Supplementary material to paper

"Cluster-based measures of regional concentration. Critical overview"

Katarzyna Kopczewska
University of Warsaw, Faculty of Economic Sciences
Ul. Dluga 44/50, 00-241 Warszawa, Poland
e-mail: kkopczewska@wne.uw.edu.pl
ORCID: http://orcid.org/0000-0003-1065-1790

The supplementary file contains the description of the measures of sectoral and geographical concentration.

Definitions for all measures:

$s_{ij}^s = \frac{y_{ij}}{y_j} = \frac{y_{ij}}{\sum_i y_{ij}}$	is the ratio of activity in sector *i* in region *j* to total activity (all *i* sectors) in the region j
$s_i = \frac{y_i}{y} = \frac{\sum_j y_{ij}}{\sum_i \sum_j y_{ij}}$	is the ratio of activity in sector *i* in all *j* regions to total activity (all *i* sectors in all *j* regions)

A) Sectoral concentration – indices calculated by industries for given region

The Gini Index - traditional sectoral index, introduced by Gini (1909, 1936) based on the transformed empirical distribution of the economic activity:

$$GINI_j^s = \frac{2}{n^2 \bar{R}} \sum_{i=1}^{n} \Lambda_i |R_i - \bar{R}|$$

where n is number of industries, $R_i = \frac{s_{ij}^s}{s_i}$ is ratio of activity , $\bar{R} = \frac{1}{N}\sum_{i=1}^{n} R_i$ is the average of activity ratio (by industries), Λ_i is the rank of the industry's position in descending order of R_i, s_{ij}^s and s_i defined as above. This Gini index is formulated for single regions and all sectors, and the reference points are all sectors in all regions. Thus it measures the average under/over representation of the sectors in a given region in comparison with the benchmark given by all regions (a kind of average regional structure). It takes values from 0 to 1, where 0 is for the uniform distribution of activity among the sectors within the region, while 1 is for a full sectoral concentration in the region.

The Relative Specialisation Index (RSI), is calculated as a **maximum of Location Quotient (LQ)** in the region (by sectors). The index introduced by Duranton and Puga (2000), based on the transformed empirical distribution of economic activity, is calculated as $RSI_j = max_i(LQ_{ij})$, while $LQ_{ij} = {S_{ij}^s}/{S_i}$,

where S^s_{ij} and S_i are as defined above. The higher the over-representation of the sector in the region, the higher the regional RSI. It takes the values from 0 to max, where 0 is for the underrepresentation of all sectors in the region and max is for the degree of over/under-representation of the sector in the region.

The Location Quotient - one of the most popular and simplest measures of concentration, introduced by Hoover (1936). It is calculated as the relation of the local sectoral activity to the regional sectoral activity:

$$LQ = \frac{s_{ij}^{s}}{s_i} = \frac{y_{ij} \Big/ \sum_i y_{ij}}{\sum_j y_{ij} \Big/ \sum_i \sum_j y_{ij}}$$

where the counter $s_{ij}^{s}$and denominator $s_i$are defined as above.. Its construction is similar the components of the Gini index. Sectoral LQ's summed up are often called a measure of *specialisation*, even if they measure sectoral concentration only. It is the measure of relative regional activity, comparing the distributions of activity by industry, having all regions together (country) as the reference area. It gives the values for each cell.

The Hachman Index of Economic Diversification – is based on the Location Quotient and empirical distribution of the economic activity. It is calculated as an inverse total of the sectoral LQs, weighted with the share of regional-sectoral activity. It is expressed as follows:

$$HI = \frac{1}{\sum_{i=1}^{n}[(\frac{s_{ij}^{s}}{s_i}) \cdot s_{ij}^{s}]} = \frac{1}{\sum_{i=1}^{n}[LQ \cdot s_{ij}^{s}]}$$

where $s_{ij}^{s}$and $s_i$are defined as above. It is the measure of similarity of regional and national industrial structures. It is limited between 0 and 1, where 0 is when the region has a completely different structure than the country and 1 is for the same industrial structure on a regional and national level.

