UNIVERSITÀ POLITECNICA DELLE MARCHE
DIPARTIMENTO DI ECONOMIA

The Balassa Index Meets the Dissimilarity Theil Index: a Decomposition Methodology for Location Studies

Eleonora Cutrini

QUADERNI DI RICERCA n. 274

November 2006

Abstract

It is well known that the broad concept of localisation has two connotations, namely, regional specialisation and industrial concentration. The main purpose of this paper is to introduce an entropy index of overall localisation suitable to conceptualise specialisation and concentration as the two sides of the same medal in a nested geographical perspective. The system of dissimilarity entropy measures is potentially applicable to the assessment of the spatial distribution of several economic phenomena when a twofold geographical level of analysis is considered. In the specific case, the decomposition provides an accurate method to quantify the cross-country divergence in localisation from the agglomeration within countries.

JEL Class.: C43, L16, O18, R12

Keywords: Localisation, Concentration, Specialisation, Dissimilarity entropy index, Within-between country decomposition

Indirizzo: Dipartimento di Economia, Università Politecnica delle Marche, Piazzale Martelli 8, 60121, Ancona, e.cutrini@univpm.it

The Balassa Index Meets the Dissimilarity Theil Index: a Decomposition Methodology for Location Studies*

Eleonora Cutrini

Introduction

In the last decades the pattern of change in the territorial distribution of economic activities has become a prominent topic in the political debate and in the academic research since the increasing international integration is deemed to engender agglomeration of economic activities and rising specialisation of national and regional economies.

The development of region-based empirical studies has been hindered until recent years not only because of the shortage of regional data but also for the lack of a methodology able to disentangle the geographical clustering internal to countries from cross-country location patterns as claimed by Combes e Overman (2004). So far the different basic unit of analysis (region or country), the different geographical benchmarks (country or Europe as a whole), and the different measures (absolute or relative) have been the main variations on the methodology adopted to measure specialisation and concentration. Economists have continued to assess the location patterns mainly at a single geographical level despite the coexistence of different institutional levels calls for rigorous methodologies to analyse structural changes at different spatial scales, simultaneously. Evaluating if the coalescence of economic activities is occurring mostly within countries or instead at wider spatial distances helps understanding how and to what extent each supranational, national or regional policy makers has to be involved in designing appropriate policies.

As for the geographic concentration side, some recent developments have been done in this

I am grateful to Eckhardt Bode, Laura De Dominicis, Riccardo Lucchetti, Massimo Tamberi and the participants of the International Workshop on Spatial Econometrics and Statistics held in Rome (25-27 May 2006) for their useful comments and suggestions on earlier drafts. The usual disclaimers apply.

direction. Brülhart e Traeger (2005) presented a nested analysis exploiting the decomposability of entropy measures across geographic subgroups for Europe, while Marcon e Puech (2003) introduced distance-based methods and provided descriptions of the spatial distribution of French manufacturing firms at different geographic scales.

Besides, the empirical evidence was provided focusing either on specialisation or on concentration trends with a limited number of works looking at both the two sides of localisation in a rigorous analytical framework. The adoption of absolute entropy measures helps Aiginger e Davies (2004) in giving support of the opposing trends in specialization of countries (increasing) and in concentration of industries (decreasing) in Europe. Although, results are strongly affected by the methodology used, and Aiginger e Davies (2004) show that the expected identity arise when unweighed relative measures are aggregated. Assessing the changing spatial distribution across the United States, Mulligan e Schmidt (2005) show that, adopting the relative mean deviation (or Krugman index) to measure specialisation and concentration, it is possible to define an unweighted aggregate localisation index. Recently, a general methodology for the construction of polarisation measures which assess concentration and specialisation simultaneously is presented in Bickenbach e Bode (2006).

Yet, an integrated analysis of overall localisation -with concentration on one side and specialisation on the other side- combined with the adoption of a twofold geographical perspective is still a novelty in the literature.

More specifically, the present work goes beyond the existing literature in two respects. First, by computing the dissimilarity entropy index to assess relative specialisation it is possible to define a region-based specialisation measure for each country relative to a supranational benchmark which is decomposable by an inner specialisation component- a weighed average of regional specialisation indices relative to the reference country- and a country bias. Secondly, the overall localisation index introduced here allows to conceptualise specialisation and concentration as the two specular sides of the same concept. The aggregate identity between the two sides of localisation is complemented by the twofold geographical decomposition (into *between country* and *within country* components) underpinning the entire work.

The remainder of the paper is organized as follows. Section 1 reviews the commonly used indicators to assess specialisation and concentration with a specific focus on entropy measures. Section 2 introduces the relative measures of concentration and specialisation are

derived from the original dissimilarity Theil index. Section 3 presents the main methodological contribution: the identity between relative concentration and relative specialisation and its geographical decomposition. Finally, section 4 gives some conclusions and further developments.

1 From absolute to relative measures of localisation

Traditionally, the most frequently used absolute indices in regional studies have been the Gini coefficient, the Herfindhal index (used in Sapir (1996), Aiginger e Pfaffermayr (2004)) and the entropy index (used Aiginger e Pfaffermayr (2004), Aiginger e Davies (2004)). One of the main drawback in the use of the Gini is that it places higher importance on changes in the middle parts of the distribution (See Cowell (1995)). Besides, although the Herfindhal measure makes use of all the information, it turns out to be heavily influenced by the largest shares. The use of entropy indices, which allows to account for minimal distributional changes, stems also in their desirable decomposition properties into subgroups.

