Taran, Zina; Mirkin, Boris

Article

Exploring patterns of corporate social responsibility using a complementary K-means clustering criterion

Business Research

Provided in Cooperation with:
VHB - Verband der Hochschullehrer für Betriebswirtschaft, German Academic Association of Business Research

Suggested Citation: Taran, Zina; Mirkin, Boris (2020) : Exploring patterns of corporate social responsibility using a complementary K-means clustering criterion, Business Research, ISSN 2198-2627, Springer, Heidelberg, Vol. 13, Iss. 2, pp. 513-540, https://doi.org/10.1007/s40685-019-00106-9

This Version is available at:
https://hdl.handle.net/10419/233153

Business Research (2020) 13:513–540
https://doi.org/10.1007/s40685-019-00106-9

ORIGINAL RESEARCH

Exploring patterns of corporate social responsibility using a complementary *K*-means clustering criterion

Zina Taran[1] · **Boris Mirkin**[2,3]

Received: 14 February 2018 / Accepted: 5 December 2019 / Published online: 18 January 2020

Abstract Companies' objectives extend beyond mere profitability, to what is generally known as Corporate Social Responsibility (CSR). Empirical research effort of CSR is typically concentrated on a limited number of aspects. We focus on the whole set of CSR activities to identify any structure to that set. In this analysis, we take data from 1850 of the largest international companies via the conventional MSCI database and focus on four major dimensions of CSR: Environment, Social/Stakeholder, Labor, and Governance. To identify any structure hidden in almost constant average values, we apply the popular technique of *K*-means clustering. When determining the number of clusters, which is especially difficult in the case at hand, we use an equivalent clustering criterion that is complementary to the square-error *K*-means criterion. Our use of this complementary criterion aims at obtaining clusters that are both large and farthest away from the center. We derive from this a method of extracting anomalous clusters one-by-one with a follow-up removal of small clusters. This method has allowed us to discover a rather impressive process of change from predominantly uniform patterns of CSR activities along the four dimensions in 2007 to predominantly single-focus patterns of CSR activities in 2012. This change may reflect the dynamics of increasingly interweaving and

✉ Zina Taran
ztaran@deltastate.edu

Boris Mirkin
bmirkin@hse.ru; mirkin@dcs.bbk.ac.uk

[1] Department of Management, Marketing and Business Administration, Delta State University, 1003 W Sunflower Road, Cleveland, MS 38733, USA

[2] Department of Data Analysis and Artificial Intelligence, National Research University Higher School of Economics, 20 Miasniskaya, Moscow, RF 101000, Russia

[3] Department of Computer Science, Birkbeck University of London, Malet Street, WC1E 7HX London, UK

structuring CSR activities into business processes that are likely to be extended into the future.

Keywords Corporate social responsibility · Quantitative patterns · Cluster analysis · *K*-means · Anomalous cluster · CSR trends

1 Introduction

Issues of Corporate Social Responsibility (CSR) and sustainability have become increasingly important for both academic research and business practice (Chen and Chen-Hsun 2017; Hult 2011). As society becomes more and more concerned with environmental and social issues, the public increasingly expects companies to behave in environmentally and socially responsible ways. Business communities have responded to these expectations (Sethi et al. 2017). Business-school accrediting bodies have begun to add ethics and sustainability to their accreditation standards (AACSB International 2017; IACBE 2017), and many companies have established sustainability-officer positions. The perceived urgency and importance of CSR at times have escalated to a race between organizations to launch initiatives, whether or not benefits of such actions materialize (Wirl 2014).

The subject of CSR has drawn considerable interest for research. Most of CSR research studies its effect on company performance and the factors moderating and mediating that effect (see for example: Chen et al. 2013; Chih et al. 2008; Hong and Andersen 2011; Heltzer 2011; Jo and Harjoto 2011; Jo and Na 2012; Luo and Bhattacharya 2006; McGuire et al. 2012; Moura-Leite and Padgett 2014; Mulyadi and Anwar 2012; Nelling and Webb 2009; Park et al. 2014; Peters and Mullen 2009; Sun 2012; Sun and Stuebs 2013). Researchers also have attended to related subjects, such as how CSR affects and is affected by business and society, the role of administration, and the role of government and law makers (Carroll 1999; Cochran 2007; Bosch-Badia et al. 2013; Li et al. 2017). Altogether, the common opinion is that CSR is no longer a collateral activity of the companies, but rather part of the core corporate strategy aimed at streamlining and improving imperfections of narrowly defined market goals (Porter and Kramer 2011; Bosch-Badia et al. 2013). Failure to engage or adequately engage in CSR may have negative consequences, potentially resulting in bad publicity, lowered reputation, or diminished value of a brand (Peloza and Shang 2011).

Researchers have described a number of multidimensional systems for engaging in CSR (see, for example, Schreck 2011). One of the most popular multidimensional systems for engaging in CSR is what is sometimes referred to as the "4D List" (Albinger and Freeman 2000; Peloza and Shang 2011):

- Social: directed at the local community and society at large;
- Labor: directed at own employees;
- Environment: directed at the habitat and natural environment; and
- Governance: directed at ensuring transparent and just corporate governance.

These dimensions are considered the most important and general, covering almost all possible CSR activities. Measurement scales for these and some other multidimensional systems mostly have been unified, standardized, and maintained by the Kinder, Lydenberg, and Domini (KLD) database, currently handled, in a modified form, by the MSCI. This database serves as a popular and prominent source for research projects exploring empirical evidence of CSR activities along the dimensions in the 4D List. Typically, published research concerns the interrelation between individual CSR dimensions and company activities in various areas, such as strategy, control, or performance (see, for example, Martinez 2014; Block and Wagner 2014; Krüger 2015; Michelon et al. 2013; Sen and Bhattacharya 2001). To date, no publication has analyzed the distribution of corporate efforts across dimensions of CSR as a whole. Identified patterns could provide companies or other bodies with a context and reference point in the analysis, planning, and assessment of CSR activities.

There can be different structural patterns of CSR activity along its dimensions. Whereas some companies may focus on just one dimension, such as the environment ("going green") or labor (staff development), others might prefer to split their efforts by contributing equally to two or more dimensions. One would expect that a large company's CSR policies in this regard would be more or less consistent.

This paper analyzes the largest international companies in the MSCI database to determine whether any consistent multivariate profiles of CSR activities exist. If such a profile or a set of profiles do exist, a related question then would be, what changes in profile types occur over time? The method of *K*-means clustering, arguably the most popular clustering method because of its computational and intuitive simplicity (Hartigan and Wong 1979, Hennig et al. 2015, Lord et al. 2017, Mirkin 2019), will enable us to identify patterns across profiles.

The remainder of this article has the following organizational structure. Section 2 outlines the most popular view of CSR according to its roots in stakeholder theory, discusses the multidimensional nature of CSR, and outlines research questions. Section 3 describes data and methods. We use the square-error criterion of the popular *K*-means clustering as a benchmark and change it for an equivalent (complementary) clustering criterion. The complementary criterion requires finding big anomalous clusters. This gives a substantiation to a method that extracts anomalous clusters one-by-one and then leaves only those largest of them. This strategy mitigates a common drawback of *K*-means: the need for a user-defined number of clusters and the initial location of them—a real issue for an uninitiated user, especially in the global analysis of CSR activities. Section 4 describes and analyzes thus identified clusters and discusses methods, findings, and corresponding possible future developments in CSR activities. Section 5 concludes the paper with a brief account of the results and directions for future work.

2 Corporate social responsibility and its patterns

2.1 Defining CSR

The usual understanding of corporate social responsibility (CSR) is typified by McWilliams and Siegel definition: as a company's "actions that appear to further some social good, beyond the interests of the firm and that which is required by law" (McWilliams et al. 2011). Despite the popular view epitomized by Milton Friedman's famous pronouncement that "the social responsibility of business is to increase its profits" (Friedman 1970), society, business, and academia have moved in the direction of a wider perspective, whereupon businesses do have additional responsibilities to society (Carroll 1979; Porter and Kramer 2011).

Although scholars somewhat differ in their definitions of CSR and "sustainability" (e.g., Chan and Cheung 2015), a general consensus exists that "sustainability" is essentially CSR plus efforts to remain profitable (for example, Hult 2011). Considering John Elkington's description of the triple bottom line of sustainability as "people, planet, and profit" (Elkington 1998, p. 73), CSR can be assigned to the double bottom line of "people and planet."

