PRZEGLĄD STATYSTYCZNY
NUMER SPECJALNY 1 – 2012

JAN W. OWSIŃSKI

# ON THE OPTIMAL DIVISION OF AN EMPIRICAL DISTRIBUTION (AND SOME RELATED PROBLEMS)[1]

## 1. THE BACKGROUND

It is quite frequent that a set of entities observed and analysed is divided up into subsets. This is done so as to treat the entities within the subsets as similar, or even "equivalent" in terms of the feature(s) considered, and to be able to correctly distinguish between the subsets, i.e. entities put in different subsets being sufficiently dissimilar with respect to these features.

There may be various reasons for such divisions, examples being: differentiation of policy measures with respect to regions featuring various levels of GDP per capita, membership fees for countries featuring various economic power, social policy measures for households characterised by different levels of poverty, etc. There may also simply be the reasons of pragmatism in addressing intensities of respective phenomena ("highly developed", "developed",. . . , countries, or "marginally poor", "poor", "very poor", etc. households). These reasons are usually closely coupled with a cognitive interest, in that we would like to make the divisions possibly "objective" and that we would like to know whether there exist any mechanisms that drive the divisions – and what they are, if they do exist.

This problem is very closely associated with what is called "categorisation" – i.e. representation of a set of values, taken by some variable, characterising the entities analysed, by a limited number of "categorical" values, to be then used instead of the original (measured) values. Another aspect is associated with the possibility of representing these categories through some natural language expressions.

In very general terms one might classify the approaches to determination of the distribution division here meant in the following manner:

– *statistical*, most often based on some quantile-type criteria, with the respective levels selected on the basis of either tests, or, more often, expert-based assessments;
– *political*, when expert-based assessment is directly used to set the dividing lines, over which a political decision process takes place;

[1] The work, whose results are reported here, was partly financed from the projects: TIROLS, N N516 195237, financed by the Polish Ministry of Science and Higher Education, and the modeling project in the framework of the "Development Trends of Masovia", MBPR/W/TRM-0712-08-2011, co-financed from the EU funds.

– *substantive*, when analysis is carried out of the phenomenon, giving rise to the observed values, and the dividing lines are determined as meaningful from the point of view of the phenomenon itself.

In reality, of course, the three often appear in conjunction, with varying emphasis. We shall illustrate these approaches and the related questions with the academic example of Fig. 1[2]. This figure shows monthly remuneration in Polish zlotys in a small company.

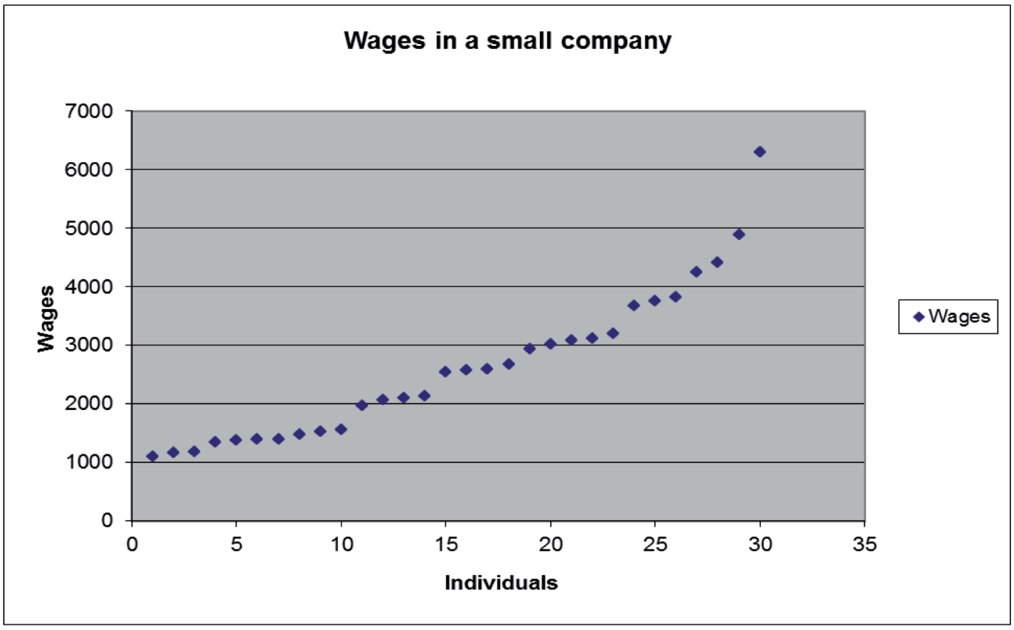

Figure 1. An academic example of a distribution: wages in a small company

If we wanted to split the distribution like the one of Fig. 1 into few meaningful categories, the premises could be as follows:

– statistical (or, actually, in this case, ”statistical”): for instance, given that the descriptive statistics for this case are min = 1 100 PLN per month, max = 6 300 PLN, mean = 2621.67 PLN, and standard deviation = 1273.64 PLN, it would seem reasonable to divide the distribution into “low wages” (up to mean – standard deviation), “middle wages” (within the standard deviation away from the mean), and “high wages” (above mean + standard deviation); the 30 observations would then be split up into three subsets, composed of, respectively, 4, 20 and 6 observations; otherwise, this could be done on the basis of quantiles; the problem with this division is that it takes into account the shape of the distribution to a very limited degree – actually, it cuts through one of the distinct plateaus, visible in Fig. 1, while associating in one category a lot of different wage levels;

[2] All figures in his text result from the research of the author.