Entropy indices have become customary to evaluate income distribution[1] while they have been less used in the analysis of the spatial distribution of economic activities.

Their recent application to industrial location across countries and regions has been mainly based on the generalised entropy measures. Therefore, entropy indices have hitherto been adopted mainly in the Shannon's basic form that can be traced back to the early contributions of information theory (Shannon e Weaver (1949)) (Aiginger e Pfaffermayr (2004), Aiginger e Davies (2004)). Even when they are normalized (Troutt e Acar (2005)), they can only be referred to as absolute concentration indices measuring the distance between the geographical spreading of a specific economic activity and the theoretical uniform distribution across spatial units, without taking into account that the sectoral localisation could be simply the result of a localised manufacturing sector. Relative concentration indices are corrected for the overall spatial distribution of the aggregate activity. In other words, they gauge the extent to which the distribution across spatial units of a specific sector departs from the regional distribution of aggregate manufacturing. Brülhart e Traeger (2005) proposed weighted gen-

[1] In the inequality literature the individual is taken as basic unit of analysis and the country represents the reference *meso-level* to disentangle the two components of the overall world income inequality (Sala-I-Martin (2002)). Instead, in regional studies per capita income inequality across geographical units has been assessed relying on a hierarchical spatial structure (country-region-province levels) (Akita (2000)).

eralised entropy indices which turn out to be conceptualised as dissimilarity measures. The use of the *dissimilarity Theil index* (Theil (1967), Maasoumi (1993)) (the discrete observable analogue of the Kullback-Leibler divergence in probability theory) to assess relative concentration is subject to a straightforward economic interpretation. Indeed, splitting the overall relative concentration into its different components allows to disentangle the contribution of national borders in defining comparative advantages from the magnitude of internal regional agglomeration which may be the result of external economies or intra-firm increasing returns to scale.

Shannon's entropy (Shannon e Weaver (1949)) is an inverse measure of absolute concentration, the higher the entropy the lower the degree of concentration and viceversa (see the Appendix for details on the notation used throughout the paper).

$$E_k = -\sum_{i=1}^{m}(s_{ik})\ln(s_{ik}) \tag{1}$$

Specularly, absolute specialisation is defined as:

$$E_i = -\sum_{k=1}^{n}(v_{ik})\ln(v_{ik}) \tag{2}$$

The *normalised Shannon index*, a deviation of the observed entropy from the uniform distribution (maximum entropy) (used in Aiginger e Pfaffermayr (2004) and Aiginger e Davies (2004) among others), is defined as follows:

$$C_k = \ln(m) + \sum_{i=1}^{m} s_{ik}\ln(s_{ik}) \tag{3}$$

A different measure of normalised entropy can be constructed as follows (Troutt e Acar (2005)):

$$C_k = -(\ln(m))^{-1}\sum_{i=1}^{m} s_{ik}\ln(s_{ik}) \tag{4}$$

Although the territorial implications of industry-specific shock are best assessed by looking at absolute measures, relative indices are more appropriate to capture the degree of localisation of industries and the level of specialisation of countries and regions. While the set of absolute indices drawn on the merely statistical meaning of concentration, their benchmark

being independent from the aggregate employment pattern, relative measures are subject to a more straightforward economic interpretation since they can be envisaged as dissimilarity indicators strictly connected to the concept of comparative advantage. Relative measures are indeed frequently based on the Hoover-Balassa index. Traditionally, in many country-based empirical studies the most commonly used index for measuring relative concentration and relative specialisation has been the Gini location quotient based on the Hoover-Balassa Index[2] (Kim (1995), Amiti (1999), Haaland *et al.* (1999), Brülhart e Torstensson (1996), Brülhart (1998), Brülhart (2001)).

Besides, relative specialisation based on the Manhattan distance (used in Krugman (1991), Hallet (2000), Midelfart *et al.* (2004), Mulligan e Schmidt (2005) among others) view each of the relationship between the numerator and the denominator of the Balassa index as a difference instead of a ratio, and then it is obtained summing these differences across all sectors in the case of specialisation (equation 5) and across all the regions in the case of concentration (equation 7).

$$RS_i = \sum_{k=1}^{n} abs(v_{ik} - v_k) \tag{5}$$

Where $v_k = \frac{\sum_{j \neq i} x_{ik}}{\sum_k \sum_{j \neq i} x_{ik}}$. The index RS_i ranges from 0 (nil relative specialisation with respect to the rest of the EU) to 2 (maximum relative specialisation)[3].

Analogously, a relative concentration based on the Manhattan distance can be easily derived:

$$RC_k = \sum_{i=1}^{m} abs(s_{ik} - s_k) \tag{7}$$

A different relative entropy measure based on location quotient have been proposed to measure specialisation (equation 8) and concentration (equation 9) at the country level (Aiginger e Davies (2004)):

[2]A different version of the location Gini coefficient is used in Midelfart *et al.* (2004) which is based on the ratios $\frac{v_{ik}}{\frac{1}{m}\sum_{i=1}^{m} v_{ik}}$ instead of the usual Balassa indices.

