Finding the Odd-man-out in Development Scoring and Classifications

Jan W. Owsiński¹, Artur Wieclaw²

Abstract

The paper presents a simple application, meant mainly for classifying and ranking of the administrative units, which is being developed so as to constitute a module in a broader system, under development within the Polish project ANAGMIS. The project aims at creation of an evaluation-oriented model for assessing the quality of development of the administrative units, with emphasis on sustainability of development. The application presented in the paper is devoted to joint ranking of units for more than one variable, and to identification of common (aligned) sequences of orders and of the non-aligned ones. Its important aspect is visualisation of the results, as well as orientation at relatively large sets of data (up to several thousand and beyond).

1. Introduction: the setting for the work reported

The paper presents a simple application for classifying and ranking administrative units, which is being developed so as to constitute a module in a broader system, under development within the Polish project called ANAGMIS. The project aims at creation of an evaluation-oriented model for assessing the quality of development of the administrative units, with special emphasis on sustainability of development. Resulting from the application of the ultimate model, coupled with appropriate ordering techniques, there should be a ranking of the administrative units, based on some – hopefully simple – scoring system. Both the application itself and the results, in terms of scoring, will be available on the net to the subscribing administrative units, with adequate provisions for interactive use (including corrections of erroneous data, proposals concerning the content and presentation, as well as FAQs).

The model proposed presupposes a passage from the complete substitutability of the “development factors”, such as, for instance, broadly conceived “economy”, “society” and “ecology”, when these remain in a certain “balance”, towards (requisite) complementarity, when farther away from such a balance. This kind of approach must, of course, account for the diversity of the development paths or, at least, initial conditions. Thus, one of the tasks within the project consists in the identification of the “development types” or “paths”.

It is namely the prerequisite of the model that the units, which feature diversity of values attached to the respective aggregate categories (like the three mentioned above) within certain limits, such that substitutability between them holds, can be classified according to the “generic conditions”, equivalent to the “initial points” or “development types”. Within each type a proper Pareto front can, in principle, be empirically identified (similarly to the DEA methodology), along with the units that fall behind in terms of “quality of development”. The distinction into types is, of course, related to both the formal status of the administrative units (see below) and their actual resource and development situation. It is further assumed that within the particular types of units, once the limits of substitutability are exceeded (the values of particular variable aggregates for a unit being beyond the limits defined for the individual types), then it is expected that the quality of development may fall, unless accompanied by appropriate shift in other variable values (which may take the form of proper complementarity, though not necessarily so).

¹ Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447 Warszawa, Poland, owsiinski@ibspan.waw.pl (the work reported by this author has been partly supported within the project NT1400/31/1404, ANAGMIS, financed by the Ministry of Science and Higher Education of the Republic of Poland)
² student, Warsaw School of Information Technology, wieclaw@wit.edu.pl
The project accounts for all the basic administrative self-governmental units in Poland (gminas – communes), altogether some 2500 of them (classified into some 350 counties – powiats, and these, in turn, into 16 provinces – voivodships), for which some 20 variables and the indicators based on them are used in the analysis. Communes are classified into three formal categories: urban, rural (some 1500 of those) and rural-urban. This formal classification is not sufficient, though, to warrant appropriate distinction of the development situations of particular communes. Therefrom the task mentioned before.

Within the so defined domain of research endeavour the paper presents a simple technique, mainly oriented at visualisation of orderings and at visually identifiable outliers (“odd-men-out”). It is in view of the ultimate use of the methods and techniques developed within the project ANAGMIS that emphasis is placed on the visualisation aspect. Given also the number of objects, the respective techniques have to account for the number of objects of the order of $10^5 - 10^4$. This implies a specific approach to visualisation (e.g. showing of the shape of the curves of ordered values and comparison of these shapes).

Thus, in particular, the analysis of multicriteria ordering is performed with visualisation of the partial and consecutive results of the procedure. So, when looking for the multicriteria ordering of objects (whether in toto or in particular formal categories) and the respective outliers, visualisation allows for an on-hand inspection, both of the nature of orderings involved and their aggregates, and the resulting distortions.

2. The modalities and purposes of use

Within the outlined framework it is assumed that a user is primarily interested in the subject matter, and much less in the intricacies of mathematical rigour of the potentially applied methods. Thus, in view of the orientation at the specific methodology of assessing the degree of sustainability of development of the administrative units, which is based on a concept derived from the utility / production function domain, a user is supposed to be looking at the pairs of variables, which remain in a definite relation. This relation may either be a substitutability relation and a complementarity relation, depending upon the perspective taken.

A good example for the complementarity case is the one of the water supply vs. sewage and wastewater treatment indicators. If an administrative unit has a high share of population served by water supply system, this being a reflection of common provision of service, i.e. of somehow understood “welfare” and the associated expenditures, then it is interesting to look at the sewage and wastewater treatment side. A similarly symmetric, though less technical is the case of car ownership indicator vs. the existing road infrastructure and the current expenditures into it.