– political: a division could be based on the “minimum wage” definition, along with other criteria, linked, e.g., with taxation, social security etc.; such criteria do not, of course, account for the shape of the distribution;

– substantive: wages may depend upon certain regulations, involving education, length of work experience, position in the company etc.; this division would seem to correspond the best to the respective reality, but, especially in this particular case, the multiplicity of criteria and the corresponding partial subdivisions would yield either too narrow categories or would require some additional aggregating operator to be applied.

All in all – we would like to have, therefore, a methodology that could lead to reasonable division into categories on the basis of actual reality in the form of the given empirical distribution.

## 2. NOTATION AND INITIAL FORMULATION

We are given a univariate empirical distribution of a quantity $x$ taking values from $\boldsymbol{R}_+$. The distribution consists of $n$ observations, indexed $i$, $i = 1,\dots,n$. This set of indices shall be denoted $I$. Without any loss to any sort of reasoning, we assume that the values taken by $x$ in this distribution, denoted $x_i$, are ordered in a non-decreasing sequence, i.e. $x_{i+1} \geq x_i$ for all $i$.

Next, assume we consider, instead of the sequence $\{x_i\}_I$, the corresponding sequence, formed by the corresponding cumulative distribution, i.e. the values $z_i$ defined as $z_i = \sum_{i'=1,\dots,i} x_i$. We deal, therefore, with the sequence $\{z_i\}_I$ that is increasing, and, moreover, also convex. This means that a straight line, joining any two points of the sequence $\{z_i\}_I$, say $z_i$ and $z_{i+\Delta I}$, where $\Delta i$ is any integer number contained in the interval $[2,n\text{-}i]$, has values above those of the corresponding $z_i$, i.e. $z_{i+1},\dots,z_{i+\Delta i-1}$.

Having such data, very much like a Lorenz curve, we would like to construct a sort of piece-wise linear approximation that is exactly based upon the segments of straight lines, which is in some sense “optimal”. Namely, we would like to determine a set of segments such that the resulting error (sum of absolute differences between the actual values of $z_i$ and the corresponding values of the approximating function) is possibly low, while the number of segments we distinguish is also kept reasonably low.

If we could obtain such a “more objective” division, based only on the shape of the sequence of $z_i$, without predefining the number of classes, then the assignment of labels, i.e. class names, such as “highly developed” etc., would be done a posteriori on the basis of characteristics of the classes obtained, rather than as the result of an arbitrary and largely subjective perspective on how the classes “should” be defined or named.

The present analysis was motivated by exactly such a proposal, forwarded by Nielsen, (Nielsen, 2011), with respect to the indicators of country development levels. Another domain of interest with similar features is the one of distribution of wealth within a society, with the $i$’s corresponding to the successive, somehow defined, wealth classes. The proposal of Nielsen (2011), is analysed here and significantly extended on the basis of the "bi-partial" approach, as exemplified, for instance, in the papers by the present author, Owsiński (1990, 2011, 2012).

## 3. SOME PROPERTIES OF THE PROBLEM AND AN OBJECTIVE FUNCTION

A way to solve the problem as formulated before might consist in minimising the error for subsequent numbers of classes (segments) and finding the "most appropriate" solution in terms of the error value and the number of classes. The weak point of such a procedure consists in the necessity of finding a "proper" trade-off between the differences in error and the count of the number of segments.

Namely, obviously, the error for the optimum approximation decreases monotonically as a function of the number of classes. Thus, as the number of classes increases by one, the (optimum, i.e. minimum) error decreases by a certain number, in general – different for each addition of one class. Hence, it is easy to see that the essential problem persists – namely that of indicating the "most appropriate" number of classes.

It would therefore appear as "natural" to look for a different form of the objective function we try to optimise (minimise), rather than, in very general terms, the previously discussed, "total error + number of classes".

For this purpose we shall introduce some further notation. Thus, let us denote with $q$ the index of the subsequent classes, ranging from 1 to $p$, i.e. $p$ is the overall number of classes distinguished. Denote with $A_q$ the set of indices $i$ of observations $x_i$ (and therefore also $z_i$), classified in class $q$. Note that if, as proposed before, the piece-wise linear approximation is based on straight line segments, defined by values of $z_i$, i.e. the lines always go through two points $z_i$, then the sets $A_q$ may either overlap on one observation (the maximum index in $A_q$ being the same as the minimum index in $A_{q+1}$) or be disjoint (the maximum index in $A_q$ being equal the minimum index in $A_{q+1}$ minus 1[3]). We shall denote by $z^{q\,\min}$ and $z^{q\,\max}$, respectively, the minimum and maximum values of cumulative distribution, corresponding to the set $A_q$. These values, in turn, correspond to the indices $i^{q\,\min}$ and $i^{q\,\max}$, respectively. Additionally, we denote with $q(i)$ the index of the class, to which $i$-th observation belongs. (Note that in the case the classes overlap on "border" observations, $q(i)$ are the sets of two consecutive values.) Let us denote the set of $i$ index values, defining the partition of the sequence $1,\ldots,n$ into the subsets $A_q$, i.e. the sequence composed of the values $1 = i^{1\,\min}, i^{1\,\max}, i^{2\,\min}, i^{2\,\max}, i^{3\,\min}, \ldots, i^{p\,\max} = n$, by $\boldsymbol{iq}$. Of course, by specifying $\boldsymbol{iq}$, for a given sequence $\{z_i\}$, we define $\{A_q\}$ and the entire solution. When referring to the explicit set of subsets $\{A_q\}$ we may also use the notation $P$, for partition of the set of observations.