[3]The normalised version

$$RS_i = \frac{1}{2} \sum_{k=1}^{n} abs(v_{ik} - v_k) \tag{6}$$

ranges from o to 1.

$$RSPEC_i = \sum_{k=1}^{n} B_{ik} \ln(B_{ik}) \tag{8}$$

$$RCONC_k = \sum_{i=1}^{m} B_{ik} \ln(B_{ik}) \tag{9}$$

The Theil index was instead adopted by Brülhart e Traeger (2005) as a weighted version of the generalised entropy measure to assess relative concentration:

$$T_k = \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ijk} \ln(\frac{s_{ijk}}{s_{ij}}) \tag{10}$$

The partition properties of relative entropy measures have been exploited to study the spatial concentration of economic activities (Brülhart e Traeger (2005), Mori *et al.* (2005)) but the literature lacks a methodology for dealing with specialisation at different geographical levels. Instead, the aggregation properties of relative entropy indices have been already explored by Aiginger e Davies (2004) and Mulligan e Schmidt (2005). Mulligan e Schmidt (2005) provided numerical examples on the identity between specialisation and concentration through the use of the measures defined in equation 5 and 7 while Aiginger e Davies (2004) found the identity between relative specialisation and relative concentration when they rely on the unweighted aggregation of the relative measures defined in equation 8 and 9.

The contribution of this work is therefore twofold and addresses two methodological issues. Firstly, the dissimilarity Theil index is used to evaluate relative specialisation at different geographical levels. Secondly, an aggregation of weighted relative entropy measures is introduced to derive an overall localisation index which is decomposed into between and within country components.

2 Dissimilarity Theil indices of concentration and specialisation: basic definitions

The information content of an *indirect message*, as defined in the information theory, provided economists with a particularly valuable tool once it was introduced into economics (Theil (1967)). The expected information of an indirect message[4] is:

$$T_{im} = \sum_{i=1} q_i \ln(\frac{q_i}{p_i}) \tag{11}$$

The entropy index in the general form of indirect message is interpretable as a dissimilarity index and it can be used to assess the divergence between two distributions. It may be considered only as a 'distance index' (Maasoumi (1993)), as it does not satisfy some properties necessary to be defined as a proper metric[5].

The dissimilarity version of the Theil index is particularly useful for the study of the spreading of comparative advantages across space. It may be used in two different versions: as a measure of relative concentration, and as a measure of relative specialisation.

The relative concentration index used in Brülhart e Traeger (2005) (equation 10 can be obtained by substituting s_{ijk} for q_i and s_{ij} for p_i in equation 11. Similarly, the two geographical components of each concentration index are obtained (See the appendix for details on the decomposition).

Turning the attention to the specialisation side of localisation, it is possible to compute several measures varying the basic unit of analysis and/or the territorial benchmarks[6]. Clearly, the choice of the b asic unit and the benchmark depends upon the purpose of the analysis.

[4]The indirect message is a generalisation of the direct message. In the latter case, the rationale of the message is noticing that one of the events has actually occurred while in the former case the message modify the probabilities of each event of the distribution to take place. p_i represents the *ex-ante* probability and q_i corresponds to the *ex-post* probability that the event i will occur. The direct message index may be categorised as a dissimilarity index between the actual distribution and the theoretic uniform distribution.

[5]A metrics or a distance function, defined on X*X (X is a not empty set), with $x, y, z \subseteq X$, satisfies the following axiomatic principles :
1)$d(x, y) \geq 0$;
2)$d(x, y) = d(y, x)$;
3)If $d(x, y) = 0$ then $x = y$;
4) $d(x, y) + d(y, z) \geq d(x, z)$. The dissimilarity Theil index satisfies only the principle 1) which says that the distance should be a positive real number, and the principle 3) which states that the minimum distance should be 0.

[6]The superscripts of relative specialisation measures indicate the benchmark distribution: c is country, $^\circ$ is the supranational geographical unit.

If one is interested in assessing the divergence between the regional manufacturing allocation among sectors and the country structure, one has to rely on the regional specialisation index relative to the reference country which is obtained by substituting v_{ik} for p_i and v_{ijk} for q_i in equation 11:

$$T_{ij}^{c} = \sum_{k=1}^{n} v_{ijk} \ln(B_{ijk}) \tag{12}$$

Similarly, if one aims at evaluating the dissimilarity between the allocation across sectors of one region compared to a supranational geographical unit (Europe, USA, world), one has to refer to the following index:

$$T_{ij}^{\circ} = \sum_{k=1}^{n} v_{ijk} \ln(BI_{ijk}^{*}) \tag{13}$$

Finally, if one substitute v_k for p_i and v_{ik} for q_i in equation 11 a measure of the distance between the sectoral composition at the country level and the manufacturing structure of the wider territorial benchmark can be derived:

$$T_{i}^{\circ} = \sum_{k=1}^{n} v_{ik} \ln(B_{ik}) \tag{14}$$

If T_{ij}° takes the value 0, then region j located in country i has a manufacturing structure which is identical to the European average. Similarly, if $T_{ij}^{c} = 0$ then the manufacturing distribution across sectors of region j located in country i mirrors the manufacturing structure of country i. Similarly, if $T_{i}^{\circ} = 0$ then the manufacturing distribution of country i matches the EU distribution across sectors. The higher the index, the more the regional manufacturing structure is dissimilar from the wider geographical unit choosen as a benchmark.

Relative measures are frequently based on the Balassa index. In this setting the dissimilarity Theil index is a weighted sum of the log of the Balassa indices. A noteworthy advantage in the use of the logarithmic transformation is that it allows to deal with the skewness in the original distribution of the Balassa indices[7]. Besides, thanks to its partition properties, the distinction between the evolution of the country comparative advantages and the one of the internal regional localisation turns out to be technically feasible for each sector.