The Ogive Index – is introduced by Tress (1938) to measure industrial diversity. It is mainly a measure of the export structure, but it is sometimes applied to the regional production structure. It is based on the uniform (equal) distribution of the export shares treated as a benchmark, and it captures the deviations from it. In the regional version, the equal distribution of activity (e.g. employment) is a benchmark. The formula for the regional Ogive index is expressed as follows:

$$Ogive = \sum_{i=1,j}^{n} \frac{(s_{ij}^{s} - \frac{1}{n})^2}{\frac{1}{n}}$$

where s_{ij}^{s} is defined as above (empirical share of activity in given sector), and $1/n$ is the ideal share of activity in a given industry, resulting from equal distribution. Values of the Ogive index (between 0 and max) are for the whole region, and define diversification or sectoral concentration of activity in the region analysed. When the activity shares among sectors are equal, the Ogive index is 0, which is interpreted as perfect diversity. The more diversified (different, unequal) the values, the higher the Ogive measure.

The Refined Diversification Index – is one the earliest popular measures of the economic diversity of the regions, introduced by Tress (1938) and developed by Isard (1960). The "*crude index of diversification*" is defined as follows: one starts with shares of activity in all *k* sections in a given region (denoted as x (%)), ranks them decreasingly and sums up cumulatively (each crude share is associated with the cumulative one, which is the sum of all bigger shares and itself). The crude diversification index is the sum of all those cumulative sums. If all activity is concentrated in one sector, the diversification index has a value of *k**100. The refined diversification index is defined as: $(crude.div.index_{region.1}\text{-}crude.div.index_{regions.all})/(crude.div.index_{max}\text{-}\ crude.div.index_{regions.all})$. This is an absolute measure, with no reference to neighbours and other regions. It gives one single value for all

sectors in one region. This index ranges from 0 to 1, where 0 is for the full diversification of the region (equal shares), and 1 is for complete non-diversification.

The Krugman Dissimilarity Index - also referred as to the Krugman specialisation index, introduced by Krugman (1991a, p.76), is based on the standard error concept and measures the standard error of the industry shares. It is based on an empirical distribution and is expressed as follows:

$$K_j = \sum_{i=1}^{n} |s_{ij}^{s} - s_i|$$

which means it is the total by *n* industries, summing up the differences between the share of activity in a given industry *i* in a given region *j* and the share of activity in a given industry in all regions (or reference area). The total is for the absolute values (| |) of the differences. The minimum value is 0, the maximum is $2 \cdot (n-1)/n$, where 0 is fully consistent with the referential one for the industrial

structure and the more dissimilar the structure, the higher the Krugman measure. The higher the Krugman dissimilarity K_j index value, the stronger the deviation of the regional economic structure from the average reference structure. The maximum asymptotic value is 2 (e.g. for 1 000 000 sectors this limit is 1,9999999998), and for a small *n* it is less (as $1{,}333_{n=3}$, $1{,}75_{n=8}$, $1{,}9_{n=20}$ etc.). This overrepresentation of activity in a given industry is often treated as a specialisation, but in fact this is only the regional structure of the industries.

The Relative Diversity Index (RDI) - introduced by Duranton and Puga (2000), is calculated as an inverse Krugman Dissimilarity Index, which compares the regional and national structure by summing up (by industries for one region) the absolute values of the differences between the regional share of the industry and the national one. The RDI is as follows:

$$RDI_j = \frac{1}{K_j}$$

The more similar the regional and national economies, the smaller the Krugman dissimilarity index and the higher the RDI.

The Hallet Index of industrial concentration – introduced by Hallet (2000) resembles Krugman's dissimilarity index after slight modifications – it is in fact the half of Krugman's index. It is as follows:

$$S_i = \frac{1}{2}\sum_{i=1}^{n} |y_{i,j} - \overline{y_i}|$$

where $y_{i,j}$ is originally the share of the Gross Value Added (GVA) in a given region in a given sector (and may be substituted with e.g. employment) and $\overline{y_i}$ is the national average summed over all sectors. This indicator compares the absolute difference between the shares in activity (e.g. GVA or employment) delivered in region in sector, to the over-regional average of this value in the sector, summed up across all sectors. The minimum value is 0, and appears when the activity structures in the region are the same as in the over-regional distribution (like in a country or macro-region). The maximum value is 0.5 when the structures differ significantly (or even completely). The higher the value, the stronger the sectoral concentration. A zero value means equal shares of industries (no over-representation). The unit value represents the extreme, single sector concentration.