[3] We shall stick to the second option in what follows. This assumption is, of course, just a choice with no deeper consequences: in the examples further on we deal with both cases.

Given that we assume the piece-wise linear form of the approximation, we can write down the general expression for the $q$-th piece as

$$z^q(i) = a^q i + b^q, \tag{1}$$

where, in fact, we can no longer care whether $i$ is discrete or continuous, but we observe the respective values only for natural $i$. This is an essential assumption for the further calculations. The values of the coefficients $a^q$ and $b^q$ are determined in a natural manner from the standard formulae, where we assume, formally, that each segment is composed of at least two consecutive observations, i.e. $i^{q\max}{>}i^{q\min}$:

$$a^q = \frac{z^{q\max} - z^{q\min}}{i^{q\max} - i^{q\min}}, \tag{2}$$

and

$$b^q = z^{q\min} - \frac{z^{q\max} - z^{q\min}}{i^{q\max} - i^{q\min}} i^{q\min}. \tag{3}$$

Note that after differentiating $z^q(i)$ as in (1) we obtain the increasing sequence of levels $a^q$, corresponding to classes in terms of values of $x_i$.

In view of the convexity of the sequence of $z_i$, the sequence of $a^q$ is non-decreasing (the differences $i^{q\max}$-$i^{q\min}$ are accompanied by appropriately increasing differences $z^{q\max}$-$z^{q\min}$), while the sequence of $b^q$ is non-increasing.

We can now formulate the "minimum approximation error" problem, with the respective objective function, denoted $C_D(\{A_q\})$, as follows:

$$\min\nolimits_{\boldsymbol{iq}}\left(C_D\left(\left\{A_q\right\}\right) = \sum\nolimits_q \sum\nolimits_{i\in A_q} \left(z^q(i) - z_i\right)\right), \tag{4}$$

where minimisation is performed with respect to the sequence $\boldsymbol{iq}$, defining the approximation. We shall denote the optimum sequence, corresponding to the minimum in (4), by $\boldsymbol{iq}^*$.

Since in conditions of convexity there is just one optimum sequence $\boldsymbol{iq}^*$ for each consecutive value of $p$ (quite in line with Nielsen, 2011), we can denote the minimum value of $C_D(\{A_q\})$ for a given $p$ by $C^*_D(p)$. Thus, we have $C^*_D(p) \geq C^*_D(p+1)$. Obviously, equality can only occur, when there exist sequences of $x_i = x_{i+1} = \ldots$, so that corresponding $z_i, z_{i+1}, \ldots$, are indeed situated on a straight line. Otherwise, any increase of $p$ leads to the decrease of $C^*_D(p)$. Actually, one could go, with the above formulation of the problem, to the extreme of $p = n$, when $C^*_D(n) = 0$, an "ideal approximation"! Each observation would then constitute a separate "class" with just one representative. Of course, formulae (2) and (3) would become obsolete (division by zero). (Note, though, that we do not deal here with the proper "approximation", as it is understood in numerical analysis, where problem formulation is altogether different.)

Obviously, when the previously mentioned sequences of $x_i = x_{i+1} = \ldots$, occur, so that the corresponding $z_i, z_{i+1}, \ldots$, are situated on a straight line, $C^*_D(p)$ shall remain at the absolute minimum of 0 also for $p < n$, down to the value, determined by the total length of such uniform sequences.

While construction of approximating segments for such sequences is not a question, the issue that we address here is related to the possibility of finding a way to tell "how different the successive observations have to be in order to assign them to different segments (classes)".

## 4. CONSTRUCTION OF A BI-PARTIAL OBJECTIVE FUNCTION

If the problem were formulated as "minimise the error with as low number of segments as possible", we would face the following minimisation problem:

$$\min(C_D(\{A_q\}) + w(p)), \tag{5}$$

where $w(p)$ expresses the weight we attach to the number of segments in the potential solution. For consecutive values of $p$ the minimum of $C_D(\{A_q\})$ would be found, and then the minimum of the function from (5) determined. Although this procedure might seem altogether cumbersome, but, since we expect not too many segments to correspond to optimum, it would still be numerically feasible. We shall not go into the technical details of such an approach, for reasons given below.

Namely, now, the essence of the problem is transferred to the determination (or knowledge) of the weight function $w(.)$. In some cases, when $p$ has a concrete interpretation, carrying with it, for instance, some cost, while error minimisation leads to definite benefits, like in technical applications, or in operational research, then determination of $w(.)$ is quite feasible, even if still charged with difficulties. This is not, however, the case with our problem, where we look for some possibly "natural" division of the cumulative distribution, and no cost / benefit, except for the facility of use of appropriate linguistic labels ("very highly developed", "highly developed",...), is involved.