[7]For a detailed analysis of the properties of the Balassa indices of revealed comparative advantage, see de Benedictis e Tamberi (2004).

While in concentration studies two-levels geographical decomposition are emerging (Brülhart e Traeger (2005), Mori *et al.* (2005)), the literature produced so far lacks of a rigorous methodology to disentangle specialisation at regional and national levels (Combes e Overman (2004)). Such a tool would be useful for a better understanding of the European patterns of specialisation and to investigate the mixed trend in regional specialisation which varies with geographical scale. To fill this gap, a way of condensing into a single index of specialisation two descriptive levels (regions and countries) is presented. When the dissimilarity logic is adopted, the country specialisation relative to a supranational unit (T_i°) can be envisaged as a residual of the averaged regional specialisation relative to the same benchmark, once the divergence of the regional manufacturing structures with reference to the country has been accounted for.

Adopting a regional standpoint, the country specialisation is best defined as the averaged regional specialisation indices relative to supranational unit (aRS_i°) and it turns out to be constituted by two elements: an inner country component (aRS_i^c), which accounts for the internal regional specialisation with respect to the country, and the country bias, in other words the country specialisation relative to the whole area (T_i°). The following relation holds:

$$aRS_i^\circ = aRS_i^c + T_i^\circ \tag{15}$$

where:

$$aRS_i^\circ = \sum_{j=1}^{r_i} T_{ij}^\circ s_{ij}^* \tag{16}$$

and

$$aRS_i^c = \sum_{j=1}^{r_i} T_{ij}^c s_{ij}^* \tag{17}$$

In this setting, each country relative specialisation to the supranational unit (T_i°) is simply the difference between the two country-based averaged regional specialisation measures:

$$T_i^\circ = \sum_{j=1}^{r_i} (T_{ij}^\circ - T_{ij}^c) s_{ij}^* \tag{18}$$

To conclude, the weighted average of regional specialisation indices of each country relative to the supranational unit (aRS_i°) is decomposable into a within country component of regional

specialisation (aRS_i^c) and a 'between country' component (T_i°).

3 Overall localisation: concentration and specialisation as the two sides of the same medal

In addressing regional specialisation (paragraph 2), I looked at the structural diversity between the region and the country (T_{ij}^c) or the dissimilarity between the region and the wider territorial benchmark (T_{ij}°). Similarly, the geographical decomposition of relative entropy measures allows to assess the spatial distribution of one industry compared to that of all industries combined, at two geographical level of analysis (Brülhart e Traeger (2005)).

Although, if one aims at evaluating the entire distribution of economic activities across regions and sectors it is better to refer to aggregate measure of localisation.

Intuitively, specialisation and concentration should be strictly connected. Statistically, the relationship between the two dimensions of localisation depends on the nature of the local indices adopted. If one relies on relative measures, the reason why the identity at the aggregate level should always apply is that in the construction of each relative measures the researcher makes use of the information of the entire matrix where the rows refer to sectors and the columns refer to countries[8]. Instead, absolute measures are constructed through the use of the limited information provided by each row (column) of a matrix. In this case the only benchmark is the theoretical uniform distribution of each row (column) with any connection with the rest of the matrix distribution. As a matter of fact, in the literature, the expected identity has been found only when unweighted relative entropy measures (Aiginger e Davies (2004)) or the commonly used relative mean deviation (Mulligan e Schmidt (2005)) were adopted.

In the present contribution the recognition of the aggregate identity between the two sides of localisation is complemented by a twofold geographical scale perspective. The analytical framework presented allows to consider concentration and specialisation as the specular manifestations of the same concept, both across and within countries.

Overall localisation is assessed through a condensed dissimilarity index in which the log

[8]Specialisation is observed by reading down each column, whilst concentration is computed along each row.

of Balassa indices are weighted by sectoral regional shares of the aggregate manufacturing (v^*_{ijk}):

$$L = \sum_{k=1}^{n} \sum_{i=1}^{m} \sum_{j=1}^{r_i} v^*_{ijk} \ln(B^*_{ijk}) \tag{19}$$

Since $v^*_{ijk} = v_k s_{ijk} = s_{ij} v_{ijk}$

it is possible to rewrite the average dissimilarity measure (L) as follows:

$$L = \sum_{k=1}^{n} \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ij} v_{ijk} \ln(B^*_{ijk}) = \sum_{k=1}^{n} \sum_{i=1}^{m} \sum_{j=1}^{r_i} v_k s_{ijk} \ln(B^*_{ijk}) \tag{20}$$

Substituting equations 13 and 37 into equation 20 we derive a first intuition of the twofold interpretation of the total averaged dissimilarity index:

$$L = \ell(S) = \ell(C) \tag{21}$$

where S is the matrix of the sectoral contributions to specialisation relative to the supranational economy ($v_{ijk} \ln(B^*_{ijk})$) and C is the matrix of regional contributions to relative concentration ($s_{ijk} \ln(B^*_{ijk})$).