2.2 Stakeholders

The understanding of what exactly a corporation is responsible for and to whom it is responsible has been maturing from ad hoc, hodge-podge early views to more systematic approaches (Fassin 2009). Central to the operationalization of CSR is the idea of stakeholders as groups that have interest in the way a company does business, in addition to its outcomes (Freeman 1984). Traditionally, CSR literature treats any responsibility to stakeholders other than the shareholders (or, oftentimes, customers) as a social one. However, some researchers argue for making a distinction between stakeholders and social issues (Clarkson 1995). Categorized according to their role vis-à-vis the corporation (as employees, customers, etc.), different stakeholder groups have different concerns. These "functional" groupings of stakeholders are not uniform or homogeneous in their interests or concerns, sometimes to the point of conflict (Betts and Taran 2011).

CSR is then understood as a multidimensional construct (Schreck 2011; Weber and Gladstone 2014) that includes a variety of actions and principles directed at satisfying society-related concerns of non-shareholder stakeholders. These concerns include helping the environment, community (local or global), society, and people (including employees), while pursuing just and ethical governance.

2.3 Measuring CSR

Practitioners and scholars have made considerable efforts to measure CSR (Jones 2017), ranging from a CSR ranking based on numeric values for each possible "good practice" of a firm (as identified by some authoritative body) with an accompanying summative score (Chen et al. 2013), to counting the number of

company policies for each identified area of concern (Welford 2005). Consulting organizations have emerged with guidelines on measuring and reporting of CSR-related practices and their outcomes, usually referred to as CSP (Corporate Social Performance). For example, London Benchmarking Group (LBG) provides guidelines regarding charitable community contributions, including their impact on business and society (London Benchmarking Group 2015).

Practitioners and scholars evaluate and rank organizations based on different areas of concern for different groups of stakeholders (see Clarkson 1995), resulting in several data sets, such as the KLD Socrates database, which has been superseded by the MSCI ESG database (Jo and Na 2012; MSCI 2011).

Kinder, Lydenberg, and Domini rated companies based on a number of strengths and concerns (Lougee and Wallace 2008; Mattingly and Berman 2006; See Exhibit 1).

As the data service changed hands and evolved, some variables and methods of measurement have changed as well; however, the general spirit of KLD ratings has been somewhat retained in the MSCI historical ratings data (See Exhibit 2, cf MSCI 2011). Nevertheless, the database did undergo further changes in the newer MSCI rankings (MSCI, 2011). We note that scorings in the MSCI are treated as continuous-valued variables, so that the operation of score averaging is not out of scope for the MSCI data.

The final IVA score is determined as a weighted sum of the four major factors from 4D list. For example, FedEx faced huge labor disputes and issues with the carbon emissions of its fleet, which led to a CSR rating of CCC (the lowest grade). FedEx improved its rating by significantly changing its carbon emissions. In the same industry, Deutsche Post AG received excellent scores on all four CSR factors: environmental efforts, labor relations, stakeholder relations, and governance. Deutsche Post AG earned a rating of AAA (the highest grade).

2.4 CSR profile

Let us refer to the set of grades of a corporation across the four dimensions in the 4D list above as its CSR profile. Different patterns emerge for company CSR profiles. These patterns may vary on a continuum between an "even" pattern and a "focused" one. We consider that a company exercises an even CSR profile if its performance along each of the four dimensions is about the same. A company could strive to be responsible on all CSR fronts and "be a good citizen." Alternatively, a company that pays no attention to CSR at all has an even pattern as well. A company that concentrates its CSR efforts on just one dimension has a focused CSR pattern. Discerning such patterns from the aggregate data is all but impossible in most cases.

Even more complicated would be trying to predict possible directions of changing CSR profiles from the aggregate data over time. An important factor to consider here is the fact that practitioners have been strongly advised to be strategic in their CSR by choosing directions that would most matter to their reputations. According to Peloza and Shang (2011),

> "The opportunity to differentiate through various CSR activities means that managers should not simply look to outspend their competition on CSR, or

assume that greater levels of CSR investment will lead to improved consumer perceptions of value…. Brands that will be most successful are those that use CSR activities to provide incremental consumer value matched to product category salience of those values" (Peloza and Shang 2011 p. 130).

One would expect that after the initial push in the 1990s to "be good in general," businesses would have become more strategic with their CSR choices. Although that may be true, "being strategic" means different things to different companies and makes prediction difficult from aggregate data. Specifically, each of the following scenarios—(SA), (SB), (SC), (SD), and (SE)—is compatible with the strategic-behavior approach.

(SA) Companies might strive to become better "global citizens" along every dimension. Then, one would see increase in total ratings as well as increase in the number of even profiles

(SB) Corporations may keep losing interest in social responsibility, so that CSR ratings go down and even profiles with low average ratings proliferate

(SC) Businesses could pursue policies oriented at just one or two CSR dimensions and display focused profiles

(SD) Businesses might start with even profiles, but narrow their efforts to only those dimensions that have the greatest impact on their stakeholders. This would show an increase in focused profiles in the data by the end of the period

(SE) Businesses could start with focused profiles, but expand their efforts to other dimensions. This would show an increase in unfocused profiles

2.5 Research questions

It stands to reason that over time, companies would change their CSR profiles. There are several possible scenarios of such change: they could become more committed to CSR in general; they can shift away from the even profiles and more toward the focused profiles; they can shift away from focused profiles to even profiles; they can shift among focused or unfocused profiles. Here are our research questions:

Research Question 1 What are the patterns in CSR activities?
Research Question 2 What are the dynamics of patterns of CSR activities?

3 *K*-means clustering: classic and complementary criteria

3.1 *K*-means: an introduction

The task is to discern patterns prevailing in 2007 and in 2012, and compare and contrast these two time points. The most appropriate multidimensional statistics technique for this type of analysis is clustering, because it is specifically oriented

towards finding different patterns in a data set. A popular clustering method, *K*-means partitioning, seems especially suitable for our goals (for more information about *K*-means partitioning, see Hartigan and Wong 1979, Hennig et al. 2015, Mirkin 2019). It partitions a data set represented by multidimensional points corresponding to observations (companies over the four CSR dimensions, in our case) into non-overlapping clusters. Each cluster is a bunch of points around its center, a pattern computed as a mean across objects in the cluster (calculated separately for each variable). Usually, the user has to guess the right number of clusters and specify some hypothetical cluster representatives, the "seeds."

The number, *K*, and the initial location of "seeds" constitute the input to *K*-means algorithm. The algorithm outputs a partition of the set of objects in subsets, clusters, and cluster centers, points in the multivariate space that correspond to within-cluster averages. The *K*-means algorithm runs a sequence of iterations, each consisting of two steps: a cluster update and center update.

This technique has two big advantages. First, it locally minimizes a natural criterion, the sum of squared Euclidean distances between the objects and their cluster centers. Second, it computationally makes a typology. It is intuitive and computationally convenient. However, the method has limitations, as well. It requires the user to pinpoint initial cluster seeds or, if the user cannot, generates them randomly, thus leading to possibly inadequate results. The number of clusters may be difficult for the user to specify, as well. In the literature, scholars have proposed ideas for how to automate this process (see, for example, Mirkin 2019; Rodriguez and Laio 2014).

This paper uses a complementary criterion for *K*-means. The complementary criterion is mathematically equivalent to the original *K*-means criterion, but it provides a very different rationale for the clustering process. According to this complementary criterion, the goal is to find big anomalous clusters. Although finding a globally optimal solution is as computationally intensive as minimizing the original *K*-means criterion, the complementary criterion leads to a simple heuristic for building "anomalous" clusters one-by-one, thus making the choice of the number of clusters much easier. In this way, the complementary criterion serves as a substantiation of the so-called anomalous-cluster initialization heuristic (Chiang and Mirkin 2010; de Amorim et al. 2016).

