Kiel Institute for the World Economy
Duesternbrooker Weg 120
24105 Kiel (Germany)

Kiel Working Paper No. 1276

Disproportionality Measures of Concentration, Specialization, and Polarization

by

Frank Bickenbach, Eckhardt Bode

May 2006

Disproportionality Measures of Concentration, Specialization, and Polarization

Frank Bickenbach and Eckhardt Bode*

Abstract:

The paper extends the methodological toolbox of measures of industrial concentration and regional specialization. First, a taxonomy is proposed which gives rise to a modular construction system for disproportionality measures based on three characteristic features: the projection function, the reference distribution, and the weighting scheme. The taxonomy helps reduce the mismatch between research purpose, data and statistical measure which has been one of the major obstacles to reliable inferences in the literature. Second, the taxonomy is extended (i) to measures of polarization which evaluate specialization and concentration simultaneously and allow for a nested analysis at different spatial and industrial scales, and (ii) to spatial concentration measures which deal with the checkerboard problem and MAUP by taking into account information from neighboring regions.

Keywords: statistical measures, regional specialization, industrial concentration, economic polarization, geographical weighting

JEL classification: C43, F15, R12

Frank Bickenbach
Kiel Institute for the World Economy
24100 Kiel, Germany
Telephone:+49 431 8814 278
E-mail: frank.bickenbach@ifw-kiel.de

Eckhardt Bode
Kiel Institute for the World Economy
24100 Kiel, Germany
Telephone:+49 431 8814 462
E-mail: eckhardt.bode@ifw-kiel.de

* The authors would like to thank Christiane Krieger-Boden, Serge Rey and participants of the WRSA annual meeting 2006, Santa Fe, for helpful comments and suggestions. Financial support by the EU Commission, project EURECO, is gratefully acknowledged.

1. Introduction

Inspired by the new economic geography, concerns have been raised that economic integration at the regional or international level may further the polarization of economic activities, i.e., the spatial concentration of industries (henceforth [industrial] concentration for short) and/or the industrial specialization of regions ([regional] specialization). Innovative, dynamic industries may concentrate in the core regions, leaving the periphery with mature industries which face fierce competition from the world markets. As a consequence, peripheral regions may grow slower in terms of income and employment, and may be more vulnerable to adverse macroeconomic shocks.

In search of stylized facts on economic polarization in Europe various studies have explored the evolution of regional specialization or industrial concentration using measures borrowed directly or indirectly from the income inequality literature.[1] The results are remarkably inconclusive for several reasons (Combes and Overman 2004): First, many of the studies lack an unambiguous research focus and a clear test hypothesis. Second, the results are sensitive to the sectoral and spatial scales of the available data. With aggregate data available from public sources, the modifiable areal unit problem (MAUP)[2] cannot be avoided or effectively solved. In addition, the sectoral and/or spatial aggregates are treated as "anonymous" units which gives rise to the checkerboard problem.[3] And third, the choice of the measure has been largely ad hoc in most studies. The interdependencies between the research purpose (test hypothesis), the available data, and the statistical measures have largely been neglected, not least so because a taxonomy of measures which allows for systematically assessing the properties of the measures on the background of the available data and the research purpose has not been available. The sensitivity of results on the evolution of regional specialization to the choice of the measure is illustrated in Table 1: The specialization in terms of employment across 88 manufacturing industries is found to have increased in almost all of the 18 Spanish NUTS2 regions, or in hardly any region, depending on the measure employed.

---

1 Examples are the Theil index, the Gini index, the coefficient of variation, or the so-called Krugman index (relative mean deviation). See Bode et al. (2003), Combes and Overman (2004), and Nijkamp et al. (2003) for recent reviews.

2 The modifiable areal unit problem (MAUP; see, Openshaw and Taylor 1979; Arbia 1989) arises from discretizing heterogeneous continuous variables. It comes under two guises: Discretizing space averages away heterogeneity, such that results are sensitive to the scale of aggregation (*scale* problem), and the boundaries between the discrete spatial units may be misplaced (*arbitrary boundary* problem). MAUP is actually a manifestation of the general aggregation problem inherent to any micro-founded macroeconomic analysis. It is, however, particularly relevant for aggregate studies focusing directly on the heterogeneity at a more disaggregate level.

3 The checkerboard problem arises from neglecting relevant information on the locations of or distances between regions or industries (Arbia 2001).

Table 1 — Sensitivity of results to the choice of a measure: Counts of Spanish NUTS2 regions for which measures of regional specialization indicate an increasing specialization during the period 1978–1992[a]

	"absolute"		"relative"	
measure	no of regions	%	no of regions	%
Theil index	9	50.0	2	11.1
Coefficient of variation	7	38.9	16	88.9
Relative mean deviation ("Krugman index")	10	55.6	2	11.1
Gini index	9	50.0	1	5.6

[a] Dataset: Employment by 18 Spanish NUTS2 regions and 88 manufacturing industries. "Absolute" measures evaluate the deviation of the distribution of employment across industries from the uniform distribution; "relative" measures the deviation of the distribution of employment across industries from the corresponding distribution at the national level.

To solve the problem of ambiguous inferences, Combes and Overman (2004) set up a catalog of "baseline criteria" for a "perfect" measure. In a nutshell, the criteria request the measure to

(i) be comparable across industrial and spatial units and scales, and unbiased by MAUP and the checkerboard problem in both the regional and the sectoral dimension;

(ii) allow for specifying an unambiguous and meaningful null hypothesis of no concentration or specialization which may capture both systematic variation, suggested by economic theory, and random variation; and

(iii) be suitable for statistical testing.

Rather than trying to develop a perfect measure, the purpose of the present paper is to improve upon existing "inequality" measures (and their selection) by extending them to "disproportionality" measures that better fit the baseline criteria. While "inequality" measures use as the reference the mean of the variable of main interest, "disproportionality" measures allow to choose the reference from a wide variety of possible distributions. This will allow a more rigorous exploratory analysis of industrial concentration and regional specialization processes. The paper, first, proposes a taxonomy for disproportionality measures which gives rise to a modular system of three characteristic features of any measure: the projection function, the reference distribution, and the weighting scheme. For a given, well-specified research purpose, and for known characteristics of the available data, the taxonomy can be used for defining the appropriate disproportionality measure as a combination of the most appropriate realization of each of the three characteristic features.[4] For an vaguely specified research purpose, and unknown characteristics of the available data, the taxonomy helps sharpen the focus and explore the data. In doing so, it helps specify the null hypothesis properly (baseline criterion ii), identify the data requirements in terms of the preferred level of sectoral and/or spatial disaggregation and take into account the limitations of the available data as far as possible (baseline criterion i), and perform sensitivity tests (baseline criterion iii).[5]

Second, the paper introduces measures of polarization of an economy ([economic] polarization) which assess industrial concentration and regional specialization simultaneously. These polarization measures are straightforward generalizations of the disproportionality measures discussed in the first part; they can be defined from a generalized modular system of characteristic features, with a joint region/industry reference distribution and two sets of weighting schemes. These measures can be used for a nested comparison of industrial concentration and regional specialization patterns (baseline criterion i).

And third, the paper introduces spatial concentration measures which take into account the spatial ordering of the basic units by allowing to consider the neighborhood structures of the regions as well as additional empirical or theoretical information about the unobserved intra-regional distributions of the variable of interest. As generalizations of the concentration measures discussed in the first part, the spatial measures can be defined from a generalized modular system of characteristic features, with a spatial weights matrix as an additional feature. In dealing with the checkerboard problem and the MAUP (baseline criterion i), these measures are alternatives to the distance-based concentration statistics based on Ripley's K functions proposed recently by Duranton and Overman (2005) and Marcon and Puech (2003; 2005). The two generalizations benefit from the methodological discussion of segregation measures in the sociological literature.

For expositional convenience, the discussion of the taxonomy is limited to a few selected measures: the Gini coefficient, and the relative mean deviation as examples of intuitive ad-hoc measures, and the Generalized Entropy (GE) class of measures and its two most prominent members, the Theil index and the GE(2), as examples of axiom-based measures (Cowell

---

4 The bulk of the literature so far has taken measures as a fixed combination of two of these features, the projection function and the reference distribution. An important step towards a more flexible combination of features is Brülhart and Träger (2005) that introduces different references for the same projection function. Varying the weighting scheme independently of the reference distribution has not been considered an option.

5 A detailed discussion of the conceptual and technical issues of testing for the statistical significance of changes of a measure over time, or of the differences between regions or industries, is beyond the scope of the present paper. Among the few studies performing rigorous statistical tests are Brülhart and Träger (2005) and Mori et al. (2005). While the tests proposed by Mori et al. (2005) are specific to the Kullback-Leibler D statistic which is similar to a Theil index, the bootstrap tests proposed by Brülhart and Träger (2005) for the Theil index and the CV can, in principle, be applied to other disproportionality measures discussed in this paper.

1995). The taxonomy may be extended easily to other inequality measures. The taxonomy does not, however, cover measures which differ conceptually from these measures, such as the so-called “dartboard” measures (Ellison and Glaeser 1997; Maurel and Sédillot 1999) which are actually estimates of correlation coefficients, or the distance-based statistics based on Ripley’s K function.

