A Stochastic Multi-criteria Divisive Hierarchical Clustering Algorithm

ALESSIO ISHIZAKA\textsuperscript{a}, BANU LOKMAN\textsuperscript{b}, and MENELAOS TASIOU\textsuperscript{b}

\textsuperscript{a}NEOMA Business School, 1 rue du Maréchal Juin · BP 215 · 76825 Mont-Saint-Aignan Cedex · France
\textsuperscript{b}Operations and Systems Management, Centre for Operations Research and Logistics (CORL), University of Portsmouth, Portsmouth, PO1 3DE, United Kingdom

ABSTRACT

Clustering is a long and widely-used technique to group similar objects based on their distance. Recently, it has been found that this grouping exercise can be enhanced if the preference information of a decision-maker is taken into account. Consequently, new multi-criteria clustering methods have been proposed. All of our proposed algorithms are based on the non-hierarchical clustering approach, in which the number of clusters is known in advance. In this paper, we propose a new hierarchical multi-criteria clustering that is based on PROMETHEE, where the number of clusters does not need to be specified. Because the outcome is dependent on the parameters of PROMETHEE, we take into account uncertainty and imprecision by enhancing our approach making use of the Stochastic Multiobjective Acceptability Analysis (SMAA) and cluster ensemble methods. SMAA is used to generate a large number of solutions by randomly varying the PROMETHEE parameters, followed by the use of ensemble clustering, which reaches a consensus solution. Our new approach is illustrated in a clustering study of the performance evaluation of US banks according to a set of financial and non-financial (i.e. environmental, social and corporate governance; ESG) criteria. We find that established banks appear in the overall best-performing clusters, with more contemporary banks following suit. In additional analysis we compare financial and overall (financial and non-financial) performance, and find a mixed appreciation of the ESG aspects in this industry in the middle clusters.

Keywords: OR in Banking · Multi-criteria Clustering · PROMETHEE · SMAA · Cluster Ensembles.

\textcopyright\ Alessio Ishizaka
alessio.ishizaka@neoma-bs.fr

Banu Lokman
banu.lokman@port.ac.uk

Menelaos Tasiou
menelaos.tasiou@port.ac.uk
1 Introduction

“Decision-making is as old as man” (Köksalan et al., 2013, p.87). Decision-making can be found in decisions made as early as the first forms of human life, assumed in tasks such as strategically choosing one’s survival (Hastie and Dawes, 2009). The art of decision-making has significantly expanded in the last 60 years, perhaps partly in response to the amount of calculated decisions being taken around the world. The scientific foundations of decision making have been established by researchers, providing a set of tools and techniques dealing with a wide variety of types of problems. The Multi-criteria Decision Aiding (MCDA) is a scientific discipline encapsulating these tools (Greco et al., 2016). Several types of decision problems exist in MCDA, namely choice, ranking, sorting and description (Roy, 1981). The alternatives that the decision-maker (DM) are faced with are evaluated with respect to a number of criteria to find a preference order (for ranking problems) or the most preferred solution (in choice problems). Description-type problems aid in providing a full overview of the problem (Ishizaka and Pereira, 2016), while sorting-related problems assign alternatives into predefined classes (López and Ishizaka, 2017). However, another interesting problem has surfaced in the MCDA domain, which is the clustering problem.

The aim of the classical clustering problem is to group the objects into homogenous clusters in such a way that the alternatives in the same cluster are more similar to each other compared to those in other clusters. However, the grouping is done solely based on the distances between the objects that aim to measure the similarities without establishing priority relations. This intrinsic feature of the object is of quite poor information and is often inappropriate when extrinsic information is available in the form of (for example) the preference function of the DM. Therefore, multi-criteria relational clustering approaches have been proposed to solve this problem (De Smet and Guzmán, 2004).

While classification and sorting also utilise the preference information as input, the corresponding literature has concentrated efforts on non-hierarchical multi-criteria clustering, which refers to the assignment of decision alternatives into predefined groups. However, the number of clusters may not be known in advance when one takes an exploratory analysis route. Towards a solution, De Smet (2014) proposes a hierarchical clustering, where all objects are assigned into a single cluster. This cluster is then further split recursively (as one goes down the hierarchy) into more clusters, based on their positive and negative net flows. These authors build their approach on the preference matrix generated through PROMETHEE, and hence the output is sensitive to the parameter settings. However, there is a plethora of situations where the parameters may be unknown (e.g. the choice of a representative DM is infeasible) or uncertain (e.g. the DM is not able to give a precise value).

In contrast from the existing studies, in this paper we design a new hierarchical clustering algorithm that simulates the feasible space of potential combinations in a stochastic environment (see e.g. Corrente et al., 2014, proposing the fusion of SMAA and PROMETHEE). Because a large number of preference matrices (each corresponding to a different combination) is gener-
ated, a large number of clustering solutions is obtained at the end of the process. To reach a consensus solution, we make use of cluster ensembles (Fred and Jain, 2005; Topchy et al., 2005; Iam-On et al., 2010). In fact, a cluster ensemble approach provides more robust and stable solutions across different datasets compared to a single clustering approach because multiple clustering solutions are combined into a consolidated single clustering solution (Ghosh and Acharya, 2011; Boongoen and Iam-On, 2018). Therefore, this paper proposes a new hierarchical clustering that is coupled with SMAA-PROMETHEE and cluster ensembles to eventually reach a consensus, final solution. To illustrate the applicability of this approach, we use the proposed method to cluster the performance of 208 US banks on financial and non-financial attributes, the latter complementing a long-forgotten agenda of banks that is starting to get traction in the financial markets (Tett, 2019).

The rest of this paper is structured as follows. Section 2 reviews the literature of multi-criteria clustering methods. Section 3 describes the newly proposed methodology. Section 4 presents the illustrative case study and compares the results to the current state of the art algorithm. Finally, Section 5 concludes the paper and discusses future avenues for research.