As we try to find the "natural" division of the cumulative distribution (that is – provided it exists, and the method we aim at ought to tell us whether it does), therefore, we should refer to some "counterweight", analogous to that of $w(p)$ in (5), but having the same sort of meaning and kind of measurement as $C_D(\{A_q\})$. In this way we might be able to try to define the proper $p$ and at the same time the ***iq***, or, otherwise, the $\{A_q\}= P$.

Thus, similarly as in (5), we would like to add to $C_D(\{A_q\})$ a component that would penalize, in this case, for the division into segments that are in some way "too similar", especially in terms of subsequent $a^q$. In general terms the respective bi-partial objective function and the corresponding problem would look like

$$\min(C_D(\{A_q\}) + C^S(\{A_q\})), \tag{6}$$

where $C^S(\{A_q\})$ is exactly the component corresponding to the similarity between the consecutive segments, based primarily on the differences of consecutive $a^q$. The more detailed form of $C^S(\{A_q\})$ might be constructed as follows:

– first, the value of a kind of difference between the two consecutive segments, $q$-1 and $q$, could be measured, from the point of view of the succeeding segment, $q$, as, for instance,

$$z_{iq\,\min} - a^{q-1} i^{q\,\min} - b^{q-1} \tag{7}$$

expressing the difference between the actual value of $z_i$ at the beginning of the next, $q^{th}$ subset of observations, and the "approximation" of the same, resulting from the previous segment. This difference is always non-negative, due to convexity of $\{z_i\}$, and can be interpreted as a "distance" between the two consecutive segments in the approximation;

– now, as we wish to penalize (minimize) with $C^S(.)$ the *similarity* rather than distance (or difference), in order to convert (7) into similarity, we should subtract it from some kind of upper bound; this upper bound might be constituted by the maximum of a similar difference for a given data set, namely as defined by the biggest difference of tangents along the curve of $z_i$, i.e. between the beginning and the end of the curve; we shall denote the two extreme tangents by $a^{(1)}$ and $a^{(n)}$, and they are defined as:

$$a^{(1)} = z_1/i_1; \quad \text{and} \quad a^{(n)} = (z_n - z_{n-1})/(i_n - i_{n-1});^4 \tag{8}$$

in order, however, to calculate the proper difference, we must have the complete expressions for the lines, corresponding to $a^{(1)}$ and $a^{(n)}$, satisfying the conditions that allow for their use with respect to consecutive subsets $A_q$; we assume, namely, that all four lines involved here, corresponding to the tangents $a^q$, $a^{q-1}$, $a^{(1)}$ and $a^{(n)}$ cross at the point, defined otherwise by the crossing of the lines, corresponding to subsets $A_{q-1}$ and $A_q$, i.e. at the point, denoted $i^{q*}$ in Fig. 2; from this condition we can derive the values of $b$, to be used in conjunction with $a^{(1)}$ and $a^{(n)}$ (denoted, respectively, $b^{*(1)}$ and $b^{*(n)}$) in the appropriate expression, namely:

$$b^{*(1)} = b^q - (a^{(1)} - a^q)(b^{q-1} - b^q)/(a^q - a^{q-1}), \tag{9a}$$

$$b^{*(n)} = b^q - (a^{(n)} - a^q)(b^{q-1} - b^q)/(a^q - a^{q-1}). \tag{9b}$$

Now, the expression for $C^S(.)$ for a single $q$, following the reasoning here presented, can be written down as

$$a^{(n)} i^{q\,\min} + b^{*(n)} - (a^{(1)} i^{q\,\min} + b^{*(1)}) - (z_{iq\,\min} - a^{q-1} i^{q\,\min} - b^{q-1}), \tag{10}$$

where the second term in brackets is equivalent to the difference, given by (7), while the preceding terms define the reference for the given $q$. Now, then, the proposed $C^S(P)$ is the sum over $q$ of (10). Altogether, the minimised objective function we wished to have, takes on the form:

[4] This notation with respect to $i$ is meant to emphasise the generality of formulae; actually, if $i$ are just the natural numbers, starting with $i = 1$, and ending with $i = n$, these formulae get much simpler: $a^{(1)} = z_1$; and $a^{(n)} = z_n - z_{n-1}$.