The plan of the paper is as follows: Section 2 introduces the taxonomy and illustrates how disproportionality measures can be defined using the modular system of characteristic features. For expositional convenience, the chapter focuses on measures of industrial concentration as an example. The corresponding measures of regional specialization can easily be obtained by just switching indices. Chapter 3 extends the taxonomy to measures of polarization. Chapter 4 extends the taxonomy to spatial concentration measures. Chapter 5 concludes, and discusses issues of further research. An empirical illustration for selected measures is given in Bickenbach et al. (2006). A more detailed tabulation of the various measures discussed in the paper is available at http://www.uni-kiel.de/ifw/staff/bode/measures.htm.

2. The taxonomy

This chapter proposes a taxonomy of measures of concentration and specialization. Since the available data is discrete in most applications, the discussion is limited to discrete versions of the measures. All measures covered can be characterized as measures of “disproportionality” of the distribution of a population across a finite set of mutually exclusive characteristics and a pre-determined reference distribution. For expositional convenience, the following discussion will exemplify measures of industrial concentration. Thus, population is *workers within an industry*, their characteristics are the *regions* of their workplaces.[6]

Disproportionality measures of industrial concentration determine, for each region, a value for the “region-specific deviation” that is the ratio between the number of workers in the industry and a pre-determined reference. By applying an inequality measure to these region-specific deviations, they are aggregated over all regions to a scalar, the concentration measure. One important difference between the traditional inequality measures and the disproportionality measures is that the former implicitly use as the reference the mean of the variable of main interest[7] while the latter allow to choose the reference from a wide variety of possible

---

[6] For measures of regional specialization, the population is *workers within a region* and their characteristics are the *industries*. Specialization measures can thus be obtained from concentration measures by just switching the indices for regions and industries. For measures of polarization, the population is *workers within an economy*, made up by all industries and regions under investigation, and their characteristics are their region-industry affiliations. See Chapter 3 for the formal definitions of polarization measures.

[7] Comparing the values of the variable of main interest for the different regions to each other is equivalent to comparing each value to the population mean.

distributions. The mean of the variable of main interest is just a scaling factor. The possibility to choose a reference other than the population mean is important for the measurement of concentration because frequently the population mean is not the reference that best matches the specific research purpose at hand.[8]

Formally, the disproportionality measures covered by the taxonomy can be characterized as a function:

$$M_i^{W\Pi} = f_M\left(\mathbf{W}_{(r)}, \frac{\mathbf{L}_{i(r)}}{\mathbf{\Pi}_{(r)}}\right); \tag{1}$$

where

$M_i^{W\Pi}$: concentration measure (a scalar) for industry *i*: *M* reflects the projection function, the superscripts the choice of the weights and the reference;
f_M: projection function;
$\mathbf{L}_{i(r)}$: vector of the numbers of workers of industry *i* observed for region *r*, *r* =1, …, *R*;
$\mathbf{\Pi}_{(r)}$: vector of regional references;
$\mathbf{W}_{(r)}$: vector of region-specific weights.[9]

The projection function is such that the region-specific deviations, $X_{ir} = L_{ir}/\Pi_r$, (r =1, …, R), are always scaled by their weighted average across all regions *r*, i.e., by $\overline{X}_i = \sum_{r=1}^{R} w_r X_{ir} = \sum_{r=1}^{R} \frac{W_r}{\sum_r W_r} \frac{L_{ir}}{\Pi_r}$. Technically, M_i is a function of w_r and $X_{ir}/\overline{X}_i$ only, similar to inequality measures. A disproportionality measure describes the *inequality* across regions of the *proportions* of the variable of main interest and its reference. The measure assumes its minimum value, zero, if the proportions are the same in all regions, i.e., $L_{ir}/\Pi_r = L_{is}/\Pi_s$ $\forall\ r, s = 1, \ldots, R$.

The taxonomy builds on the three characteristic features of the measures in (1): (i) the region-specific weights, $\mathbf{W}_{(r)}$, (ii) the reference distribution, $\mathbf{\Pi}_{(r)}$, and (iii) the projection function, f_M.[10] Together with the variable of main interest, $\mathbf{L}_{i(r)}$, the three features unambiguously define a measure. In any meaningful empirical investigation, the specification of each characteristic feature should follow directly from the research purpose or the test hypothesis at hand.

---

8 The definition and scale of the reference may differ from that of the variable of main interest, as long as both refer to the same region.

9 Uppercase letters denote absolute numbers, lowercase letters shares. Bold characters denote vectors; their subscripts in parentheses indicate their elements and dimensions. The Appendix summarizes the notational details.

10 For measures of regional specialization, the weights are industry-specific; for measures of polarization, the weights are industry- and region-specific.

(i) The *region-specific weights*, $\mathbf{W}_{(r)}$, reflect the researcher's choice of the basic units of the analysis (Brülhart and Träger 2005): For measures of concentration, the basic units are spatial units, such that the variable of main interest is defined as, say, the number of industry i workers *per basic spatial unit*.[11] Disproportionality measures allow for specifying a variety of different geographical basic units, provided the variable of main interest as well as the reference variable can be measured consistently in terms of the basic units. Only three types of basic units have, nonetheless, been used in the literature so far:

– Choosing the regions themselves as basic units implies assigning all regions the same weight, $\mathbf{W}_{(r)} = \mathbf{1}_{(r)} = (1, \ldots, 1)$,[12] independent of their actual sizes or any other characteristics.

– Choosing square kilometers (km²) as basic units implies weighting each region by its geographical size (A_r), $\mathbf{W}_{(r)} = \mathbf{A}_{(r)} = (A_1, \ldots, A_R)$.[13]

– Choosing the average size of the area attributed to a worker of any industry in the region as basic units implies weighting each region by its total employment, $\mathbf{W}_{(r)} = \mathbf{L}_{(r)} = (L_{\bullet 1}, \ldots, L_{\bullet R})$. $L_{\bullet r}$ [$= \Sigma_i L_{ir}$] denotes the sum of workers over all industries in region r. Each worker in region r is taken to represent a share of $1/L_{\bullet r}$ of the region's area.

Measures using regions as basic units will be labeled "*unweighted measures*"; those using non-uniform region-specific weights "*weighted measures*". Weighted measures are invariant to dividing a region into sub-regions, provided the weights represent the sizes of the regions, and the sub-regions exhibit, or are assumed to exhibit, identical concentration patterns.[14]

(ii) The *reference distribution*, $\mathbf{\Pi}_{(r)}$, reflects the researcher's choice of the benchmark, resp. the null hypothesis of "no" or "no unusual concentration". As Combes and Overman (2004) emphasize, economically meaningful inferences require any deviation of the observed from the reference distribution to be attributable to what the researcher is actually willing to label "concentration". Similarly, anything the researcher wants to label "concentration" must show up as a deviation. The reference distribution should pick up any systematic components in the

---

[11] For measures of specialization, the basic units are units of (sectoral) activities, such that the variable of main interest is defined as, say, the number of region r-workers per sectoral unit.

[12] After standardizing the sum of weights to one, each region is assigned the relative weight $w_r = W_r/\Sigma_r W_r = 1/R$.

[13] As the spatial distribution of workers within the regions cannot be observed in most cases, all worker are assumed to occupy areas of the same size, and be distributed uniformly across space within the region. Each areas may be shared by workers from all industries. Although this simplifying assumption does not avoid the scale problem resulting from the unobserved intra-regional heterogeneity, preferring the smallest observable entities as basic units reduces the scale bias as far as possible (Brülhart and Träger 2005).

[14] Haaland et al. (1998) attribute this feature to relative measures. The taxonomy here makes clear that it is solely due to the choice of the weights.

observed sectoral employment patterns which the researcher is *not* willing to label "concentration" for the research purpose at hand. Disproportionality measures allow for specifying a great variety of reference distributions, provided they are defined over the basic units. Only three types of references have, nonetheless, been used in the literature so far:

– The uniform distribution is represented in (1) by $\mathbf{\Pi}_{(r)} = \mathbf{1}_{(r)}$: All regions are assumed to be of the same size under the H_0.[15]

– The topographical distribution is represented in (1) by $\mathbf{\Pi}_{(r)} = \mathbf{A}_{(r)}$. The employment in the industry under investigation is assumed to be distributed evenly across space under the H_0.

– The distribution of employment observed at a higher-level sectoral aggregate such as total regional employment is represented in (1) by $\mathbf{\Pi}_{(r)} = \mathbf{L}_{(r)} = (L_{\bullet 1}, \ldots, L_{\bullet R})$. The spatial distribution of the industry under investigation is assumed to equal that of total employment across all industries under the H_0.[16]

Measures based on the uniform reference will henceforth be labeled "*absolute measures*", those based on a non-uniform reference "*relative measures*".[17] The measures are actually rather sensitive to the choice of the reference. Changing the reference may easily reverse the inferences on the spatial concentration of an industry (see, e.g., Brülhart and Träger 2005, or several of the country studies in Traistaru et al. 2003). Simulations not reported in detail here indicate that the differences between the "absolute" and "relative" measures in Table 1 do in fact result mainly from the difference in the reference rather than the difference in the region-specific weights.