2 Literature Review

There are two broader types of grouping techniques: classification and clustering. Classification refers to a family of supervised techniques, which means that alternatives are assigned to a set of predefined classes. In MCDA, classification can be further distinguished between ordinal classification (sorting) and nominal classification. In nominal classification, classes are defined but no preference order amongst them exists (see for example Belacel, 2000; Furems, 2013; Costa et al., 2018, 2020). In ordinal classification, the classes are ordered and are defined by limiting or central profiles. Limiting profiles essentially separate classes in a form of interval, while central profiles play the role of a typical reference alternative of a given class. The classification is done via means of comparing the performance of the alternatives to that of the reference profiles that act as a representation of that very class. Several MCDA techniques have been developed to solve sorting problems; see, for example, UTADIS (Jacquet-Lagrèze, 1995), Electre-Tri (Yu, 1992), Electre-Tri-C (Almeida-Dias et al., 2010), Electre-Tri-nC (Almeida-Dias et al., 2012), FlowSort (Nemery and Lamboray, 2008), PromSort (Araz and Ozkarahan, 2007), Theseus method (Fernandez and Navarro, 2011), AHPSort (Ishizaka, 2012), MACBETHSort (Ishizaka and Gordon, 2017), DEASort (Ishizaka et al., 2018), ELECTRESort (Ishizaka and Nemery, 2014), VIKORSORT (Demir et al., 2018).

Turning to the second broad category of grouping techniques, clustering refers to a set of unsupervised techniques that group alternatives of similar calibre into the same object (called cluster). There are two families of clustering algorithms, described forthwith:

• Hierarchical clustering is a family of methods that seeks to build a hierarchy of clusters (Murtagh and Contreras, 2017). There are two broad approaches in hierarchical clustering:
the agglomerative (which is often called a bottom-up approach), according to which each alternative starts in its own cluster, and pairs of clusters are formed as one moves up the hierarchy until a final, single cluster is eventually formed; and divisive hierarchical clustering (which is often called a top-down approach), where all of the alternatives start in one single cluster, which is further split recursively as one moves down the hierarchy.

- **Non-hierarchical (or partitioning) clustering** is a family of methods according to which the alternatives are assigned into clusters after the DM has decided upon the number of clusters to be eventually formed (Saxena et al., 2017).

In MCDA, the idea behind clustering is to group alternatives that are similar, in the sense that alternatives belonging to the same cluster are not strongly preferred to each other. In other words, the metric distance used in the traditional clustering techniques is replaced by a ‘preference’ type of distance. However, only a few algorithms have been developed to solve the multi-criteria clustering problem. The first was presented by De Smet and Guzmán (2004) based on the partitioning clustering method $k$-means. The main difference is that the distance matrix is based on the binary preference relations between the alternatives. Essentially, two alternatives are considered to be similar if they are indifferent, incomparable or are preferred to the same alternatives. Consequently, an element is related to the others that determines in which cluster it is assigned to. This method is called relational multi-criteria clustering. Lidouh and De Smet (2016) applied this method on territory partitioning. The main difference in this application is that only neighbouring alternatives can be clustered together. In another extension, De Smet and Eppe (2009) identified the preference relationship between the clusters (i.e. preferred, indifferent, or incomparable). Then, Eppe et al. (2014) valued the distance based on their outranking relation and proposed an adapted $k$-means algorithm to cluster the alternatives. On the same note, Rouba and Bahloul (2014) proposed several similarity indices between alternatives and used the $k$-medoids algorithm to cluster the alternatives instead. Because different similarity indices yield different results, the ensemble clustering (which is a technique combining multiple clusters into a new final clustering) is used.

Based on the $k$-means idea, PROMETHEE CLUSTER (Figueira et al., 2004), P2CLUST (De Smet, 2013) and PCLUST (Sarrazin et al., 2018) have been developed. In all three cases, central profiles are randomly generated. Then, PROMETHEE II for PROMETHEE CLUSTER, FLOWSORT for P2CLUST and PROMETHEE I for PCLUST are used to assign the alternatives into clusters. The central profiles are updated and the procedure is repeated until the assignment does not change, exactly like in $k$-means clustering. In PROMETHEE CLUSTER, all clusters are assumed to be incomparable; in P2CLUST, they are completely ordered; and in PCLUST, they accept partial ordered ranking. Regardless of their properties, the result depends on the initial central profiles that are generated. To have the perfect partitioning, one would need to enumerate all possible partitionings and find the one that minimises the distance between alternatives of the same cluster and maximises the distance to alternatives.
from different clusters. As the number of possibilities grow exponentially, Meyer and Olteanu (2013) proposed to use a data mining technique to solve this problem.

De Smet et al. (2012) proposed an algorithm to build \( k \) ordered clusters based on a lexicographic comparison of inconsistency matrices. Rocha and Dias (2013) modified an agglomerative hierarchical clustering method to have a classification with partially ordered classes on the basis of a binary outranking matrix. In this case, the number of clusters is not known in advance.

A divisive hierarchical clustering has been proposed by De Smet (2014), which uses the simple idea of divisive alternatives with positive and negative net flows in different clusters. This is a rather interesting idea in this stream, yet it is simple in its assumptions because it is not based on any optimisation function (e.g. maximising the compactness of the cluster, minimising the dissimilarity of the cluster) attached to the mechanism. The following example shows how a non-optimal result could be obtained following the application of this algorithm.

**Example 1.** Consider the six actions to be divided into two clusters (Table 1).

<table>
<thead>
<tr>
<th>Actions</th>
<th>Criterion 1</th>
<th>Criterion 2</th>
<th>Net Flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>a₁</td>
<td>100</td>
<td>100</td>
<td>0.315</td>
</tr>
<tr>
<td>a₂</td>
<td>95</td>
<td>100</td>
<td>0.285</td>
</tr>
<tr>
<td>a₃</td>
<td>95</td>
<td>95</td>
<td>0.255</td>
</tr>
<tr>
<td>a₄</td>
<td>80</td>
<td>75</td>
<td>0.045</td>
</tr>
<tr>
<td>a₅</td>
<td>70</td>
<td>75</td>
<td>-0.015</td>
</tr>
<tr>
<td>a₆</td>
<td>0</td>
<td>0</td>
<td>-0.885</td>
</tr>
</tbody>
</table>

Table 1: Actions to be clustered and their net flow

If we rank them using PROMETHEE with equal weights for both criteria (whose values are considered such that the larger they are the better) and a linear function with indifference \((q)\) and preference \((p)\) thresholds \(q = 0\) and \(p = 100\), then we obtain the net flow given in the last column of Table 1. The algorithm of De Smet (2014) would form the two following clusters: \((a₁, a₂, a₃, a₄)\) with the positive net flows and \((a₅, a₆)\) with the negative ones. However, action \(a₅\) is arguably very far from action \(a₆\) and it would be certainly more reasonable to group it with actions \(a₁, a₂, a₃, a₄\). In this paper, we propose a new divisive hierarchical clustering to split clusters according to the preference distance.

