the loss of information associated with the sectoral aggregation of an input-output table. This loss can be defined on the basis of the difference in the entropy before and after the aggregation, and the result is the following:

$$\begin{aligned} \Delta(H) &= H_0(X, Y) - H_1(X, Y) = \\ &= -\sum_{i=1}^n \sum_{j=1}^m p_{ij} \log p_{ij} + \\ &= \left[\left(\sum_{i=1}^m \sum_{j=1}^m p_{ij} \right) \log \left(\sum_{i=1}^m \sum_{j=1}^m p_{ij} \right) + \left(\sum_{i=1}^m p_{i,m+1} \right) \log \left(\sum_{i=1}^m p_{i,m+1} \right) + \dots + \left(\sum_{i=1}^m p_{i,n} \right) \log \left(\sum_{i=1}^m p_{i,n} \right) + \left(\sum_{j=1}^m p_{m+1,j} \right) \log \left(\sum_{j=1}^m p_{m+1,j} \right) + \dots \right. \\ &+ \left. \left(\sum_{j=1}^m p_{nj} \right) \log \left(\sum_{j=1}^m p_{nj} \right) + \sum_{i=m+1}^n \sum_{j=m+1}^n p_{ij} \log(p_{ij}) \right] = -\sum_{i=1}^n \sum_{j=1}^m p_{ij} \log p_{ij} + \left[\left(\sum_{i=1}^m \sum_{j=1}^m p_{ij} \right) \log \left(\sum_{i=1}^m \sum_{j=1}^m p_{ij} \right) + \right. \\ &\left. \sum_{j=m+1}^n \left(\sum_{i=1}^m p_{i,j} \right) \log \left(\sum_{i=1}^m p_{i,j} \right) + \sum_{i=m+1}^n \left(\sum_{j=1}^m p_{i,j} \right) \log \left(\sum_{j=1}^m p_{i,j} \right) + \sum_{i=m+1}^n \sum_{j=m+1}^n p_{ij} \log(p_{ij}) \right] \quad (9) \end{aligned}$$

Observe that the first term is positive as it is affected by a negative sign. Likewise, the rest of the summands have a negative sign because, even though there are sums of probabilities, these sums always happen to be, as a result of their construction, less than one and, therefore, their logarithm is negative.

The first term of this expression $\left(\sum_{i=1}^n \sum_{j=1}^m p_{ij} \log p_{ij} \right)$ can be decomposed in the way that is shown below

$$\sum_{i=1}^n \sum_{j=1}^m p_{ij} \log p_{ij} = \sum_{i=1}^m \sum_{j=1}^m p_{ij} \log p_{ij} + \sum_{i=m+1}^n \sum_{j=m+1}^n p_{ij} \log p_{ij} + \sum_{i=1}^m \sum_{j=m+1}^n p_{ij} \log p_{ij} + \sum_{i=m+1}^n \sum_{j=1}^m p_{ij} \log p_{ij}$$

If we denote by $\Delta(H) = H_0(X, Y) - H_1(X, Y)$, we obtain

$$\Delta(H) = \left[\left(\sum_{i=1}^m \sum_{j=1}^m p_{ij} \right) \log \left(\sum_{i=1}^m \sum_{j=1}^m p_{ij} \right) \right]^T + \left[\sum_{j=m+1}^n \left(\sum_{i=1}^m p_{i,j} \right) \log \left(\sum_{i=1}^m p_{i,j} \right) \right]^T + \left[\sum_{i=m+1}^n \left(\sum_{j=1}^m p_{i,j} \right) \log \left(\sum_{j=1}^m p_{i,j} \right) \right]^T$$

where

$$\left[\left(\sum_{i=1}^m \sum_{j=1}^m p_{ij} \right) \log \left(\sum_{i=1}^m \sum_{j=1}^m p_{ij} \right) \right]^T = \left(\sum_{i=1}^m \sum_{j=1}^m p_{ij} \right) \log \left(\sum_{i=1}^m \sum_{j=1}^m p_{ij} \right) - \sum_{i=1}^m \sum_{j=1}^m p_{ij} \log p_{ij} \quad (10)$$

Also

$$\left[\sum_{j=m+1}^n \left(\sum_{i=1}^m p_{i,m+1} \log \left(\sum_{i=1}^m p_{i,m+1} \right) \right) \right]^T = \left[\sum_{j=m+1}^n \left(\sum_{i=1}^m p_{i,m+1} \log \left(\sum_{i=1}^m p_{i,m+1} \right) \right) \right] - \sum_{i=m+1}^n \sum_{j=1}^m p_{ij} \log p_{ij} \quad (11)$$