3.2 *K*-means square-error and complementary criteria

Given *K*, the problem is to find such a partition $S = \{S_1, S_2, \ldots, S_K\}$ and cluster centroids $c_k = (c_{k1}, c_{k2}, \ldots, c_{kV})$, $k = 1, 2, \ldots, K$, that minimize the square-error criterion:

$$D(S, c) = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v \in V} (y_{iv} - c_{kv})^2 = \sum_{k=1}^{K} \sum_{i \in S_k} d(y_i, c_k), \quad (1)$$

where $d(y_i, c_k)$ is the squared Euclidean distance between data point y_i and cluster center c_k.

The *K*-means algorithm follows the so-called alternating minimization scheme for criterion (1). Starting with some set of *K* centers *c*, it finds an optimal partition *S*, minimizing $D(S, c)$ at the given *c*, and then finds c' minimizing $D(S, c)$ at just found *S*. The procedure is repeated until convergence—that is, until c' coincides with *c*. In practice, the method converges fast to a local minimum, which is dependent on the choice of initial *c*. The issues related to choice of *K* and initial *c* are well known and subject to ongoing debate (for a sample of literature, see de Amorim and Hennig 2015; Mirkin 2019; Mur et al. 2016, and references therein).

Let us consider $T(Y) = \sum_{i=1}^{N} \sum_{v \in V} y_{iv}^2$, referred to as the data scatter, and

$$F(S, c) = \sum_{k=1}^{K} |S_k| \sum_{v \in V} c_{kv}^2 = \sum_{k=1}^{K} |S_k| \langle c_k, c_k \rangle, \tag{2}$$

where $\langle c_k, c_k \rangle$ is the inner product of c_k by itself, the squared Euclidean distance from c_k to 0. These are related to *K*-means criterion in (1) via equation:

$$T(Y) = F(S, c) + D(S, c). \tag{3}$$

Equation (3) implies that the complementary criterion in (2) is to be maximized to minimize $D(S, c)$.

Provided that the origin preliminarily is shifted into the point of "norm", i.e., the gravity center, the meaning of the complementary criterion is as follows: find as numerous and as anomalous clusters as possible, to maximize $F(S, c)$. In contrast to the square-error criterion $D(S, c)$, which does not depend on the location of the space origin, 0, the criterion $F(S, c)$ pertains to the origin, as its items $\langle c_k, c_k \rangle$ heavily depend on that, which is used in the one-by-one greedy optimization approach below. To sharpen the structure of a data set, we counterpose it to a point of "norm", the grand mean of the data set. Therefore, when using the complementary criterion, we do not skip a data preprocessing option, the subtraction of the point of "norm" from all data points.

3.3 Maximizing the complementary criterion: one-by-one approach

An option for finding big and anomalous clusters would be to begin by building anomalous clusters independently, so that each cluster *S* and its center *c* maximize the contribution:

$$f(S, c) = |S| \langle c, c \rangle. \tag{4}$$

An exact solution to this non-polynomial problem cannot be found easily. Therefore, we consider a locally optimal solution. Assume, for the start, an initial cluster to be a singleton, so that $|S| = 1$. To maximize (4) then, one has to put it into the point that is furthest away from the origin, 0. This, unlike the conventional *K*-means, gives us a reasonable initialization for the clustering process. To move further, we attend to the same alternating minimization scheme that is utilized in the conventional *K*-means algorithm. Given cluster *S*, its center *c* is computed as the average:

$$c = c(S) = \frac{\sum_{i \in S} y_i}{|S|},$$

where y_i is a row of the data matrix corresponding to observation $i \in I$. Given c, an optimal update of cluster S should be computed according to the following rule CUR:

Cluster update rule (CUR):

Given a cluster S, remove $i \in S$ from S if $f(S, c) > 2|S| \langle c, y_i \rangle - \langle y_i, y_i \rangle$ and add $i \notin S$ to S if $f(S, c) < 2|S| \langle c, y_i \rangle + \langle y_i, y_i \rangle$.

A proof, that thus updated S indeed maximizes criterion (4) at a given c, is in Appendix to the paper. This update rule gives rise to the following algorithm for building an anomalous cluster.

Algorithm EXTAN (EXTracting an ANomalous cluster)

Input: A data matrix.

Output: List of observations S and its center c.

1. Initialization: Find an observation maximally distant from 0 and make it the initial center, c, of the anomalous cluster being built.
2. Anomalous cluster update: Given c, update S according to CUR rule above.
3. Anomalous center update: Given S, update the center as the within-S mean c'.
4. Test: If $c' \neq c$, assign $c = c'$ and go back to step 2. Otherwise, move on to Step 5.
5. Output: Output the list S and its center c.

EXTAN works according to the alternate minimization principle. One could use an incremental approach by making only one point to move at a time: adding to or removing from S that one object i in the CUR rule which gives the maximum increase in the value of criterion $f(S, c)$, and halting the process when no move can increase the criterion.

Of course, the result of EXTAN depends on the location of 0, as already mentioned.

Using EXTAN as a subroutine, we can propose the following one-by-one algorithm for greedily maximizing the complementary criterion $F(S, c)$ in (2).

Algorithm BANCO (Big Anomalous Clusters One-by-one)

Input: A data matrix and a user-defined integer t—the minimum cluster size (and, possibly, the point of norm, g).

Output: A partition of the set of observations S in K clusters (K is determined by t) and cluster centers $c_1, c_2, \ldots, c_K$.

1. Data preprocessing: *Centering* Take the input point g if provided or, if not, compute the grand mean, the vector of average values of the features, and take it as g. Subtract g from all the data matrix rows. Optionally, normalize data features. Set counter of clusters $k = 1$. Define I_k the set of all the observations.
2. Iterative EXTAN: At a given k, apply EXTAN to the data matrix over the set of observations I_k to output cluster S_k and its center c_k. Define $I_{k+1} = I_k - S_k$. If $I_{k+1} \neq \varnothing$, set $k = k + 1$ and start step 2 again. Otherwise, move to the next step.

3. Large cluster's centers' selection: Consider all the sets S_k obtained, and determine those satisfying $|S_k| > t$. Define K the number of these clusters S_k (k = 1, 2, …, K).
4. Output the K anomalous clusters and their centers.

The BANCO's output serves as the input to a run of K-means.

According to the algorithm BANCO, the number of clusters K is determined by another user-defined quantitative parameter, t, the minimum number of observations in a cluster. Both K and t relate to the level of granularity of an aggregate representation of the data set structure by clusters. However, the same value of K usually corresponds to an interval of t values. Therefore, specifying a t value imposes on the user somewhat less pressure: first, the minimum cluster size requires less data knowledge, and, second, the resulting K is not that sensitive to errors over the value of t. Sometimes, the choice of t can be operationalized as follows. Take all the anomalous clusters built by BANCO, and look at the sequence of their cardinalities sorted in the descending order. Take that t which is defined by a large drop in the values if there is a drop. Also, BANCO can be easily adapted to the case when K is known to the user: just take the K largest clusters at the Step 3.

3.4 Previous research on the issue of the right number of clusters

The issue of finding the right number of clusters is attracting considerable research efforts (see reviews in Mirkin 2019 and Chiang and Mirkin 2010 and for some later attempts, de Amorim and Hennig 2015, Lord et al. 2017, Zhou et al. 2017, and references therein). Generally speaking, all of the efforts can be divided in three groups:

a. those based on exploring the data set before clustering,
b. those based on values of some indexes while clustering, and
c. those based on exploration of the set of partitions found at multiple runs of a clustering algorithm.

Approaches falling in the group (c) are most popular. Some of them—such as Gap and Jump statistics, Hartigan and Calinski–Harabasz indexes, and consensus-based indexes—have been tested by Chiang and Mirkin (2010) in their extensive computational experiments. In these experiments, the Hartigan index appeared superior to the others in terms of recovery of the number of clusters K, although not in terms of recovery of clusters themselves.