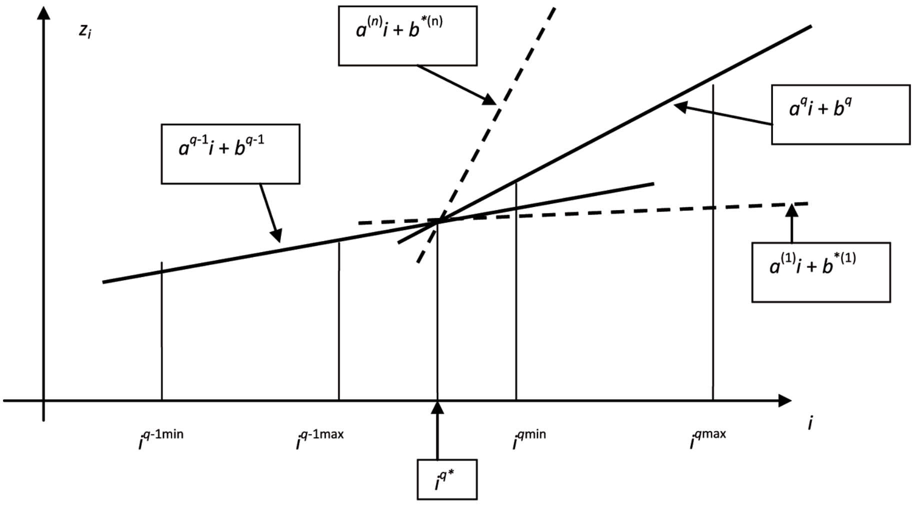

Figure 2. Schematic presentation of the way, in which parameters of $C^S(.)$ are determined

$$C^S_D(\{A_q\}) =$$
$$C_D(\{A_q\}) + C^S(\{A_q\})) =$$
$$\Sigma_q\Sigma_{i\in Aq}(a^q i + b^q - z_i) + \Sigma_q(a^{(n)}i^{q\,\mathrm{min}} + b^{*(n)} - (a^{(1)}i^{q\,\mathrm{min}} + b^{*(1)}) - (z_{iq\,\mathrm{min}} - a^{q-1}i^{q\,\mathrm{min}} - b^{q-1})). \tag{11}$$

in which we take $a^0 = 0$ (which is quite natural), and $b^0 = 0$ (which is somewhat artificial).

Let us note at this point that the objective function defined above implies for the example of Fig. 1 a solution composed of five segments, the divisions into less and more segments, minimizing $C_D(.)$, yielding higher values of $C^S_D(.)$.

At the end of this section let us also indicate that the approach, in which fit or precision of some kind of rendition or representation is counterbalanced by a sort of "cost" incurred by it, e.g. in the form of the number of parameters (like in the polynomial approximation, for instance), is quite common to statistics, especially empirical statistics, taking the concrete forms in such criteria as, say, AIC or BIC. Yet, in these criteria we deal with two quite different 'parts', each one expressing a completely different aspect of the problem at hand, while in the objective function here proposed we try to reflect possibly faithfully the actual duality of the respective problem, and to express the two aspects in the same kind of terms.

## 5. SOME GENERAL PROPERTIES AND POTENTIAL ALGORITHMS

Generally, the construction of the bi-partial objective function follows only the prerequisites of "global' rationality – namely that we oppose two measures that individually represent a one-sided rationality (here, in particular, error minimization), and that together imply some sort of compromise, based on their joint minimization or maximization (see Owsiński, 2011, 2012). In this, we do not enforce neither the concrete structure (value of $p$), nor any weight – although weights can, of course, be applied, and even may be effectively used, as this shall be seen later on. Actually, the methodology based on the bi-partial objective function applies equally to the multidimensional distributions.

In this particular case, we constructed the bi-partial objective function out of two components, one, $C_D(\{A_q\})$, corresponding to the error, resulting from the "approximation" of the sequence of $z_i$ with a limited number of line segments, and the second, $C^S(\{A_q\})$, corresponding to the penalty for the too small change of the angle of the line between two consecutive segments. Although we have not shown this with respect to the second component, the two components display opposite monotonicity along the number of segments, $p$, that is – minimum $C_D(\{A_q\})$, or $C^*_D(p)$, decreases along $p$, while $C^S_D(\{A_q\})$ increases (we refer here only to the sequence of $\boldsymbol{iq}$ minimizing $C_D(\{A_q\})$).

The above remark indicates one of the fundamental principles of construction of the bi-partial objective function, namely the *opposite monotonicity* of the two components.

One might indicate, in this context, that the two components in this example do not quite correspond to each other (are not "symmetric"): there is only one element per segment in $C^S(\{A_q\})$, while there are card$A_q$-2 elements per segment in $C_D(\{A_q\})$, which, definitely, introduces a bias (in this realisation of the bi-partial objective function for the distribution division problem the segments obtained cannot be too big, i.e. card $A_q$ too high). Indeed, we have provided here only an example – actually, the entire formulation of the problem, also involving the "error function", is arbitrary (we could use, e.g. the sum of squares formulation). The primary purpose of the exercise presented was to show the way the bi-partial function can be constructed and how it functions.

Concerning the optimisation algorithms, the off-the-shelf choice for a single-dimensional problem is the dynamic programming approach, similarly as in the classical categorisation problem (see, e.g., Gan, Ma and Wu, 2007), or that proposed by Thierry Gafner (1991). For the example at hand, though, we can also consider the approach by the present author, which is closely associated with the idea and the properties of the bi-partial objective function. We shall provide here only the basic precepts of this approach.

Actually, we shall illustrate the approach with the concrete form of $C^S_D(\{A_q\})$, considered here. Thus, assume we consider, instead of (11), a parameterised form:

$$C^S_D(\{A_q\}, r) = (1 - r)C_D(\{A_q\}) + rC^S(\{A_q\})) \qquad (12)$$

with $r \in [0,1]$, and we look for the minimum of $C^S_D(\{A_q\},r)$ over $\boldsymbol{iq}$, i.e. $C^{S*}_D(r)$.

