DP2: PENA DISTANCE METHODOLOGY

The following text is a revision of Montero et al. (2013), section 2.3.

Pena Distance (DP2) is an iterative method that assigns weights to partial indicators based on their correlation with a global index, which is the quantification of the **latent variable** we want to measure (Pena 1977). Its main strength lies in leveraging all the valuable information contained in the indicators while eliminating redundant variance (i.e., avoiding multicollinearity). Traditionally, this technique has been applied to measure quality of life and other social indicators, but it is equally suitable for fields such as environmental assessment thanks to its strong statistical properties: **multidimensionality, comparability, and interpretability**.

First, DP2 is a multidimensional indicator, capable of aggregating diverse variables expressed in different measurement units. Second, it is a quantitative distance measure, enabling comparisons across spatial and/or temporal units by referencing an ideal or baseline state. Finally, DP2 is an exhaustive indicator that does not rely on reducing information (like Principal Components, for instance). Instead, it captures all meaningful statistical content—excluding false or duplicate variance—allowing interpretation on ordinal or, preferably, cardinal scales. This feature supports the inclusion of numerous variables, as redundant variance is automatically removed during computation, thus preventing multicollinearity. Following Ivanovic (1974), the more relevant variables included, the more comprehensive the synthetic index becomes, since each variable contributes unique information not found in others. DP2 effectively filters out superfluous common variance, retaining only original, non-redundant information.

These properties make DP2 particularly useful for integrating multiple dimensions, such as combining air and noise pollution indicators with subjective measures into a single synthetic index. Even when these data differ in units or contain overlapping information, DP2 converts them into comparable abstract units, focusing exclusively on useful variance and discarding the rest.

DP2 involves several iterations or matrix rearrangements. The **point of departure** of the whole process is a matrix V of order (K,m), in which m is the number of observations (e.g. census tracts) and K is the number of partial indicators. Each element of this matrix, v_{kj}, represents the state of the partial indicator k in the observation j. In this matrix, those partial indicators negatively connected with the latent variable must undergo a change of sign (i.e. all their data must be multiplied by –1). Conversely, the variables that are positively linked with the latent variable remain unchanged. As a result, an increase/decrease in the values of any partial indicator will correspond to an improvement/worsening of the latent variable content.

In a second stage, we compute a distance matrix D such that each element, d_{kj}, for each observation j, is defined as:

$$d_{kj} = \left| v_{kj} - v^*_{kj} \right| \qquad (1)$$

where v^*_{kj} is the k^{th} component of the reference base vector $v^*_j = \left\{ v^*_{1j} \quad v^*_{2j} \quad \ldots \quad v^*_{Kj} \right\}$ in the observation j. A reference value must be defined for each partial indicator so as to compare different observations in terms of the global index of the latent variable. It is quite common to consider the minimum value as the reference (e.g., Sánchez et al. 2022, Chasco 2024). As a result, a higher value in DP2 (which will always take positive values) will imply a higher level of the global index, since it implies a greater distance with respect to a theoretical 'undesired' situation. In addition, this property allows the observations to be ranked in terms of the global index. Therefore, d_{kj} measures the distance between the partial indicator k in the observation j and its reference value.

In a third stage, in order to express all the indicators in comparable abstract units, we compute a first global index, the **Frechet Distance** (DF), which is defined as:

$$DF(j) = \sum_{k=1}^{K} \frac{d_{kj}}{\sigma_k} = \sum_{k=1}^{K} \frac{\left| v_{kj} - v_{kj}^* \right|}{\sigma_k} \quad ; \quad j = 1,2,...,m \tag{2}$$

where σ_k is the standard deviation of partial indicator *k*. For each partial indicator, the distance between two observations d_{kj} is weighted by the inverse of σ_k. That is to say, the contribution of each d_{kj} to the global indicator is inversely proportional to the standard deviation of its corresponding indicator. This weighting scheme, which is similar to those used in heteroskedastic models, accords less importance to those distances with more variability, and vice versa.