### 3 Methodology

#### 3.1 The PROMETHEE methods

The PROMETHEE methods I & II (Brans and Vincke, 1985; Brans et al., 1986) have been broadly popular in the MCDA community, with a huge variety of applications crossing many
different disciplines (see e.g. Behzadian et al., 2010, for a review of the literature). This section discusses only parts of these methods required for the proposed algorithm. For a more comprehensive discussion of the whole PROMETHEE family of methods and their use in different circumstances, we refer the interested reader to the study of Brans and De Smet (2016).

Given a set of alternatives \( A = \{a_1, \ldots, a_n\} \) that will be evaluated based on criteria \( G = \{g_1, \ldots, g_m\} \), where \( g_j : A \to \mathbb{R}, j \in J = \{1, \ldots, m\} \), for each criterion \( g_j \in G \), PROMETHEE methods use a function \( P_j(a_i, a_{i'}) \), \( i \neq i' \) that shows the degree of preference of \( a_i \) over \( a_{i'} \) on criterion \( g_j \) being a non-decreasing function of \( d_j(a_i, a_{i'}) = g_j(a_i) - g_j(a_{i'}) \). The DM can choose from six functions for each criterion. However, for reasons of simplicity we will hereby use the commonly-used piece-wise linear function, which is defined as:

\[
P_j(a_i, a_{i'}) = \begin{cases} 0 & \text{if } d_j(a_i, a_{i'}) \leq q_j \\ \frac{d_j(a_i, a_{i'}) - q_j}{p_j - q_j} & \text{if } q_j < d_j(a_i, a_{i'}) < p_j \\ 1 & \text{if } d_j(a_i, a_{i'}) \geq p_j \end{cases}
\]

where \( q_j \) and \( p_j \) are the indifference and preference thresholds set by the DM for each criterion \( g_j \in G \). Each criterion \( g_j \) is assigned a weight \( w_j \), with \( w_j > 0 \) and \( \sum_{j=1}^{m} w_j = 1 \). For each pair of alternatives \((a_i, a_{i'}) \in A \times A\), PROMETHEE methods compute how much \( a_i \) is preferred over \( a_{i'} \) taking into account all criteria \( g_j \in G \), as follows:

\[
\pi(a_i, a_{i'}) = \sum_{j=1}^{m} w_j P_j(a_i, a_{i'}),
\]

with \( \pi(a_i, a_{i'}) \in [0, 1] \). Higher values show greater preference of \( a_i \) over \( a_{i'} \), and vice versa. Computing \( \pi(a_i, a_{i'}) \) for every \( i \) and \( i' \) forms the \( n \times n \) matrix \( \Pi \) (Equation 3), the diagonal of which is equal to zero, and shows how much an alternative \( a_i \) in row \( i \) is preferred to its counterpart \( a_{i'} \) in column \( i' \).

\[
\Pi = \begin{pmatrix} \pi(a_1, a_1) & \pi(a_1, a_2) & \ldots & \pi(a_1, a_n) \\ \pi(a_2, a_1) & \pi(a_2, a_2) & \ldots & \pi(a_2, a_n) \\ \vdots & \vdots & \ddots & \vdots \\ \pi(a_n, a_1) & \pi(a_n, a_2) & \ldots & \pi(a_n, a_n) \end{pmatrix}.
\]

To evaluate each action with respect to all the other actions, two scores are computed (Brans and Vincke, 1985):

- The positive flow: \( \phi^+(a_i) = \frac{1}{n-1} \sum_{a_{i'} \in A} \pi(a_i, a_{i'}) \);
- The negative flow: \( \phi^-(a_i) = \frac{1}{n-1} \sum_{a_{i'} \in A} \pi(a_{i'}, a_i) \).
The positive and negative preference flows are then aggregated into the net preference flow:
\[ \phi(a_i) = \phi^+(a_i) - \phi^-(a_i). \]
In our case, we will use the matrix \( \Pi \) to cluster the alternatives, which will be explained in the following sections.

### 3.2 Divisive hierarchical clustering

The aim of multi-criteria clustering is to group actions that have similar preferences. In our newly proposed algorithm, the preference similarity is given by the preference matrix \( \Pi \) in Equation 3. The preference degree may take values between 0 and 1, where 1 means that an action \( a_i \) is absolutely preferred over \( a_j \). Because the matrix \( \Pi \) is not symmetric, we choose the highest preference degree between the two pairwise comparisons:
\[
\max\{\pi(a_i, a'_i), \pi(a'_i, a_i)\}.
\]

As in any divisive hierarchical clustering algorithm, all objects are first included in a single large cluster. At each iteration, a cluster is further divided into two. The key point is the rule of how one divides or splits the cluster. Our algorithm proposes to find those two actions, \( a_u, a_v \in A \), that are the most opposed (i.e. \( \pi(a_u, a_v) = \max_{a_i, a_i' \in A} \{\pi(a_i, a_i'), \pi(a_i', a_i)\} \)) to form the centroids of the two newly-formed clusters. If there are two or more equal highest preference degrees, then one can chose randomly or the first one found can be arbitrarily selected (as is done here for reasons of simplicity).

To allocate the remaining \( n - 2 \) alternatives \( a_i \in A, i \neq u, v \), we use the preference degree as a comparison with the two cluster centroids \( a_u \) and \( a_v \). In particular, the alternatives are assigned to that cluster with the closest preference distance to the centroid. This is accomplished by comparing the value of the preference degree between the action to classify and the centroid. Because the matrix is not symmetric, we need to take the highest value of the preference degree, i.e. \( \max\{\pi(a_i, a_u), \pi(a_u, a_i)\} \) and \( \max\{\pi(a_i, a_v), \pi(a_v, a_i)\} \). The action is then allocated to the cluster with the smallest preference degree. In the case of equality, we look the symmetric preference degree value; that is, we take the \( \min\{\pi(a_i, a_u), \pi(a_u, a_i)\} \) and the \( \min\{\pi(a_i, a_v), \pi(a_v, a_i)\} \). Once again, the cluster is then allocated to the cluster with the smallest preference degree. If the equality in the preference degree persists, then a random procedure can be used or the less populated cluster receives the action. These two new created clusters are then divided into two new clusters in the same way. The division procedure is then repeated until only one action exists in each cluster or when a stopping condition is met, such as when a preference degree threshold is met.
Algorithm 1 Multicriteria divisive hierarchical clustering algorithm

Input: Alternatives $A = \{a_1, \ldots, a_n\}$, preference threshold value $\varepsilon \in [0,1]$ and preference matrix $\Pi_{n \times n}$.