Finally

$$\left[\sum_{i=m+1}^n \left(\sum_{j=1}^m p_{i,j} \log \left(\sum_{j=1}^m p_{i,j} \right) \right) \right]^T = \left[\sum_{i=m+1}^n \left(\sum_{j=1}^m p_{i,j} \log \left(\sum_{j=1}^m p_{i,j} \right) \right) \right] - \sum_{i=m+1}^n \sum_{j=1}^m p_{ij} \log p_{ij} \quad (12)$$

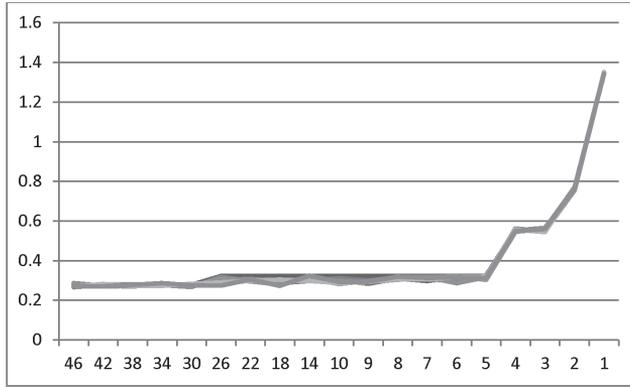
The expression (10) contains the changes in the amount of information caused by the aggregation in rows and columns, that is to say, the loss of information due to those sectors of activity that are aggregated. The expression (11) reflects the loss in the amount of information in the sectors which are not the ones intended to be directly joined, but those that are aggregated in a collateral way, as a result of the aggregation carried out with regard to other sectors. It reflects the loss of information caused by the aggregation carried out in columns. The expression (12) is analogous to (11) but it refers to the collateral aggregations in rows.

To conclude, the term $\sum_{i=m+1}^n \sum_{j=m+1}^n p_{ij} \log p_{ij}$ that is shown in (10) refers to those sectors that have not been aggregated and, therefore, it becomes null when it is subtracted from the corresponding terms of $H_o(X,Y)$.

Given the properties of the logarithmic function, we can deduce that the entropy before the aggregation is higher than the entropy after having carried out this operation. Thus, and considering what was previously indicated with regard to the signs of the different addends, we reach the conclusion that if we take all the terms into account, the aggregation leads to a reduction in the information that an input-output table provides. It is a decrease in absolute terms, that is, the indicator which quantifies the information contained in the matrix is lower.

If the distribution is intermediate, the loss will be lower when the aggregation does not contribute drastically to move the matrix coefficients away from a uniform distribution.

Below we present a graph derived from the performed simulations that reflects the loss of information associated with different aggregations with non-degenerate intermediate distributions.



Graph 3: Loss of information. Intermediate distributions

Once more, the horizontal axis of the graph is arranged from the highest to the lowest when it comes to the number of sectors (from a lower level of aggregation to a higher). The loss of information is reflected in the vertical axis.

As it can be observed, a kind of inverted L is represented, where the lowest values are reached in the initial and central parts, while the highest are in the opposite end. When the table is very aggregated, a new aggregation generates a sharp reduction in the amount of information; this performance coincides with the one that was seen in Graph 1.

3.5. Relative loss

The loss of information due to the aggregation depends both on the changes undergone by the matrix coefficients and on the variation in the number of sectors. In other words, even though the distribution of the flows between sectors can be more uniform as a result of aggregation, the reduction in the number of sectors results in a decrease of the final value of the measure, since a lower number of sectors will provide less information regarding the economy. In order to try to eliminate the impact that the reduction in the number of sectors has on the loss of information, we propose a relative measure of loss of information by dividing

the loss that was previously defined between the number of sectors that form it, but in a “softened” way, that is, starting from the logarithm of referred number ($\log n$).

As it has been indicated before, Shannon entropy is delimited at the top by $\log n$ and at the bottom by zero. If the loss of information is expressed in relative terms (with regard to the bound) the effect caused by the change in the number of sectors can be avoided and the loss of information will only depend on the distribution.

A relative loss of information can be defined in the following terms:

$$\Delta^R(H) = \frac{H_0(X,Y)}{\log r} - \frac{H_1(X,Y)}{\log s} \quad (13)$$

where n is the number of coefficients before the aggregation and s is their number after this aggregation has taken place. That is to say