A genuine method in group (a) is a precursor to the proposed algorithm EXTAN, an anomalous clustering method proposed by Mirkin (1990) as an extension of the Principal Component Analysis to binary factor scores. Given an $N \times V$ data matrix $\boldsymbol{Y}$, the first principal component consisting of $N \times 1$ vector of factor scores $\boldsymbol{z}$ and $V \times 1$ loading vector $\boldsymbol{c}$ minimizes the square error $D = \|\boldsymbol{Y} - \boldsymbol{z}\boldsymbol{c}^T\|^2$ where $\|\cdot\|^2$ is the sum of squared differences between the corresponding $N \times V$ matrices. As is well known, the optimal product $\boldsymbol{z}\boldsymbol{c}^T$ is equal to $\mu\boldsymbol{v}\boldsymbol{w}^T$, where μ is the maximum singular value of $\boldsymbol{Y}$, and $\boldsymbol{v}$ and $\boldsymbol{w}$ are its corresponding normed singular vectors. The method of principal clusters takes the same square error $D = \|\boldsymbol{Y} - \boldsymbol{z}\boldsymbol{c}^T\|^2$ criterion, but

Table 1 Illustrative example of eight objects

Object	x	y
A	− 1	3
B	0	3
C	2	2
D	1	1
E	− 1	1
F	0	1
G	− 2	− 1
H	0	− 1

constraints the solution to binary 0/1 vectors $\boldsymbol{z}$ only, so that the binary $\boldsymbol{z}$ corresponds to cluster $S = \{i: z_i = 1\}$ and the optimal c, to S within-cluster mean vector. Geometrically, this is equivalent to finding cluster S and V-dimensional vector $\boldsymbol{c}$ minimizing criterion (Mirkin 1990):

$$A(S, c) = \sum_{i \in S} d(y_i, \mathrm{c}) + \sum_{i \notin S} d(y_i, 0), \tag{5}$$

Criterion $A(S, c)$ in Eq. (5) is similar to K-means criterion in Eq. (1), except that the number of clusters here is 2 and one of them has a non-varying center, 0. The principal cluster analysis method, also referred to as the anomalous-cluster method in Chiang and Mirkin (2010), is alternating minimization method, in some aspects similar to EXTAN above, except that the goal of maximizing the cluster sizes, so much prominent in EXTAN, is absent from the former method.

To illustrate the difference between the two methods, consider, for example, the set of eight two-dimensional observations A–H in Table 1. Assume that the "norm" here is specified as the origin, point 0 = (0, 0), and no normalization is required. Then, repeated applications of EXTAN bring two non-trivial clusters $S1 = \{A, B, C\}$ and $S2 = \{D, E, F\}$, leaving G and H singletons. In contrast, anomalous-cluster method outputs only one non-trivial cluster, $S1 = \{A, B, C\}$ while leaving the other five points singletons. Obviously, no threshold t can reconcile these results.

4 Empirical analysis

4.1 Ratings database: description of MSCI

Having superseded the KLD data, MSCI historical data provide the opportunity to observe patterns in CSR over time. The variables available are in Exhibit 2. Ratings are in a letter-grade format from AAA, AA, A, etc. to CCC, converted into a seven-point scale, with AAA being a 7 and CCC a 1. The universe of companies included in this data has kept expanding over time, so that by March, 2007, the number of companies grew to provide grounds for a global comparison. As many as 1850

Table 2 Basic statistics

	2012			2007		
	Mean	St. Dev	St. Dev/mean (%)	Mean	St. Dev	St. Dev/mean (%)
Strategic governance factor						
v17	5.27	1.64	31.10	5.44	1.88	34.63
Human capital						
v18	5.66	1.90	33.60	5.53	1.72	31.13
Environment factor						
v19	5.06	1.90	37.50	4.90	1.72	35.17
Stakeholder capital						
v20	4.93	1.67	33.90	5.29	1.85	34.91

companies have ratings in both March, 2012 and March, 2007 data. These companies form the set of observations used in this study.

A number of companies rated in 2012, but not in 2007 (904 altogether), are not included in the analysis, since they cannot be used for comparison across time.

We included into our analysis all of the major components in 4D List. These variables form the Social and Eco ratings: Strategic Governance Factor v17, Human Capital Factor v18, Environment Factor v19, and Stakeholder Capital Factor v20 (see the Exhibit 2 for components of the factors). Table 2 provides descriptive statistics for the variables. Altogether, the data set used in our analysis includes the above variables along the scores for each of the major dimensions for the 1850 companies that have been in both 2007 and 2012 lists.

4.2 Means and correlations at aggregate data

We first try to look at the data structure using the means of the grades over dimensions under consideration and correlation coefficients between the dimensions. The feature means at the entire data set are, as usual, referred to as grand means to distinguish them from the means at clusters.

When looking at the levels of the values over the entire data set (see Table 2), one can see that all of the mean levels are close to 5.0, whereas the levels of variation are about 30–35%. This pattern does not change much from 2007 to 2012, which we attribute to the existence of various pattern types that are hidden behind the average values. Therefore, grand means alone provide no insights that are useful for our analysis.

Looking at the pattern of correlations between the CSR components (Table 3), we can see a great difference between the time points. In 2007, correlations between all of the four dimensions are much greater than those in 2012. The level of correlations in 2007 is at the level of 0.65–0.70. That means, on average, if a company earned a high score on one of the CSR dimensions, then it was very likely to score high over the other dimensions as well. Similarly, the companies scoring low over one dimension were likely to score low over the other dimensions.

Table 3 Correlation coefficients between the variables

$N = 1850$	2007		
	v17	v18	v19
Strategic governance factor			
v17			
Human capital			
v18	0.680***		
Environment factor			
v19	0.692***	0.563***	
Stakeholder capital			
v20	0.724***	0.640***	0.628***
	2012		
Strategic governance factor			
v17			
Human capital			
v18	0.252**		
Environment factor			
v19	0.193**	0.192**	
Stakeholder capital			
v20	0.141**	0.160**	0.245**

**Correlations statistically significant at 0.01 level

***Correlations statistically significant at 0.001 level

Table 3 shows that the starting point of the CSR activity process is not something that happened a long time ago. The starting period, however, early CSR discussions started, did cover year 2007 and probably later. However, we can safely claim that a new phase already started by 2012 because of the significant changes in the pattern of correlations in 2012 (see bottom of Table 3). In 2012, the correlations are not high anymore, but rather small, around 0.15–0.25. The shift to low correlation illustrates dramatic changes in the patterns of CSR activities: there is little agreement between the dimensions now.

There is no point to claim that the prevailing pattern in 2012 is a focused one, because small correlations can hide contradictory processes in different subsamples. However, one cannot deny that the even patterns of CSR activity are not prevailing anymore in 2012. At a first glance, these two sentences may seem to contradict each other. Yet, they do not contradict each other, because there can be plenty possibilities between the even and focused pattern types. Say, while pattern $a = (5, 5, 5, 5)$ is even and $b = (7, 2, 2, 2)$ is focused, patterns $c = (1, 7, 3, 7)$ and $d = (9, 9, 2, 2)$ are neither even nor focused.

Probably, these changes have occurred because of a great global recession 2008/9, which affected the economic behavior significantly. The recession required companies to restructure their CSR activities as innate parts of the businesses, rather than as purely external activities.

Developing better understanding of the structure of CSR activity patterns requires our finding and analyzing CSR similarity clusters of corporations.

4.3 The structure of clusters: comparative analysis of CSR activities from 2007 to 2012

We applied method BANCO to both data sets, 2007 and 2012. As described later in Sect. 5.2, the method produced six clusters at each of the datasets. Characteristics of the found clusters are shown in Tables 4 and 5.

Let us first take a look at the patterns exhibited in the 2007 clusters (Table 4). In Table 4, there are six clusters represented by their centers expressed in the natural scales of MSCI scores. Additionally, the centers are presented in their relation to the grand mean as the relative difference $(m - \mathrm{Gm})/\mathrm{Gm}$, per cent, where m is the within-cluster central value and Gm is the grand mean of the same feature. The column on the left shows the cluster sizes.

One can see from the Relation to Grand Mean on the right of Table 4, the clusters from 2007 indeed manifest more or less uniformly even patterns. In particular, clusters 1 and 2 (totaling to about 600 companies) perform much better than the grand mean values, by about 45% and 22% on average, respectively. In contrast, clusters 5 and 6 (totaling to 579 companies) exhibit profiles that are lower than the grand mean values. Clusters 5 and 6 are uniformly underperforming by about 15–30% and 40–50%, respectively. Middle clusters 3 and 4 are less uniform: cluster 3 slightly outperforms the averages in all but the Environment; this pattern is reversed in cluster 4.

