Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 41 (2019) 383–390

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

8th Manufacturing Engineering Society International Conference

Big Data and Advanced Analytics in Industry 4.0: a comparative analysis across the European Union

Luca Greco[a,*] , Piera Maresca[b], Jesús Caja[b]

[a]DEMM Department, University of Sannio, Piazza Guerrazzi, Benevento 82100, Italy
[b]Universidad Politécnica de Madrid, Ronda de Valencia, 3, Madrid 28012, Spain

Abstract

Digital economy and Factory (Industry) 4.0 are among the main challenges in the era of Big Data Analytics. A digital transformation is required since factories and enterprises need to face quick changes in the technological process and continuous flows of massive data, in order to improve the decision making process. First, we want to investigate the actual scenario in Europe. For instance, the data from the Digital Economy Society Index (DESI) by the European commission (available at https://ec.europa.eu/digital-single-market/en/countries-performance-digitisation) gives on overview on countries' performance digitization. The report measures such factors as connectivity, digital skills, digital public services, and more. In particular, it is of interest to analyze the different incidence of employers with digital skills (Analytics, Cybersecurity, Cloud, for instance) on competitiveness and efficiency among countries and industrial sectors, with a privileged eye on Spain and Italy, and to investigate the profile of those industrial sector that are actually investing in digital skills and those that are not.

Keywords: Big data; Digitalization; Industry 4.0

1. Introduction

Digital transformation policies are quickly transforming the nature of Industry and society. Digital economy and Factory (Industry) 4.0 are among the main challenges in the era of Big Data Analytics. A digital transformation is

* Luca Greco. Tel.: +39-0824-3052-45
E-mail address: luca.greco@unisannio.it

2351-9789
10.1016/j.promfg.2019.09.023

required since factories and enterprises need to face quick changes in the technological process and continuous flows of massive data, in order to improve the decision making process. Nowadays, it is well acknowledged that any industry, in every field, can benefit from the acquisition of digital skills. A comprehensive analysis of the progress made by the EU members in terms of digital transformations can be found at the Digital Transformation Scoreboard 2018 by the European Commission [1]. In particular, it is of interest to analyze the different incidence of employers with digital skills (Analytics, Cybersecurity, Cloud, Internet-of-Things,High performance computing) on competitiveness and efficiency among countries and industrial sectors. Progress in such technologies is growing but there is still large room for improvement across Europe.

In this contribution, we aim at monitoring the state of play and evolution of digital transformation in Europe through several indicators that are used to assess Europe's digital performance in digital competitiveness. To this end, the analysis stems from number of national macro indicators that form the Digital Economy and Society Index (DESI) developed by the European commission (data and reports are available at https://ec.europa.eu/digital-single-market/en/countries-performance-digitisation).

The DESI is composed of 5 principal indices (say dimensions): connectivity (CO), human capital (HC), use of internet (UI), integration of digital technologies (IDT) and digital public services (DPS), measured for the 28 EU members. These five dimensions are assumed to be the principal policy areas of concern for a digital economy and society.

Connectivity is the ability to connect to the internet, meant as a possibly high speed digital infrastructure. The human capital index measures the level of basic and advanced skills. Use of internet takes into account the nature of consumed online contents. Integration of digital technologies denotes the ability to enhance efficiency and economic growth through digitalization. Digital public services are a dimension concerning business and citizen interaction with the Public Sector. The DESI at country level is obtained as a weighted mean of the five indexes described above, with weights, 0.25, 0.25, 0.15, 0.20 and 0.15, respectively. According to the obtained ranking, the 20 EU members are divided into three main clusters, where a cluster is meant as a group of countries with similar characteristics. In particular, the first nine ranked countries are profiled as high performing countries (Denmark, Sweden, Finland, the Netherlands, Luxemburg, Ireland, UK, Belgium, Estonia), the subsequent ten in the ranking constitute the group of medium performing countries (Spain, Austria, Malta, Lithuania, Germany, Slovenia, Portugal, Czech Republic, France, Latvia), whereas the last nine form the cluster of low performing countries (Romania, Greece, Bulgaria, Italy, Poland, Hungary, Croatia, Cyprus, Slovakia).

These data have been augmented by also considering the Digital Transformation Enablers' Index (DTEI) from [1], used to reflect enabling conditions in the considered countries.