(iii) The *projection function*, f_M, reflects the researcher's relative emphasis on region-specific deviations of different magnitude. Some measures, such as the Theil index, emphasize variations in the range of lower values of the region-specific deviations (e.g., industry is strongly underrepresented in a region), others, such as the coefficient of variation, emphasize variations in the range of higher values of the region-specific deviations (e.g., industry is strongly overrepresented). Again others, such as the relative mean deviation, emphasize changes in the

---

15 The choice of the uniform reference may reflect the researcher's emphasis on the qualitative characteristics of regions, or on administrative or political issues. Any systematic quantitative differences, including those resulting from the available regional classification schemes, are interpreted purposefully as concentration.

16 Higher-level aggregates as a reference allow to control for systematic differences between the regions in their size or some external determinants such as their attractiveness for firms and workers, their regulatory frameworks and other institutional or political factors. They may, however, be endogenous: Bigger industries, or major industry-specific shocks may affect them significantly. They may, in addition, filter out aggregate determinants of industrial concentration the researcher actually would like to attribute to concentration.

17 Brülhart and Träger (2005) introduce the term "*topographic measures*" for measures using the area as a reference. It should be noted that this reference is just one out of many possible non-uniform references.

balance between regions with over- and underrepresented industries and incidences of regions jumping across the reference.[18] As a consequence, the results may easily differ even qualitatively if different projection functions are used to evaluate differences in industry concentration between regions or points in time, depending on the regions, e.g. over- or underrepresented regions, where the differences are observed (see Table 1 for an illustration).

In the literature, concentration and specialization measures have so far been classified by their projection function and their reference distributions (Haaland et al. 1998). The reference and the weights have always been assumed to be the same: “Absolute” measures have used the uniform distribution both as weights and reference; “relative” measures have used the distribution of total regional employment, “topographic” measures the distribution of area. Varying the reference independent of the region-specific weights has not been considered an option.[19] The present paper argues, and illustrates below, that this is unnecessarily restrictive. By distinguishing carefully between the reference and the weights, the taxonomy adds one additional degree of freedom to the opportunities for choosing the measure which best fits the specific research purpose.

Disentangling reference and weights is useful for two reasons: First, the research purpose or test hypothesis may request using a weighting scheme that differs from the reference: One example is studies of political decisions or other public administrative issues which requires choosing the spheres of influence of local governments or administrations like counties or states as the basic units while the aggregate regional employment or the employment in another industry may be the proper benchmark (reference). The proper region-specific weights will be $\mathbf{W}_{(r)} = \mathbf{1}_{(r)}$, the proper reference $\mathbf{\Pi}_{(r)} = \mathbf{L}_{(r)}$. Or the research purpose may request to compare the spatial distribution of an industry to that of total employment while controlling for differences between regions in their geographical sizes. The proper basic units will be km² which gives rise to region-specific weights $\mathbf{W}_{(r)} = \mathbf{A}_{(r)}$, the proper reference will be $\mathbf{\Pi}_{(r)} = \mathbf{L}_{(r)}$ in this case.[20] None of these applications have been considered an option so far. They are, however, possible and perfectly consistent with the taxonomy discussed here. The corresponding Theil measures will be given below.

---

18 In addition to theoretical aspects, the choice of the projection function may be subjected to practical considerations. A projection function which does not put too much emphasis on extreme (positive and/or negative) region-specific deviations may be preferred to reduce the effects of indivisibilities in firm sizes or “outliers” on the measure. Alternatively, or in addition, the sensitivity of the results can be assessed by comparing the results for different projection functions.

19 This is true not only for the literature on regional specialization and industrial concentration but also for those on income inequality and segregation.

20 As will be discussed in more detail below, the resulting measure requires the additional assumption that the fraction of industry-*i* workers in the total workforce does not vary across space within a region.

The second reason is that, by clarifying the distinct functionality of the weights and the reference, the taxonomy facilitates sensitivity tests. Changing selectively the region-specific weights or the reference will help assess the sensitivity of the preferred measure to a variation of the basic units or the null hypothesis.

Using the taxonomy, four different measures can be defined for each projection function: An *unweighted absolute*, an *unweighted relative*, a *weighted absolute*, and a *weighted relative* measure. If defined along the lines discussed above, these measures share virtually all properties of their conventional counterparts. Table 2, which will be discussed below in more detail, gives an overview of the general principle of defining concentration measures for selected projection functions: the generalized entropy (*GE*) class of measures, the Theil index (*T*), the coefficient of variation (*CV*),[21] the relative mean deviation (*RMD*), and the Gini coefficient (*G*).[22] The first column of Table 2 gives, for each projection function, a general form that can be used to derive all related measures. Given region-industry employment, $\mathbf{L}_{i(r)}$, a measure may be unambiguously defined by choosing a reference distribution, region-specific weights and a projection function. The remaining three columns of Table 2 give three examples of measures obtained for different combinations of weights and references: the *unweighted absolute*, an *unweighted relative* and a *weighted relative* measure.[23] In order to compare the values of different measures directly, it may be useful to normalize the measures to the (0, 1) interval by dividing them by their upper bounds.[24]

All three variants of measures exemplified in Table 2 have actually been employed in studies of industrial concentration or regional specialization, though not for all the projection functions: Among the *weighted relative* measures are (i) the so-called “Krugman index” (weighted relative *RMD*) used, e.g., by Krugman (1991), Hallet (2002), Dohse et al. (2002), or Traistaru et al. (2003); (ii) the so-called “relative” Theil index (Brülhart and Träger 2005, Bode and Krieger-Boden 2005), (iii) the “relative” CV (Brülhart and Träger 2005), and (iv) the “locational” Gini coefficient, as used by Krugman (1991), Amiti (1998) or Brülhart (2001). An *unweighted relative* measure is the “locational” Gini coefficient, as used by Südekum (2006). And among the *unweighted absolute* measures are (i) the traditional Gini coefficient (e.g., Aiginger and Leitner 2002, Midelfart-Knarvik et al. 2002), as well as the (ii) Theil index as

---

21 The Theil index and (a transformed version of) the CV are actually members of the GE class of measures. Owing to their popularity in the literature, they are nonetheless depicted separately in Table 2.

22 Table A1 gives the corresponding measures of regional specialization. Tables 2 and A1 can easily be extended to projection functions based on other inequality measures discussed in the literature (see, e.g., Cowell 1995; Silber 1999).

23 To save space, weighted absolute measures are omitted, and the relative and the weighted measures are exemplified only for total regional employment as a reference or as weights.

24 For details on the calculation of upper bounds see http://www.uni-kiel.de/ifw/staff/bode/measures.htm.

Table 2 — Concentration measures for selected projection functions, reference distributions and basic units[a]

projection function	general form (reference: / weights:)	unweighted absolute $\Pi_r := 1$, $W_r := 1 \Rightarrow w_r = 1/R$	unweighted relative $\Pi_r := L_{\bullet r}$, $W_r := 1 \Rightarrow w_r = 1/R$	weighted relative $\Pi_r := L_{\bullet r}$, $W_r := L_{\bullet r} \Rightarrow w_r = \lambda_r$
GE(α)	$GE_i(\alpha) = \omega \sum_{r=1}^{R} w_r \left[\left(\frac{L_{ir}/\Pi_r}{\sum_r w_r L_{ir}/\Pi_r} \right)^{\alpha} - 1 \right]$	$GE_i^{UA}(\alpha) = \omega \left[\frac{1}{R} \sum_{r=1}^{R} (R\lambda_{ir})^{\alpha} - 1 \right]$	$GE_i^{UR}(\alpha) = \omega \left[\frac{1}{R} \sum_{r=1}^{R} \left(R \frac{l_{ir}}{\sum_r l_{ir}} \right)^{\alpha} - 1 \right]$	$GE_i^{WR}(\alpha) = \omega \left[\sum_{r=1}^{R} \lambda_r (LC_{ir})^{\alpha} - 1 \right]$
Theil = GE(1)	$T_i = \sum_{r=1}^{R} w_r \frac{L_{ir}/\Pi_r}{\sum_r w_r L_{ir}/\Pi_r} \ln \left(\frac{L_{ir}/\Pi_r}{\sum_r w_r L_{ir}/\Pi_r} \right)$	$T_i^{UA} = \sum_{r=1}^{R} \lambda_{ir} \ln(R \cdot \lambda_{ir})$	$T_i^{UR} = \sum_{r=1}^{R} \frac{l_{ir}}{\sum_r l_{ir}} \ln \left(R \frac{l_{ir}}{\sum_r l_{ir}} \right)$	$T_i^{WR} = \sum_{r=1}^{R} \lambda_{ir} \ln(LC_{ir})$
CV = 2GE(2)$^{0.5}$	$CV_i = \left[\sum_{r=1}^{R} w_r \left[\left(\frac{L_{ir}/\Pi_r}{\sum_r w_r L_{ir}/\Pi_r} \right)^2 - 1 \right] \right]^{0.5}$	$CV_i^{UA} = \left(R \sum_{r=1}^{R} \lambda_{ir}^2 - 1 \right)^{0.5}$	$CV_i^{UR} = \left[R \sum_{r=1}^{R} \left(\frac{l_{ir}}{\sum_r l_{ir}} \right)^2 - 1 \right]^{0.5}$	$CV_i^{WR} = \left(\sum_{r=1}^{R} \lambda_r LC_{ir}^2 - 1 \right)^{0.5}$
RMD / Krugman	$RMD_i = \frac{\sum_{r=1}^{R} w_r \left\| \frac{L_{ir}}{\Pi_r} - \sum_r w_r \frac{L_{ir}}{\Pi_r} \right\|}{\sum_r w_r L_{ir}/\Pi_r}$	$RMD_i^{UA} = \sum_{r=1}^{R} \left\| \lambda_{ir} - \frac{1}{R} \right\|$	$RMD_i^{UR} = \sum_{r=1}^{R} \left\| \frac{l_{ir}}{\sum_r l_{ir}} - \frac{1}{R} \right\|$	$RMD_i^{WR} = \sum_{r=1}^{R} \left\| \lambda_{ir} - \lambda_r \right\|$
Gini	$G_i = \frac{\sum_{r=1}^{R} \sum_{s=1}^{R} w_r w_s \left\| \frac{L_{ir}}{\Pi_r} - \frac{L_{is}}{\Pi_s} \right\|}{2 \sum_r w_r L_{ir}/\Pi_r}$	$G_i^{UA} = \frac{1}{2R} \sum_{r=1}^{R} \sum_{s=1}^{R} \left\| \lambda_{ir} - \lambda_{is} \right\|$	$G_i^{UR} = \frac{1}{2R \sum_r l_{ir}} \sum_{r=1}^{R} \sum_{s=1}^{R} \left\| l_{ir} - l_{is} \right\|$	$G_i^{WR} = \frac{1}{2} \sum_{r=1}^{R} \sum_{s=1}^{R} \lambda_r \lambda_s \left\| LC_{ir} - LC_{is} \right\|$