In general, the 2007 clusters manifest a more or less a simplistic structure of CSR activities: corporations form a continuum of CSR activity effort ranging from approximately 2–3 to 7–8 on the KLD grading scale; at any point on this continuum, the total company effort is more or less uniformly even across the four CSR dimensions. The number of companies that either overperform on all of the dimensions or underperform on all of the dimensions is 1184, about two-thirds of the total number of companies in the sample.

The 2012 clusters (see Table 5, which is formatted similarly to Table 4) show a rather different picture. The difference can be seen even in the distribution of companies over the clusters. That was rather uniform in 2007, with cluster sizes varying between 219 and 307. The size differences are much sharper in 2012: they range from 170 to 492. Moreover, the spread over the grading scale is smaller by a grade on each side of the range, from 3–4 to 6–7 (versus 2–3 to 7–8 in 2007) on all but the Stakeholder Capital scales. The thoroughly overperforming and underperforming clusters still are present, but they cover a much smaller part of the set. Specifically, only cluster 1 (258 companies), on the plus side, and cluster 6 (287 companies), on the minus side, fall within this category. This represents a sharp decline of the balanced effort: from about 1250 companies in 2007 to 545 companies in 2012.

Also, the levels of deviation from the grand mean values are smaller in 2012. They reach about − 27% in cluster 1 and − 33% in cluster 6 on average, whereas

Table 4 Clusters 2007

K	N_k	Cluster center				Relation to grand mean (% over/under grand mean)			
		Environment	Strat. governance	Human capital	Stakeholder capital	Environment	Strat. governance	Human capital	Stakeholder capital
1	258	6.99	8.11	7.8	7.97	42.5	49	41	50.7
2	347	6.37	6.74	6.29	6.35	29.9	23.9	13.7	20.1
3	329	4.38	5.86	6.33	5.77	− 10.6	7.8	14.4	9.1
4	337	5.45	4.98	4.75	4.57	11.2	− 8.4	− 14.1	− 13.6
5	333	3.38	4.1	4.78	4.3	− 31	− 24.6	− 13.6	− 18.7
6	246	2.65	2.69	3.1	2.67	− 45.9	− 50.6	− 43.9	− 49.5
Total	1850	4.90	5.44	5.53	5.29	0	0	0	0

Table 5 Clusters 2012

K	N_k	Cluster center				Relation to grand mean (% over/under grand mean)			
		Environment	Strat. governance	Human capital	Stakeholder capital	Environment	Strat. governance	Human capital	Stakeholder capital
1	258	6.15	7.36	6.74	6.28	21.6	39.5	19.1	27.3
2	297	5.74	4.82	8.16	5.07	13.6	− 8.5	44.1	2.8
3	346	7.09	5.02	4.82	5.35	40.1	− 4.8	− 14.9	8.4
4	492	3.91	5.21	5.42	5.66	− 22.7	− 1.1	− 4.3	14.7
5	170	4.22	6.44	5.92	2.07	− 16.5	22.1	4.6	− 58
6	287	3.38	3.58	3.39	3.53	− 33.1	− 32.1	− 40.1	− 28.3
Total	1850	5.06	5.27	5.66	4.93	0	0	0	0

these differences in 2007 are approximately 46% and − 48% on average, respectively.

Other clusters do not manifest even patterns at all. In particular, clusters 2 and 3 in Table 5 pertain to single-focus profiles. Cluster 2 exhibits a high-level activity over Human Capital, which is here 44% greater than the grand mean, while the other three components are closer to the average level, varying from − 8.5 to 13.6%. Similarly, Cluster 3 exhibits a high level of activity over Environment, 40% greater than the grand mean. Cluster 4 can be also counted as a single-focus group, this time at the Stakeholder capital as the focus dimension. Moreover, cluster 4 significantly underperforms over another dimension, Environment. Cluster 5 also may be considered as exhibiting a single-focus CSR pattern, assuming that it includes businesses that have less equity to devote to CSR activities. It underperforms on most dimensions, most notably by almost 60% on the Stakeholder capital front, but is 22% above average on Strategic governance.

Therefore, clusters 2012 illustrates a great change in patterns of CSR activities during the period 2007–2012. In the beginning of the period, the prevailing pattern for companies was to uniformly split their efforts, whichever they were, to each of the components of CSR activities. About 600 of them were overperforming and 600 underperforming in terms of the MSCI grades on each dimension. By 2012, this majority shrunk to only clusters 1 and 6 staunchly maintaining the even pattern on either overperforming or underperforming ends. The total number of companies in these two clusters, 545, accounts for just about 30% of the company set. The others have noticeably switched to single-focused profiles of CSR activities. Each of the clusters 2–5 overperforms noticeably on one dimension and has varying performance over other dimensions, at least some of which they underperform on. The change of CSR activities can be seen on the level of individual companies, too: most of them have changed their CSR pattern from 2007 to 2012.

4.4 Discussion of trends in CSR evolution over the four dimensions

Overall, the found clusters provide the following answers to our Research Questions: in 2007, prevailing pattern was to uniformly outperform or underperform on all four dimensions. By 2012, while a little less than 1/3 of the companies still exhibited the even pattern, the prevailing pattern changed to single focus. Also, it appears; the early period of CSR developments did not occur in the remote past, but rather was still running as recently as 2007. One can also see that the cluster structure of the CSR activities in the set of 1850 of the largest global companies has considerably changed between 2007 and 2012—probably yet another consequence of the Great Recession in 2008/9. We see a dramatic turn of the largest companies from a more or less uniform pattern of CSR efforts (or lack of such) in 2007 towards single-focus patterns in 2012, which goes in line with the scenario (SD) out of four scenarios outlined above. This dramatic turn shows a tendency that is likely to continue in future developments. Probably, the process of building more focused profiles is going to grow further, provided that the business environment framework in the world does not change much.

Most likely, the clusters of "staunchly uniform" overperformer and underperformer will remain, although at further reduced sizes. The clusters with contrast single-focus patterns will be further enhanced. Probably, while CSR activities mature and further consolidate as parts of the business process, the single-focus groups could expand to embrace two or even more dimensions. Therefore, a greater number of CSR clusters should be expected, with those displaying a double-focus activity, featuring more prominently in the future data.

5 Discussion

We discuss our findings in relation to two perspectives: (a) the data and (b) the method.

5.1 Issues related to the data

This paper uses the MSCI database as the main data source.

The MSCI ratings are not perfect even beyond the obvious limitation of any ranking system in so far as they are rooted in their creators' sets of assumptions and simplifications. For example, they do not account for "whether a company's political activities support or undermine environmental regulation" (Schendler and Toffel 2011). Also, many relevant companies appear to have been excluded from the database (Adam and Shavit 2008). The choice of lower level input variables of a greater granularity, which MSCI aggregates into the four indexes, and the particular method of aggregation may need further peer-review validation.

Additionally, the data aggregate CSR performance across the subsidiaries and strategic business units, which might obscure the picture considerably. More precisely, MCSI provides only aggregate measures showing only the resulting balance between positive and negative aspects of a company's social performance and concealing the raw scores. Other flaws in MSCI methodologies are also a limiting factor. The database is rather expensive and not available to some researchers or organizations, which limits the scope of scientific and practical discourse that can rely on it.

For our purposes, what is important is that the 10-grade scoring system utilized in MSCI is robust and well serves various application purposes, which implies that the scores, as well as conventional statistics—the means and the like, derived from them can be used for reasoning about phenomena which they relate to.

5.2 Issues related to the method

Here, we discuss the specifics of our method related to (a) the choice of the number of clusters and (b) to the cluster anomality concept. Also, we are going to support our conclusions using a more conventional form of *K*-means clustering.

(a) The number of clusters and granularity

There have been a lot of research efforts devoted to the issue of the "right" number of clusters (see Mirkin 2019, Mur et al. 2016, Zhou et al. 2017 and references therein). No approach proposed so far can be considered universally adequate. Our view is that clustering transforms data from one level of granularity to a coarser one. The level of granularity should be specified either externally or based on some preliminary data analysis. The number of clusters K is a characteristic of granularity that goes well with our intuition. Unfortunately, our intuition is not so good for choosing the number of clusters. When we cannot determine that the K-clusterings under consideration are optimal, we cannot decide which K is better. The K-means clustering criterion, as is, may provide no guidance on this, because the criterion's value monotonely decreases when K grows. In this aspect, the other characteristic of granularity with which the EXTAN algorithm operates—the minimum number of objects in a cluster—suggests a more convenient tool for choosing K. One might be able to take a look at the distribution of anomalous-cluster cardinalities to make a reasonable choice.