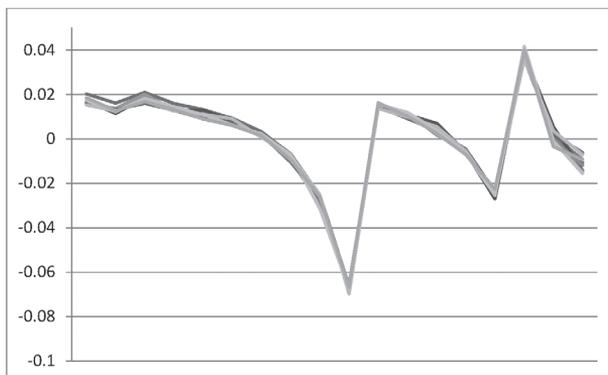
$$\Delta^R(H) = \frac{-\sum_{i=1}^m \sum_{j=1}^m p_{ij} \log p_{ij}}{\log r} + \frac{\left[\left(\sum_{i=1}^m \sum_{j=1}^m p_{ij} \right) \log \left(\sum_{i=1}^m \sum_{j=1}^m p_{ij} \right) + \sum_{j=1}^n \left(\sum_{i=1}^m p_{i,m+1} \right) \log \left(\sum_{i=1}^m p_{i,m+1} \right) + \sum_{i=m+1}^n \left(\sum_{j=1}^m p_{i,j} \right) \log \left(\sum_{j=1}^m p_{i,j} \right) + \sum_{i=m+1}^n \sum_{j=m+1}^n p_{ij} \log(p_{ij}) \right]}{\log s}$$

$\log n > \log s$, while $H_0(X,Y) > H_1(X,Y)$, therefore the sign of the relative loss is inconclusive, in other words

$$\frac{H_0(X,Y)}{\log r} < \frac{H_1(X,Y)}{\log s} \quad (14)$$

depending on the situation.

When $(\log s) H_0(X,Y) < (\log n) H_1(X,Y)$ the amount of initial information, which is higher, is compensated by the reduction in the number of sectors, whereas if it is the case that $(\log s) H_0(X,Y) > (\log n) H_1(X,Y)$ this compensation does not take place.



Graph 4: Relative loss of information. Intermediate distributions

In this graph, it can be observed that for certain aggregations the relative loss of information is negative, which means that the information before the aggregation is less than the information contained in the table after having performed this operation. This happens in those situations where the aggregation leads to a result in which the distribution approximates more to the uniform one.

Also, it can be noticed that in very aggregated tables a slightly negative loss takes place, due to the great homogeneity derived from the resulting small number of sectors.

4. Amount and loss of information in the input-output table for Asturias

As mentioned earlier, we are going to apply the concepts which were previously explained to the Input-Output Table for Asturias corresponding to the year 2010, since this one was the latest to be published at the time of writing this study.

Different aggregations have been used in order to observe its impact on the amount of associated information. In particular, we have used as a starting point the aggregation proposed by SADEI⁸ in 66 sectors; af-

⁸ The classification proposed by SADEI is a modification of the CNAE 2009.

terwards, we have used the one that appears in EUROSTAT with regard to the input-output tables, in 60 sectors⁹. Moreover, we have taken into consideration the aggregation performed in the WIOD database in 32 sectors for each country, as referred database is often used by input-output analysts. In this application, we have used the classification in 21 sectors proposed by the CNAE (National Classification of Economic Activities), although as a consequence of homogeneity problems, the resulting aggregated Input-Output Table for Asturias is formed by 20 sectors in the end. Another classification that has been taken into account is the International Standard Industrial Classification of All Economic Activities (ISIC), which considers 16 sectors. Finally, aggregations in 10 sectors of EUROSTAT and in 4 of SADEI have been used; both aggregations have a very small number of sectors, but thanks to this, we have at our disposal a wide range of possibilities to analyse. All the classifications that have been used here can be found in the annex.

Table 1: Amount of information associated with the Input-Output Table for Asturias corresponding to the year 2010

SOURCE	Sectors	H(X,Y)	Upper bound	Percentage of bound
SADEI	66	5,8537	8,3793	69,859
EUROSTAT	60	5,6938	8,1551	69,819
WIOD	32	5,1830	6,9315	74,775
CNAE	21	3,9775	5,9915	66,386
CIU	16	3,4658	5,5452	62,501
EUROSTAT	10	2,9605	4,6052	64,286
SADEI	4	1,8819	2,7726	67,874

From the previous table we can deduce that the highest amount of information is presented by the most disaggregated tables, as had been indicated before. However, if we observe the last column of the table, we notice that not always is the number of sectors directly related to the amount of information contained in it. Observe, for instance, the classification proposed in the WIOD database: even though it is in 32

⁹ Given the differences with the one proposed by SADEI, the final aggregation which has been used had to be in 59 sectors.

sectors, in relative terms, it is better than the initial one proposed by SADEI or that of EUROSTAT in 59. This is due to the fact that referred aggregation leads to structures which are more uniformly distributed, where the groupings of sectors have a more homogeneous performance.

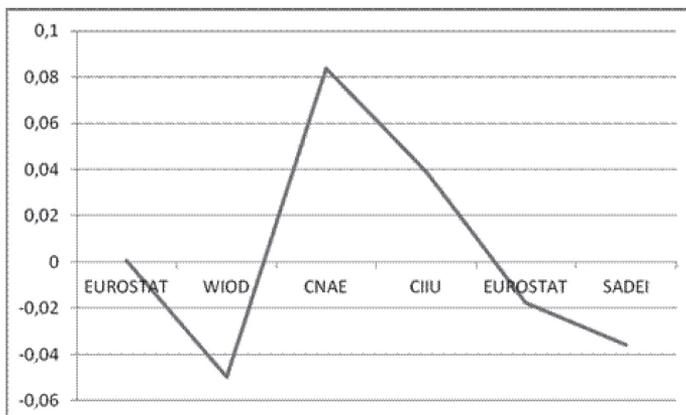
Below, we proceed to quantify the relative loss derived from the performed aggregations.