Our aim is to cluster countries according to both well established and promising brand new clustering techniques from the statistical theory rather than just group countries in an arbitrary way based on the ranking stemming from the DESI index. Furthermore, we aim at evaluating the discriminating power of the six indices under study in order to track the state of digital progress in Europe. In addition, the analysis could lead to the detection of countries showing anomalous patterns with respect to the others, hence unveiling peculiar attitudes and policies that are worth of special attention. Such countries are denoted outliers.

This contribution is structured as follows: in Section 2 some methodological notes are given about the adopted clustering strategies; Section 3 is devoted to the analysis of the data described in the Introduction.

2. Cluster analysis: classical background and new developments

Cluster analysis denotes the set of statistical techniques aimed at grouping similar observations. A general introduction can be found in [3]. The original data set is partitioned into several subsets such that units within the same set are similar to each other and different from the units belonging to different sets. After clusters of units have been composed, one could explore them and focus on cluster profiles, that are summaries of the units within each group, rather than on the original set of raw measurements. For instance, a common practice consists in using the cluster means (also called centroids). Actually, this is a sample reduction strategy, since cluster profiles will be used as new observations, whose number is expected to be considerably lower than the original sample size. There are several approaches to cluster sample units according to their degree of similarity or proximity. When the data have a

quantitative nature, the proximity of units and groups is commonly measured by using some metrics as the Euclidean distance or the Mahalanobis distance. The kind of distance determines the shape of the clusters. By using the Euclidean distance we obtain spherical clusters, whereas the employ of the Mahalanobis distance leads to more flexible elliptically shaped clusters.

The most popular clustering algorithms can be summarized into hierarchical and non hierarchical methods. Hierarchical algorithms can be agglomerative or divisive, depending on the fact that one keeps on aggregating the single units and groups formed along the path or splitting half the data in one group at each step. For instance, in an agglomerative fashion, at each step the pair of clusters (or units) with minimum distance are merged. The distance between clusters can be measured by summarizing the distances between units from the different groups. There are several ways to measure the distance between groups, each giving rise to a different technique: one could use the maximum (complete linkage), the minimum (single linkage), the average (average linkage) distance between couple of units belonging to different clusters but also minimize the total within-cluster variance (Ward method).

Non hierarchical algorithms, on the contrary, do not evolve through successive aggregations or splits, but look for the partition of the data in a fixed number of clusters that minimizes some objective function (more in general that fulfills some criterion). A very popular approach is represented by K-means, according to which groups centroids are found in order to minimize the sum of the squared Euclidean distances within each group. In order to get more flexibly shaped clusters, one could resort to model based clustering techniques characterized by the use of the Mahalanobis distance and the assumption that the data are a random sample from a finite mixtures of (multivariate) Gaussian components.

An open issue in cluster analysis concerns the selection of the optimal number of clusters: thirty indices and several strategies for determining the number of clusters exist. A common approach is to find the best clustering scheme by comparing the different results stemming from varying all combinations of number of clusters, distance measures, and clustering methods and selecting the one chosen by the majority of the available indices and measures of cluster accuracy.

2.1 Robustness issues

Classical clustering procedures can be badly affected by the occurrence of outliers. Outliers are unexpected anomalous values in that they could show some unusual patterns with respect to the bulk of the data, or even no pattern at all. The occurrence of such data inadequacies could make difficult to recover the underlying clustering structure of the data at hand: spurious clusters may be found and/or genuine separate clusters may be forced to be merged. Outliers are not meant to be classified into any of the clusters the genuine data are partitioned in**.** The aim of a robust clustering analysis is twofold: from the one hand we want to identify the correct number of genuine clusters, from the other hand we want to detect possible outliers, since it could be useful and interesting to classify them somehow and explore possible sources of outlyingness. In particular, one could distinguish among two complementary kinds of outliers: structural and component-wise. In the first case, outliers are characterized by their own random mechanism and exhibit very different features with respect to all the considered dimensions; in the second case, on the contrary, each dimension could be contaminated separately.