Output: Cluster identification matrix $D_{n \times 1}$ and the number of clusters $k$.

**Step 1.** Initially set the number of clusters to 1 $k = 1$ and allocate all alternatives into this cluster, $D(a_i) = 1$ for all $a_i \in A$. Go to Step 2.

**Step 2.** Find the highest preference degree between the alternatives in the same cluster: $x = \max \{\pi(a_i, a_{i'}) : a_i, a_{i'} \in A, D(a_i) = D(a_{i'}), i \neq i'\}$. If $x \geq \varepsilon$, then go to Step 3 to apply a division procedure. Otherwise, Stop.

**Step 3.** To form the centroids of the two newly-formed clusters, find the two actions, $a_u$ and $a_v$, that are currently in the same cluster $D(a_u) = D(a_v)$ but the most opposed to each other: $\pi(a_u, a_v) = x$. If there are two or more equal highest preference degrees, then choose randomly. Go to Step 4.

**Step 4.** Define a new cluster $k \leftarrow k + 1$ with a centroid $a_v$: $D(a_v) = k$. Allocate the remaining alternatives $a_i \ i \neq u, v$ that are originally in the same cluster $D(a_i) = D(a_u)$ using the preference degree as a comparison with the two cluster centroids:

- If $\max \{\pi(a_i, a_u), \pi(a_v, a_i)\} < \max \{\pi(a_i, a_u), \pi(a_u, a_i)\}$, then $a_i$ is allocated to the cluster of $a_v$, $D(a_i) = D(a_v)$.
- Otherwise, if $\max \{\pi(a_i, a_u), \pi(a_v, a_i)\} > \max \{\pi(a_i, a_u), \pi(a_u, a_i)\}$, $a_i$ is allocated to the cluster of $a_u$, $D(a_i) = D(a_u)$.

If $\max \{\pi(a_i, a_u), \pi(a_v, a_i)\} = \max \{\pi(a_i, a_u), \pi(a_u, a_i)\}$, allocate $a_i$ randomly to the same cluster with $a_u$ or $a_v$: $D(a_i) = D(a_u)$ or $D(a_i) = D(a_v)$.

Go to Step 2.

We will use Example 1 to illustrate the algorithm. Let preference threshold value be $\varepsilon = 0.5$.

From the data of the Table 1, we can calculate the preference degrees given in matrix II.

$$
\Pi = \begin{pmatrix}
0 & 0.025 & 0.05 & 0.225 & 0.275 & 1 \\
0 & 0 & 0.025 & 0.2 & 0.25 & 0.975 \\
0 & 0 & 0 & 0.175 & 0.225 & 0.95 \\
0 & 0 & 0 & 0 & 0.05 & 0.775 \\
0 & 0 & 0 & 0 & 0 & 0.725 \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.
$$

In Step 1, we initially place all actions in the same cluster and set the number of clusters to $k = 1$. In Step 2, we find the highest preference degree $x = \max_{a_i, a_{i'} \in A} \{\pi(a_i, a_{i'}), \pi(a_{i'}, a_i)\}$, which in our case is equal to 1. Because $x \geq \varepsilon = 0.5$, we go to Step 3, define a new cluster $k = 2$ and apply the division procedure. Because $\pi(a_1, a_6) = 1$, actions $a_1$ and $a_6$ are the two most far apart and are the base of the two centroids of the two clusters. While $a_1$ and $a_6$ are placed into clusters 1 and 2, respectively, and we allocate the remaining actions as explained in Step 4:
• \( a_2 \) is assigned to the first cluster with \( a_1 \), \( D(a_2) = 1 \) because \( \max\{\pi(a_2, a_1), \pi(a_1, a_2)\} = 0.025 < \max\{\pi(a_2, a_6), \pi(a_6, a_2)\} = 0.975. \)

• \( a_3 \) is assigned to the first cluster with \( a_1 \), \( D(a_3) = 1 \) because \( \max\{\pi(a_3, a_1), \pi(a_1, a_3)\} = 0.05 < \max\{\pi(a_3, a_6), \pi(a_6, a_3)\} = 0.95. \)

• \( a_4 \) is assigned to the first cluster with \( a_1 \), \( D(a_4) = 1 \) because \( \max\{\pi(a_4, a_1), \pi(a_1, a_4)\} = 0.225 < \max\{\pi(a_4, a_6), \pi(a_6, a_4)\} = 0.775. \)

• \( a_5 \) is assigned to the first cluster with \( a_1 \), \( D(a_5) = 1 \) because \( \max\{\pi(a_5, a_1), \pi(a_1, a_5)\} = 0.275 < \max\{\pi(a_5, a_6), \pi(a_6, a_5)\} = 0.725. \)

Then, we go to Step 2 and find the highest preference degree between the alternatives in the two newly-formed clusters, \( x = \max\{\pi(a_i, a_i'): a_i, a_i' \in A, D(a_i) = D(a_i'), i \neq i'\} = \pi(a_1, a_5) = 0.275. \) Because \( x \leq \varepsilon = 0.5 \), the algorithm stops. The first cluster is composed of \( a_1, a_2, a_3, a_4, a_5 \), while the second cluster is composed of \( a_6 \). This separation seems more logical than the one proposed by the algorithm of De Smet (2014) in Example 1.

As with other MCDA methods, PROMETHEE is unavoidably subjective because the DM needs to specify the weights, the indifference and the preference thresholds for each criterion. The DM is sometimes unable to specify exactly these values because they are unsure that information might be missing or the data are imprecise, and so on. Sometimes, one is confronted with a situation where there is no single DM but a large number of DMs with heterogeneous preferences. Several uncertainties are linked to the determination of the parameters and the clustering depends on these specific parameters. To solve this problem, the Stochastic Multiobjective Acceptability Analysis (SMAA) (Lahdelma et al., 1998; Lahdelma and Salminen, 2001; Pelissari et al., 2020) is used to take into account the full spectrum of solutions, as explained in the next section.