Equation 21 refers to the twofold connotation of the concept of localisation. From the specialisation point of view, the aggregation gives an idea of the average dissimilarity between the regional distribution across sectors and the manufacturing structure of the supranational economy selected as benchmark. Similarly, from a concentration standpoint, the composite measure of localisation informs about the average dissimilarity between the distribution across geographical units of sectors and the location across geographical units of overall manufacturing. As a matter of fact, typical dissimilarity is a summary statistics of both relative specialisation indices and relative concentration ones, weighted by regional shares (s_{ij}) and sectoral shares (v_k) of aggregate manufacturing of the whole area, respectively:

$$L = \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ij} T^{\circ}_{ij} = \sum_{k=1}^{n} v_k T_k \tag{22}$$

4 Decomposing overall localisation

The entropy measure of overall localisation introduced in the previous paragraph meets several baseline principles outlined by Combes e Overman (2004) as 'the perfect measure' requirements:

- It is comparable across industrial sectors;
- It comparable across spatial units and scales;
- It specifies an unambiguous and meaningful null hypothesis of no localisation (when the index has a nil value);
- It is suitable for statistical testing (bootstrap).

Although it is biased by the modifiable areal unit problem and the checkerboard problem (Arbia (1989)), it is endowed with a set of desirable decomposition properties by geographical subgroups.

The decomposition of the overall dissimilarity index is obtained by means of the relation between the location quotients used so far. Since the following relation holds:

$$B^*_{ijk} = B_{ik}B_{ijk} \tag{23}$$

Then localisation index defined in equation 19 becomes:

$$\sum_{k=1}^{n}\sum_{i=1}^{m}\sum_{j=1}^{r_i} v^*_{ijk}\ln(B_{ijk}B_{ik}) \tag{24}$$

$$\sum_{k=1}^{n}\sum_{i=1}^{m}\sum_{j=1}^{r_i} v^*_{ijk}\ln(B^*_{ijk}) = \sum_{k=1}^{n}\sum_{i=1}^{m}\sum_{j=1}^{r_i} v^*_{ijk}\ln(B_{ijk}) + \sum_{k=1}^{n}\sum_{i=1}^{m} v^*_{ik}\ln(B_{ik}) \tag{25}$$

The between country component (L^b) of total averaged dissimilarity (L) is:

$$L^b = \sum_{k=1}^{n}\sum_{i=1}^{m} v^*_{ik}\ln(B_{ik}) \tag{26}$$

$$= \sum_{k=1}^{n}\sum_{i=1}^{m} s_i v_{ik}\ln(B_{ik}) = \sum_{k=1}^{n}\sum_{i=1}^{m} v_k s_{ik}\ln(B_{ik}) \tag{27}$$

$$= \sum_{i=1}^{m} s_i T_i^{\circ} = \sum_{k=1}^{n} v_k T_k^b \tag{28}$$

The within country component (L^w) of total localisaation (L) is:

$$L^w = \sum_{k=1}^{n} \sum_{i=1}^{m} \sum_{j=1}^{r_i} v_{ijk}^* \ln(B_{ijk}) \tag{29}$$

$$= \sum_{k=1}^{n} \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ij} v_{ijk} \ln(B_{ijk}) = \sum_{k=1}^{n} \sum_{i=1}^{m} \sum_{j=1}^{r_i} v_k s_{ijk} \ln(B_{ijk}) \tag{30}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ij} T_{ij}^c = \sum_{k=1}^{n} v_k T_k^w \tag{31}$$

The overall localisation index (L) is an averaged dissimilarity index and it is interpretable as a summary statistics of regional specialisation indices (T_{ij}°) weighted by the manufacturing regional shares (s_{ij}):

$$L = \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ij} T_{ij}^{\circ} = \sum_{i=1}^{m} s_i T_i^{\circ} + \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ij} T_{ij}^c \tag{32}$$

Similarly, overall localisation can be seen as a summary statistics of relative concentration Theil indices (T_k) weighted by the industry shares (v_k):

$$L = \sum_{k=1}^{n} v_k T_k = \sum_{k=1}^{n} v_k T_k^b + \sum_{k=1}^{n} v_k t_k^w \tag{33}$$

The twofold identity- with specialisation on one side and concentration on the other side- still holds for each component (both equation 32 and equation 33 correspond to 34):

$$L = L^b + L^w \tag{34}$$

where:

$$L^b = \ell(S^b) = \ell(C^b) \tag{35}$$

S^b is the matrix of the sectoral contributions to country specialisation relative to the supranational spatial unit ($v_{ik} \ln(B_{ik})$) and C^b is the matrix of the country contributions to

between country concentration ($s_{ik} \ln(B_{ik})$).

and:

$$L^w = \ell(S^w) = \ell(C^w) \tag{36}$$

S^w is the matrix of the sectoral contributions to inner-country regional specialisation ($v_{ijk} \ln(B_{ijk})$) and C^w is the matrix of regional contribution to within country relative concentration ($s_{ijk} \ln(B_{ijk})$).

The overall dissimilarity indicator introduced here allows to conceptualise relative specialisation and relative concentration as two sides of the same medal, as in Aiginger e Davies (2004) but in the present case the aggregation is obtained applying to each concentration index the corresponding industry shares of the aggregate manufacturing. Similarly, the overall dissimilarity index is a weighted dissimilarity index of regional specialisation indices weighted by the regional shares of aggregate manufacturing.