Specifically, let us take a look at Table 6 presenting EXTAN results including the cardinalities of ten largest anomalous clusters. As one can see, for both data sets, there are six relatively large anomalous clusters, with a significant drop in the cardinality of the next anomalous cluster: from more than a hundred to 56 or less. This leads us to accept the number 6 as the number of clusters to build. We apply K-means clustering starting from the centers of the six largest anomalous clusters.

(b) The cluster anomality concept

The concept of cluster anomality is not postulated here, but rather derived from the K-means square-error criterion (1). To give intuition to the concept of anomalous cluster, consider a uniformly distributed one-feature data set whose range is between a and b reals. If a need in dividing the set in two clusters emerges, it would be reasonable to divide the interval $[a, b]$ in two halves separated by the mid-point $(a + b)/2$. By contrast, the EXTAN algorithm would first find an anomalous cluster around a, constituting a chunk between a and half-distance between a and the grand mean, which should be near the mid-point $(a + b)/2$—that is, the interval $[a, (3a + b)/4]$. The next anomalous cluster would be on the right end, constituting an interval about $[(a + 3b)/4,b]$. These are the largest anomalous clusters. A run of bisecting K-means starting from these cluster centers would converge into the original two clusters separated by the mid-point. Therefore, even in such a difficult case as the uniform distribution, the concept of anomalous cluster as the starting point of K-means is consistent with our intuition. The anomalous-

Table 6 Parameters of the results from the EXTAN algorithm to the data 2007 and 2012

Year	Total number of anomalous clusters	Ten largest anomalous-cluster sizes									
		1	2	3	4	5	6	7	8	9	10
2007	14	511	502	224	193	171	127	34	26	24	17
2012	20	555	438	225	173	115	114	56	54	47	31

Table 7 Contingency table of cluster membership in our results vs. K-means

	K-means generated cluster					
	1	2	3	4	5	6
2007 clusters described in the manuscript						
1	0	0	0	243	3	0
2	0	0	250	0	0	8
3	38	286	1	0	0	4
4	9	84	0	0	238	6
5	0	36	1	0	0	310
6	260	0	0	41	32	0

cluster approach was experimentally validated in Chiang and Mirkin (2010) and de Amorim et al. (2016).

We compared our results with those found via a more conventional method. Specifically, we applied a conventional version of *K*-means implemented in the popular computational environment Matlab (Matlab 2018). This version uses the so-called K ++means method by Arthur and Vassilivitskii (2007). The program still needs a user-defined number of clusters K. It generates a random set of K seeds using the principle that the next seed must be far away from the previous ones. This introduces a structure in the process of randomly generating initial centers, but leaves the selection random, thus each time different. In the Matlab version, initial centers are generated five times, and from the resulting five partitions, the program generates the minimum value of criterion (1).

We ran the Matlab operation *K*-means at $K = 6$ on our data, standardized as usual by subtraction of the grand mean from each feature and division over the feature range.

We obtained 6-cluster solutions for both standardized data sets: 2007 and 2012. Below is a contingency table for two partitions, one with clusters-2007 described above, and the other found by the Matlab command *K*-means on the same dataset. The former corresponds to rows, the latter to columns (Table 7):

Essentially, most objects are stable under two clusterings. Hypothetically, if 243 out of 246 objects of our cluster 1 go to a same cluster in the Matlab's partition, 250 out of 258 objects of cluster 2 remain stable, etc. Altogether, 1587 objects, which comprise 85.8% of the total 1850, are stable. A popular measure of agreement exists between partitions: Cohen's kappa (see Fleiss et al. 1969). Largely, the Cohen's kappa is a measure of stability. In our case, kappa = 0.8299, which is an "almost perfect match", according to Landis and Koch (1977). When considering the average profiles within Matlab clusters, we found very similar values; so our interpretation remains valid here.

It should be mentioned, though, that the results of Matlab's *K*-means are rather random. Therefore, we generated a hundred 6-cluster partitions with that program; the average kappa for them is 0.7977 with a standard deviation of 0.1565 which is about 20% of the average.

Table 8 Contingency table of cluster membership in our results vs. *K*-means

	K-means generated cluster					
	1	2	3	4	5	6
2007 clusters described in the manuscript						
1	5	8	274	0	0	0
2	0	0	0	2	14	242
3	166	0	0	0	2	2
4	1	49	9	6	281	0
5	4	1	0	285	7	0
6	1	468	2	6	0	15

Regarding the data 2012, we obtained a Matlab *K*-means partition whose contingency table with our clustering is in Table 8 giving an even better match, with kappa = 0.9135.

Further computations, with a hundred runs of Matlab kmeans, lead to somewhat more modest, but still rather high, average kappa equal to 0.7858 with the standard deviation of 0.1290.

These results support our conclusions.

6 Conclusion

In this paper, we assume that the four dimensions—Environmental, Social/Stakeholder, Labor, and Governance—in the 4D list appropriately represent the structure of the CSR process, albeit in a somewhat aggregate form. To analyze the status and tendencies of the CSR activities among the most advanced global companies, we use a set of companies that covers two separate moments in time. To find out structures of prevailing patterns, we develop a version of *K*-means method, creating more flexibility in the number of clusters than the conventional *K*-means has. To this end, we consider a clustering criterion emerging in the context of data-scatter decomposition in two items: the square error and complementary criterion. This complementary criterion explicitly states the goal of clustering as that of finding big anomalous clusters. A method EXTAN for one-by-one finding the clusters to maximize separate items in the complementary criterion is developed to more reasonably determine the right number of clusters. The method is applied as a precursor to the conventional *K*-means run over the MSCI data sets.

The cluster structures found for 2007 and 2012 data sets lead us to conclusions that can be stated, in brief, as follows.

An unexpected phenomenon is observed. As recently as of 2007, the overall CSR process on the level of a single corporation was as yet at an early stage of the CSR efforts, more or less uniformly distributed over all the dimensions in the 4D List, both on high and low levels. We can see a process of change from the predominantly uniform patterns of CSR activities in 2007 to the predominantly

single-focus patterns of CSR activities in 2012. This process is probably influenced by the Great Recession of 2008/9. This leads us to hypothesize that this process probably will continue to cover not only the Stakeholder Capital dimension in full, as is currently the case, but also in the directions of focusing at other dimensions, as well as to double-focus and triple-focus efforts in the near future. These increased efforts should lead to a more complex cluster structure of the CSR activities in the future.

Limitations of this study lie with the limited scope of the MSCI database and empirical nature of cluster analysis.

Directions for further research involve, on one hand, increasing the granularity of the inputs and, on the other hand, investigating interrelation between managerial attitudes, financial and reputational outcomes, and CSR processes within the clusters to formulate prescriptions for both the managers and the policy-makers.

The results from this study have implications for marketing practitioners. Companies striving to do well along each of the four CSR dimensions belong to an elite group of the best performing companies. At companies with aims to enter this elite milieu, marketing practitioners would be wise to raise their stakeholders' awareness of that fact. In companies with a single-focus CSR profile, marketing practitioners would benefit from tracking and following the communication efforts of other companies with a similar profile, both inside and outside the industry. This, by no means, implies that the companies' CSR efforts should be somehow just an insincere "marketing ploy". Rather, these are suggestions for framing the dialog with the consumers with the view that business ethics is undergoing a deep transformation to include the social responsibility as its immanent and important part.

For scholars and analysts studying the impact of various CSR activities, this paper may lay the groundwork for further empirical investigation of CSR as a multidimensional process. In particular, CSR profiles may provide important variables mediating and moderating the financial and reputational impact of CSR. Additionally, this paper describes a useful and practical modification of the *K*-means method that provides a different angle to efforts in finding the right number of clusters and that may serve as an example of how not to determine the number of clusters ad hoc, but rather to derive them from the data.

Policy-makers should note that these results suggest that companies may increasingly concentrate on only some CSR dimensions at the expense of others.