DF is a valid concept of distance only in a theoretical situation of uncorrelated indicators. When there is a direct relationship between the partial indicators (as is usual), DF will include some duplicated information. Therefore, DF must be corrected so as to eliminate this dependence effect (i.e. the redundant information existent in other variables), which is assumed to be linear. This is why, for each observation *j*, DF is the maximum value that DP2 can reach, which is defined as:

$$DP2(j) = \sum_{k=1}^{K} \frac{d_{kj}}{\sigma_k} \left(1 - R^2_{k \cdot k-1,k-2,...,1}\right) \quad ; \quad j = 1,2,...,m \tag{3}$$

where $R^2_{k \cdot k-1,k-2,...,1}$ is the determination coefficient of the regression of each partial indicator *k* on the others (*k*–1, *k*–2…,1). It expresses the proportion of the variance of *k* that is linearly explained by the remainder of the partial indicators.[1] As a result, the correction factor $(1 - R^2_{k \cdot k-1,k-2,...,1})$ deducts the proportion of the variation of the observed values that is explained by the linear dependence. Notice that R^2 is an abstract concept unrelated to the measurement units of the indicators.

DP2 implies a decision about the order of entry of the partial indicators into the computation process. That is to say, it must be decided which partial indicator *k* comes first in contributing its variance to the global index, which comes second, etc. In this process, the first indicator (*k*=1) will contribute all its information to the global index (d_1/σ_1). However, the second indicator (*k*=2) will only add that part of its variance that is not correlated with the first indicator: $(d_2/\sigma_2)(1 - R^2_{2 \cdot 1})$. Similarly, the third indicator will contribute to DP2 the part of its variance that is not correlated with either the first or the second indicators: $(d_3/\sigma_3)(1 - R^2_{3 \cdot 2,1})$ and so forth.

Obviously, DP2 will adopt different values depending on the decision. Thus, it is important to find an **objective hierarchical method** that leads to a unique entrance order of the partial indicators. If DF is a compendium of all the partial indicators, it seems logical to make the selection considering the correlation between each partial indicator and DF. The indicator most closely correlated to DF will be the leader given that it is the most informative, i.e. the indicator that contributes most variance to the global index.

The whole process is a four-step procedure that can be summarized as follows:

- First, we compute the DF values for each observation using expression (2); i.e. considering the reference base vector v^* of minimum values.
- Second, we calculate the correlation coefficients of the partial indicators and DF to order the former in accordance with their degree of dependence on the latter.
- Third, we compute DP2 (expression 3) considering the previously determined entrance order of the partial indicators. This first global index is called DP-1.
- Fourth, we make a new ranking with the partial indicators in accordance with their correlation degree with DP-1 with the aim of re-computing DP2. We call this second global index DP-2.
- We repeat this iterative process until a convergence is reached, i.e. the difference between two contiguous DP indexes is zero. In the case of non-convergent DP values, we can choose the first DP index (or even the average of the final two).

[1] If all the partial indicators are uncorrelated, R^2=0 and DP2=DF.

The numerical value of the DP index has no real meaning, but it is useful for comparing the state of different observations in terms of the latent variable. We can rank observations according to this criterion. If we use the same variables and method, we can compare our results for a place (e.g., Madrid) with those obtained for other places or even at other points of time. DP2 can be used to compare changes in relative positions and even to detect their causes.

REFERENCES

Chasco, C (2024). Objective Index of Quality of Life for Urban Areas. In: Maggino, F. (eds) *Encyclopedia of Quality of Life and Well-Being Research*. Springer, Cham, pp. 4779-4783.

Ivanovic B (1974) Comment établir une liste des indicateurs de development *Revue de Statistique. Apliquée* 22(2): 37-50

Montero JM, C Chasco and B Larraz (2010) Building an environmental quality index for a big city: a spatial interpolation approach combined with a distance indicator, *Journal of Geographical Systems* 12-4, 435-459. https://link.springer.com/article/10.1007/s10109-010-0108-6

Pena JB (1977) *Problemas de la medición del bienestar y conceptos afines (Una aplicación al caso español)*. Presidencia del Gobierno, Instituto Nacional de Estadística, Madrid

Sánchez B, J Velázquez, I Gómez, E Sánchez, F Herráez and C Chasco (2022) A well-being index for housing in the central districts of Madrid, *Journal of Urban Planning and Development*, 148(2). https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29UP.1943-5444.0000793