The National Averages Index (NAI) - is based on a concept of squared difference, as it is as follows (w.g. Wund, 1992):

$$NAI = \sum_{i=1}^{n} \frac{(s_{ij}^{s} - s_i)^2}{s_i}$$

which means it is total by *n* industries, summing up squared differences between the share of activity in a given industry *i* in given region *j* (s_{ij}^{s}) and share of activity in a given industry in all regions (or reference area) (s_i). It is based on the same concept as Krugman's dissimilarity index. It ranges from 0 to max. When the economic structure of a region is the same as in the country (all regions), then NAI=0, which is interpreted as a low disparity the between the national and regional economy. The higher the disparity, the higher the value of the NAI.

Shannon's H, Theil's H and Relative H – are based on the entropy concept, which is to measure the deviation of the analysed distribution from full concentration (minimum of entropy) or from full dispersion (maximum of entropy). Full dispersion is mostly given with the uniform distribution, where the probabilities of all events are equal. Similarly to the Ogive index, it refers the empirical distribution to the uniform benchmark distribution[1]. On the basis of the Shannon entropy (Shannon, 1948), Horowitz and Horowitz (1968) developed the regional entropy measure of competition H, expressed as:

$$H = -\sum_{n=1}^{N} s \cdot \ln s$$

where *s* is the probability of the point (discrete) event, and *N=1,2,…,n* is the number of events. In regional studies in the regional concentration measurement, s is the share of activity in a given sector in a given region with reference to the full regional activity ($s = s_{ij} = \frac{y_{ij}}{\sum_{i=1}^{n} y_{ij}}$) and the number of events *n* is the number of industries inside the region. The maximum value of Shannon's H is for equal probabilities of all events (uniform distribution), is $s = 1/n$ and takes the value $H_{max} = -n \cdot (1/n) \cdot ln\, 1/n = \ln n$. The minimum value of Shannon's H (H_{min}=0) is available only in the case of a full concentration of activity in one single industry, $s = 1/1$. If there are two equally likely events with s=1/2 then H=1. One should note that the entropy is directly log-proportional to the number of industries, the higher the number of sectors, the higher the entropy measure.

It can be easily transformed from information theory to competitiveness. When there are many firms in the sector or many sectors in a region, the uncertainty grows and the entropy increases. The market with an equal share of all firms has the highest degree of competitiveness, and conversely the more diversified the shares of a company (with dominating firms) the lower the competitiveness and the lower the risk of operating. An extreme point of a single firm in the industry is a monopoly which operates without competition and risk. A high entropy is then for a high competitiveness with high *n*. In regional applications, *N* is the number of industry classes and *s* is the share (proportion) of each industry (i.e. in the employment). Then the maximum H is obtained at full industrial diversification (equal shares of all industries) and full sectoral concentration (single industry in a region) for a minimum entropy H.

The formula above is transformed to give the **relative entropy** R:

$$R = \frac{H}{H_{max}} = \frac{H}{\ln n}$$

where H is the measured entropy and $\ln n$ is the maximum entropy for a finite number of *n* events. It gives the missing gap between the observed and potential entropy, and thus the degree of getting to the highest competitiveness, assuming a given number of sectors. The interpretation is as follows: R=1 for equal shares of industries within the region, R=0 is for a full concentration of the industry.

[1] Entropy designates a "*measure of the disorder of a system*", the "*measure of unpredictability of information content*" as well "*the uncertainty associated with a random variable*". In information technology it is understood as the expected value of the information in the message. In terms of predictability, the lower the entropy, the lower the risk and the higher the predictability.

Theil's entropy is a measure built on the Shannon entropy. By relativising the input, it measures the disorder within the measure, namely the degree to which Theil's entropy deviates from the maximum Shannon entropy. It is expressed as:

$$T_{Theil} = H_{max} - H_{Theil}$$

where H_{max} is the maximum Shannon entropy (for equal distribution) and H_{Theil} is the Shannon entropy for the observed data. Thus Theil's entropy is the gap between the observed and the maximum entropy, and it is called *redundancy*.