Then, assume we start the procedure from $r = 0$. Thus, we have

$$C^S_D(\{A_q\}, 0) = C_D(\{A_q\}), \qquad (13)$$

and, of course, the "optimum" partition for this situation is the one with $p = n$, or, at most (according to our form of the "error function"), $p = \text{int}[n/2]+1$, where int[$v$] is the highest integer number lower than $v$ (this fact resulting from the zeroing of elements of $C_D(\{A_q\})$ at the endpoints of each segment). Yet, we are not, in general, interested in such a trivial solution. As we increase $r$ from 0, non-zero weight starts to be assigned to $C^S(\{A_q\})$, and in order to obtain $C_D^{S*}(r)$ for such $r$, it will "pay" at some definite value of $r$, say $r^1$, to join two segments, for which the difference of angle is the smallest, and hence the penalty in $C^S(\{A_q\})$ is the biggest. This value of $r^1$ can indeed be easily determined on the basis of the formulae here provided.

As we increase the parameter $r$ further, we find its next value, $r^2$, for which merging of another pair of consecutive segments "pays" in terms of $C_D^S(\{A_q\},r)$. And so on. The proper solution is found for the last $r^t$ that is not bigger than $^1/_2$ – i.e. for the equal weights of the two components of $C_D^S(\{A_q\},r)$. This, of course, is a sub-optimisation algorithm, as it does not guarantee reaching of the proper minimum of $C_D^S(\{A_q\})$, since the sole operations involved are the aggregations of segments. Yet, experience shows that it either actually reaches the minimum, or is very close to it.

No matter which method is used (the dynamic programming or the one proposed by the author), the basic rationale consists in forming the segments, $A_q$, composed of sequences of possibly similar $x_i$, i.e. $z_i$ forming a curve possibly close to a straight line. An adequately pronounced "jump" over one or several consecutive $i$'s would then correspond to a change from $q$ to $q$+1 in the optimum solution.

## 6. PROBLEMS WITH REAL-LIFE EMPIRICAL DISTRIBUTIONS

The here outlined treatment of the problem of dividing the empirical distributions, though, encounters in many instances an essential hindrance in the form of the actual form of such empirical distributions. The essential question relates to two, often closely interconnected issues:

(a) a very regular – or very close to very regular – shape of the empirical distribution; examples of such situations are provided in Figs. 3 and 4, showing, first, the annual value of own revenue per capita of the more than 300 municipalities of the Polish province of Masovia, based on the data from the Central Statistical Office (GUS), ordered in decreasing sequence; although the data are just for one year, and not for a very big "sample", the regularity is, indeed, striking; then, Fig. 4 shows – for the same set of municipalities – the shares of population of definite age segment, and, again, the regularity defies any attempt of "natural" division of the distribution; these figures show also the second of the issues that ought to be considered, namely

(b) the shape of the empirical distribution that runs "counter" to the intuitive precepts we have outlined before, i.e. that a "category" ought to be formed by the values of $x_i$ that are possibly similar; the difficulty in the case of Fig. 3 is with the upper end of the distribution – how many categories one would (intuitively) assign there? The really striking cases of distributions where similar problems appear at their both ends were considered in Owsiński (2012) on the basis of QOL (2005) and QOL (2007).

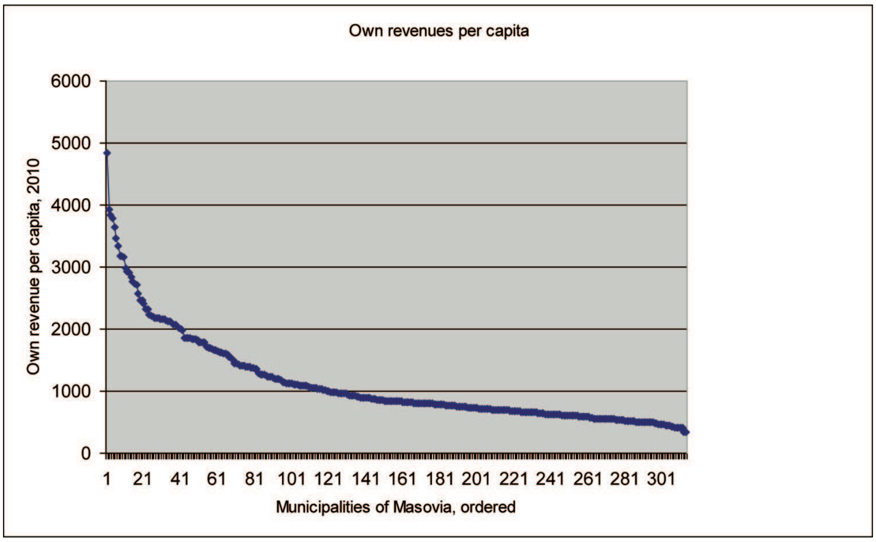

Figure 3. An example of an empirical distribution: own revenues per capita of the municipalities of Masovia in Poland in 2010 (in Polish zlotys per annum)