Two important characteristics of these two examples can easily be seen: (1) one expects a relatively high correlation of the two variables forming such a pair; (2) the two variables often belong to – at least superficially understood – “different domains”, e.g. the one of life quality, or (urban) civilisation, or economic development, on the one hand, and environmental quality, nature protection, and – beware! – sustainable use of natural resources, on the other hand.

This is, indeed, the area in which the application here presented is meant to serve. It is, namely, intended to be used to:

- order objects according to a (“primary” or “lead”) variable $x_\lambda$, one of the variables describing objects $x_i = \{x_{i1}, \ldots, x_{ik}, \ldots, x_{im}\}$,
- visualise the ordering through the values of the variable (“the primary order shape”),
- order another variable, $x_\nu$, and show the respective values against the background of the “primary order shape”
- parameterise the variables, i.e. form the variable $x_r(k, k') = rx_\nu + (1-r)x_\lambda$, where $r \in [0,1]$, and order objects according to the parameterised variable for the values of $r$ starting from $r=1$ down to $r=0$,
- visualise the segments of the orderings, which are analogously ordered according to $x_\lambda$ and $x_r(k, k')$, and which are not analogously ordered according to these variables.
The idea behind is outright simple: it is assumed that the two selected variables ought to have “something in common”, as expressed through the relation of complementarity, in particular. There should exist a high degree of correlation between these two variables, and so the respective orderings should also be rank-correlated. The parameterisation, whether starting with the value of \( r=1 \) or with \( r=0 \), will show, with the help of the application, the segments of the orderings, which are “parallel”, i.e., the sequences of objects identically ordered according to the two variables, and/or the segments which are not parallel. All this is, of course, sensible only for the variables of which we expect that they behave in a definite, correlated manner.

Of special interest are these segments, which remain not parallel for the extreme values of the parameter \( r \), that is – the “last” non-aligned segments of the orderings. These are exactly the “odd-men-out” with respect to the two variables. A simple procedure would consist in identification of the arg(min, \( r^{kk'}(i) \)), where \( r^{kk'}(i) \) is the value of the parameter, for which the object \( i \) starts to be included in a sequence aligned with the one for the variable \( k' \) when starting from the ordering for the variable \( k \) (note that for high rank correlation coefficient values between the orderings for \( k \) and \( k' \) there may exist a high share of \( i \), for which \( r^{kk'}(i) = 1 \)). Along with arg(min, \( r^{kk'}(i) \)) a sequence (ordering) of objects with appropriately low values of \( r^{kk'}(i) \) would be identified. This operation would be performed for a number of pairs \((k,k')\) of interest to an analyst, and the objects consistently appearing in the identified “potential-odd-men-out” sequences would constitute the true-to-life “odd-men-out”. As additional criteria the rank correlation coefficient and the sum of variable value differences along the two orderings are calculated at each step of parameterisation.

3. An illustration: the working of the application

The set of variables considered within the ANAGMIS project encompasses some 20 items for all the municipalities of Poland, for the consecutive years. Even though the official statistics for this level of resolution are limited in terms of sectors (while being very ample in terms of sheer numbers, especially in the domain of demography), they allow for quite sufficient treatment of the development issue.

The application here presented allows for a versatile visualisation-oriented presentation of the orderings according to two and three variables. In particular, it allows for the realisation of the procedure outlined at the end of the preceding section.

A user can select
- the set of variables analysed,
- the set of objects analysed,
- the directions of ordering,
- the pairs (or triples) of the variables, for which the orderings are to be shown,
- the “lead” variable in a pair or in a triple,
- the mode of visualisation (simple ordering, aligned sub-sequences, non-aligned sub-sequences, ...),

and then, once the visualisation already shown,
- the value of the parameter \( r \) for the next step,
- the range of objects to be shown (“zooming in”), so that the particular sub-sequences may be read out as to their content.

An illustration for the functioning of the application is provided by the two diagrams below. The first of them shows the image of joint ordering according to two variables (respective variable values). A user always selects one “lead variable”, the second one being in a way a background for it. In the illustration below it is interesting to note the generally similar character of the two curves, notwithstanding the “wiggles” of the background variable.
The second illustration shows the snapshot of the mode, in which common segments of orderings and the non-common ones made explicit. The illustration shows the blown-up fragment of the total ordering of the 2500 units, for the subsequence of some 50 units, selected by the user.

Because of the “magnification” the names of the units can appear on the screen, and so the “aligned” and “non-aligned” sequences can be pinpointed.
Altogether, an effective application was developed, with the intention of interactive use within a broader system. The idea behind is that the “odd-men-out” would be subject to a more detailed analysis for the causes of their specific “behaviour”. On this, though, a knowledgeable user should make appropriate decisions.

Bibliography