Here, we will consider those robust clustering techniques stemming from the idea of trimming and snipping. Trimming is based on discarding a fixed proportion of entire observations, whereas snipping is concerned with the elimination of a fixed proportion of single cells. The main feature of both approaches is that trimming and snipping are impartial, that is they are performed simultaneously with the process of forming clusters. These robust techniques introduce the level of trimming/snipping as a further element to be tuned in the analysis. The trimming/snipping level can be chosen by monitoring the change in the estimates or in the objective function as a function the contamination level. The reader is pointed to [2] for a recent account on robustness issues in cluster analysis and other multivariate techniques.

3. Clustering the 28 EU members

The data at hand give the six measures outlined in the Introduction for the 28 EU members. The data are available upon request to the authors. The overall DESI index is also reported and we want to investigate its ability to discriminate among countries in terms of digitalization technologies. According to the DESI index, Denmark is ranked first, Spain is at the eleventh place, whereas Italy stays towards the bottom in the twenty-fifth position.

First of all, an explorative graphical analysis is performed: we look at the marginal distribution of each variable featuring the data at hand, which are all given in Figure 1. The inspection of the box-plots does not unveil any anomalous pattern but for a suspicious low value of connectivity for Greece. However, the nature of outliers is more underhanded since they are prone to masking effects that become more serious with growing dimensions. Then, one could argue that there are not apparent anomalous values in the marginal distributions but nothing more can be told about outlying features that could affect the joint distribution of the data and the dependencies structers among the different variables. On the contrary, this task can be accomplished by using appropriate robust clustering methods

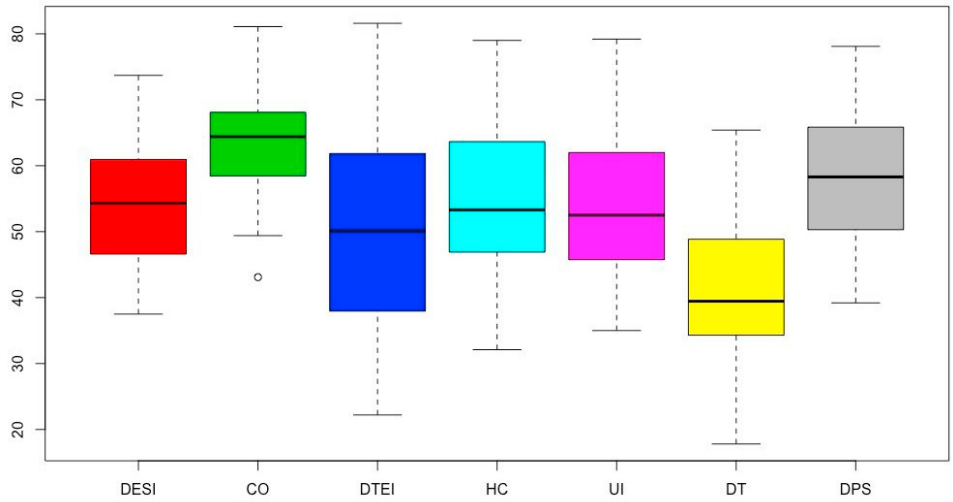

Fig. 1. Box-plots of the six measurements composing the DESI dataset.

Let us start considering different clustering strategies and exploit the differences among results stemming from different methods. Figure 2 displays the hierarchy of clusters (named a dendrogram) corresponding to the average linkage and the Ward method. The first agglomerative method suggests three separate clusters, whereas the Ward method leads to a less straightforward partition. Actually, in the right panel of Figure 2, two clusters are a plausible choice. Formal criteria applied to both clustering techniques lead to select a solution characterized by three and two groups, respectively. Hence, different methods lead to (slightly) different solutions. Figure 3 also gives the agglomerative coefficients corresponding to each method. The agglomerative coefficient is a measure of the strength of the clustering structure [3]: values closer to 1 suggest strong clustering structure. Then, according to this measure, there is more support for a more simple structure composed by two clusters rather than three.

Despite the different solution, the dendograms in Figure 3 exhibit evident common features: the first countries that are joined are Sweden and the Netherlands, in the bottom left, and Austria and Germany, in the bottom-middle, for instance. The height given on the y-axis is a measure of the distance at which units and clusters are merged. We see that several couples are formed in the first steps of the hierarchical procedures, then it may happen that a new country joins the couple, as Denmark with Sweden and the Netherlands in both panels, or two couples are merged to form a more complex cluster of four, as with Slovenia and Czech Republic with Portugal and Cyprus, in the left panel. Furthermore, independently from the agglomerative strategy, it is possible to observe that Finland, Latvia and Romania

are the last to join a group. It is worth to stress a couple of basic rules: at each step only one new group is formed, then, each cluster assignment is irrevocable. Selecting the number of clusters can be described as choosing the height at which cutting the dendrogram by drawing an horizontal line.