Table 2: Relative loss due to the aggregation

	H66	H59	H32	H20	H16	H10
H59	0,040					
H32	-4,915	-4,956				
H20	3,473	3,433	8,389			
H16	7,358	7,318	12,273	3,885		
H10	5,574	5,533	10,489	2,100	-1,785	
H4	1,985	1,945	6,900	-1,488	-5,373	-3,589

The previous table must be read in the following way: the initial entropies are situated in the first row, while in the first column the final aggregations are shown. In other words, $H^R(66)-H^R(59)= 0.040$, that is to say, the relative loss when aggregating from the 66 sectors of the initial table to the 59 is of 0,04, and so on. Consequently, if the initial table is aggregated in 32 sectors (according to the classification found in the WIOD database), a negative relative loss takes place, that is to say, a relative gain of information. Therefore, it can be noticed that the aggregations in 32 sectors (used in the WIOD database) and in 4 (SADEI) are the ones that, in relative terms, would produce a gain of information at the time of aggregating, since they lead to more uniform classifications of sectors. Nevertheless, it is also true that using a classification in 4 sectors in an analysis is not particularly helpful.

The graph below represents the relative loss when considering an initial table in 66 sectors and aggregating it according to the different classifications previously indicated. As it can be observed, the aggregations found in the WIOD database (32) and in SADEI (4) generate a relative gain of information, despite having a smaller number of sectors. In the case of the groupings of SADEI, we consider that it is due to the homogeneity derived from the small number of sectors.



Graph 5: Relative loss of information due to different aggregations

5. Aggregation and linkages

Other effects of aggregation are the ones referred to the input-output structural analysis. One of the main strengths of Input-Output analysis is its ability to evaluate the significance of the different productive sectors within the economic structure of a country (region). It is relevant to consider whether the level of segregation can influence the value of these multipliers and the results of structural analysis performed.

There are a great number of studies that deal with how the aggregation affects the results of a structural analysis. We can highlight, among them, those of Hatanaka (1952), Ara (1959), Tilanus and Theil (1965), Morimoto (1970), Blin and Cohen (1977), Lauritzen (1989) or Russo (2001). In this part of the article, we are going to study the impact that the aggregation has on the linkages.

The extraction method has its origins in the studies carried out by Strassert (1968), who presents it as a structural evaluation alternative with regard to the classic methods. This author proposes to quantify the effect that would occur in an economy if a certain sector was hypothetically extracted from it. In order to do this, the productive sector under consideration is eliminated from the matrix \mathbf{A} of technical coefficients.

We start with Leontief demand model shown in the equation (2) and it is assumed that one sector is extracted from the economy. In this case, this equation can be rewritten as

$$\bar{\mathbf{x}}(k) = \left[\mathbf{I} - \bar{\mathbf{A}}(k) \right]^{-1} \bar{\mathbf{y}}(k) \quad (15)$$

where $\bar{\mathbf{A}}(k)$ is a matrix of order $(n-1)(n-1)$, as the row and the column of the k th sector have been eliminated, $\bar{\mathbf{x}}(k)$ represents a vector of total output and $\bar{\mathbf{y}}(k)$, a demand vector, both of dimension $n-1$.

Therefore, and given the values that $\mathbf{y}(k)$ and $\bar{\mathbf{y}}(k)$ reach, $x_i(k)$ will be smaller than x_i , that is, $\bar{x}_i(k) < x_i \forall i=1, 2, \dots, k-1, k+1, \dots, n$. Where $\bar{\mathbf{x}}(k)$ is obtained as if the k th sector did not exist in the economy and, consequently, this sector does not generate relationships with other productive sectors. Then, the sum of the differences between the elements of \mathbf{x} and $\bar{\mathbf{x}}(k)$ can be considered as the measure of the interrelationships of this sector extracted from the rest. In this way, the following equation is presented:

$$L(k) = \sum_{i=1, i \neq k}^n \left[x_i - \bar{x}_i(k) \right] \quad (16)$$

where $L(k)$ is a global measure of the importance of the k th sector.

This initial idea that Strassert proposed has undergone extensions, some of which are those presented by Cella (1984), Sonis, et al (1995) and Dietzenbacher and Van der Linden (1997). In the present study, we are going to apply the extension proposed by Sonis et al. (1995).