The Kullback-Leibler Divergence (KLD) – is the so called relative entropy measure. It was developed by Kullback and Leibler (1951) and it assesses two distributions and the direct divergence between them. In the information theory it is understood as the information lost when A is approximating B. It can be expressed as follows:

$$KLD_j = \sum_{i=1}^{n} s_{ij} ln \frac{s_{ij}}{q_i}$$

where q_i stands for the share of activity in sector *i* in the country and s_{ij} is defined as above as the share of activity in a given sector in a given region with reference to the full regional activity ($s = s_{ij} = \frac{y_{ij}}{\sum_{i=1}^{n} y_{ij}}$). The KLD is thus the sum by *n* sectors (*i=1,2,..,n)* for a given region *j*. It is to measure the difference in economic stucture at the regional and national level. Because of the construction, the KLD is always non-negative. A minimum value of KLD=0 is for a full similarity of the distributions. The higher the divergence of the two distributions, the higher the value of the KLD. It might be indefinite for the mono-industry and situations with an absent industry in the region. This is because *ln(0)* is indefinite. Thus, the precondition of the KLD analysis would be that all national industries are represented in a region. Mori et al. (2005) introduced it in regional science to compare two distributions of economic structure: regional and national ones. They apply the rule that *0·ln(0)=0*, which makes the KLD definite in all situations (see Mori et al., 2005 for more properties of the KLD).

The Lilien Indicator – introduced by Lilien (1982), measures the dynamics of sectoral reallocation of activity (e.g. employment). It is expressed as follows:

$$\sigma = [\sum_{i=1}^{n} \frac{y_{ijt}}{\sum_i y_{ijt}} \cdot (\Delta \log y_{ijt} - \Delta \log \sum_i y_{ijt})^2]^{1/2}$$

where y_{ijt} is the activity in industry *i* in region *j* in period *t*, $\sum_i y_{ijt}$ is the total activity in region *j* in period *t* (total of *i*), Δ is the first-difference operator. It compares the changes of the sectoral regional activity to the full regional activity over time overs sectors. High values of this index indicate a relatively strong shift between the industries. Zero implies structural stability over time.

The Herfindahl Index – based on the data of the firm's size, applied simply to single firms can measure the monopolistic position of firms and assess the market organisation. In a regional context, it is applied to the distribution of the industry shares within a region (or between regions) to cover economic diversity. It is expressed as follows:

$$H = \sum_{i=1}^{n} s_{ij}^2$$

where $s_{ij} = \frac{y_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij}}$ is the share of activity ($\sum s_{ij} = 1$).A high value of H results from an uneven distribution and thus indicates at a high degree of concentration (in the case of a firm's monopoly), while a low value of H is for an even distribution and/or for a high level of competition. The H index may be between 0 (when many units are evenly distributed) and 1 (one significant share covering most

of the activity). Decreasing values prove increasing diversification, and conversely increasing values of H stand for monopolisation or extreme concentration.

B) Geographical concentration – indices calculated by regions for given industries

The Krugman Concentration Index – is the version of Krugman's dissimilarity index in the inter-regional (not inter-sectoral) cross-section. It measures industrial structure by regions, thus comparing the shares of activity in a given industry across regions. It is expressed as follows:

$$K_i = \sum_{j=1}^{m} \left| \frac{y_{ij}}{\sum_{j=1}^{m} y_{ij}} - \frac{\sum_{i=1,j}^{n} y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}} \right|$$

which means it is the total by *m* regions, summing up the differences between the share of activity in a given region *j* in a given industry *i* and the share of activity in a given region in all industries (or in the reference industry). The minimum value is 0 (industrial structure fully consistent with the reference one), the maximum is $2 \cdot (m-1)/m$.