A solution with only two genuine clusters is also preferred by the K-means algorithm and the model based technique.

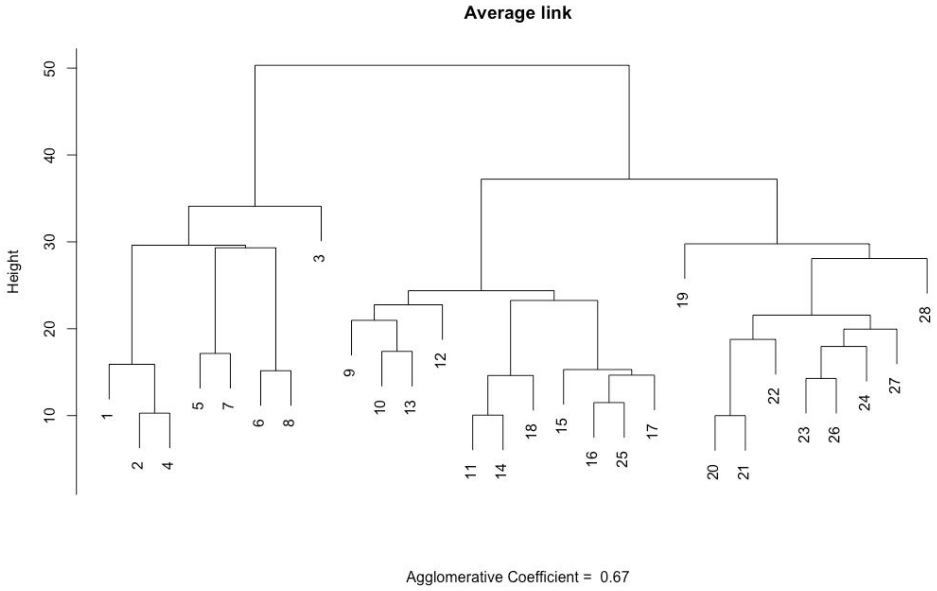

Fig. 2. Dendrogram from Average linkage.

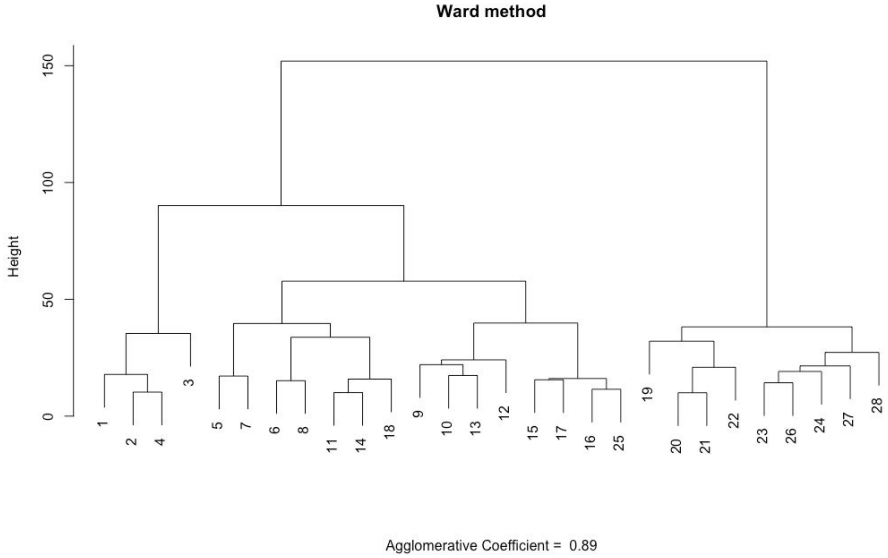

Fig. 3. Dendrogram from Ward method.