To have a closer look to the way in which the twofold identity works see table 1. Smaller sectors are likely to be more geographically concentrated than larger sectors. Specularly, the regional size affects its degree of specialisation, smaller regions/countries usually are more specialised than larger regions/countries. It is worth noting that the overall localisation index is less biased with respect to sectoral scale and regional scale than the specific measures of concentration and specialisation (see table 1).

	s_i (a)		aRS_i^{EU} (b)	S_i^{EU} (a)*(b)	T_i^{EU} (c)	S_i^b (a)*(c)	aRS_i^c (d)	S_i^w (a)*(d)
Belgium and Luxembourg	0.030		0.169	0.005	0.035	0.001	0.133	0.004
Germany	0.266		0.112	0.030	0.050	0.013	0.063	0.017
Spain	0.110		0.154	0.017	0.044	0.005	0.110	0.012
Finland	0.013		0.137	0.002	0.090	0.001	0.047	0.001
France	0.159		0.096	0.015	0.015	0.002	0.081	0.013
Greece	0.009		0.318	0.003	0.166	0.002	0.152	0.001
Italy	0.194		0.129	0.025	0.049	0.009	0.080	0.016
Netherlands	0.039		0.098	0.004	0.054	0.002	0.044	0.002
United Kingdom	0.181		0.128	0.023	0.012	0.002	0.116	0.021
Total	**1**	**L**		**0.124**	L^b	**0.038**	L^w	**0.086**

	v_k (a)		T_k (b)	C_k (a)*(b)	T_k^b (c)	C_k^b (a)*(c)	T_k^w (d)	C_k^w (a)*(d)
Food	0.119		0.112	0.013	0.031	0.004	0.081	0.010
Textiles	0.074		0.287	0.021	0.140	0.010	0.147	0.011
Wood	0.027		0.167	0.004	0.062	0.002	0.105	0.003
Paper	0.081		0.128	0.010	0.041	0.003	0.087	0.007
Chemicals	0.064		0.160	0.010	0.021	0.001	0.139	0.009
Rubber and plastics	0.054		0.073	0.004	0.010	0.001	0.063	0.003
Other non-metallic mineral products	0.046		0.167	0.008	0.036	0.002	0.132	0.006
Basic metals and fabricated metal products	0.150		0.070	0.010	0.008	0.001	0.062	0.009
Machinery and equipment	0.114		0.088	0.010	0.040	0.005	0.048	0.006
Electrical and optical equipment	0.117		0.090	0.011	0.027	0.003	0.063	0.007
Transport equipment	0.101		0.169	0.017	0.051	0.005	0.118	0.012
Manufacturing nec	0.052		0.095	0.005	0.029	0.002	0.065	0.003
Total	**1**	**L**		**0.124**	L^b	**0.038**	L^w	**0.086**

Note: C_k is the weighted Theil index of concentration where the weight v_k is the sectoral share of E-10 manufacturing employment. Similarly, weighted measures of concentration are obtained in the successive columns for within country indices (C_k^w) and between-country ones (C_k^b).

As for the specialisation side, S_i^{EU} is the weighted average of regional specialisation relative to E-10 for each country i where the weight s_i is the country's proportion of all E-10 manufacturing employment. Similarly, S_i^b is the weighted country-based relative specialisation with respect to E-10; and S_i^w is the weighted sum of regional specialisation relative to the country manufacturing structure. $aRS_i^{EU} = \sum_{j=1}^{r_i} T_{ij} s_{ij}^*$ and $aRS_i^c = \sum_{j=1}^{r_i} T_{ij}^c s_{ij}^*$

Source: SBS-Region dataset, Eurostat

Table 1: Twofold decomposition of the entropy measure of overall localisation, 2001

5 Concluding remarks and further developments

This paper provides a decomposition methodology based on the use of dissimilarity entropy indices. The decomposition introduced allows to overcome a typical shortcoming of the existing empirical literature, namely the focus on a single geographical level of analysis. Thanks to the partition properties of dissimilarity entropy measures, it is possible to handle two relevant hierarchical grids assessing the evolution of their relative importance in the pattern of localisation of economic activities.

The paper presents a taxonomy of concentration and specialisation measures derived from the original dissimilarity Theil index. On the specialisation side, the country index is defined as a mere difference between the weighted average of the regional specialisation relative to the supranational geographical unit and the inner country regional specialisation.

The concept of localisation with its connotations of specialisation and concentration finds an explicit statistical counterpart here. In fact, the paper contends that, when the dissimilarity logic is adopted, specialisation and concentration can be seen as two side of the same medal both across and within country. The indicator of overall localisation allows to contemplate the full structure of industries and regions in a single analysis since it is constructed through the aggregation of the specific indices used to assess, on one side, regional (and country) specialisation and, on the other side, sectoral localisation.

Further developments of the methodology are possible. The decomposition in the sectoral dimension, i.e. between high-tech and low-tech sectors is one promising possibility. Moreover, the methodology might be improved through several refinements to overcome the MAUP and the checkerboard problem both in the sectoral and spatial dimension.