The Agglomeration V – proposed by Franseschi, Mussoni and Pelloni (2009), is based on a comparison of the industrial dispersion within and between regions. Its construction is as follows:

$$V_i = \frac{\frac{1}{\overline{y_i}}\sqrt{\frac{\sum_j (y_{ij} - \overline{y_i})^2}{m}}}{\frac{1}{\overline{y_j}}\sqrt{\frac{\sum_j (y_j - \overline{y_j})^2}{m}}}$$

where y is the share of activity (in sector *i* and/or region *j*), m is the number of regions. The coefficient V_i is calculated for each sector. In fact, the counter is the dispersion of the regional sectoral shares compared with the average sectoral share, summed up by regions, and the denominator is the dispersion of a region's share compared with the average region's share, summed up by regions. The values of the agglomeration index falling below 1 ($V_i<1$) appear when the differences in the sector are smaller than the differences in the country, which indicates that the given sector is less geographically concentrated than the overall economy. On the contrary, values of V_i higher than 1 ($V_i>1$) are for bigger regional rather than national differences, which proves there is an increased geographical concentration.

Clustering index – introduced by Bergstrand (1985), relates the shares of activity in the sector and in the region, and weights this with the distance between regions, using the gravity model style. The clustering index is calculated for each sector *m*. It is expressed as follows:

$$C_n = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\frac{y_i^n y_j^n}{d_{ij}}\right)}{\sum_{i=1}^{m}\sum_{i=1}^{m}\left(\frac{y_i y_j}{d_{ij}}\right)}$$

where *m* is the sector, *i* and *j* represent the pairs of regions, y_i^n, y_j^n are the sectoral regional shares of activity measured in the region's total activity, y_i, y_j are the regional shares of activity measured in the national total activity and d_{ij} is the distance between the regions. Values $C_n=1$ are in the case of a similar distribution of activity in the sector and in the whole economy, weighted with the distance. High values of C_n suggest that neighbouring regions have a similar share of a given activity. As dividing by 0 is not possible, one should correct the distances by epsilon, by adding a small value to all pair distances. This impacts strongly on the individual components, but totals stay relatively stable.

Relative H, Theil's H, Shannon's H – entropy measures are calculated in the inter-regional cross-sections to assess the geographical concentration of the industries for each sector separately. One

compares the empirical distribution of activity among regions for a given sector with the benchmark one, which assumes an equal distribution of firms. The formulas are almost the same as for the sectoral concentration case, with the difference that *s* is the share of activity in a given sector in a given region with reference to the full activity in the sector ($s = s_{ij} = \frac{y_{ij}}{\sum_{i=1}^{m} y_{ij}}$) and where the number of events *n* is the number of regions in the sample. This could reveal the spatial pattern of business allocation. Shannon's H would be 0 if all the firms from the sector are in one region, and the maximum value if the firms were allocated equally to the regions. This geographical concentration of business in the sector may indicate mechanisms of over-regional agglomeration.

The Kullback-Leibler Divergence (KLD) – represents the relative entropy calculated by regions for the given industries as proposed by Mori et al. (2005). It is then sought to measure the "*complete spatial dispersion*" or the "*degree of localisation*". The benchmark distrubution is then the uniform distribution of firms among regions, which gives an equal probability of a business location of a given industry in the regions analysed. It is expressed as follows:

$$KLD_i = \sum_{j=1}^{m} s_{ij} ln \frac{s_{ij}}{q_j}$$

where q_j is the expected value of activity in region *j* in the industry analysed, and s_{ij} is the share of activity in a given sector in a given region with reference to the full industrial activity ($s = s_{ij} = \frac{y_{ij}}{\sum_{j=1}^{m} y_{ij}}$). The properties of the index are the same as in the SC approach. The interpretation is conducted as an inter-regional comparison of a given industry. KLD_i close to 0 means a *complete spatial dispersion* of the business. The higher the value of KLD_i the higher the degree of localisation, which can be understood as the degree of regional concentration. It is worth noting, that a KLD with uniform benchmark as above gives the same result as Theil's H.