Let us investigate, now, to what extent the partitions we have found are affected by the occurrence of some structural or component-wise outliers. To this end we run trimmed K-means and snipped K-means for k=2,3 and different levels

of contamination. It is worth to notice that at least in the trimming approach, a clustering structure with two clusters looks more appropriate with growing number of trimmed countries, otherwise we will end with at least one cluster composed by very few observations that is not meant to be informative. The trimmed k-means algorithm ends up with 2 trimmed observations and two clusters. Namely, the trimmed units correspond to Finland and Romania. Table 1 gives the cluster centroids and the raw measurements corresponding to Finland and Romania. Cluster 1 is composed by the most advanced countries in terms of digital skills, whereas cluster two is composed by countries where there is still large room for improvement and the need to apply new policies aimed at digitalizing industry and society. For what concerns the nature of the two outliers, Finland is no doubt a member of the first group composed by the more digitalized economies since it exhibits the largest values for the Human Capital Use of Internet and Digital Technologies indices but also a surprising low value for the DTEI index that is worth of particular investigation to better assess its policies. Romania, on the contrary, is detected as outlying since it shows the lowest value for the majority of the considered indices, hence standing on the bottom of the ranking. For what concerns the ability of the DESI index to discriminate among advanced and developing countries in terms digitalization, we observe that almost all the countries are well ranked according to the simple clustering structure based on two groups stemming from the majority of the involved techniques. The only exception is represented by France, ranked eighteenth, but an an element of cluster 1. A spatial map of the clusters is displayed in Figure 4, where outlying countries are denoted by a zero label. Spain is an element of cluster 1, whereas Italy is in cluster 2.

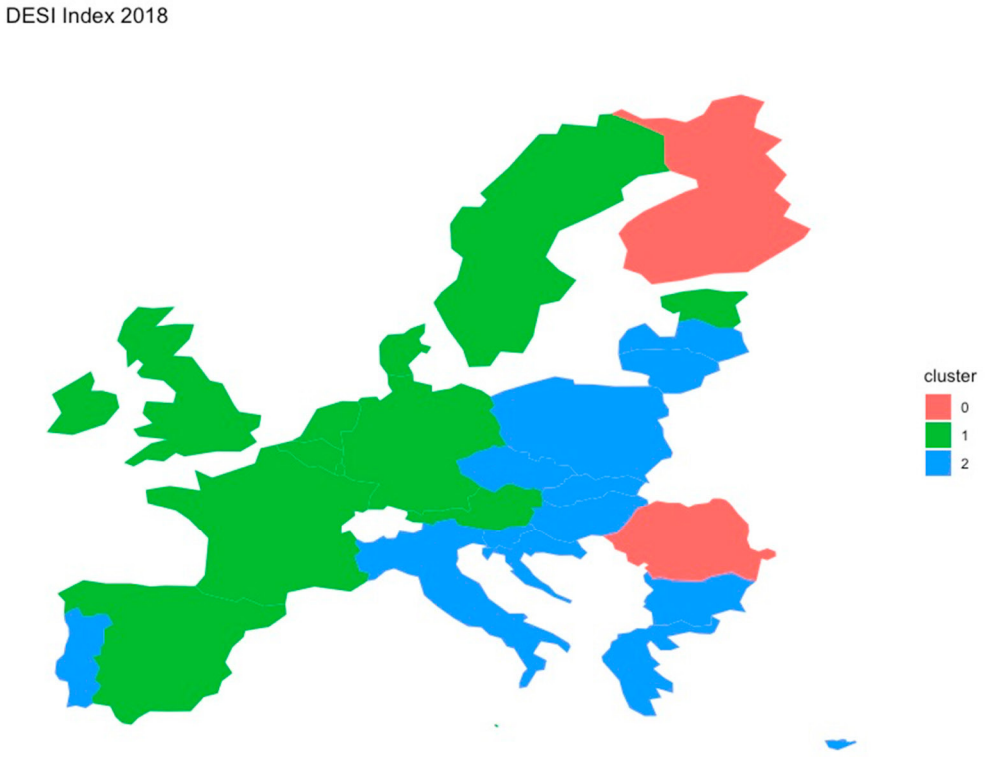

Fig. 4. Clusters and trimmed units from trimmed k-means

The quality of clustering can be measured in different ways. The agglomerative coefficient is a solution but a very popular approach consists in evaluating the average silhouette. This is an index that compares the distances of each

unit from the other units belonging to the same group and the distances of each unit with those units assigned to different clusters. The larger the silhouette, the more applicable is the quality of clustering. A silhouette plot is shown in Figure 5 for the considered trimmed K-means.