[a] $\omega := (\alpha^2 - \alpha)^{-1}$. See the Appendix for notational details. The lower bound of all measures is 0. For the upper bounds see http://www.uni-kiel.de/ifw/staff/bode/measures.htm.

used by Aiginger and Davies (2004), and (iii) the CV in the version of Aiginger and Leitner (2002).[25]

To illustrate the taxonomy, consider, first, the so-called "Krugman" index which, for the concentration of industry i, is defined as

$$K_i = \sum_{r=1}^{R} \left| \lambda_{ir} - \lambda_r \right| := \sum_{r=1}^{R} \left| \frac{L_{ir}}{L_{i\bullet}} - \frac{L_{\bullet r}}{L_{\bullet\bullet}} \right|. \tag{2}$$

Being calculated as the unweighted sum of the absolute region-specific differences in employment shares for industry i, λ_{ir}, and the reference, λ_r, its interpretation is easy and intuitive: A value of, say, $K_i = 0.5$ indicates that a share of at least one fourth ($^1/_2 K_i$) of the industry's total workforce has to move to another region in order to get an employment distribution that exactly corresponds to the reference distribution. The "Krugman" index has traditionally been classified as a "relative" measure. The taxonomy proposed in the present paper suggests looking at K_i in a slightly different way: By rearranging (2), the Krugman index can be shown to be a *weighted relative* relative mean deviation (RMD_i^{WR}):

$$\begin{aligned} RMD_i^{WR} &= f_M\left(\mathbf{L}_{i(r)}, \mathbf{L}_{(r)}, \mathbf{L}_{(r)}\right) = \sum_{r=1}^{R} \frac{L_{\bullet r}}{L_{\bullet\bullet}} \left| \frac{\dfrac{L_{ir}}{L_{\bullet r}}}{\sum_r \dfrac{L_{\bullet r}}{L_{\bullet\bullet}} \dfrac{L_{ir}}{L_{\bullet r}}} - 1 \right| \\ &= \sum_{r=1}^{R} \lambda_r \left| LC_{ir} - 1 \right|. \end{aligned} \tag{3}$$

$LC_{ir} = \lambda_{ir}/\lambda_r$ is the location coefficient for region-industry (ir). By setting $\mathbf{\Pi}_{(r)} = \mathbf{W}_{(r)} = \mathbf{L}_{(r)}$, (3) can alternatively be derived directly from the general definition of the RMD given in the first column of Table 2.

The first line of (3) clarifies the constructive principle of all measures discussed in the present paper: The measure first evaluates, for each region, the region-specific deviation by comparing the value for the region-industry, L_{ir}, to the corresponding reference value, $L_{\bullet r}$. Second, the region-specific deviations are converted into the metric of the measure by the projection function. The projection function of the *RMD* stipulates to (i) scale this ratio by the weighted mean across all region-specific ratios, $\Sigma_r \lambda_r L_{ir}/L_{\bullet r} = L_{i\bullet}/L_{\bullet\bullet}$; (ii) subtract 1 from the region-specific deviation; (iii) take the absolute value of the resulting difference; and (iv) take the weighted average over all regions.

25 Another intuitive measure, the Herfindahl index can be shown to be closely related to the unweighted absolute *CV* and *GE*(2) measures. The Herfindahl index, employed, among many others, in several of the contributions to Traistaru et al. (2003), is defined as the squared sum of region-industry employment shares.

Following the same general procedure, any of the three characteristic features of the specialization measure may be varied separately. Setting $\mathbf{W}_{(r)} = \mathbf{1}_{(r)}$. and $\mathbf{\Pi}_{(r)} = \mathbf{L}_{(r)}$, an *unweighted relative RMD* is obtained from (4) as

$$RMD_i^{UR} = f_M\left(\mathbf{L}_{i(r)}, \mathbf{1}_{(r)}, \mathbf{L}_{(r)}\right) = \sum_{r=1}^{R} \left| \frac{l_{ir}}{\sum_r l_{ir}} - \frac{1}{R} \right|,$$

where $l_{ir} = L_{ir}/L_{\bullet r}$. Setting $\mathbf{\Pi}_{(r)} = \mathbf{W}_{(r)} = \mathbf{1}_{(r)}$ gives the *unweighted absolute RMD*,

$$RMD_i^{UA} = f_M\left(\mathbf{L}_{i(r)}, \mathbf{1}_{(r)}, \mathbf{1}_{(r)}\right) = \sum_{r=1}^{R} \left| \lambda_{ir} - \frac{1}{R} \right|. \tag{4}$$

Comparing the Krugman index, or weighted relative RMD in (2) to the unweighted absolute RMD in (4) clarifies the usefulness of the taxonomy vis-à-vis the traditional distinction of absolute and relative measures: According to the traditional distinction the two measures differ in just one characteristic, namely the reference. The taxonomy makes clear that the two measures actually differ in two characteristics, the reference *and* the region-specific weights.[26]

As for the *RMD* which is frequently characterized as an ad-hoc measure, the taxonomy does also work for measures such as the General Entropy (GE) class of measures which have several useful properties defined by a set of axioms (see, e.g., Cowell 1995; Litchfield 1999). One useful property is decomposability, which means that for any set of subgroups of a population total inequality within the population can be decomposed into the inequality within the subgroups and that between the subgroups. Following Brülhart and Träger (2005), the decomposition is used to derive unweighted and weighted relative GE measures.

The GE class of measures is generally defined as

$$GE(\alpha) = f_{GE}^{\alpha}\left(\mathbf{Y}_{(n)}\right) = \frac{1}{\alpha(\alpha-1)} \frac{1}{N} \sum_{n=1}^{N} \left[\left(\frac{Y_n}{\overline{Y}} \right)^{\alpha} - 1 \right] \qquad (-\infty \le \alpha \le \infty) \tag{5}$$

for the vector of some characteristics $\mathbf{Y}_{(n)} = (Y_1, \ldots, Y_N)$ of a population. The members of the population are the "basic units", and $\overline{Y} = \frac{1}{N}\sum_n Y_n$ is the mean across the basic units. The parameter α governs the sensitivity of the projection function to changes in the ranges of high and low values of the $Y_n / \overline{Y}$ ratios. With $\alpha < 2$ the measure is more sensitive to (mean-

---

[26] For testing the sensitivity of, e.g., the weighted relative RMD in (3) to changes in the weighting scheme, it may in addition be informative to selectively change the weights of the region-specific deviations while keeping the scaling factor of the region-specific deviations unchanged, i.e., compare (3) to $\Sigma_r|LC_{ir}-1|$. Although the latter expression is not an RMD in terms of the present taxonomy the comparison may still yield valuable information on the sensitivity of the preferred RMD to the weighting scheme.

preserving) changes in observations with low values of $Y_n / \overline{Y}$; with $\alpha > 2$ it is more sensitive to (mean-preserving) changes in observations with high values of $Y_n / \overline{Y}$ (e.g., Cowell 2000). The most prominent GE measures are those given by $\alpha = 2$ which is a simple monotonic transformation of the coefficient of variation, $GE(2) = ½CV^2$; and by $\alpha \rightarrow 1$ which is the Theil index, i.e., $GE(1) = T$.

$GE(\alpha)$ in (5) can be decomposed into a within-groups and a between-groups component such that $GE(\alpha) = GE_w(\alpha) + GE_b(\alpha)$. For H subgroups with N_h basic units in subgroup h ($h = 1, \ldots, H$), the between group component is given by

$$GE_b(\alpha) = f_{GE}^{\alpha}\left(\overline{Y}_{(h)}, \mathbf{N}_{(h)}\right) = \frac{1}{\alpha(\alpha-1)} \sum_{h=1}^{H} \frac{N_h}{N}\left[\left(\frac{\overline{Y}_h}{\overline{Y}}\right)^{\alpha} - 1\right], \tag{6}$$

where $\overline{Y}_h = \frac{1}{N_h}\sum_{n=1}^{N_h} Y_{nh}$ is the unweighted mean of subgroup h, and $\overline{Y} = \sum_h \frac{N_h}{N}\overline{Y}_h \left[= \frac{1}{N}\sum_n Y_n\right]$ the weighted average of all subgroup means. X_{nh} denotes the characteristic of the n^{th} member of the h^{th} subgroup.

