EMPLOYING SIMILARITY MEASURES TO EXAMINE THE DEVELOPMENT OF TECHNICAL INFRASTRUCTURE IN POLISH COUNTIES

Kesra Nermend, Ph.D.

*Faculty of Economics and Management
University of Szczecin
Mickiewicza 64, 71-101 Szczecin
kesra@szafir.univ.szczecin.pl*

Received 25 February 2008, Accepted 17 April 2008

Abstract

The article presents a comparison made with reference to three similarity measures and Minkowski measures, namely urban measure, the Euclidean measure and Chebyshev measure, as well as the variants of the Euclidean measure and square of the Euclidean measure. Their properties were compared and synthetic measures were created on this basis. Infrastructural-technical variables referring to Polish counties were used as data in the creation of measures.

Keywords: similarity measures, square of the Euclidean measure.

JEL classification: C10, C19, O18.

1. Analysis of Theoretical Properties of Distance Measures

There are many different similarity measures, which can be used for creating synthetic measures. The choice of a specific measure often affects the position of analyzed objects in the formed ranking, therefore the selection of proper distance measures is a very important task. The aim of the article is the analysis of theoretical properties of Minkowski's similarity measure and the square of the Euclidean measure. There are also discussed the practical properties of the mentioned measures on the basis of created synthetic measures on the example of technical infrastructure in the Polish counties.

Similarity measures, which are variants of Minkowski distances, are very often used to group data and create synthetic measures. Minkowski distance is expressed in the following formula¹:

$$m_M(i, k) = \sqrt[p]{\sum_{j=1}^m |z_{ij} - z_{kj}|^p}, \quad (1)$$

where:

$m_M(i, k)$ – value of Minkowski measure between object i and object k .

$p \geq 1$ – parameter determining the type of distance;

z_{ij} – variable j determining object i ;

m – number of variables.

Invariability with regard to a shift is the most prominent feature of all Minkowski distances. Let us shift an object i and an object k by the same value Δ_j along each coordinate:

$$m_M(i, k) = \sqrt[p]{\sum_{j=1}^m |(z_{ij} + \Delta_j) - (z_{kj} + \Delta_j)|^p} = \sqrt[p]{\sum_{j=1}^m |z_{ij} + \Delta_j - z_{kj} - \Delta_j|^p} = \sqrt[p]{\sum_{j=1}^m |z_{ij} - z_{kj}|^p}. \quad (2)$$

It can be noticed that the values of the shift have canceled each other out and do not affect the value of measure. Hence, the value of measure does not depend on the position of objects in space with regard to chosen system of coordinates but only on their positions relative to one another.

Let us multiply values of variables of the object i and the object k by the same s value determining the scale of objects:

$$m_M(i, k) = \sqrt[p]{\sum_{j=1}^m |sz_{ij} - sz_{kj}|^p} = \sqrt[p]{\sum_{j=1}^m |s|^p |z_{ij} - z_{kj}|^p} = |s| \sqrt[p]{\sum_{j=1}^m |z_{ij} - z_{kj}|^p}. \quad (3)$$

Hence, if the scale of the compared objects increases s times, the similarity measure will increase s times as well.

Urban measure of similarity, which is obtained if $p = 1$, is the simplest similarity measure among Minkowski distance measures²:

$$m_m(i, k) = \sum_{j=1}^m |z_{ij} - z_{kj}|, \quad (4)$$

where:

$m_m(i, k)$ – value of urban measure between object i and object k .

Figure 1 shows a connection between the urban measure of similarity and the distance from the pattern. Lines have been drawn for a steady increase in the comparison coefficient.

The urban measure is characterized by reduced sensitivity to distinctive difference in the value of one variable. This can be easily noticed thanks to lines of identical values of measure. In a two-dimensional space, these lines are squares turned by 45 degrees (from the horizontal position). As far as the Euclidean space is concerned, the largest distance between the center of the square and points on its edges are found at its angles. However, here the distance is the same everywhere, which results in slighter sensitivity to objects located at the angles of this square.

Fig. 1. Values of urban measure of similarity depending on the position against the pattern lying in the point P(5,5)

Source: own compilation.

The Euclidean measure (for which $p = 2$) is one of the most commonly employed Minkowski distances, namely³:

$$m_e(i, k) = \sqrt{\sum_{j=1}^m |z_{ij} - z_{kj}|^2}, \quad (5)$$

where:

$m_e(i, k)$ – value of the Euclidean measure between object i and object k .

The Euclidean measure is in accordance with the measure of distance in the surrounding space, hence results obtained with the use of this measure can be interpreted very easily. Lines of identical values of measures in a two-dimensional space form circles (Figure 2).