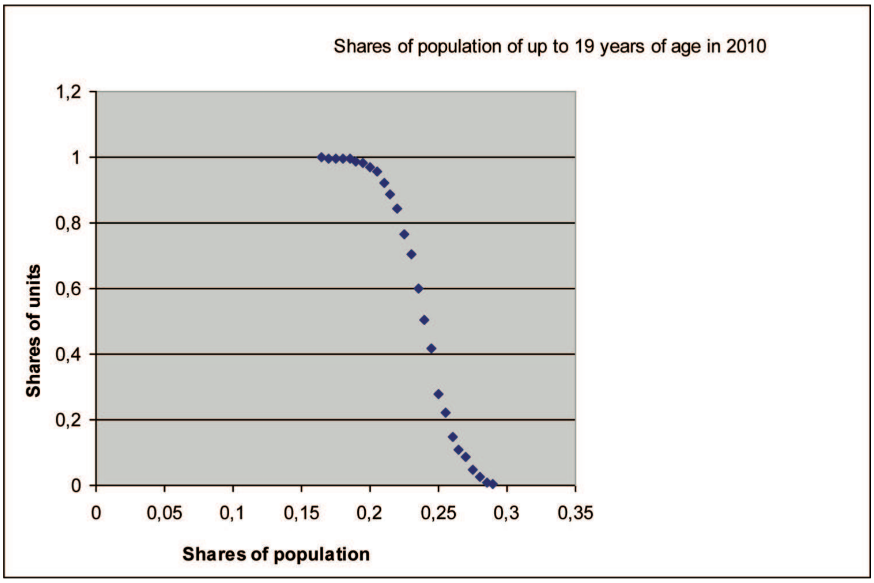

Figure 4. An example of an empirical distribution: shares of municipalities of Masovia in Poland in 2010 with definite shares of population below 19 years of age

## 7. WHAT CAN BE DONE TO "SPITEFUL" DISTRIBUTIONS?

With respect to such regular (and yet – characteristically shaped) distributions the following questions appear to be justified:

- for a (very) regular distribution, which could relatively easily be approximated by a certain functional shape (notwithstanding its potentially established "theoretical-statistical" foundations) is any division at all, of statistical or any other character, reasonable?
- if, however, a methodology, like the one outlined here, were applied to such a regular distribution, what should be the interpretation of the results obtained?

The first of these questions could be answered through the following procedure:

i. definitely, it must be established what functional form, and, if possible, what theoretical statistical distribution fits the given empirical distribution with a sufficient degree of exactness and/or significance;

ii. this ought to be analysed against the background of the substantive knowledge of the processes behind the empirical distribution; it is feasible that in some cases an explanation for the shape obtained could be provided, or at least an interpretation of the values, represented in the distribution, and that possibly in intuitive terms;

iii. unless the functional and/or theoretical fitted shape provides an obvious basis for a division, there is, indeed no well-founded "computational" ground for performing it;

iv. the sole basis for any potential division remains, therefore, with the substantive analysis of the respective processes; in many cases, though, given the frequent purposes of such an analysis, the risk appears of a political influence on the results (like in case of classification of regions, countries, or setting of poverty lines); in any case, though, point **ii** above preserves its validity.

Now, regarding the second question, it might also be answered through a procedure like the one outlined above, in its variant, consisting of the following steps:

i'. assume it is established what functional form, and, if possible, what theoretical statistical distribution fits the given empirical distribution with a sufficient degree of exactness and/or significance;

ii'. this form ought to be analysed in the perspective of the substantive knowledge of the processes behind the empirical distribution; possibly, in some cases an explanation for the shape obtained could be provided, or at least an interpretation of the values, represented in the distribution, and that possibly in intuitive terms;

iii'. division is performed with the methodology outlined, based on the functional form, identified under step i', meaning that the *approximation is carried out not with respect to the piece-wise linear function, like in the illustrative case here considered, but with respect to the concrete form identified, with its parameters shifting between the segments of the division*;[5]

[5] The very first step might consist in approximation with the simplest quadratic polynomial segments, forming a continuous approximating function. Thereby, some of the segments – representing, e.g. the middle part of the distribution, could be reduced to linear ones.

iv'. the segments, forming the division obtained, ought to be analysed against the background of substantive knowledge, in analogy to (or: in the framework of) the analysis, carried out in point ii' above; with this, it ought to be remembered that the segments obtained are founded directly on the shape of distribution.

In the case of distribution, shown in Fig. 3, the results obtained indicated the optimum number of segments, $p$, as equal between 8 and 15, with relatively small differences of value of $C_D^S(.)$ for these $p$. This confirms the hesitation, concerning such forms of distributions, as concerns application of the dividing approaches. Yet, analysis of the segments obtained for this series of similarly good solutions, ought to be quite educative also in deeper substantive sense. In particular, one might try to assess the "quality" of the empirical distribution against both the substantive precepts and the degree of fitting to the functional shape identified (like in the case of the QoL rankings, analysed in Owsiński, 2012).

*Instytut Badań Systemowych PAN*

## LITERATURA

[1] Gafner Th., (1991), *Mathematical programming approach to classification*, Ph. D. dissertation, Institute of Statistics, Faculty of Economics and Business, University of Neuchatel.

[2] Gan G., Ma Ch., Wu J., (2007), *Data Clustering*, Theory, Algorithms and Applications, SIAM & ASA, Philadelphia.