A Appendix

A.1 Notation

x_{ijk} variable of main interest: number of workers in sector k (k=1,...,n) in region j(j=1,...,r_i) located in country i (i=1,...,m)

x_{ij} total employment in region ij

x_{ik} total employment in sector k in country i

x_i total employment in country i

x_k total employment in sector k at the higher level of spatial aggregation

x total employment at the higher level of spatial aggregation

v_{ijk} $:= \frac{x_{ijk}}{x_{ij}}$ share of sector k in total employment of region ij

v_{ik} $:= \frac{x_{ik}}{x_i}$ share of sector k in total employment of country i

v_k $:= \frac{x_k}{x}$ share of sector k in total EU-10 employment

v^*_{ijk} $:= \frac{x_{ijk}}{x}$ share of sector k of region ij in total employment

v^*_{ik} $:= \frac{x_{ik}}{x}$ share of sector k of country i in total employment

s_i $:= \frac{x_i}{x}$ share of country i in total employment

s_{ij} $:= \frac{x_{ij}}{x}$ share of region ij in total employment

s_{ik} $:= \frac{x_{ik}}{x_k}$ share of sector k of country i in total employment of sector k

s_{ijk} $:= \frac{x_{ijk}}{x_k}$ share of sector k of country i in total employment of sector k

s^*_{ij} $:= \frac{x_{ij}}{x_i}$ share of region ij in total employment of country i

s^*_{ijk} $:= \frac{x_{ijk}}{x_{ik}}$ share of sector k of region ij in total employment of sector k and country i

B^*_{ijk} $:= \frac{v_{ijk}}{v_k}$ regional location quotient relative to the supranational economy

B_{ijk} $:= \frac{v_{ijk}}{v_{ik}}$ regional location quotient relative to the country

B_{ik} $:= \frac{v_{ik}}{v_k}$ country location quotient relative to the supranational economy

A.2 Decomposing the relative concentration index

As already pointed out relative concentration refers to the dissimilarity in the localisation of each sector k with respect to the spreading of the overall manufacturing sector across the spatial units considered (countries, regions). If a sector k spreads exactly proportionally to total manufacturing employment the relative concentration index will exhibits a nil value.

$$T_k = \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ijk} \ln \frac{v_{ijk}}{v_k} \tag{37}$$

Adding and subtracting the term $\sum_{i=1}^{m} s_{ik} \ln(v_{ik})$ to equation 37 the following equation is obtained:

$$T_k = \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ijk} \ln \frac{v_{ijk}}{v_k} + \sum_{i=1}^{m} s_{ik} \ln(v_{ik}) - \sum_{i=1}^{m} s_{ik} \ln(v_{ik}) \tag{38}$$

$$T_k = \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ijk} \ln(v_{ijk}) - \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ijk} \ln(v_k) + \sum_{i=1}^{m} s_{ik} \ln(v_{ik}) - \sum_{i=1}^{m} s_{ik} \ln(v_{ik}) \tag{39}$$

and because $\sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ijk} = \sum_{i=1}^{m} s_{ik}$

$$T_k = \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ijk} \ln(v_{ijk}) - \sum_{i=1}^{m} s_{ik} \ln(v_k) + \sum_{i=1}^{m} s_{ik} \ln(v_{ik}) - \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ijk} \ln(v_{ik}) \tag{40}$$

Combining the second and the third elements the *between country* component is obtained:

$$T_k^b = \sum_{i=1}^{m} s_{ik} \ln \frac{v_{ik}}{v_k} \tag{41}$$

instead, the *within country* component is obtained combining the first element of 40 with the forth one:

$$T_k^w = \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ijk} \ln(v_{ijk}) - \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ijk} \ln(v_{ik}) \tag{42}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ijk} \ln \frac{v_{ijk}}{v_{ik}} \tag{43}$$

so that

$$T_k = T_k^{bc} + T_k^{wc} \tag{44}$$

$$T_k = \sum_{i=1}^{m} s_{ik} \ln \frac{v_{ik}}{v_k} + \sum_{i=1}^{m} \sum_{j=1}^{r_i} s_{ijk} \ln \frac{v_{ijk}}{v_{ik}} \tag{45}$$

The Theil within countries (T_k^w) is interpretable as a weighted average of the Theils indices between regions inside each country, where the weights are the shares of the countries in total employment in sector k (s_{ik}).

$$T_k^w = \sum_{i=1}^{m} \frac{X_{ik}}{X_k} T_{ik}^{br} = \tag{46}$$

$$= \sum_{i=1}^{m} \frac{X_{ik}}{X_k} \sum_{j=1}^{r_i} \frac{X_{ijk}}{X_{ik}} \ln \frac{\frac{X_{ijk}}{X_{ik}}}{\frac{X_{ij}}{X_i}} = \tag{47}$$

$$= \sum_{i=1}^{m} s_{ik} \sum_{j=1}^{r_i} s_{ijk}^* \ln \frac{s_{ijk}^*}{s_{ij}^*} \tag{48}$$

The relative concentration index has no upper bound and the lower limit is 0.

- If $T_k = 0$ it must be that $BI_{ijk}^* = 1$ for each regions in the area; in this case the localisation of manufacturing sector k overlaps the distribution of the overall manufacturing sector so that in sector k no region shows neither a comparative advantage nor a comparative disadvantage with respect to the overall area (in our case the European Union), $T_k = 0$ occurs when $s_{ijk} = s_{ij}$ for each region, so that the sector k is distributed across the European regions in the same way as the total manufacturing sector span across the same regions. $T_k = 0$ is thus a benchmark of no relative concentration. The higher the value of total relative concentration index (T_k) is, the more the allocation of regional comparative advantages in sector k compared with the EU is uneven. An increasing total relative concentration index over time denotes a process of regional specialisation in that sector somewhere in Europe.

- If $T_k^w = 0$ then $BI_{ijk} = 1$ for each region; no internal region exhibits a comparative advantage (disadvantage) in sector k compared with the national counterpart($s_{ijk}^* = s_{ij}^*$), $T_k^w = 0$ defines a benchmark of *no-relative concentration within countries*.