Chebyshev measure (for which $p \rightarrow \infty$) is an extreme case of Minkowski measures, namely⁴:

$$m_c(i, k) = \max_j |z_{ij} - z_{kj}|, \quad (6)$$

where:

$m_c(i, k)$ – value of Chebyshev measure between object i and object k .

Fig. 2. Values of the Euclidean measures of similarity depending on the position against the pattern lying in the point P(5,5)

Source: own compilation.

Chebyshev measure is so to speak the opposite of the urban measure. The greatest differences between variables determine the value of Chebyshev measure (Figure 3). If two objects differ in one value of variable a lot, this value will determine the value of the measure discussed.

Fig. 3. Values of Chebyshev measure of similarity depending on the position against the pattern lying in the point $P(5,5)$

Source: own compilation.

Sensitivity to the value of a maximum difference between variables can be regulated through changing coefficient p . The slightest sensitivity is observed when $p = 1$, whereas the greatest one is found when $p \rightarrow \infty$. Coefficient p can be regulated in the bracket $\langle 1; \infty \rangle$ via selecting sensitivity depending on needs. Figure 4 shows lines of identical values of measure given for certain values of coefficient p .

Fig. 4. Values of Minkowski measure of similarity depending on the position against the pattern lying in the point $P(5,5)$ for different values of coefficient p

Source: own compilation.

On the basis of the Euclidean distance, the square of the Euclidean distance has been created, namely⁵:

$$m_{e2}(i, k) = \sum_{j=1}^m |z_{ij} - z_{kj}|^2, \quad (7)$$

where:

$m_{e2}(i, k)$ – value of square of the Euclidean measure between object i and object k .

The square of the Euclidean measure is not Minkowski measure. Hence, the influence of shift and the influence of change in a scale on the value of measure must be examined individually. Let us change the value of all variables by the coefficients of shift Δ_j :

$$m_{e2}(i, k) = \sum_{j=1}^m (z_{1j} + \Delta_j - z_{2j} - \Delta_j)^2 = \sum_{j=1}^m (z_{1j} - z_{2j})^2 = d_{1,2}. \quad (8)$$

Thus, the value of measure does not depend on the location of objects in space but only on their positions relative to one another.

In the case of all the aforementioned measures, s -fold increase in the value of all variables of the compared objects brought about s -fold change in the value of measure. Let us multiply the value of all variables of the compared objects by s for the square of the Euclidean measure, namely:

$$m_{e2}(i, k) = \sum_{j=1}^m (sz_{1j} - sz_{2j})^2 = s^2 \sum_{j=1}^m (z_{1j} - z_{2j})^2 = s^2 d_{1,2}. \quad (9)$$

It turns out that s -fold increase in the value of coordinates results in s^2 -fold increase in measure. Thus, this measure is extremely sensitive to major differences in the values of variables of the compared objects, and slightly sensitive to minor differences (cf. Figure 5). Employing this measure to classify objects may result in the fact that some classes will include just one object. Such a situation may be observed when there are several objects classified in the case of which the coordinate of one object differs a lot from coordinates of other objects.

Fig. 5. Values of the square of the Euclidean measure of similarity depending on the position against the pattern lying in the point $P(5,5)$

Source: own compilation.

2. Empirical Research

In order to compare the above presented measures of similarity, statistical measures have been created on the basis of each measure. Four infrastructural-technical variables describing 314 Polish counties in 2006 were employed. Data was derived from the Regional Data Base (Central Statistical Office). Some counties were not taken into account because there was no data for them. Furthermore, counties representing large urban areas such as Warsaw or Wrocław were not taken into consideration as well. These counties should be analyzed separately due to their distinctive character.

Initially, the following seven variables were taken into account:

- the length of working sewage system to 100 kilometres,
- the length of working water supply system to 100 kilometres,
- improved county hard surface roads to 100 square kilometres,
- county hard surface roads to 100 square kilometres,
- total average monthly gross earnings,
- average floor surface of a flat per one occupant,
- the number of living quarters per one person.

Nonetheless, three variables were rejected (because of their slight variability). The following variables remained: the length of a working sewage system to 100 kilometres, the

length of a working water supply system to 100 kilometres, improved county hard surface roads to 100 square kilometres and county hard surface roads to 100 square kilometres. Weights (connected with variability coefficient) have been ascribed to the aforementioned variables, namely:

- the length of working sewage system to 100 kilometres (weight 44.63%),
- the length of working water supply system to 100 kilometres (weight 26.67%),
- improved county hard surface roads to 100 square kilometres (weight 14.6%),
- county hard surface roads to 100 square kilometres (weight 14.1%).