Acknowledgements The authors are indebted to the anonymous referees and Editor for valuable comments taken into account in the current version

Appendix 1

Consider first an $i \in S$ and $S - i$, that is, S without i. How one could warrant that $f(S - i,\ c') > f(S,\ c)$, where c' is the center of $S - i$? According to definition, $f(S - i,c') = (|S| - 1)$:

$$\left\langle \frac{\sum_{j\in S} y_j - y_i}{|S|-1}, \frac{\sum_{j\in S} y_j - y_i}{|S|-1} \right\rangle =$$

$$\frac{1}{|S|-1}\left\langle \sum_{j\in S} y_j - y_i, \sum_{j\in S} y_j - y_i \right\rangle = \frac{1}{|S|-1}\left(\left\langle \sum_{j\in S} y_j, \sum_{j\in S} y_j \right\rangle - 2\left\langle y_i, \sum_{j\in S} y_j \right\rangle + \langle y_i, y_i \rangle \right).$$

Similarly, $f(S,c) = \frac{1}{|S|}\left\langle \sum_{j\in S} y_j, \sum_{j\in S} y_j \right\rangle = \frac{1}{|S|-1}\left(1 - \frac{1}{|S|}\right)\left\langle \sum_{j\in S} y_j, \sum_{j\in S} y_j \right\rangle = \frac{1}{|S|-1}$

$$\left(\left\langle \sum_{j\in S} y_j, \sum_{j\in S} y_j \right\rangle - f(S,c) \right).$$

By subtracting the last expression from the first formula, one obtains:

$f(S - i,\ c') - f(S,\ c) = \frac{1}{|S|-1}\left(f(S,c) - 2\left\langle y_i, \sum_{j\in S} y_j \right\rangle + \langle y_i, y_i \rangle \right)$. This is positive if and only if:

Exhibit 1 KLD ratings of variables and criteria

Strengths	Concerns
	Community
Charitable giving	Investment controversies
Innovative giving	Negative economic impact
Non-US charitable giving	Tax disputes
Support for education	Other concerns
Support for housing	
Volunteer programs	
Other strengths	
	Corporate governance
Compensation	Compensation
Ownership	Ownership
Political accountability	Political accountability
Transparency	Transparency
Other strengths	Accounting
	Other concerns
	Diversity
Board of directors	Controversies
CEO	Non-representation

Exhibit 1 continued

Strengths	Concerns
Employment of the disabled	Other concerns
Promotion	
Women and minority contracting	
Work/life benefits	
Gay and lesbian policies	
Other strengths	
	Employee relations
Health and safety	Union relations
Retirement benefits	Health and safety
Union relations	Retirement benefits
Cash profit sharing	Workforce reductions
Employee involvement	Other concerns
Other strengths	
	Environment
Beneficial products and services	Agricultural chemicals
Clean energy	Climate change
Pollution prevention	Hazardous waste
Recycling	Ozone depleting chemicals
Other strengths	Regulatory problems
	Substantial emissions
	Other concerns
Human rights	
Labor rights	Labor rights
Relations with indigenous peoples	Relations with indigenous peoples
Other strengths	Myanmar
	Other concerns
	Product
Benefits the economically disadvantaged	Antitrust
Quality	Marketing/contracting controversy
RandD/innovation	Safety
Other strengths	Other concerns

$f(S,c) - 2\left\langle y_i, \sum_{j\in S} y_j \right\rangle + \langle y_i, y_i\rangle > 0$, that is, if $f(S, c) > 2|S|\ \langle c, y_i\rangle - \langle y_i, y_i\rangle$. This proves one part of CUR rule. The other part, for $i \notin S$, is proved analogously by considering the difference $f(S + i, c') - f(S, c)$ where c' is the center of $S + i$.

Appendix 2. Tables and exhibits

Exhibit 2 MSCI historical data

Total rating, v11		
Eco value 21 rating, v12		
Social rating, v14		
Strategic governance factor, v17	(1) SG strategy, v21	
	(2) Strategic capability/adaptability, v22	
	(3) Traditional governance concerns, v23	
Human capital factor, v18	(1) Employee motivation and development, v 24	
	(2) Labor relations, v25	
	(3) Health and safety, v26	
Stakeholder capital factor, v20	Stakeholder capital subfactors	(1) Customer/stakeholder partnerships, v27
		(2) Local communities, v28
		(3) Supply chain, v29
	Products and services subfactors	(1) Intellectual capital/product development, v30
		(2) Product safety, v31
	Emerging markets subfactors	(1) EM strategy, v32
		(2) Human rights/child and forced labor, v33
		(3) Oppressive regimes, v34
Environment factor, v19	Overall environmental scores	Risk factors, v35
		Environmental management capacity, v36
		Opportunity, v37
	Environmental risk factors	(1) Historic liabilities, v38
		(2) Operating risk, v39
		(3) Leading/sustainability risk indicators, v40
		(4) Industry specific risk, v41
	Environment management capacity	(1) Environmental strategy, v42
		(2) Corporate governance, v43
		(3) Environmental management systems, v44
		(4) Audit, v45
		(5) Environmental accounting/reporting, v46
		(6) Env. training and development, v47
		(7) Certification, v48
		(8) Products/materials, v49
	Opportunity factors	(1) Strategic competence, v50
		(2) Environmental opportunity, v51
		(3) Performance, v52

References

AACSB International. 2017. Eligibility Procedures and Accreditation Standards for Business Accreditation. http://www.aacsb.edu/-/media/aacsb/docs/accreditation/standards/business-2017-update.ashx?la=en. Accessed 23 Sep 2017.

Adam, A.M., and T. Shavit. 2008. How can a ratings-based method for assessing corporate social responsibility (CSR) provide an incentive to firms excluded from socially responsible investment indices to invest in CSR? *Journal of Business Ethics* 82 (4): 899–905.

Albinger, H.S., and S.J. Freeman. 2000. Corporate social performance and attractiveness as an employer to different job seeking populations. *Journal of Business Ethics* 28 (3): 243–254.

Arthur, D., and Vassilivitskii, S. 2007. K-means ++: The advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. *Society for Industrial and Applied Mathematics*, 1027–1035.

Betts, S.C., and Z. Taran. 2011. Conflicting issues and corporate social responsibility: aligning organizational efforts with stakeholder interests. *Journal of International Management Studies* 11 (3): 39–46.

Block, J.H., and M. Wagner. 2014. The effect of family ownership on different dimensions of corporate social responsibility: Evidence from large us firms. *Business Strategy and the Environment* 23 (7): 475–492.

Bosch-Badia, M.T., J. Montllor-Serrats, and M. Tarrazon. 2013. A: Corporate Social Responsibility from Friedman to Porter and Kramer. *Theoretical Economics Letters* 3 (3A): 11–15.

Carroll, A.B. 1979. A three-dimensional conceptual model of corporate performance. *Academy of Management Review* 4: 497–505.

Carroll, A.B. 1999. Corporate social responsibility: evolution of a definitional construct. *Business & Society* 38 (3): 268–295.

Clarkson, M.B.E. 1995. A stakeholder framework for analyzing and evaluating corporate social performance. *The Academy of Management Review* 20 (1): 92–117.

Chan, A.K., and S.Y.L. Cheung. 2015. Special issue on corporate social responsibility and sustainability: An introduction. *Journal of Business Ethics* 130 (4): 753–754.

Chen, R.Y., and L. Chen-Hsun. 2017. Assessing whether corporate social responsibility influence corporate value. *Applied Economics* 49 (54): 5547–5557. https://doi.org/10.1080/00036846.2017.1313949.

Chen, R.C.Y., H. Tang, and S. Hung. 2013. Corporate social responsibility and firm performance. *Journal of American Business Review, Cambridge* 2 (1): 181–188.

Chiang, M., and B. Mirkin. 2010. Intelligent choice of the number of clusters in K-Means clustering: an experimental study with different cluster spreads. *Journal of Classification* 27 (1): 3–40.

Chih, H., C. Shen, and F. Kang. 2008. Corporate social responsibility, investor protection, and earnings management: Some international evidence. *Journal of Business Ethics* 79 (1): 179–198.

Cochran, P.L. 2007. The evolution of corporate social responsibility. *Business Horizons* 50 (3): 449–454.

de Amorim, R.C., and C. Hennig. 2015. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences* 324: 126–145.

de Amorim, R.C., V. Makarenkov, and B. Mirkin. 2016. A-Wardpβ: Effective hierarchical clustering using the Minkowski metric and a fast K-means initialisation. *Information Sciences* 370: 343–354.