  If the within component of the relative concentration index is 0 then sector k is proportionally distributed to total manufacturing employment in the internal regions of each country. Put it differently, a nil value of 'within' relative concentration suggests that no internal region comparative advantage with respect to the reference country exists. The higher the domestic component is, the more the inner allocation of comparative advantages of each country is uneven. An increasing value of the 'within country' factor is related to a process of rising diversification internal to the countries.

- If $T_k^b = 0$ it must be that $BI_{ik} = 1$ for each i; no country has a comparative advantage (disadvantage) in sector k with respect to the overall area ($s_{ik} = s_i$), $T_k^b = 0$ defines a benchmark of *no-relative concentration between countries*;

- $T_k = 0$ implies that both $T_k^w = 0$ and $T_k^b = 0$

- Either $T_k^w = 0$ or $T_k^b = 0$ does not imply that $T_k = 0$.

  The 'across country' factor embodies the importance of national comparative advantages with respect to the supranational area considered as a benchmark. A nil value of T_k^{bc} implies that the across countries distribution of sector k overlaps perfectly the allocation across countries of manufacturing as a whole. In other words, countries reveal neither a comparative advantage nor a comparative advantage in the specific sector k analysed. Accordingly, the higher the 'between country' component is, the more the allocation of national comparative advantages in sector k is unbalanced. An increase in the 'between country' component of relative concentration indicates an increasing unequal allocation of comparative advantages, associated to a process of country specialisation.

References

Aiginger, K. e Davies, S. W. (2004). *Industrial specialisation and geographic concentration: two sides of the same coin? not for the European Union.* Journal of Applied Economics, VII(2): 231–248.

Aiginger, K. e Pfaffermayr, M. (2004). *The Single Market and geographic concentration in Europe.* Review of International Economics, 12(1): 1–11.

Akita, T. (2000). *Decomposing regional income inequality using two-stage nested Theil decomposition method.* Paper presented at the 6th World Congress of the Regional Science Association International, Lugano, May.

Amiti, M. (1999). *Specialisation patterns in Europe.* Weltwirtschaftliches Archiv, (135): 573–593.

Arbia, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems.*

Bickenbach, F. e Bode, E. (2006). *Disproportionality Measures of Concentration, Specialization, Polarization.* Kiel Working Paper N. 1276, Kiel Institute for the World Economy.

Brülhart, M. (1998). *Trading places: industrial specialisation in the European Union.* Journal of Common Market Studies, 36(3): 329–346.

— (2001). *Evolving geographical concentration of European Union.* Weltwirtschaftliches Archiv, 137(2): 215–243.

Brülhart, M. e Torstensson, J. (1996). *Regional integration, scale economies and industry location in the European Union.* CEPR Discussion Paper N. 1435.

Brülhart, M. e Traeger, R. (2005). *An account of geographic concentration patterns in Europe.* Regional Science and Urban Economics, (35): 597–624.

Combes, P.-P. e Overman, H. (2004). *The spatial distribution of economic activities in the European Union*, vol. Handbook of Urban and Regional Economics, vol. 4. Elsevier-North Holland, Amsterdam.

Cowell, F. A. (1995). *Measuring Inequality*. Prentice Hall, Hemel Hempstead, 2a ed.

de Benedictis, L. e Tamberi, M. (2004). *Overall Specialization Empirics: Techniques and Applications*. Open Economies Review, 15(4): 323–346.

Haaland, J. I., Kind, H. J., Midelfart-Knarvik, K. H. e Torstensson, J. (1999). *What determines the economic geography of Europe?* CEPR Discussion Paper N. 2072.

Hallet, M. (2000). *Regional specialisation and concentration in the EU*. Economic papers n. 141, The European Commission, Directorate General for Economic and Financial Affairs.

Kim, S. (1995). *Expansion of markets and the geographic distribution of economic activities: the trends in US regional manufacturing structure, 1860-1987*. The Quarterly Journal of Economics, pp. 881–908.

Krugman, P. (1991). *Geography and Trade*. MIT Press, Cambridge.

Maasoumi, E. (1993). *A compendium to information theory in economics and econometrics*. Econometric reviews, 12(2): 137–181.

Marcon, E. e Puech, F. (2003). *Evaluating the geographic concentration of industries using distance-based methods*. Journal of Economic Geography, (3): 409–428.

Midelfart, K. H., Overman, H. G., Redding, S. J. e Venables, A. J. (2004). *The Location of European industry*, vol. European Integration and the functioning of product markets. Edwarg Elgar.

Mori, T., Nishikimi, K. e Smith, T. E. (2005). *A divergence statistic for industrial localization*. The Review of Economics and Statistics, 87(4): 635–651.

Mulligan, G. F. e Schmidt, C. (2005). *A Note on Localisation and Specialization*. Growth and Change, 36(4): 565–576.

Sala-I-Martin, X. (2002). *The disturbing 'rise' of global income inequality*. NBER Working Paper N° 8904.

Sapir, A. (1996). *The effects of Europe's internal market program on production and trade: a first assessment*. Weltwirtschaftliches Archiv, (132): 457–475.

Shannon, C. E. e Weaver, W. (1949). *The Matematical Theory of Communication*. Urbana.

Theil, H. (1967). *Economics and Information Theory*. North-Holland, Amsterdam.

Troutt, M. D. e Acar, W. (2005). *A Lorenz-Pareto measure of pure diversification*. European Journal of Operational Research, 167: 543–549.