The Bruelhart-Traeger Index – proposed by Bruelhart and Traeger (2005), is based on the entropy concept, but is decomposable, and supported by the boostrap test. It allows for the between- and within-country comparisons as well quantifying the contribution of the sectors to the overall concentration. Starting from the generalised entropy, they develop the basic entropy indices, GE as the *Theil index* and CV as the *coefficient of variation*, which are as follows:

$$GE(1)_i = \sum_{j=1}^{m} \frac{n_j}{N} \frac{\overline{y_{ij}}}{\overline{y_i}} log \frac{\overline{y_{ij}}}{\overline{y_i}}$$

and

$$CV_i = \frac{1}{\overline{y_i}} [\sum_{j=1}^{m} \frac{n_j}{N} (\overline{y_{ij}} - \overline{y_i})^2]^{1/2}$$

where n_j is the weighting variable (i.e. total regional activity, $n_j = \sum_{i=1}^{n} y_{ij}$), N is total national activity, $\overline{y_{ij}} = y_{ij}/n_r$ is the share of regional (*j*) sectoral (*i*) activity in regional activity, $\overline{y_i} = \widetilde{Y_i}/N$ (with $\widetilde{Y_i} = \sum_j y_{ij}$) is the share of sectoral activity in full national activity. The GE index reflects the sectoral concentration, with regard to the differences in regional activity (employment).

The Gini Index – for geographical concentration (different than for the sectoral concentration), calculated for the *i* sector is as follows:

$$GINI_i^c = \frac{2}{m^2 \bar{C}} \sum_{j=1}^{m} \Lambda_j |C_j - \bar{C}|$$

where m is number of regions, $C_j = \frac{s_{ij}^c}{s_j}$ is ratio of activity, $\bar{C} = \frac{1}{m} \sum_{j=1}^{m} C_j$ is the average of activity ratio (by regions), Λ_j is the rank of the region's position in descending order of C_j, $s_{ij}^c = \frac{y_{ij}}{y_i} = \frac{y_{ij}}{\sum_j y_{ij}}$ is

the ratio of activity in sector *i* in region *j* to total activity (all *j* regions) in sector *i*, $s_j = \frac{y_j}{y} = \frac{\sum_i y_{ij}}{\sum_i \sum_j y_{ij}}$ is the ratio of activity in region *j* in all *i* sectors to total activity (all *i* sectors in all *j* regions). This Gini index is formulated for single industries and all regions, and the reference point are all the industries in one region. Thus, it measures the average under/over representation of one sector in all regions in comparison with the benchmark given by all sectors in all regions (a kind of sectoral average structure). The interpretation of the traditional Gini index is straightforward. It can take values from 0 to 1. Gini=0 means a uniform distribution of activity among sectors/regions, thus the studied and benchmarked distributions are equal. Gini=1 is in the case of a full concentration (whole sectoral activity in one region only / full activity of the region in one sector only). The higher the value of the Gini index, the lower the similarity between industries and regions.

The Locational Gini Index – introduced by Kim et al. (2000) and Guillain & LeGallo (2010), is simpler than the traditional Gini, and its values for concentration in *n* sectors can be compared. It is expressed as follows:

$$Gloc_n = \frac{\Delta}{4\overline{\mu_x}}$$

and

$$\Delta = \frac{1}{m(m-1)} \sum_{j=1}^{m} \sum_{i=1}^{m} |x_i - x_j|$$

which is the total of matrix of the absolute values of the differences in the share's proportion for all pairs of regions, and

$$x_{i(j)} = \frac{region\ i's\ (or\ j's)\ share\ of\ activity\ in\ sector\ m}{region\ i's\ (or\ j's)\ share\ of\ total\ activity}$$

which is the proportion of shares: sector in region and region in country, and

$$\overline{\mu_x} = \frac{1}{m} \sum_{j=1}^{m} x_j$$

which is the average proportion x_j, where *i,j* are the subscripts of regions, *m* is the number of regions, and *n* is the number of sectors (industries). This index takes values from 0 to 0.5, and Gloc=0 is for equal distributions (between regions) of activity in the sector and whole economy, and Gloc=0.5 indicates an extreme concentration of full activity in a single region only.