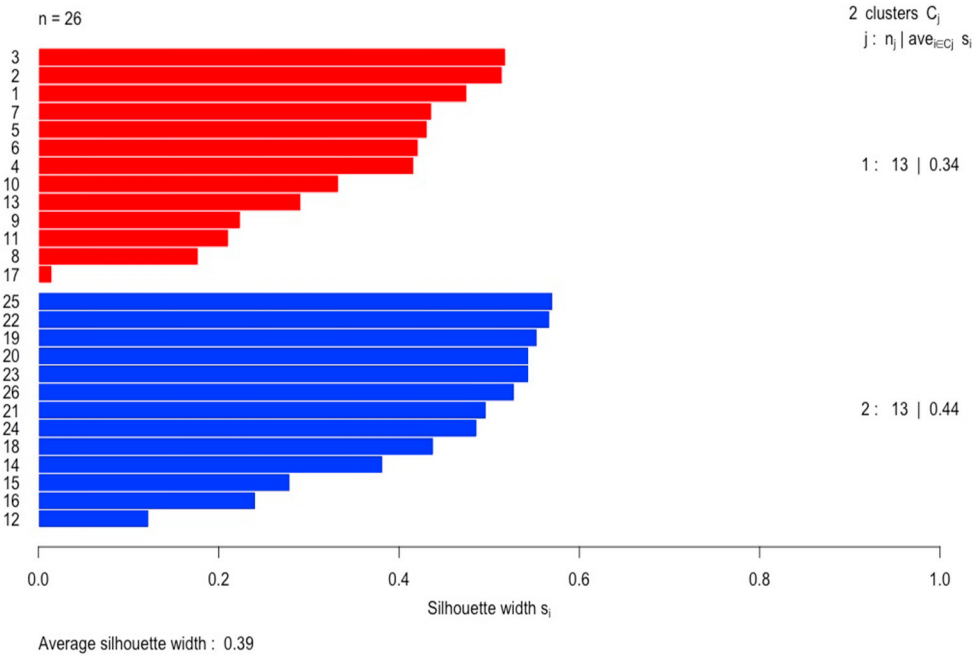

Fig. 5. Silhouette plot from trimmed k-means with two clusters

Let now consider the possibility of snipping and looking for cell wise rather than structural outliers. The snipping algorithm has been run starting with the same level of contamination of trimmed K-means, that is twelve outlying cells, and then by reducing it progressively. It is interesting to stress the strong agreement between the fitted centroids and final cluster assignments with those dispalyed in Table 1, that we do not report here.

Table 1. Cluster profiles and raw measurements for Finland and Romania.

	CO	DTEI	HC	UI	DT	DPS	size
Cluster 1	70.1	65.3	64.2	58.9	46.7	65.1	13
Cluster 2	58.4	38.4	46.2	48.1	34.8	52.3	13
Finland	66.1	66.1	79	79.2	65.4	60.9	
Romania	58.1	22.2	32.1	35	17.8	41.4	

4. Conclusions

All clustering methods under study give a two large group structure. The robust techniques unveil some anomalous patterns, namely for Romania and Finland. Spain is in the group of advanced countries, Italy is not. This result tells that a simpler clustering structure than that suggested in the reports published by the European Commision is feasible to describe the digitalization process of Industry 4.0 in EU. It comes clear that there are two large blocks of countries developing at different rates. From the one hand there are the high performing northern and western countries, from the other side we have the low performing southern and eastern countries. In this scenario, Spain is a positive surprise since it is the only southern country where digital transformation policies are real and effective in order to fill the gap with more digitalized realities.

References

[1] Digital Transformation Scoreboard 2018: EU businesses go digital: Opportunities, Outcomes and Uptake, available at: https://ec.europa.eu/growth/toolsdatabases/dem/monitor/sites/default/files/Digital%20Transformation%20Scoreboard%202018_0.pdf.
[2] Digital Economy and Skills (Unit F.4), Countries' performance in digitization, available at: https://ec.europa.eu/digital-single-market/en/countries-performance-digitisation.
[3] L. Kaufman, P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, 1st ed. New York: Wiley; 1990.
[4] A. Farcomeni, L. Greco. *Robust methods for data reduction*, 1st ed. Boca Raton: CRC Press; 2015.