### 3.3 The SMAA-PROMETHEE method

The SMAA-PROMETHEE method was proposed by Corrente et al. (2014) to deal with uncertainty and imprecision in real world decision-making situations. Similarly to the PROMETHEE methods, SMAA has a wide gamut of applications (see e.g. Pelissari et al., 2020, for a recent review of the literature), thus their fusion offers a successful method for the PROMETHEE family to deal with uncertainties and imprecision.

This is achieved by considering distributions on the parameters where uncertainty may arise. These could regard, for instance, the ill-determination of weights \( (w_i) \), the indifference \( (q_j) \) or preference \( (p_j) \) thresholds or even the criteria \( (g_j) \). Handling of the distributions \( f_w, f_q \) and \( f_p \) relates to whether the DM can provide information to shape these distributions (i.e. any linear inequalities, thresholds etc.) or not. In the latter case, these distributions are uniformly ranging in the feasible spaces \( W, Q \) and \( P \), which are accordingly defined as:
\[ W = \left\{ \mathbf{w} = [w_1, \ldots, w_m] : w_j > 0, j = 1, \ldots, m, \sum_{j=1}^{m} w_j = 1 \right\} , \]

\[ P = \{ \mathbf{p} = [p_1, \ldots, p_m] : \min d_j \leq p_j \leq \max d_j, j = 1, \ldots, m \} , \text{ and} \]

\[ Q = \{ \mathbf{q} = [q_1, \ldots, q_m] : 0 \leq q_j \leq p_j, j = 1, \ldots, m \} . \]

If, however, the DM is able to provide any kind of information (e.g. assurance regions (see e.g. Allen et al., 1997; Allen and Thanassoulis, 2004)) in certain or all the indicators, or a more appropriate distribution than the uniform can be used, then the spaces \( W, Q \) and \( P \) can reasonably be adjusted accordingly.

Understandably, in the real world, the use of Monte Carlo simulations simplifies the computational procedure of Equations (1) and (2) (compared to numerical integration) in the aforementioned spaces \( W \), \( Q \) and \( P \). In particular, assuming the following \( m \times s \) matrices:

\[
\mathbf{RW}_{m \times s} = \begin{pmatrix}
w_{11} & w_{12} & \cdots & w_{1s} \\
w_{21} & w_{22} & \cdots & w_{2s} \\
\vdots & \vdots & \ddots & \vdots \\
w_{m1} & w_{m2} & \cdots & w_{ms}
\end{pmatrix}, \quad
\mathbf{RP}_{m \times s} = \begin{pmatrix}
p_{11} & p_{12} & \cdots & p_{1s} \\
p_{21} & p_{22} & \cdots & p_{2s} \\
\vdots & \vdots & \ddots & \vdots \\
p_{m1} & p_{m2} & \cdots & p_{ms}
\end{pmatrix}, \quad
\mathbf{RQ}_{m \times s} = \begin{pmatrix}
q_{11} & q_{12} & \cdots & q_{1s} \\
q_{21} & q_{22} & \cdots & q_{2s} \\
\vdots & \vdots & \ddots & \vdots \\
q_{m1} & q_{m2} & \cdots & q_{ms}
\end{pmatrix},
\]

where \( mc = 1, \ldots, s \) denotes the \( mc \)th Monte Carlo simulation, collect the sets of preferences that summarise the spaces \( W \), \( Q \) and \( P \) in an allegedly representative sample. Then, \( s \times n \times n \) matrices \( \Pi_{mc} \), \( mc = 1, \ldots, s \) can be formed by computation of Equations 1 and 2 for every set of preferences, as follows:

\[
\Pi_{mc} = \begin{pmatrix}
\pi_{mc}(a_1, a_1), & \pi_{mc}(a_1, a_2), & \ldots & \pi_{mc}(a_1, a_n) \\
\pi_{mc}(a_2, a_1), & \pi_{mc}(a_2, a_2), & \ldots & \pi_{mc}(a_2, a_n) \\
\vdots & \vdots & \ddots & \vdots \\
\pi_{mc}(a_n, a_1), & \pi_{mc}(a_n, a_2), & \ldots & \pi_{mc}(a_n, a_n)
\end{pmatrix}, \quad (4)
\]

These matrices \( \Pi_{mc} \) are input to the clustering algorithm. By changing the values of thresholds and weight parameters, SMAA - PROMETHEE generates \( s \) different preference matrices, each of which may result in a different clustering of alternatives. Because we have several different solutions, we need an algorithm to combine them into a final consensus solution. This is done with ensemble clustering.

\footnote{While there exists no standard benchmark in the choice of an appropriate number of simulations (i.e. parameter \( s \)), Tervonen and Lahdelma (2007) suggest a value of \( s = 10,000 \) to provide for robust results.}
3.4 Consensus Generation Method

Ensemble Clustering (EC) refers to the process of combining multiple clustering solutions into a final consensus solution. Given a library of $s$ clustering solutions $L = \{C^1, \ldots, C^s\}$, where each solution has a different number of $k^{mc}$ clusters, EC applies a consensus function $\Phi(L, k^*)$ to generate a final clustering solution $C^*$ with $k^*$ clusters. In this approach, we define $k^*$ value based on the library and set it to the $k$ number that occurs most frequently in the set. Then, Hybrid Bipartite Graph Formulation (HBGF)\footnote{Source Code is available at \url{https://web.engr.oregonstate.edu/~xfern/HBGF_spec.m}} of Fern and Brodley (2004) is used to reduce the cluster ensemble problem to a graph partitioning problem.

**Description of HBGF.** Given a cluster ensemble $L = \{C^1, \ldots, C^s\}$, HBGF constructs a weighted graph $G = (V, \lambda)$ where $V = V^L \cup V^A$ and:

- $V^L$ is a set of $t$ vertices, each representing a cluster of the ensemble (i.e. $t = \sum_{mc=1}^s k^{mc}$), and $V^A$ is a set of $n$ vertices each representing an alternative in $A$.