Let us consider now an initial situation, previous to the aggregation, represented on the basis of a matrix \mathbf{A} of technical coefficients.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \quad (17)$$

where $a_{ij} = \frac{x_{ij}}{x_j}$

That is to say, the proportion of purchases that the sector j makes from the sector i , with regard to the totality of the purchases of sector j .

The Pure Linkage Method is based on the idea of the need to isolate one sector (or group of sectors) to determine its importance. In order to achieve this objective, it proposes the following decomposition of the matrix \mathbf{A} of technical coefficients:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{1r} \\ \mathbf{A}_{r1} & \mathbf{A}_{rr} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{1r} \\ \mathbf{A}_{r1} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{rr} \end{bmatrix} = \mathbf{A}_1 + \mathbf{A}_r \quad (18)$$

The sector (or group of sectors) 1 is the one which is going to be isolated from the rest of the economy. The Leontief inverse matrix (\mathbf{L}) can be decomposed in the following way: $\mathbf{L} = \mathbf{P}_2 \mathbf{P}_1$ or as $\mathbf{L} = \mathbf{P}_1 \mathbf{P}_3$. Where $\mathbf{P}_1 = [\mathbf{I} - \mathbf{A}_r]^{-1}$, $\mathbf{P}_2 = [\mathbf{I} - \mathbf{P}_1 \mathbf{A}_1]^{-1}$ and $\mathbf{P}_3 = [\mathbf{I} - \mathbf{A}_1 \mathbf{P}_1]^{-1}$.

The equation $\mathbf{L} = \mathbf{P}_2 \mathbf{P}_1$ isolates the interaction in the rest of the economy of the sector 1. \mathbf{P}_2 shows the impacts derived from the demand of the sector 1 in the economy as a whole ($\mathbf{P}_1 \mathbf{A}_1$).

The Leontief inverse matrix can be expressed on the basis of the previous expressions in the following manner:

$$\mathbf{L} = \begin{bmatrix} \tilde{\Delta}_1 & \tilde{\Delta}_1 \mathbf{A}_{1r} \\ \Delta_r \mathbf{A}_{r1} \tilde{\Delta}_1 & \mathbf{I} + \Delta_r \mathbf{A}_{r1} \tilde{\Delta}_1 \mathbf{A}_{1r} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \Delta_r \end{bmatrix} = \mathbf{P}_2 \mathbf{P}_1 \quad (19)$$

$$\text{where } \tilde{\Delta}_1 = [\mathbf{I} - \mathbf{A}_{11} - \mathbf{A}_{1r} \Delta_r \mathbf{A}_{r1}]^{-1}, \Delta_r = [\mathbf{I} - \mathbf{A}_{rr}]^{-1} \text{ and } \Delta_1 = [\mathbf{I} - \mathbf{A}_{11}]^{-1}. \quad (20)$$

On the other hand,

$$\mathbf{P}_2 = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \Delta_r \mathbf{A}_{r1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \tilde{\Delta}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{A}_{1r} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \quad (21)$$

$$\text{where } \mathbf{P}_2 = [\mathbf{I} - \mathbf{B}_1]^{-1} \text{ and } \mathbf{B}_1 = \mathbf{P}_1 \mathbf{A}_{11} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{1r} \\ \Delta_r \mathbf{A}_{r1} & \mathbf{0} \end{bmatrix}. \quad (22)$$

On the basis of this last equation, we can define the Pure Backward Linkage (PBL) as $\text{PBL} = \mathbf{i}_{11} \Delta_r \mathbf{A}_{r1} \mathbf{x}_{11}$, where q_{11} is the total output of sector 1, that is to say

$$\mathbf{q}_{11} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_m \end{bmatrix} \quad (23)$$

Our aim is to compare the variation that the linkages experience when the aggregation is performed. To do so, we are going to consider that the group 1 is formed by those sectors that will be aggregated afterwards. Before the aggregation takes place, the expression of \mathbf{A} is

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{1r} \\ \mathbf{A}_{r1} & \mathbf{A}_{rr} \end{bmatrix} \quad (24)$$

\mathbf{A}_{11} represents a matrix of technical coefficients of m rows and m columns (which will be aggregated in a second stage). \mathbf{A}_{r1} is formed by n - m rows and m columns and it presents the amounts of purchases of the group 1 from the rest of the sectors. \mathbf{A}_{1r} is composed of m rows and n - m columns; it represents the sales of the group 1 to the rest of the economy. Finally, \mathbf{A}_{rr} consists of n - m rows and columns and it shows the flows of the rest of the sectors.

If we apply the expressions indicated before, the Pure Linkage associated with the group of sectors (1) will be $PBL_1 = \mathbf{i}_{11} \Delta_r \mathbf{A}_{r1} \mathbf{q}_{11}$.