Traditionally, the [unweighted] absolute $GE(\alpha)$ measures of industrial concentration have been derived from (5), the [weighted] relative measures from the between-group component (6) of (5), assuming the unobservable within-group component to be zero (Brülhart and Träger 2005).[27] The taxonomy suggests to rather use a general form of the between-group component (6) as the unique basis for all GE measures of regional concentration (see Table 2, first row for examples):

$$GE_i(\alpha) = f_{GE}^{\alpha}\left(\mathbf{L}_{i(r)}, \mathbf{W}_{(r)}, \mathbf{\Pi}_{(r)}\right) = \frac{1}{\alpha(\alpha-1)} \sum_{r=1}^{R} w_r \left[\left(\frac{\frac{L_{ir}}{\Pi_r}}{\sum_r w_r \frac{L_{ir}}{\Pi_r}}\right)^{\alpha} - 1\right]. \tag{6a}$$

All these measures will have the usual properties of GE measures, provided the variables of main interest, weights and references are related to the basic units in a consistent way.

For illustration, consider the two examples given above, and assume the Theil index, GE(1), to be the appropriate projection function. The study of political decisions or other public administrative issues which suggests choosing $\mathbf{W}_{(r)} = \mathbf{1}_{(r)}$ and $\mathbf{\Pi}_{(r)} = \mathbf{L}_{(r)}$ should be based on

---

[27] The [weighted] relative $GE(\alpha)$ measures are obtained from (6) by assuming (i) the basic units to be the area accounted for on average by one industry i-worker, (ii) re-defining the characteristics as some ratio so that, e.g., $Y_n = L_{in}/L_{\bullet n}$, and decomposing it into within- and a between-region components. Assuming the unobservable within-region components to be zero, the between-regions component is left which is the relative measure (Brülhart and Träger 2005). Notice that the decomposition is formally equivalent to aggregating the unobservable basic units to observable industries, if the within component is zero.

the unweighted relative Theil index $T_i^{UR} = \sum_{r=1}^{R} \lambda_{ir} \ln(R\lambda_{ir})$. And the study comparing the spatial distribution of an industry to that of total employment for each km² ($\mathbf{W}_{(r)} = \mathbf{A}_{(r)}$; $\mathbf{\Pi}_{(r)} = \mathbf{L}_{(r)}$) should be based on the weighted relative Theil index $T_i^{WR} = \sum_{r=1}^{R} \frac{A_r}{A} \frac{l_{ir}}{\sum_r \frac{A_r}{A} l_{ir}} \ln\left(\frac{l_{ir}}{\sum_r \frac{A_r}{A} l_{ir}} \right)$. Notice that the decomposition of the corresponding GE measure or Theil index for the basic units requires assuming that the fraction of industry-*i* workers in the total workforce does not vary across space within a region, i.e., $L_{in}/L_{\bullet n} = L_{ir}/L_{\bullet r}$ $\forall\ n = 1, \ldots, A_r$. L_{in} and $L_{\bullet n}$ denote industry-*i* and total employment on the *n*-th km² in region *r*.

A third and final group of measures to be exemplified here are measures based on the Gini projection function.[28] The Gini coefficient is generally defined as two times the area between the Lorenz curve and the 45° line (shaded area in Figure 1) in a box plot of cumulated shares of individuals in the population on the horizontal axis and the cumulated shares of their characteristics on the vertical axis. In terms of the taxonomy of the present paper, the population, depicted on the horizontal axis, is made up by the basic units, the shares of which are the (relative) region-specific weights, w_r. The characteristics, depicted on the vertical axis, are the weighted region-specific deviations, the shares of which are $w_r \frac{L_{ir}}{\Pi_r} \Big/ \sum_r w_r \frac{L_{ir}}{\Pi_r}$. All observations are sorted in ascending order by the region-specific deviations (L_{ir}/Π_r,).[29] This convention allows to classify the various Gini coefficients used in the literature, and facilitates comparisons to measures with different projection functions.

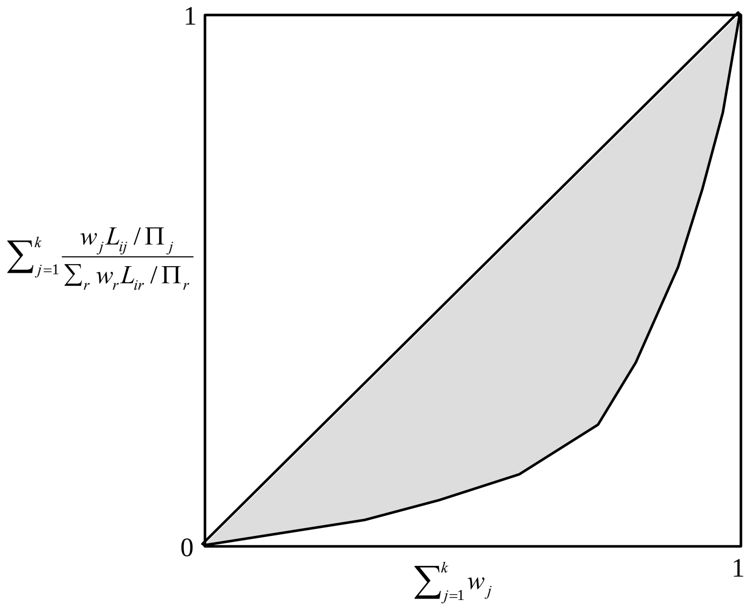
The Gini coefficients for the various choices of basic units and references can be defined similar to the definitions of the *RMD* and *GE* measures above:[30] With regions as basic units, the *unweighted absolute* Gini coefficient is defined by $\mathbf{\Pi}_{(r)} = \mathbf{W}_{(r)} = \mathbf{1}_{(r)}$; the *unweighted relative* Gini is defined by $\mathbf{\Pi}_{(r)} = \mathbf{L}_{(r)}$ and $\mathbf{W}_{(r)} = \mathbf{1}_{(r)}$. In both cases, cumulated values of $w_r = 1/R$ are plotted on the horizontal axis. The vertical axis gives the cumulated employment shares of the region-industry, $\Sigma_{j\leq k} l_{jr}$ (unweighted absolute Gini; see, e.g., Midelfart-Knarvik et al. 2002), or the cumulated normalized employment ratios between the region and the reference, $\sum_{j\leq k} \left(\frac{L_{ij}}{L_{\bullet j}} \Big/ \frac{1}{R} \sum_r \frac{L_{ir}}{L_{\bullet r}} \right) = \sum_{j\leq k} l_{ij} \Big/ \frac{1}{R} \sum_r l_{ir}$ (unweighted relative Gini; Südekum 2006). Finally, when

28 As the relative mean deviation, the Gini coefficient is an intuitively appealing ad hoc measure. It meets the requirements of the axiomatic approach only under specific conditions. It is, e.g., additively decomposable into a within and a between component only under specific conditions. In contrast to the GE measures, the resulting between components are not necessarily *weighted relative* Gini coefficients (see, e.g., Yao 1999).

29 k in Figure 1 and the subsequent formulae indexes the observation with the k^{th} lowest region-specific deviation.

30 See Table 2 for the details. There are various different ways of formally expressing the Gini coefficients. We prefer the one which does not require sorting observations by the ratio of the values depicted at the vertical and horizontal axes.

Figure 1 – Lorenz curve of industrial concentration



the basic units are areas attributed to individual workers, $\mathbf{W}_{(r)} = \mathbf{L}_{(r)}$, and the reference is total regional employment, $\mathbf{\Pi}_{(r)} = \mathbf{L}_{(r)}$, the horizontal axis depicts cumulated employment shares in the reference economy, $\Sigma_{j\leq k}\lambda_j$, and the vertical axis cumulated region-industry employment shares, $\Sigma_{j\leq k}\lambda_{ij}$, giving the weighted relative Gini coefficient (e.g., Krugman 1991).

The taxonomy introduced in the preceding chapter allows to define and classify a great variety of measures of industrial concentration, or, for that purpose, of regional specialization. The following chapters will introduce two generalizations of the taxonomy that will broaden the scope of proportionality measures, another step towards meeting the "baseline criteria" of Combes and Overman (2004).

3. Generalization 1: Measures of polarization

Polarization measures evaluate industrial concentration and regional specialization within an economy simultaneously. Formally, they are straightforward generalizations of the proportionality measures discussed above. Rather than one row or column of the (IxR) matrix of region-industry employment, $\mathbf{L}_{(ir)}$, the polarization measures evaluate all elements of the matrix. Proportionality measures for two-dimensional data have recently been discussed in the sociological segregation literature (Reardon and Firebaugh 2002). The following discussion adopts these segregation measures and generalizes them along the lines of the taxonomy.