Elkington, John. 1998. *Cannibals with Forks, Stony Creek*. Gabriola, CT: New Society Publishers.

Fassin, Y. 2009. The stakeholder model refined. *Journal of Business Ethics* 84 (1): 113–135.

Fleiss, J.L., J. Cohen, and B.S. Everitt. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin* 72 (5): 323–327.

Freeman, R.E. 1984. *Strategic management: A stakeholder approach*. Boston: Pitman/Ballinger.

Friedman, M. 1970. *The social responsibility of business is to increase its profits*. New York: New York Times Magazine.

Hartigan, J.A., and M.A. Wong. 1979. Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28 (1): 100–108.

Heltzer, W. 2011. The asymmetric relationship between corporate environmental responsibility and earnings management. *Managerial Auditing Journal* 26 (1): 65–88.

Hennig, Ch., M. Meila, F. Murtagh, and R. Rocci (eds.). 2015. *Handbook of Cluster Analysis*. Boca Raton: Chapman and Hall/CRC Press.

Hong, Y., and M.L. Andersen. 2011. The relationship between corporate social responsibility and earnings management: An exploratory study. *Journal of Business Ethics* 104 (4): 461–471.
Hult, G.T. 2011. Market-focused sustainability: Market orientation plus! *Academy of Marketing Science Journal* 39 (1): 1–6.
IACBE. 2017. Mission, vision, core values, broad based goals http://iacbe.org/about-page/mission-vision-values-governance/last. Accessed 20 Sep 2017.
Jo, H., and M.A. Harjoto. 2011. Corporate governance and firm value: The impact of corporate social responsibility. *Journal of Business Ethics* 103 (3): 351–383.
Jo, H., and H. Na. 2012. Does CSR reduce firm risk? Evidence from controversial industry sectors. *Journal of Business Ethics* 110 (4): 441–456.
Jones, E. 2017. Bridging the gap between ethical consumers and corporate social responsibility: an international comparison of consumer-oriented CSR rating systems. *Journal of Corporate Citizenship* 2016 (65): 30–55.
Krüger, P. 2015. Corporate goodness and shareholder wealth. *Journal of Financial Economics* 115 (2): 304–329.
Landis, J.R., and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33 (1): 159–174.
Li, D., C. Cao, L. Zhang, X. Chen, S. Ren, and Y. Zhao. 2017. Effects of corporate environmental responsibility on financial performance: The moderating role of government regulation and organizational slack. *Journal of Cleaner Production* 166: 1323–1334.
London Benchmarking Group. 2015. From inputs to impact: Measuring corporate community contributions through the LBG framework—A Guidance Manual. http://www.lbg-online.net. Accessed 9 Sep 2015.
Lord, E., M. Willems, F.J. Lapointe, and V. Makarenkov. 2017. Using the stability of objects to determine the number of clusters in datasets. *Information Sciences* 393: 29–46.
Lougee, B., and J. Wallace. 2008. The corporate social responsibility (CSR) trend. *Journal of Applied Corporate Finance* 20 (1): 96–108.
Luo, X., and C.B. Bhattacharya. 2006. Corporate social responsibility, customer satisfaction, and market value. *Journal of Marketing* 70 (4): 1–18.
Martinez, F. 2014. Corporate strategy and the environment: towards a four-dimensional compatibility model for fostering green management decisions. *Corporate Governance* 14 (5): 607–636.
Matlab. 2018. kmeans. https://www.mathworks.com/help/stats/kmeans.html. Accessed 26 Jul 2018.
Mattingly, J.E., and S.L. Berman. 2006. Measurement of corporate social action: Discovering taxonomy in the Kinder Lydenburg Domini ratings data. *Business and Society* 45 (1): 20–46.
McGuire, J., S. Dow, and B. Ibrahim. 2012. All in the family? Social performance and corporate governance in the family firm. *Journal of Business Research* 65 (11): 1643.
McWilliams, A., D. Siegel, and P.M. Wright. 2011. Corporate Social Responsibility: a Theory of the Firm Perspective. *Academy of Management Review* 26 (1): 117–127.
Michelon, G., G. Boesso, and K. Kumar. 2013. Examining the link between strategic corporate social responsibility and company performance: An analysis of the best corporate citizens. *Corporate Social Responsibility and Environmental Management* 20 (2): 81–94.
Mirkin, B.G. 1990. A sequential fitting procedure for linear data analysis models. *Journal of Classification* 7 (2): 167–195.
Mirkin, B. 2019. *Core Data Analysis: Summarization, Correlation, and Visualization*, 2nd ed. New York: Springer.
Moura-Leite, R., and R. Padgett. 2014. The effect of corporate social actions on organizational reputation. *Management Research Review* 37 (2): 167–185.
MSCI. 2011. User Guide and ESG Ratings Definition. http://msci.com. Accessed 25 Oct 2015.
Mulyadi, M.S., and Y. Anwar. 2012. Impact of corporate social responsibility toward firm value and profitability. *The Business Review, Cambridge* 19 (2): 316–322.
Mur, A., R. Dormido, N. Duro, S. Dormido-Canto, and J. Vega. 2016. Determination of the optimal number of clusters using a spectral clustering optimization. *Expert Systems with Applications* 65: 304–314.
Nelling, E., and E. Webb. 2009. Corporate social responsibility and financial performance: The "virtuous circle" revisited. *Review of Quantitative Finance and Accounting* 32 (2): 197–209.
Park, J., H. Lee, and C. Kim. 2014. Corporate social responsibilities, consumer trust and corporate reputation: South Korean consumers' perspectives. *Journal of Business Research* 67 (3): 295–302.

Peloza, J., and J. Shang. 2011. How can corporate social responsibility activities create value for stakeholders? A systematic review. *Academy of Marketing Science Journal* 39 (1): 117–135.

Peters, R., and M.R. Mullen. 2009. Some evidence of the cumulative effects of corporate social responsibility on financial performance. *Journal of Global Business Issues* 3 (1): 1–14.

Porter, M.E., and M.R. Kramer. 2011. Creating shared value. *Harvard Business Review* 89 (1): 2–17.

Rodriguez, A., and A. Laio. 2014. Clustering by fast search and find of density peaks. *Science* 344 (6191): 1492–1496.

Schendler, A., and M. Toffel. 2011. The factor environmental ratings miss. *MIT Sloan Management Review* 53 (1): 17–18.

Schreck, P. 2011. Reviewing the business case for corporate social responsibility: New evidence and analysis. *Journal of Business Ethics* 103 (2): 167–188.

Sen, S., and C.B. Bhattacharya. 2001. Does Doing Good Always Lead to Doing Better? Consumer Reactions to Corporate Social Responsibility. *Journal of Marketing Research* 38 (2): 225–243.

Sethi, S., T. Martell, and M. Demir. 2017. Enhancing the role and effectiveness of corporate social responsibility (CSR) reports: The missing element of content verification and integrity assurance. *Journal of Business Ethics* 144 (1): 59–82.

Sun, L. 2012. Further evidence on the association between corporate social responsibility and financial performance. *International Journal of Law and Management* 54 (6): 472–484.

Sun, L., and M. Stuebs. 2013. Corporate social responsibility and firm productivity: Evidence from the chemical industry in the United States. *Journal of Business Ethics* 118 (2): 251–263.

Weber, J., and J. Gladstone. 2014. Rethinking the corporate financial–social performance relationship: examining the complex, multistakeholder notion of corporate social performance. *Business and Society Review* 119 (3): 297–336.

Welford, R. 2005. Corporate social responsibility in Europe, North America and Asia: 2004 survey results. *The Journal of Corporate Citizenship* 17: 33–52.

Wirl, F. 2014. Dynamic corporate social responsibility (CSR) strategies in oligopoly. *OR Spectrum* 36 (1): 229–250.

Zhou, S., Xu, Z., and Liu, F. 2017. Method for determining the optimal number of clusters based on agglomerative hierarchical clustering. In *IEEE Transactions on Neural Networks and Learning Systems*, 99: 1–11, https://doi.org/10.1109/tnnls.2016.2608001.