The matrices Δ_{21} and Δ_2 are formed by the following elements

$$\mathbf{A}_{r1} = \begin{bmatrix} a_{m+1,1} & a_{m+1,2} & \dots & a_{m+1,m} \\ a_{m+2,1} & a_{m+2,2} & \dots & a_{m+2,m} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{n,m} \end{bmatrix} \quad \text{and} \quad \Delta_r = \begin{bmatrix} \beta_{m+1,m+1} & \dots & \dots & \beta_{m+1,n} \\ \beta_{m+2,m+1} & \dots & \dots & \beta_{m+2,n} \\ \dots & \dots & \dots & \dots \\ \beta_{n,m+1} & \dots & \dots & \beta_{n,n} \end{bmatrix} \quad (25)$$

Thus,

$PBL_1 = \beta_{m+1,m+1} \sum_{j=1}^m a_{m+1,j} x_j + \dots + \beta_{m+1,n} \sum_{j=1}^m a_{n,j} x_j + \dots + \beta_{n,n} \sum_{j=1}^m a_{n,j} x_j$. Let us suppose that the aggregation is performed by joining the first m sectors.

We will then obtain

$$\mathbf{A}^* = \begin{bmatrix} a^*_{11} & a^*_{12} & \dots & a^*_{1s} \\ a^*_{m+11} & a^*_{m+1,m+1} & \dots & a^*_{2s} \\ \dots & \dots & \dots & \dots \\ a^*_{s1} & a^*_{s2} & \dots & a^*_{s,s} \end{bmatrix} = (26)$$

$$\begin{bmatrix} \frac{\sum_{i=1}^m \sum_{j=1}^m X_{ij}}{\sum_{j=1}^m x_j} & \frac{\sum_{i=1}^m X_{i,m+1}}{x_{m+1}} & \dots & \frac{\sum_{i=1}^m X_{in}}{x_n} \\ \frac{\sum_{j=1}^m X_{m+1,j}}{\sum_{j=1}^m x_j} & X_{m+1,m+1} & \dots & X_{m+1,n} \\ \dots & \dots & \dots & \dots \\ \frac{\sum_{j=1}^m X_{nj}}{\sum_{j=1}^m x_j} & X_{n,m+1} & \dots & X_{nm} \end{bmatrix}$$

If we partition the matrix, we obtain

$$\begin{bmatrix} \mathbf{A}_{11}^* & \mathbf{A}_{1r}^* \\ \mathbf{A}_{r1}^* & \mathbf{A}_{rr}^* \end{bmatrix} \quad (27)$$

Now \mathbf{A}_{11}^* represents a matrix with just one element, due to the fact that it is formed by the m sectors aggregated in rows and in columns in a single sector. \mathbf{A}_{1r}^* represents the matrix of technical coefficients referred to those sectors that are aggregated in rows; its dimension is 1 row and $n-m$ columns. \mathbf{A}_{r1}^* contains the technical coefficients associated with the sectors that are aggregated in columns with a dimension of $n-m$ rows and 1 column. Last but not least, \mathbf{A}_{rr}^* includes the technical coefficients derived from the non-aggregated sectors with $n-m$ rows and columns; it coincides with \mathbf{A}_{rr} .

On the basis of both partitioned matrices \mathbf{A} and \mathbf{A}^* we are going to try to compare the effects of aggregation in the calculation of the multipliers.

Now, let us calculate the value of the Pure Backward Linkage, considering that the first m sectors are aggregated. In this case, we would have a $PBL_{11}^* = \mathbf{i}^*_{11} \Delta_r^* \mathbf{A}^*_{r1} \mathbf{q}^*_{11}$, where the terms with an asterisk refer to the aggregated magnitudes.

$$\mathbf{A}_{r1}^* = \begin{bmatrix} a_{m+1,1}^* \\ a_{m+2,1}^* \\ \dots \\ a_{n,1}^* \end{bmatrix} \quad \Delta_r^* = \begin{bmatrix} \beta_{m+1,m+1} & \dots & \dots & \beta_{m+1,n} \\ \beta_{m+2,m+1} & \dots & \dots & \beta_{m+2,n} \\ \dots & \dots & \dots & \dots \\ \beta_{n,m+1} & \dots & \dots & \beta_{n,n} \end{bmatrix} \quad \text{and} \quad q_{11}^* = x_1 + x_2 + \dots + x_m$$

Therefore

$$PBL_{11}^* = \left(\sum_{j=m+1}^n \sum_{i=m+1}^n \beta_{m+1,j} a_{i,k}^* + \dots + \sum_{j=m+1}^n \sum_{i=m+1}^n \beta_{m+2,j} a_{i,k}^* + \dots + \sum_{j=m+1}^n \sum_{i=m+1}^n \beta_{n,j} a_{i,k}^* \right) q_{11}^* \quad (28)$$

$$\text{with } q_{11}^* = \sum_{k=1}^m q_k$$

$$\text{where } a_{m+1,1}^* = \frac{\sum_{j=1}^m X_{m+1,j}}{\sum_{j=1}^m x_j}, \dots, a_{n,1}^* = \frac{\sum_{j=1}^m X_{nj}}{\sum_{j=1}^m x_j}. \quad (29)$$