In terms of the taxonomy polarization measures require specifying an (Ix1) vector of industry-specific weights ($\mathbf{W}_{(i)}$) in addition to the (Rx1) vector of region-specific weights ($\mathbf{W}_{(r)}$), and re-defining the reference. Similar to the region-specific weights, the industry-specific weights reflect the choice of basic units in the sectoral dimension. The basic units may be whole industries ($\mathbf{W}_{(i)} = \mathbf{1}_{(i)}$), or relate to the activities of individual workers ($\mathbf{W}_{(i)} = \mathbf{L}_{(i)}$). The reference is a bivariate distribution represented by an (IxR) matrix $\mathbf{\Pi}_{(ir)}$. For absolute measures it becomes $\mathbf{1}_{(ir)}$. For relative measures it may take various values. If the references are total employment by industry and region, $\mathbf{\Pi}_{(ir)} = \mathbf{L}_{(i)}\mathbf{L}_{(r)}^T$, a matrix with element (i, r) equal to $L_{i\bullet}L_{\bullet r}$. If the references for industries are total employment by industry and those for regions are the area by region, $\mathbf{\Pi}_{(ir)} = \mathbf{L}_{(i)}\mathbf{A}_{(r)}^T$. If the reference is determined from some probabilistic model, $\mathbf{\Pi}_{(ir)} = [\mathrm{E}(L_{ir})]$ is the expected region-industry employment under the null hypothesis of no polarization. Notice that there is no necessity to specify the same kinds of references for all industries, or all regions. The references for some industries may, e.g., be related to the regions' areas, while those for other industries is related to the regions' total employment.

Table 3 depicts the general forms of the measures for several projection functions, similar to the first column of Table 2. The various weighted and unweighted absolute and relative measures can be derived from these general forms in a way similar to that outlined in the preceding chapter. To give a few examples, a weighted relative GE measure of polarization for $\mathbf{\Pi}_{(ir)} = \mathbf{L}_{(i)}\mathbf{L}_{(r)}^T$, $\mathbf{W}_{(i)} = \mathbf{L}_{(i)}$ and $\mathbf{W}_{(r)} = \mathbf{L}_{(r)}$ can be derived from the general form of the GE polarization measure in Table 3 as

$$GE^{WR}(\alpha) = \frac{1}{\alpha(\alpha-1)}\sum_{i=1}^{I}\sum_{r=1}^{R} l_i\lambda_r\left(LC_{ir}^{\alpha}-1\right); \tag{7}$$

an unweighted relative polarization GE for $\mathbf{\Pi}_{(ir)} = \mathbf{L}_{(i)}\mathbf{L}_{(r)}^T$, $\mathbf{W}_{(i)} = \mathbf{1}_{(i)}$ and $\mathbf{W}_{(r)} = \mathbf{1}_{(r)}$ as

$$GE^{UR}(\alpha) = \frac{1}{\alpha(\alpha-1)}\frac{1}{IR}\sum_{i=1}^{I}\sum_{r=1}^{R}\left[\left(IR\frac{LC_{ir}}{\sum_i\sum_r LC_{ir}}\right)^{\alpha}-1\right]. \tag{8}$$

For addressing the baseline criterion (i) in Combes and Overman (2004) which requests measures to be comparable across industries and regions, the decomposition of GE measures of polarization or other decomposable disproportionality measures is a particularly useful tool. The Theil index of polarization (see Table 3), for example, can be decomposed into a within- and a between-industry component such that the within-industry component is a weighted average of the Theil indices of industrial concentration for the individual industries (see Table 2), and the between-industry component the Theil index of specialization in the economy on average (see Table A1).[31] A similar decomposition can be done in the other

[31] With the Theil index of polarization given by $T = \sum_{i,r} w_i w_r (X_{ir}/\bar{X})\ln\left(X_{ir}/\bar{X}\right)$, where $X_{ir} = L_{ir}/\Pi_{ir}$ and $\bar{X} = \sum_{i,r} w_i w_r X_{ir}$, the decomposition yields $T = \sum_i w_i(\bar{X}_i/\bar{X})T_i + T_b$, where $\bar{X}_i = \sum_r w_r X_{ir}$, $T_i = \sum_r w_r (X_{ir}/\bar{X}_i)\ln\left(X_{ir}/\bar{X}_i\right)$, and $T_b = \sum_i w_i \bar{X}_i/\bar{X}\ln\left(\bar{X}_i/\bar{X}\right)$.

dimension to obtain region-specific Theil indices of regional specialization. The between component is then the Theil index of average industrial concentration.

For the case of weighted relative polarization measures which use industry and region totals as references and weights, i.e., $\mathbf{\Pi}_{(ir)} = \mathbf{L}_{(i)}\mathbf{L}_{(r)}^T$, $\mathbf{W}_{(i)} = \mathbf{L}_{(i)}$, and $\mathbf{W}_{(r)} = \mathbf{L}_{(r)}$, the polarization measures are simply the weighted averages of the corresponding concentration or specialization measures (Reardon and Firebaugh 2002). This is true not only for the GE measures but also for those measures that do not meet the general decomposability requirement. For the RMD, e.g., one gets

$$RMD^{WR} = f_{MRD}\left(\mathbf{L}_{(ir)}, \mathbf{L}_{(i)}, \mathbf{L}_{(r)}, \mathbf{L}_{(i)}\mathbf{L}_{(r)}^T\right) = \sum_{i=1}^{I}\sum_{r=1}^{R} l_i \lambda_r \left|LC_{ir} - 1\right| = \sum_{i=1}^{I} l_i MRD_i^{WR} = \sum_{r=1}^{R} \lambda_r MRD_r^{WR}. \quad (9)$$

Table 3 — Polarization measures for selected projection functions: general forms[a]

measure	general form
GE(α)	$GE(\alpha) = \frac{1}{\alpha(\alpha-1)}\sum_{i=1}^{I}\sum_{r=1}^{R} w_i w_r \left[\left(\frac{\frac{L_{ir}}{\Pi_{ir}}}{\sum_{i,r} w_i w_r \frac{L_{ir}}{\Pi_{ir}}}\right)^{\alpha} - 1\right]$
GE(1) = Theil	$T = \sum_{i=1}^{I}\sum_{r=1}^{R} w_i w_r \frac{\frac{L_{ir}}{\Pi_{ir}}}{\sum_{i,r} w_i w_r \frac{L_{ir}}{\Pi_{ir}}} \ln\left(\frac{\frac{L_{ir}}{\Pi_{ir}}}{\sum_{i,r} w_i w_r \frac{L_{ir}}{\Pi_{ir}}}\right)$
CV = 2GE(2)$^{0.5}$	$CV = \left[\sum_{i=1}^{I}\sum_{r=1}^{R} w_i w_r \left[\left(\frac{\frac{L_{ir}}{\Pi_{ir}}}{\sum_{i,r} w_i w_r \frac{L_{ir}}{\Pi_{ir}}}\right)^{2} - 1\right]\right]^{0.5}$
RMD / Krugman	$MRD = \sum_{i=1}^{I}\sum_{r=1}^{R} w_i w_r \left\lvert\frac{\frac{L_{ir}}{\Pi_{ir}}}{\sum_{i,r} w_i w_r \frac{L_{ir}}{\Pi_{ir}}} - 1\right\rvert$
Gini	$G = \frac{1}{2\sum_{i,r} w_i w_r \frac{L_{ir}}{\Pi_{ir}}}\sum_{i=1}^{I}\sum_{j=1}^{I}\sum_{r=1}^{R}\sum_{s=1}^{R} w_i w_j w_r w_s \left\lvert\frac{L_{ir}}{\Pi_{ir}} - \frac{L_{js}}{\Pi_{js}}\right\rvert$

[a] See the Appendix for notational details. The corresponding unweighted absolute, unweighted relative and weighted relative measures are obtained from the general forms in the same way as described in Chapter 2 and Table 2. The lower bound of all measures is 0. For the upper bounds see http://www.uni-kiel.de/ifw/staff/bode/measures.htm.

4. Generalization 2: Spatial measures of industrial concentration

All disproportionality measures discussed so far are invariant to the spatial ordering of the regions under investigation. Ignoring the spatial ordering of the data gives rise to the checkerboard problem: The measures systematically understate the true degree of industrial concentration if the industry is clustered at a spatial scale larger than the spatial units under investigation. To avoid the checkerboard problem, measures need to take into account the spatial ordering of the observations. Arbia (2001) suggests combining inequality measures such as the Gini coefficient with statistics of spatial association such as Moran's I and/or the Getis-Ord statistic.[32] While the inequality measure is informative as to the a-spatial concentration, the spatial statistics give an indication of the spatial clustering.

Rather than combining a-spatial and spatial measures in an ad-hoc way, this paper suggests introducing the spatial dimension directly into the disproportionality measures. The resulting measures which may be labeled "spatial" concentration measures (spatial measures for short), are actually generalizations of the a-spatial concentration measures discussed in the preceding chapters (see Table 2). The basic idea is to complement the information on the industry in question from each region by the corresponding information on the industry from the regions close-by. Reardon and O'Sullivan (2004) suggest to do so in a way similar to a kernel density estimation, or a geographically weighted analysis.[33] More specifically, they suggest defining a measure in terms of the geographically weighted averages of the variables of main interest (L_{ir} in this paper) and the reference (Π_r). This approach will help lessen, though not completely avoid, the MAUP and the checkerboards problem inherent to any analysis of concentration based on regional aggregates.