- $\lambda$ is a symmetric matrix such that $\lambda(i, c) = \lambda(c, i) = 1$ if alternative $i$ belongs to cluster $c$. Otherwise, $\lambda(i, c) = \lambda(c, i) = 0$. If the vertices $i$ and $c$ are both clusters or both alternatives, $\lambda(i, c) = 0$.

Note that HBGF builds a matrix $Z$ that connects the alternatives to the clusters and $\lambda$ can be written as $\lambda = \begin{pmatrix} 0 & Z^T \\ Z & 0 \end{pmatrix}$. If alternative $i$ is in the cluster $c$, then the entry $Z(i, c)$ takes the value of 1, and 0 otherwise.

HBGF then partitions the graph into $k^*$ disjoint clusters of vertices using the well-appreciated spectral partitioning algorithm (SPEC) of Ng et al. (2002). Given a weighted graph $G = (V, \lambda)$ and a number $k^*$, SPEC partitions the graph into $k^*$ disjoint clusters of vertices $C^* = \{c^*_1, c^*_2, \ldots, c^*_k\}$, where $\cup_k c^*_k = V$. The partition of the alternatives can then be used as the final clustering. The goal in partitioning algorithms is to minimize the sum of the weights of the edges crossed by the partition, this is called the cut: $\sum \lambda(i, c)$ where $i$ and $c$ belong to the different clusters of vertices in the final clustering. SPEC uses the normalized cut criterion that measures both the total dissimilarity between the different groups as well as the total similarity within the groups (Shi and Malik, 2000). SPEC first constructs a diagonal degree matrix $D$ such that $D(i, i) = \sum_c \lambda(i, c)$ and computes a normalized weight matrix $\lambda' = D^{-1} \lambda$. Based on the $k^*$ largest eigenvectors of $\lambda'$, SPEC forms a new matrix $U$ by stacking the eigenvectors in columns and normalizing each row to have unit length. Each row of $U$ is then treated as a point and clustered into $k^*$ clusters using well-known $k$-means algorithm (Steinhaus, 1956; MacQueen et al., 1967). Finally, the alternative $i$ is assigned to cluster $c^*_k$ if and only if row $i$ of the matrix $U$ was assigned to cluster $c^*_k$.

Figure 1 shows how HBGF reduces the EC problem to graph partitioning problem. Given two different clustering solutions shown in Figures 1a and 1b, HBGF constructs the bipartite graph $G = (V, \lambda)$ in Figure 1c where:
Figure 1: An example of the Hybrid Bipartite Graph Formulation

- \( V = V^L \cup V^A \) where \( V^L = \{c_{11}, c_{12}, c_{21}, c_{22}\} \) and \( V^A = \{a_1, \ldots, a_9\} \).

- \( \lambda = \begin{bmatrix} Z^T & 0 \end{bmatrix} \) where \( Z = \begin{bmatrix} c_{11} & c_{12} & c_{21} & c_{22} \\ a_1 & 1 & 0 & 1 & 0 \\ a_2 & 1 & 0 & 1 & 0 \\ a_3 & 1 & 0 & 1 & 0 \\ a_4 & 1 & 0 & 1 & 0 \\ a_5 & 0 & 1 & 1 & 0 \\ a_6 & 0 & 1 & 1 & 0 \\ a_7 & 0 & 1 & 0 & 1 \\ a_8 & 0 & 1 & 0 & 1 \\ a_9 & 0 & 1 & 0 & 1 \end{bmatrix} \).}

In Figure 1c, the clusters and alternatives are represented by square and circle vertices, respectively, and zero-weight edges are omitted. Each edge connecting an alternative to cluster indicates that the alternative belongs to the cluster and has a weight of one. The dashed line in Figure 1c shows the cut and partitioning of the alternatives in \( V^A \) into \( k^* = 2 \) clusters that yields the consensus clustering solution, \( C^* = \{(a_1, a_2, a_3, a_4, a_5), (a_6, a_7, a_8, a_9)\} \).

4 Case Study: Clustering U.S. Banks beyond the CAMELS

This section presents an illustrative case study to which the proposed approach is applied. The case study regards a clustering exercise on 208 US banks based on a set of financial (CAMEL) and non-financial (ESG) criteria for the year 2018. Although this case study mainly serves as an illustration, it still presents a novelty for two reasons. First, from an application
viewpoint, clustering may be important for executive teams to look at their peers, and compare and contrast their weaknesses and strengths on the basis of a multiplicity of criteria. This could also be worthwhile for investors who can find similarly-performing (in terms of the considered criteria) banks and pick those from within the cluster that offer better premiums. Second, from a conceptual viewpoint, clustering complements standard financial factors with non-financial ones that are often found missing due to data unavailability in studies considering the CAMEL (for banks, or similar in principle for non-financial firms) assessment criteria in classification studies on credit ratings (Doumpos and Figueira, 2019; Gaganis et al., 2020). Additionally, this dataset permits us to observe whether certain banks have already adopted a more ‘inclusive’ performance that complements the classic financial maximisation agenda with a more socio-environmental agenda.

In what follows, Section 4.1 introduces the importance behind the latter argument and what this illustrative case study aims to achieve. Section 4.2 describes the data, its sources and how parameters of the proposed method are modelled. Finally, Section 4.3 discusses the results and it reports on a sensitivity analysis.

4.1 Introduction

“There are discussions of ESG [an acronym for ‘Environmental, Social and Governance’] matters have intensified in recent years as banks, investors, regulatory bodies, businesses, NGOs, and governments become more attuned to evaluating investments and corporate performance through the lens of ESG impact. There is a rapidly growing recognition that value extends beyond pure financial returns.”

Deloitte, (2019)

There is a great debate in the corporate finance world about the ultimate objectives of entities in maximising value, which can be summarised in a single question: “to whom are corporations accountable?” (Ferrel et al., 2016, p.585). Elaborating on this question, there are two sides of this argument. On the one hand, Adolf A. Berle argues that the management of a corporation should be held accountable only to the shareholders, putting at the forefront the Friedman doctrine of profit maximisation. On the other hand, Merrick Dodd is in favour of adding society as an additional judge (Macintosh, 1999). There is a large body of literature extending the latter argument and the impacts that it entails (see Ferrel et al., 2016, for a discussion), although social responsibility and sustainable aspects in the corporate world can be reasonably argued to be on the rise (Burns, 2019).