On the other hand,

$$\begin{aligned}
 a_{m+1,1}^* &= \frac{\sum_{j=1}^m X_{m+1,j}}{\sum_{j=1}^m x_j} = \frac{X_{m+1,1} + X_{m+1,2} + \dots + X_{m+1,m}}{\sum_{j=1}^m x_j} = \frac{X_{m+1,1}x_1}{\sum_{j=1}^m x_j} + \frac{X_{m+1,2}x_1}{\sum_{j=1}^m x_j} + \dots + \frac{X_{m+1,m}x_m}{\sum_{j=1}^m x_j} = \\
 &= \frac{a_{m+1,1}x_1}{\sum_{j=1}^m x_j} + \frac{a_{m+1,2}x_2}{\sum_{j=1}^m x_j} + \dots + \frac{a_{m+1,m}x_m}{\sum_{j=1}^m x_j} = a_{m+1,1}\omega_1 + a_{m+1,2}\omega_2 + \dots + a_{m+1,m}\omega_m
 \end{aligned}$$

Analogously

$$a_{n,1}^* = \frac{\sum_{j=1}^m X_{nj}}{\sum_{j=1}^m x_j} = a_{n,1}\omega_1 + a_{n,2}\omega_2 + \dots + a_{n,m}\omega_m \quad (30)$$

The rest of the addends are determined in the same way.

With

$$\omega_j = \frac{x_j}{\sum_{j=1}^m x_j} \quad \forall j = 1, 2, \dots, m \quad (31)$$

Consequently

$$a_{m+1,1}^* = a_{m+1,1}\omega_1 + a_{m+1,2}\omega_2 + \dots + a_{m+1,m}\omega_m = a_{m+1,1} \frac{x_1}{q_{11}^*} + a_{m+1,2} \frac{x_2}{q_{11}^*} + \dots + a_{m+1,m} \frac{x_m}{q_{11}^*}$$

.....

$$a_{n,1}^* = a_{n,1} \frac{x_1}{q_{11}^*} + a_{n,2} \frac{x_2}{q_{11}^*} + \dots + a_{n,m} \frac{x_m}{q_{11}^*} \quad (32)$$

Developing

$$\begin{aligned}
 PBL_1^* &= \left(\beta_{m+1,m+1} \sum_{j'=1}^m a_{m+1,j'} \frac{x_{j'}}{q_{11}^*} + \dots + \beta_{m+1,n} \sum_{j'=1}^m a_{n,j'} \frac{x_{j'}}{q_{11}^*} + \dots + \beta_{n,n} \sum_{j'=1}^m a_{n,j'} \frac{x_{j'}}{q_{11}^*} \right) q_{11}^* = \\
 &\beta_{m+1,m+1} \sum_{j'=1}^m a_{m+1,j'} x_{j'} + \dots + \beta_{m+1,n} \sum_{j'=1}^m a_{n,j'} x_{j'} + \dots + \beta_{n,n} \sum_{j'=1}^m a_{n,j'} x_{j'}
 \end{aligned}$$

That is to say, the linkages coincide if we consider in conjunction the sectors that have not been aggregated and the ones that have.

Finally, and for purposes of illustration, we are going to aggregate the agricultural, forestry and fishing¹⁰ sectors and calculate the Pure Linkage before and after having performed this operation. To start with, we will consider the symmetric matrix for Asturias corresponding to the year 2010, formed by 66 sectors, where Agriculture, livestock breeding and hunting (1), Silviculture and forest exploitation (2) and Fishing and aquaculture (3) are three separate sectors. The PBL for the block formed by those three sectors will be determined. In this case, the matrix \mathbf{A}_{r1} will be composed of 63 rows and 3 columns, \mathbf{A}_{11} will have 3 rows and columns, \mathbf{A}_{1r} will be formed by 3 rows and 63 columns and \mathbf{A}_{rr} will consist of 63 rows and columns. The value of the Pure Linkage is calculated and, if we express it in relative terms with regard to the total output of the economy, we obtain a result of 0.55%.

After that, the first three sectors of the table are aggregated and we will then obtain a matrix of 64 sectors. Now, the matrix \mathbf{A}_{11}^* will be formed by a single sector, \mathbf{A}_{r1}^* will be composed of 63 rows and 1 column, \mathbf{A}_{1r}^* will have 63 rows and columns and \mathbf{A}_{rr}^* will consist of one row and 63 columns. Once again, the Pure Linkage for the aggregated sector is determined and the result is the same 0.55%, that is to say, the value of the linkage does not change with aggregation.

6. Conclusions

The aggregation of sectors in an input-output table may be necessary when different sources of information are combined or when it is required to obtain a symmetric matrix from supply and use tables.