All the aforementioned variables have been adopted as stimulants. Synthetic measure was created for these variables, and then division into four classes was made. Figure 6 shows results obtained in the case of the Euclidean measure and the urban measure. As opposed to the Euclidean measure, the urban measure prefers mainly objects characterized by sustainable development, i.e. in the case of which proportions of the share of particular variables are similar to the model object. Objects, in the case of which the value of one variable is close to the pattern and there are considerable differences between the remaining variables, are treated worse than in the case of the Euclidean measure. Thus, if one wants as large number of variables as possible to have values similar to the pattern, the urban measure (and not the Euclidean one) should be employed. Figure 6 shows that many objects occupy lower positions in the ranking based on the synthetic measure calculated with the use of the urban measure. This particularly refers to areas near large cities such as Warsaw or Wroclaw. Such a situation results from the fact that proportions of variables differ from the model object.

Figure 7 shows results obtained by means of the Euclidean measure and Chebyshev measure. Chebyshev measure is sensitive to objects in the case of which one variable differs a lot from the pattern. This variable practically determines the value of this measure. It is particularly useful when one wants to find objects in the case of which no variable differs very much from the remaining ones. Low value of this measure given for objects indicates that these objects are under-developed in certain respect. This measure has improved a situation of counties situated in south-east Poland. Hence, this fact implies that there are no poorly developed infrastructural-technical elements. On the contrary, there are counties situated in north-west Poland that were included in a worse class. It means that their infrastructural-technical elements are not fully invested in.

Fig. 6. Comparison between the results of the classification of the Polish counties based on infrastructural-technical variables with the use of the Euclidean measure and the urban measure

Source: own compilation.

Fig. 7. Comparison between the results of the classification of the Polish counties based on infrastructural-technical variables with the use of the Euclidean measure and Chebyshev measure

Source: own compilation.

Figure 8 shows results obtained with the use of the Euclidean measure and the square of the Euclidean measure. Measures given previously brought about an increase in the value of measure for some objects, and drop in the value for other ones compared to the Euclidean measure. Hence, the order in which objects were put was subject to change. The square of the

Euclidean measure does not change the order of objects but causes non-linear change in the distance between them. The square may be employed in the case of a class that does not include many objects. Non-linear change in the distance between objects may result in the fact that some objects belonging to the most numerous class will be transferred to the neighboring classes. Such a situation may be found in the case of the classification made. Class 3 is the largest one as it includes 61% of all the counties. Class 2 includes 25% of the counties. Finally, class 1 and class 4 are the smallest ones and include 11% and 3% of the counties respectively. Employing the square of the Euclidean measure, the size of class 3 was reduced to 50% and the size of class 1 and class 4 increased to 14% and 11% respectively. The size of class 2 remained unchanged.

Fig. 8. Comparison between the results of the classification of the Polish counties based on infrastructural-technical variables with the use of the Euclidean measure and the square of the Euclidean measure

Source: own compilation.

Conclusions

Creating synthetic measure, one selects similarity measure depending on the objective he/she wants to achieve. If one wants all values of variables to be as similar to the pattern as possible (and does not want variables very similar to the pattern to determine this fact), the urban measure is the best one. On the other hand, if one wants objects that are underdeveloped in certain respect to be classified low in the ranking, Chebyshev measure

ought to be used. The Euclidean measure is universal and when one does not have any preference for the aforementioned cases, this measure will allow for finding a middle ground. When objects are concentrated in one class, one may try to transfer them to other classes by changing the Euclidean measure into its square.

Notes

¹ Walesiak (2002), p.24.

² Bock (1974).

³ Gordon (1999), pp.20-21.

⁴ Walesiak (2004), p.41.

⁵ The description of Euclides measure can be found in StatSoft Polska.

References

Bock, H. (1974). *Automatische Klassifikation*. Getynga: Vandenhoeck & Ruprecht.

Gordon, A. D. (1999). *Classification*. London: Chapman & Hall/CRC.

StatSoft Polska, <http://www.statsoft.pl/download/help6.php?sortmodul>. Retrieved February 20, 2008.

Walesiak, M. (2004). Podstawowe zagadnienia statystycznej analizy wielowymiarowej. In E. Gatnar & M. Walesiak (Eds.), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych* (pp.17-59). Wrocław: Wydawnictwo Akademii Ekonomicznej.

Walesiak, M. (2002). *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*. Wrocław: Akademia Ekonomiczna we Wrocławiu.