To extend the taxonomy introduced in Chapter 2 to spatial disproportionality measures, the L_{ir} and Π_r are to be re-defined as geographically weighted averages,

$$L_{ir}^G = \sum_{q=1}^{R} \phi_{qr} L_{iq} \text{ and } \Pi_r^G = \sum_{q=1}^{R} \phi_{qr} \Pi_q ,$$

where $\phi_{qr} := \Phi_{qr} / \sum_{q=1}^{R} \Phi_{qr}$,[34] and the superscript G denotes geographically weighted averages. The non-negative geographical weights, or spatial discount factors, Φ_{qr}, reflect the strength of the influence of any region q on region r. The strength of the influences may generally depend

---

[32] Lafourcade and Mion (2006) do essentially the same but test in addition for the effects of firm sizes by combining the dartboard measure and the Moran's I statistic.

[33] Reardon and O'Sullivan (2004) discuss this approach in the context of spatial segregation measures, and for continuous space. As in the previous chapters, the following discussion will focus on disproportionality measures for regional aggregates.

[34] $\phi_{rr} = 1$ and $\phi_{qr} = 0$ for $q \neq r$ gives the corresponding a-spatial measures (Table 2).

on geographic distances, neighborhood structures, or accessibility more generally. To meet the usual regularity conditions, the weights are row-normalized such that the weights sum up to one for each region, i.e., $\sum_{q=1}^{R} \phi_{qr} = 1$.

Extending the set of arguments of the concentration measures discussed above by an (RxR) row-normalized matrix of bilateral geographical weights, $\mathbf{\Phi}_{(r)}$,[35] and substituting L_{ir}^{G} and Π_{r}^{G} for L_{ir} and Π_r, all measures in Table 2 can be extended to spatial measures of industrial concentration. The general form of the spatial GE class of measures reads, e.g.,

$$GE_i^G = f_{GE}\left(\mathbf{L}_{i(r)}, \mathbf{W}_{(r)}, \mathbf{\Pi}_{(r)}, \mathbf{\Phi}_{(r)}\right) = \frac{1}{\alpha(\alpha-1)} \sum_{r=1}^{R} w_r \left[\left(\frac{\frac{\sum_{q=1}^{R} \phi_{qr} L_{iq}}{\sum_{q=1}^{R} \phi_{qr} \Pi_q}}{\sum_r w_r \frac{\sum_{q=1}^{R} \phi_{qr} L_{iq}}{\sum_{q=1}^{R} \phi_{qr} \Pi_q}} \right)^{\alpha} - 1 \right]. \qquad (10)$$

Due to the geographical weighting, the effect of region r on the measure is magnified if industry i is overrepresented (or underrepresented) in both the region itself and its neighbors who are assigned a comparatively high geographical weight. And it is reduced if industry i is over- (under-) represented in region r but under- (over-) represented in its neighbor regions. It should be noted that due to the interdependencies introduced by the geographical weights, decomposing the spatial GE measures in the usual way is not possible.[36] A change of concentration between subregions of one region (e.g., country) may influence the concentration within another region by affecting its subregion-specific deviations non-uniformly.

The spatial measures are capable of reducing biases resulting from the checkerboard problem and the MAUP: The checkerboard problem is reduced by taking into account the geographical ordering of the regions; the arbitrary boundary problem by geographical smoothing; and the scale problem by carefully specifying the intra- and interregional weights. If transport costs or other spatial transaction costs are considered the determinants of regional interdependencies, the geographical weights could be operationalized by some functions of the geographical or economic distance between any two regions, i.e., $\phi_{qr} = \phi(D_{qr})$ where D_{qr} denotes the distance between the regions q and r, and $\partial\phi/\partial D_{qr} < 0$.[37] One possibility of specifying the unobservable *intra*-regional distances is to assume all workers to be concentrated at a single regional center. In this case, $D_{rr} = 0$ (but $\phi_{rr} > 0$), and the inter-regional distances are just the distances

---

35 For notational convenience, $\mathbf{\Phi}_{(r)}$ denotes the matrix of row-normalized rather than absolute weights. The square matrix $\mathbf{\Phi}_{(r)}$ is actually a row-normalized spatial weights matrix with the intraregional weights on the main diagonal. All rows sum up to one. In matrix notation, $\mathbf{L}_{(r)}^{G} = \mathbf{\Phi}_{(r)} \mathbf{L}_{(r)}$ and $\mathbf{\Pi}_{(r)}^{G} = \mathbf{\Phi}_{(r)} \mathbf{\Pi}_{(r)}$.

36 See Reardon and O'Sullivan (2004) for more specific ways of decomposing spatial disproportionality measures.

37 For a discussion of alternative forms of geographical weights see, e.g., Anselin (1988).

between the regional centers. Other possibilities are to assume all workers and units to be distributed uniformly over space within each region, or to estimate the intra-regional distributions from a finer partition of regions available from, e.g., population or electoral statistics.

For geo-referenced micro-data that provide information on the distances between any pairs of establishments, Duranton and Overman (2005) and Marcon and Puech (2003; 2005) have recently proposed to describe industrial concentration by functions based on the Ripley's K function. The K-based functions, which assign each possible distance a frequency of observations,[38] arguably provide the currently most sophisticated measures of industrial concentration because they avoid the checkerboard problem and the MAUP.[39] The spatial disproportionality measures proposed in the present chapter are a promising alternative to the K-based functions. Both approaches may in principle be used for aggregate or disaggregate data. For a given level of regional aggregation, they are capable of dealing with the checkerboard problem and the MAUP to a similar extent.

5. Conclusion

On the background of the baseline criteria in Combes and Overman (2004), the paper aims at improving the methodological toolbox of disproportionality measures to allow a more rigorous exploratory analysis of industrial concentration and regional specialization. The paper, first, proposes a taxonomy of disproportionality measures which distinguishes carefully the measures' three characteristic features: the projection function, the reference distribution, and the weighting scheme. The taxonomy gives rise to a modular construction system for disproportionality measures. By adjusting each characteristic feature, a researcher may define the measure that best fits the research purpose and the limitations of the available data. The modular system is also useful for evaluating systematically the robustness of the inferences against a variation of the individual features of the measure. Although the modular system of characteristic features does not substitute for an economic theory or sufficiently detailed data, a careful reflection of the individual features on the background of the research hypothesis and the specificities of the available data may help reduce the mismatch between research purpose, data and statistical measure, that has been one of the main obstacles to consistent and reliable inferences in the literature.

---

38 The K-based functions include an "L function" which is a standardized "Ripley's K", a "K density function" which is a marginal Ripley's K smoothed by kernel density estimation, and an "M function" which is a weighted relative Ripley's K. See Marcon and Puech (2005) for a comparative survey.

39 The choice of a proper distance metric and reference in the spatial dimension, as well as the scale, arbitrary boundary and checkerboard problems in the sectoral dimension still remain an issue, however.

In two generalizations, the paper extends the specialization and concentration literature by introducing disproportionality measures of polarization, and spatial disproportionality measures. Measures of polarization may be used for evaluating regional specialization and industrial concentration simultaneously, and render possible a nested analysis of polarization, specialization and/or concentration at different spatial and industrial scales. Spatial measures of concentration help address the checkerboard problem and the MAUP, thus representing a promising alternative to K-based statistics. Using geographically weighted averages of the relevant data as an input, the spatial measures allow to take into account the specific characteristics of neighboring regions, and to account for possible intra-regional distributions of the variables of interest.

We are confident that the taxonomy and the disproportionality measures proposed in this paper will prove useful for a wide range of empirical studies on industrial concentration, regional specialization and economic polarization. There are, of course, a number of issues which warrant additional methodological research. A first issue is to generalize the taxonomy to spatial polarization measures. A second issue is to further explore opportunities for an informative decomposition of the spatial GE measures (see Reardon and O'Sullivan 2004). A third issue is to explore ways of coping with the counterparts of scale, arbitrary boundary and checkerboard problems in the sectoral dimension. Unlike the spatial dimension where geographical distance or traveling time is widely accepted as a metric for relating the locations of individual units to each other, the sectoral dimension is still lacking a widely accepted metric. A metric for the distances between industries may be based on the coefficients of input-output tables, or on proxies of the similarity of the firms' or industries' in terms of their input markets, output markets, or technologies (see Conley and Dupor 2003; Bloom et al. 2005). Based on distances between basic units in both the spatial and the sectoral dimension, the spatial polarization measures may be extended to geographically and sectorally weighted polarization measures and/or K statistics which account for MAUP and the checkerboard problem in both dimensions. A fourth issue is to investigate in more detail into the comparative pros and cons of the spatial disproportionality measures and the K-based functions for both micro and macro data.

References

Aiginger, K., and S.W. Davies (2004), Industrial Specialization and Geographic Concentration: Two Sides of the Same Coin? Not for the European Union. *Journal of Applied Economics* 7 (24): 231–248.

Aiginger K., and W. Leitner (2002), Regional Concentration in the USA and Europe: Who Follows Whom?” *Weltwirtschaftliches Archiv* 138 (4): 652–679.

Amiti, M. (1998), New Trade Theories and Industrial Location in the EU. *Oxford Review of Economic Policy* 14 (2): 45–53.

Anselin, L. (1988), *Spatial Econometrics: Methods and Models.* Dordrecht: Kluwer.

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

— (2001), The Role of Spatial Effects in the Empirical Analysis of Regional Concentration. *Journal of Geographical Systems* 3 (3): 271–281.