A table contains a certain amount of information that can be quantified through measures of statistical information, among which we can find Shannon entropy. As a table is aggregated, its amount of information de-

¹⁰ These three sectors represent the first one of the input-output table aggregated in 4 sectors (SADEI).

creases. Referred reduction is stronger when the tables are already very aggregated than when their level of disaggregation is higher.

A concept that contains this phenomenon is that of the loss of information, which presents a shape of an inverted L if intermediate distributions (neither degenerate nor uniform) are considered. In other words, the greatest loss takes place in very aggregated tables.

The distribution of the coefficients intervenes in the loss of information and so does the number of sectors. If we define the relative loss as the loss of information divided by its upper bound, we are eliminating the importance of the number of sectors. When we determine the loss in this case, we appreciate that there are situations in which an aggregation leads to a relative “gain” of information. These situations take place when the aggregated table has a more uniform behaviour than the table before the aggregation.

We have applied the previous concepts to the table for Asturias corresponding to the year 2010, which is initially aggregated in 66 sectors of activity. On the basis of this table, different aggregations have been applied: that of EUROSTAT referred to input-output tables, that of the WIOD database, the CNAE in 20 sectors, the ISIC, the aggregation of EUROSTAT in 10 sectors and that of SADEI in 4. Even though the amount of information decreases with the aggregation, if we take the relative loss into consideration, this loss is negative (relative gain) in the case of the aggregations proposed by the WIOD and by SADEI in 4 sectors.

Another analysed aspect involved studying what effects the aggregation has on the linkages. For this purpose, we have considered the Pure Linkage proposed by Sonis et al (1995). It has been theoretically proven that the value of the linkage does not vary when the sectors of activity are aggregated.

To conclude, we have calculated the PBL on the basis of the table for Asturias in 66 sectors, considering a block formed by the sectors of Agriculture, Silviculture and Fishing. We have obtained the same value as if it was calculated on the basis of the aggregated table, where the block would now be formed by the sum of the three indicated sectors.

References

- ARA, K. "The aggregation problem in input-output analysis". *Econometrica: Journal of the Econometric Society*, 27, 257-262, 1959.
- BLIN, J. M., & COHEN, C.. "Technological similarity and aggregation in input-output systems: a cluster-analytic approach". *The Review of Economics and Statistics*, 59, 82-91., 1977.
- CABRER, B., CONTRERAS, D., & MIRAVETE, E. "Aggregation in Input-Output Tables: How to Select the Best Cluster Linkage". *Economic Systems Research*, 3(2), 99-110, 1991.
- CELLA, G. "The Input-Output Measurement of Interindustry Linkages", *Oxford Bulletin of Economics and Statistics*, 46(1), 73-84, 1984.
- DIETZENBACHER, E. & VAN DER LINDEN, J. "Sectoral and Spatial Linkages in the EC Production Structure", *Journal of Regional Science*, 37(2), 235-257, 1997.
- DOBLADO J. M. *Teoría de la información: errores de agregación en las tablas input-output*. Universidad Nacional de Educación a Distancia, UNED, 1988.
- EUROSTAT: *NACE, Rev. 2*, 2007.
- HATANAKA, M. "Note on consolidation within a Leontief system". *Econometrica*, 301-303, 1952.
- INE: *CNAE 2009*, 2007.
- LEONTIEF, W. "An alternative to aggregation in input-output analysis and national accounts". *The Review of Economics and Statistics*, 49, 412-419, 1967.
- LEONTIEF, W. (Ed.). *Input-output economics*. Oxford University Press, 1986.
- MORIMOTO, Y. "On aggregation problems in input-output analysis". *The Review of Economic Studies*, 27(1), 119-126, 1970.
- PARDO, L: *Teoría de la Información Estadística*. Ed. Hespérides, 1998.
- RUSSO, G. *Input-output analysis and aggregation bias: An empirical assessment*. 84(13). Working paper Università degli Studi di Trieste, 2001.
- SADEI: *Tablas Input-Output de Asturias (2010)*, 2013.
- SONIS, M. GUILHOTO, G., HEWINGS, G. & MARTINS, E: "Linkages, Key Sectors, and Structural Change: Some New Perspectives", *The Developing Economics*, 33(3), 233-270, 1995.
- THEIL, H. "Linear aggregation in input-output analysis". *Econometrica*, 25, 111-122, 1957.
- TILANUS, C & THEIL, H. "The information approach to the evaluation of input-output forecasts". *Econometrica: Journal of the Econometric Society*, 847-862, 1965.
- TIMMER, M. (ed), *The World Input-Output Database (WIOD): Contents, Sources and Methods*, WIOD Working Paper 10, 2012.
- UNITED NATIONS: *CIIU, Rev. 3.1*, 2002.

Série Investigação

•

Imprensa da Universidade de Coimbra

Coimbra University Press

2015