The Ellison-Glaeser Index – introduced by Ellison & Glaeser (1997), compounds the effect of natural advantages as well as industry spillovers. It is expressed as follows:

$$EG_i = \frac{\sum_{j=1}^{m}(s_i - x_i)^2 - (1 - \sum_{j=1}^{m} x_i^2) \cdot H_i}{(1 - \sum_{j=1}^{m} x_i^2)(1 - H_i)}$$

where s_i is the share of activity in the industry in the region, x_i is the share of activity in the region, H is the industrial Herfindahl index on the plant level (for *X* firms of size *z*) in all regions ($H = \sum_{x=1}^{X} z_x^2$). In fact $\sum_{j=1}^{m}(s_i - x_i)^2$ reflects the similarity of the industrial and regional distributions (as a kind of taxonomy), $(1 - \sum_{j=1}^{m} x_i^2)$ is an inverse regional Herfindahl (for shares of regional activity) and H is the business industrial Herfindahl (for size of companies). The Ellison-Glaeser index might also be expressed as:

$$EG_i = \frac{\frac{G_{EG}^i}{(1 - \sum_j x_i^2)} - H_i}{(1 - H_i)}$$

where $G_{EG}^i = \sum_j (s_i - x_i)^2$ is called the spatial Gini index.

The Ellison-Glaeser index can take both negative and positive values. EG=0 is for a random distribution. Positive values prove that there is an industrial concentration. Most of the literature gives

the critical values for the EG interpretation: EG<0.02 is for the low concentration, EG values between 0.02-0.05 are for the intermediate concentration, EG>0.05 are for the high concentration of a given sector between regions.

The Maurell-Sedillot Index – introduced by Maurel and Sedillot (1999), represents an improvement of the Ellison-Glaeser index. Following Alonso-Villar et al. (2004), the MS index can be written as:

$$MS_i = \frac{\frac{(\sum_{j=1}^{m} s_i^2 - \sum_{j=1}^{m} x_i^2)}{(1 - \sum_{j=1}^{m} x_i^2)} - H}{(1 - H)}$$

and compared with EG:

$$EG_i = \frac{\frac{(\sum_{j=1}^{m} (s_i - x_i)^2)}{(1 - \sum_{j=1}^{m} x_i^2)} - H}{(1 - H)}$$

where s_i is the share of activity in an industry in a region, x_i is the share of activity in a region, H is the industrial Herfindahl index. Both indices are calculated for the industry, by summing up over regions. The interpretation of the MS index is as follows: $\frac{(\sum_{j=1}^{m} s_i^2 - \sum_{j=1}^{m} x_i^2)}{(1 - \sum_{j=1}^{m} x_i^2)}$ is treated as an excess of raw geographic concentration on productive concentration (H), and allows for controlling of the size distribution of plants. It reaches a value of 0 if the industry is located randomly across regions, without considering H. Negative values of the index (MS<0) appear when dispersion is a dominating force, and firms do not cluster. Positive values have the same threshold as the EG index: MS<0.02 is for the low concentration, MS values between 0.02-0.05 are for the intermediate concentration, MS>0.05 is for the high concentration of a given sector between regions.

C) <u>Overall concentration - indices calculated by regions and industries (single number)</u>

Geographic concentration index – used by the OECD (2009), to compare a concentration as a share of any process/activity with a share of territory (area). It is expressed as follows:

$$GC = \frac{1}{2} \sum_{j=1}^{m} |y_j - a_j|$$

where y_j is the region's share in the total activity measured, and a_j is the share of a region's area in the whole territory. Its values are: 0 in case of no concentration (full diversification) and 100% in case of full concentration.

Overall Theil's index – is calculated as the difference of Shannon's max H and Shannon's empirical H

$$T_{total} = \left[-(n \cdot m) \cdot \frac{1}{n \cdot m} \cdot \ln\left(\frac{1}{n \cdot m}\right) \right] - \left[-\sum_{i=1,j=1}^{i=n,j=m} s_{ij} \cdot \ln s_{ij} \right]$$

where s_{ij} is the ratio of activity in sector i in region j to total activity (all i=n sectors in all j=m regions) (empirical share of activity in the given sector in the region to full national activity) ($s_{ij} = \frac{y_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij}}$). Thus, the overall Theil is calculated on m x n data (all single cells of a two-dimensional table) as the difference between the maximum Shannon H and empirical Shannon H.