[3] QOL, (2005), http://www.economist.com/media/pdf/QUALITY_OF_LIFE.PDF, The Economist Intelligence Unit's quality-of-life index (as seen on September 25th, 2012).

[4] QOL, (2007), http://www.il-ireland.com/il/qofl07/2007 Quality of Life Index (as seen on September 25th, 2012).

[5] Nielsen L., (2011), *Classification of Countries Based on Their Level of Development: How it is Done and How it Could be Done*, IMF Working Paper, WP/11/31, IMF.

[6] Owsiński J.W., (1990), *On a new naturally indexed quick clustering method with a global objective function*, Applied Stochastic Models and Data Analysis, 6, 157-171.

[7] Owsiński J.W., (2011), *The bi-partial approach in clustering and ordering: the model and the algorithms*, Statistica & Applicazioni, Special Issue, 43-59.

[8] Owsiński J.W., (2012), *On dividing an empirical distribution into optimal segments*, SIS (Italian Statistical Society) Scientific Meeting, Rome, June 2012, http://meetings.sis-statistica.org/index.php/sm/sm2012/paper/viewFile/2368/229

## OPTYMALNY PODZIAŁ ROZKŁADU EMPIRYCZNEGO
## (I KILKA PROBLEMÓW Z TYM ZWIĄZANYCH)

### Streszczenie

Praca zajmuje się podziałem empirycznego rozkładu wielkości $x_i$, gdzie $i$ jest indeksem jednostki, dla której obserwujemy tę wielkość (np. $x_i$ to PKB na mieszkańca w kraju $i$-tym). Wartości $x_i$ uporządkowano niemalejąco. Analizujemy dystrybuantę rozkładu, tj. wartości $z_i = \Sigma_{i'=1,...,i}x_i$, które tworzą ciąg wypukły. Chcemy otrzymać taki podział dystrybuanty na podzbiory, by przybliżyć kształt rozkładu $\{z_i\}$ z możliwie małym błędem przy pomocy odcinków linii prostej, odpowiadających podzbiorom, a zarazem – by tych odcinków było możliwie mało. Odpowiada to kategoryzacji podobnych rozkładów (np. kraje „rozwinięte", „rozwijające się", ...), gdzie zwykle nie stosuje się metod statystycznych, tylko przesłanki „merytoryczne", bądź stosowanie metod statystycznych ogranicza się do ustalenia, np., kwantyli rozkładu, bez uwzględniania kształtu i innych przesłanek dla rozwiązania, optymalizującego wspomniane kryterium. Zaproponowano ogólną metodykę optymalizacji podziału takich rozkładów w duchu wspomnianego kryterium, funkcję celu i jej konkretną realizację, wraz z algorytmami. Na podstawie przykładów konkretnych rozkładów, zarysowano także problemy, wynikające z faktu, że rozkłady empiryczne mają często charakter, stawiający pod znakiem zapytania podstawy przyjętej metodyki i w ogóle sens podobnych zadań. Przeanalizowano możliwe pochodzenie tych rozkładów oraz skutki dla ewentualnej kategoryzacji. Zaproponowana metodyka daje podstawy do kategoryzacji empirycznych dystrybuant i narzędzie do oceny racjonalności sposobu ich otrzymywania.

**Słowa kluczowe**: rozkład empiryczny, kategoryzacja, optymalizacja, funkcja kryterium, dwustronna funkcja kryterium

## ON THE OPTIMAL DIVISION OF AN EMPIRICAL DISTRIBUTION (AND SOME RELATED PROBLEMS)

### Abstract

We consider division of an empirical distribution of $x_i$, $i$ being the index of a unit, for which we observe $x_i$ (e.g., province $i$, for which $x_i$ is the GDP per capita). Values $x_i$ are ordered non-decreasingly. We analyse the cumulative distribution, $z_i = \Sigma_{i'=1,...,i} x_i$. The sequence $z_i$ is convex. We want to divide the distribution of $z_i$ into subsets of $i$, with the shape of the distribution $\{z_i\}$ possibly well approximated by the segments of the straight line, determined for the subsets, forming a piecewise linear contour, the number of segments being possibly small. This corresponds to the frequently used categorisations for similar distributions (e.g., "developed", "developing",... countries). For such categorisations, usually no formal methods are applied but "substantive" prerequisites, or the methods applied are limited to establishing quantiles of the distribution, without considering its shape and the objective premises for determination of a different number of segments, including optimisation of the criterion mentioned before. A general approach is proposed for optimising division of such distribution conform to the criterion mentioned. A general objective function is proposed and its concrete realisation, as well as algorithms. The methodology proposed allows for obtaining the optimum divisions into categories for arbitrary distributions. Yet, on the basis of concrete empirical distributions, problems are outlined, due to the fact that the distributions obtained often display the features, leading to questioning of the foundations of the methodology proposed, and of the very sense of such categorisations. Examples of distributions of this kind, and consequences for the potential categorisations, are discussed. In summary, the methodology proposed, including the criterion function, constitutes a basis for the categorisation with respect to the cumulative distribution, and a tool for evaluating the rationality of the way, in which the distributions are obtained.

**Key words**: empirical distribution, optimum division, classes, objective function, bi-partial approach