As far as the entities’ outlook is concerned, ESG could be in some cases used as a signaling mechanism for socially responsible investors. For instance, a recent viewpoint by PIMCO (Del Andeson, 2017), a global investment management firm based in the United States, reports

---

3As ‘inclusive’ we define the analysis that takes into account socio-environmental measures to complement the sheer economic output analysis, which is a term that was used by Stiglitz et al. (2009) to criticise the singleton focus on economic output (GDP) instead of a more socio-environmental inclusive performance that complements the former.
that the global market outlook changes because analysts’ ESG views have the potential to affect investment performance, and therefore ESG compliant firms could have an edge. This happens because a portfolio manager could be confronted with entities with similar fundamental risk profiles, and therefore the choice will be made according to which entity poses a relatively higher ESG profile. While this could seem far-fetched for those sharing the Friedman doctrine of pure profit maximisation, money market fund managers have significantly increased investments in ESG money market funds in recent times (Nauman, 2019), and this trend seems to be quickly catching up around the global financial markets (Tett, 2019).

Non-financial aspects related to sustainability could be particularly important for banks (and financial entities) given the downturn of their public trust following the global financial crisis. As Del Andeson (2017) adds, “investors feel the banking industry’s reputation has been tarnished in recent years following the many high-profile breakdowns in governance and breaches of public trust. [...] we do not believe that these past transgressions should be overwhelming factors in the forward-looking ESG assessment of individual banks, particularly for banks that performed well through the financial crisis or banks that have wholly revamped their management teams and governance processes”.

Understandably, social responsibility might come at a cost for entities and, according to the literature, seminal studies on the topic can be divided into three general views, which are abridged as follows (see McGuire et al., 1988 for a first detailed discussion of the seminal studies’ views categorisation). The first view argues that corporations face a trade-off focusing both on financial and non-financial matters because the latter incur costs that put corporations striving for responsibility in a disadvantaged place to their counterparts that do not (Aupperle, Carroll and Hatfield, 1985; Ullmann, 1985; Vance, 1975). The second view states that this is dismissed because these costs are minimal and counterbalanced through better image, employee morale and productivity (Moskowitz, 1972; Parket and Eibert, 1975; Soloman and Hansen, 1985). The third view resembles the second view and argues that while the costs of better social performance are significant indeed, they are offset through a reduction in other firm costs (McGuire et al., 1988).

Regardless of the view that one shares, ESG can be regarded as a critical challenge that should be placed on top of managers’ financial responsibilities, related to corporate sustainability (Epstein et al., 2015). In fact, as the authors argue, this can be a significant task because the managers must make resource allocation trade-offs to meet these multiple goals. In our case study, we do not strive to make any connections and/or trade-offs and prove any causal effects running from non-financial aspects to financial aspects, or vice versa. Rather, by obtaining a sample of US banks, we aim to find clusters of all-round performers (in both financial and non-financial attributes) and look at their typical descriptive profiles. In our additional analysis, we compare the classification of banks’ performance depending on the notion that one shares (i.e. financial or inclusive).
4.2 Data sources, description and modelling of parameters

The overall performance of the US banks in this study will consist of a hierarchical structure that is split among financial and non-financial aspects. In the following, we give more details for each aspect, the reasons for the choices made in each, and the sources of data collection.

The financial evaluation is based on the key themes of the CAMEL rating framework, which is formally known as the Uniform Financial Rating System (UFIRS) and was introduced by US regulators in 1979. The CAMEL framework is an integral part of the assessment process carried out by central banks and regulatory bodies. Its key role is the assessment of banks’ financial conditions in the areas of capital adequacy (C), asset quality (A), management (M), earnings (E) and liquidity (L). The key output of this framework is a synthetic rating between 1 and 5, which evaluates the soundness of a bank (i.e. 1-2: low weakness, 3-4: moderate to severe weakness, 5: critical weakness). These ratings are not publicly disclosed and only the upper echelons of a bank’s management aware of the exact ratings. Therefore, prior literature suggests a gamut of financial ratios acting as proxies for the above-mentioned categories (see e.g. Berger et al. 1998; 2001; Doumpos and Zopounidis, 2010; Cole et al., 2012). In this study, we closely follow Doumpos and Zopounidis (2010). The authors provide a list of CAMEL attributes used in their case study evaluating the Greek banks, chosen in co-operation with expert analysts of the Bank of Greece, which is the key body responsible for monitoring and evaluating the performance of these banks. For reasons of simplicity and data availability, we only choose one attribute per sub-dimension, which are reported in Table 1.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Sub-dimension</th>
<th>Indicator</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial</td>
<td>Capital</td>
<td>Capital adequacy ratio</td>
<td>Authors’ elaboration on S&amp;P's Market Intelligence Platform data</td>
</tr>
<tr>
<td></td>
<td>Assets</td>
<td>Risk-weighted assets</td>
<td>Total Assets (TA)</td>
</tr>
<tr>
<td></td>
<td>Management</td>
<td>(Operating) Expenses</td>
<td>(Operating) Income</td>
</tr>
<tr>
<td></td>
<td>Earnings</td>
<td>Net income</td>
<td>Total Assets (TA)</td>
</tr>
<tr>
<td></td>
<td>Liquidity</td>
<td>Cash</td>
<td>Total Assets (TA)</td>
</tr>
<tr>
<td>Environmental</td>
<td>Resource reduction</td>
<td>Emission reduction</td>
<td>Environmental innovation</td>
</tr>
<tr>
<td>Non-Financial</td>
<td>Workforce quality &amp; diversity</td>
<td>Human rights</td>
<td>Community</td>
</tr>
<tr>
<td></td>
<td>Social</td>
<td>Product responsibility</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Corporate Governance</td>
<td>Board Governance &amp; Structure</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Governance</td>
<td>Shareholders rights</td>
<td>CSR Strategy</td>
</tr>
</tbody>
</table>

Table 2: Evaluating US banks according to financial and non-financial factors

4We opt for a US sample for two reasons: first, including institutions cross-country could introduce heterogeneity that could be subject to different regulations and macro fundamentals; and second, data on banking institutions in the United States had the highest overall availability.
For the non-financial aspects of the banks’ evaluation, we include the three key themes of an idealistic society’s civil agenda for banks’ take on sustainability: the ‘ESG’ framework (McCormick, 2011). We obtain ESG data for the banks in our sample from Refinitiv’s (formerly known as ‘DataStream’) ASSET4 database. Thomson Reuters, the host of the database, collects data points on over 400 metrics from companies’ annual reports, 178 of which were chosen after consultation with analysts and clientèle to form the three major categories (E, S and G), which are comprised of three, four and three pillars each, respectively (see Table 1). The bottom layer of 178 indicators comprise four layers of hierarchy. Thus, including all the original indicators would over-complicate this case study, whose purpose is to serve as an illustrative example of our proposed method. Therefore, we only obtain the upper level pillar composite indicator scores (shown as ‘indicators’ in Table 1) provided by Refinitiv (constructed as equal weights composite indicators) and range in the $[0, 100]$ space, with higher values denoting better performance, and vice versa.