Bickenbach, F., E. Bode, C. Krieger-Boden (2006), Economic Structures of European Regions: An Overview. Mimeo.

Bloom, N., M. Schankerman and J. Van Reenen (2005), Identifying Technological Spillovers and Product Market Rivalry. CEP Discussion Paper 675. London School of Economics.

Bode, E., J. Bradley, G. Fotopoulos, D. Kallioras, C. Krieger-Boden, N. Lubenets, E. Morgenroth, G. Petrakos, L. Resmini, R. Soltwedel and I. Traistaru (2003), European Integration, Regional Structural Change and Cohesion: A Survey of Theoretical and Empirical Literature. EURECO Working Paper No. 1. Bonn: Center of European Integration (download: http://www.zei.de/eurec/frameeureco.htm).

Brülhart, M. (2001), Evolving Geographical Concentration of European Manufacturing Industries. *Weltwirtschaftliches Archiv* 137 (2): 215–243.

—, and R. Träger (2005), An Account of Geographic Concentration Patterns in Europe. *Regional Science and Urban Economics* 35 (6): 597–624.

Combes, P.-P., and H.G. Overman (2004), The Spatial Distribution of Economic Activities in the European Union. In: J.V. Henderson and J.-F. Thisse (eds.), *Handbook of Urban and Regional Economics*. Volume 4. Amsterdam: North Holland.

Conley, T.G., and B. Dupor (2003), A Spatial Analysis of Sectoral Complementarity. *Journal of Political Economy* 111 (2): 311–351.

Cowell, F.A. (1995), *Measuring Inequality*. 2nd Edition. London: Prentice Hall.

— (2000), Measurement of Inequality. In: A.B. Atkinson and F. Bourguignon (eds.), *Handbook of Income Distribution*, Volume 1. Amsterdam: Elsevier.

Dohse, D., C. Krieger-Boden, R. Soltwedel (2002). EMU and Regional Labor Market Disparities in Euroland. In: J.R. Cuadrado-Roura, M. Parellada (eds.), *Regional Convergence in the European Union.* Berlin, Heidelberg.

Duranton, G., and H. Overman (2005). Testing for Localisation Using Micro-Geographic Data. *Review of Economic Studies* 72 (4): 1077–1106.

Ellison, G., and E.L. Glaeser (1997), Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach. *Journal of Political Economy* 105 (5): 889-927.

Haaland, J.I., H.J. Kind, K.H. Midelfart-Knarvik and J. Torstensson (1998), What Determines the Economic Geography of Europe? Discussion paper 98, 19, Norwegian School of Economics and Business Administration, Bergen.

Hallet, M. (2002), Regional Specalization and Concentration in the EU. In J.R. Cuadrado-Roura (ed.), *Regional Convergence in the European Union*. Berlin: Springer.

Krugman, P.R. (1991), *Geography and Trade*. Leuven: Leuven Univ. Press.

Lafourcade, M., and G. Mion (2006), Concentration, Agglomeration and the Size of Plants. *Regional Science and Urban Economics* (forthcoming).

Litchfield, J.A. (1999), Inequality Methods and Tools. London: London School of Economics (download: http://www1.worldbank.org/prem/poverty/inequal/methods/litchfie.pdf).

Marcon, E., and F. Puech (2003), Evaluating the Geographic Concentration of Industries Using Distance-based Methods. *Journal of Economic Geography* 3 (4): 409–428.

— (2005), Measures of the Geographic Concentration of Industries: Improving Distance-based Methods. Mimeo (download: http://e.marcon.free.fr/publications/index.htm).

Maurel, F., and B. Sédillot (1999), A Measure of the Geographic Concentration in French Manufacturing Industries. *Regional Science and Urban Economics* 29: 575-604.

Midelfart-Knarvik, K.H., H.G. Overman, S.J. Redding and A.J. Venables (2002), The Location of European Industry. *European Economy*, Special Report 2/2002. Brussels: European Commission.

Mori, T., K. Nishikimi and T.E. Smith (2005), A Divergence Statistic for Industrial Localization. *Review of Economics and Statistics* 87 (4): 635–651.

Nijkamp, P., L. Resmini and I. Traistaru (2003), European Integration, Regional Specialization and Location of Industrial Activity: a Survey of Theoretical and Empirical Literature. In I. Traistaru, P. Nijkamp and L. Resmini (eds.), *The Emerging Economic Geography in EU Accession Countries*. Aldershot: Ashgate.

Openshaw, S., and P.J. Taylor (1979), A Million or So Correlated Coefficients: Three Experiments on the Modifiable Areal Unit Problem. In N. Wrigley and R.J. Bennet (eds), *Statistical Applications in the Spatial Sciences*. London: Pion.

Reardon, S.F., and G. Firebaugh (2002), Measures of Multigroup Segregation. *Sociological Methodology* 32 (1): 33–67.

Reardon, S.F., and D. O'Sullivan (2004), Measures of Spatial Segregation. *Sociological Methodology* 34 (1): 121–162.

Silber, J. (1999) (ed.), *Handbook of Income Inequality Measurement*. Boston: Kluwer.

Südekum, J. (2006), Concentration and Specialisation Trends in Germany since Re-Unification. *Regional Studies*, forthcoming.

Traistaru, I., P. Nijkamp and S. Longhi (2003), Specialization of Regions and Concentration of Industries in EU Accession Countries. In I. Traistaru, P. Nijkamp and L. Resmini (eds.), *The Emerging Economic Geography in EU Accession Countries*. Aldershot: Ashgate.

Yao, S. (1999), On the Decomposition of Gini Coefficients by Population Class and Income Source: A Spreadsheet Approach and Application. *Applied Economics* 31 (10): 1249–1264.

Appendix

Notation:

L_{ir}	variable of main interest: number of workers in industry i ($i = 1, \ldots, I$) in region r ($r = 1, \ldots, R$);
$L_{i\bullet}$	$:= \Sigma_r L_{ir}$, number of workers in industry i at a higher-level spatial aggregate (e.g., country, EU);
$L_{\bullet r}$	$:= \Sigma_i L_{ir}$, total employment in region r;
$L_{\bullet\bullet}$	$:= \Sigma_i \Sigma_r L_{ir}$, total employment a higher-level spatial aggregate;
l_{ir}	$:= L_{ir}/L_{\bullet r}$, share of industry i in total employment of region r;
l_i	$:= L_{i\bullet}/L_{\bullet\bullet} = L_{i\bullet}/\Sigma_i L_{i\bullet}$, share of industry i in total employment at the higher-level spatial aggregate;
λ_{ir}	$:= L_{ir}/L_{i\bullet}$, share of region r in total industry-i employment at the higher-level spatial aggregate;
λ_r	$:= L_{\bullet r}/L_{\bullet\bullet} = \Sigma_i L_{ir}/\Sigma_i L_{i\bullet}$, share of region r in total employment at the higher-level spatial aggregate;
LC_{ir}	$:= l_{ir}/ l_i = \lambda_{ir}/\lambda_r$, location coefficient;
Π_z,	$z = (r, i, ir)$, absolute value of the reference;
W_z	$z = (r, i)$, absolute region- or industry-specific weight;
w_z	$:= W_z/\Sigma_z W_z$, $\Sigma_z w_z = 1$, $z = (r, i)$, relative region- or industry-specific weight;
ω	$:=(\alpha^2-\alpha)^{-1}$; α: sensitivity parameter of the Generalized Entropy measures.

Table A1 — Specialization measures for selected projection functions: general forms[a]

projection function	general form
GE(α)	$GE_r(\alpha) = \frac{1}{\alpha(\alpha-1)} \sum_{i=1}^{I} w_i \left[\left(\frac{\frac{L_{ir}}{\Pi_i}}{\sum_i w_i \frac{L_{ir}}{\Pi_i}} \right)^{\alpha} - 1 \right]$
GE(1) = Theil	$T_r = \sum_{i=1}^{I} w_i \frac{\frac{L_{ir}}{\Pi_i}}{\sum_i w_i \frac{L_{ir}}{\Pi_i}} \ln \left(\frac{\frac{L_{ir}}{\Pi_i}}{\sum_i w_i \frac{L_{ir}}{\Pi_i}} \right)$
CV = 2GE(2)$^{0.5}$	$CV_r = \left[\sum_{i=1}^{I} w_i \left[\left(\frac{\frac{L_{ir}}{\Pi_i}}{\sum_i w_i \frac{L_{ir}}{\Pi_i}} \right)^{2} - 1 \right] \right]^{0.5}$
RMD / Krugman	$RMD_r = \frac{1}{\sum_i w_i \frac{L_{ir}}{\Pi_i}} \sum_{i=1}^{I} w_i \left| \frac{L_{ir}}{\Pi_i} - \sum_i w_i \frac{L_{ir}}{\Pi_i} \right|$
Gini	$G_r = \frac{1}{2 \sum_i w_i \frac{L_{ir}}{\Pi_i}} \sum_{i=1}^{I} \sum_{j=1}^{I} w_i w_j \left| \frac{L_{ir}}{\Pi_i} - \frac{L_{jr}}{\Pi_j} \right|$

[a] See above for notational details. The corresponding unweighted absolute, unweighted relative and weighted relative measures are obtained from the general forms in the same way as described in Chapter 3 and Table 2. The lower bound of all measures is 0. For the upper bounds see http://www.uni-kiel.de/ifw/staff/bode/measures.htm.