Abridging the indicators’ meaning, starting with the environmental dimension, the three key composite indicators score an entity’s efforts to reduce the resources that it uses (resource reduction), reduce its emissions (emission reduction) and be innovative with respect to environmentally related issues (environmental innovation). These three indicators contain aspects such as whether an entity has policies and committees dedicated to reduce their environmental resources and emissions, and encourage environmental innovation (e.g. through turning to, promoting or investing in clean energy solution products). Looking at the social dimension of ESG, an entity’s score is judged on four pillars, including workforce quality and diversity (e.g. staff training, equal opportunities between genders, and across peoples’ needs, welcoming employees’ disabilities flexible working schemes etc.), community (e.g. good citizenship in many forms, including fair competition, awards and recognition from local surroundings to supporting the community), and product responsibility (e.g. policies about protecting customer health and safety, personal data, and general integrity and privacy). The last sub-dimension is corporate governance, which has three pillars: board governance and structure (e.g. policies ensuring accurate reporting, ratios of internal to independent committee members, number of meetings, % of females on board, CEO duality and compensation linked to executives’ targets), shareholders’ rights (e.g. anti-takeover defences, voting rights, election of board members with majority, as well as policies for minority shareholder protection etc.), and CSR strategy (e.g. having a committee dedicated to CSR actions, publishing related reports and having external auditors). More information on the elementary indicators can be found in Refinitiv’s dedicated website or database.

For the modelling of the parameters in the simulation, we formed a hierarchy of criteria, starting from a division between financial and non-financial aspects and following down the hierarchy into the five CAMEL components (in the first branch) and the three ESG pillars (in the second branch) (see Table 1). In the simulation environment, 10,000 weight vectors were drawn from a uniform distribution to decide upon the weights of financial and non-financial
dimensions, while simultaneously 10,000 weight vectors were randomly drawn to decide how the financial dimension’s weights are divided into the five CAMEL components, and how the non-financial dimensions are divided into the three ESG components. Finally, the three ESG components’ weights were randomly split between the three, four and three indicators of these sub-dimensions, respectively.

For reasons of simplicity, we chose the piece-wise linear PROMETHEE function (see Equation 1) for all criteria. Meanwhile, we kept the indifference \(q_j\) and preference thresholds \(p_j\) set to zero and the max of pairwise differences in \(J\) accordingly to be the constant thresholds in all 10,000 simulations.\(^5\)

4.3 Results and Discussion

This section presents the results that we obtained by using the hierarchical divisive clustering described in Section 3.2 upon the dataset described in Section 4.2. For our baseline results, we use the threshold \(\varepsilon = 0.35\).\(^6\) To define the number of clusters, we looked at the obtained results in the 10,000 simulations. The number of clusters ranged between two and seven, with the majority of simulations pointing to four clusters. A histogram is given in Figure 2. Seemingly, out of the 10,000 simulations, two clusters were found in 6.3% (630) of them, three clusters in 29.4% (2,940), four clusters in 56.8% of the simulations (5,680), five clusters in 6.8% (680), six clusters in 0.5% (50) and finally seven clusters in only 0.2% of the simulations (20).

Because the large majority of simulations (56.8%) pointed to a value of four clusters, we used this number in the HBGF approach. The 208 US banks were thus partitioned into four clusters, with 113 banks being assigned to \(c_1\), 74 to \(c_2\), 11 to \(c_3\) and 10 banks being assigned to \(c_4\).\(^7\)

Starting with some descriptive statistics to picture the typical bank in each cluster, the average net flow in the 10,000 simulations, the total assets (in mil. USD) and the year since the bank was established are shown in Figure 3. It appears that there is an order among the four clusters, with the typical bank in \(c_4\) being older, having considerably larger total assets and outperforming the typical banks of the remaining three classes.

Looking at the financial criteria, the differences between clusters are noticeable. However, there is no monotonic relations for three of the criteria (i.e., risk weighted assets, expenses and net income). Figures 4-7 show the averages of each cluster (with their 95% confidence intervals) in the elementary criteria used in the evaluation and regard financial (CAMEL) and non-financial (environmental, social and corporate governance) criteria accordingly.

Regarding the financial indicators (Figure 4), the banks in cluster 4 are more liquid and well-capitalised -compared to banks in the other clusters, which permits them to absorb losses

\(^5\)In further simulations taking into account also the parameters in question (unreported results for brevity), we find no significant differences to report in terms of cluster count or rearrangements between the alternatives in the chosen clusters.

\(^6\)Note at this point that this can be seen as an arbitrary choice of a DM that we later on (in this Section) test the robustness and propose a different alternative procedure.

\(^7\)For a full list of the banks belonging to each class, see the online supplementary appendix.
in times of distress. This is probably due to the fact that they have about 30% less risk-weighted assets on average. Additionally, they are (on average) less profitable than banks in Clusters 3 and 2, and less efficient in managing their expenses. When it comes to their performance in environmental-related criteria, the banks in Cluster 4 clearly are preferred to (see Figure 5) their counterparts in the other three clusters by a huge margin. Similar reasoning holds for social (Figure 6) and governance-related (Figure 7) criteria, aside from the board score.

The banks in Cluster 3 tend to perform similar to, or marginally better than banks in Cluster 2, but they are generally much more volatile. The criteria in which they clearly excel are liquidity, profitability and cost effectiveness (financial); emissions-free (environmental) and workforce conditions (social). In particular, they are better-equipped when it comes to holding capital for times in distress, more effective in handling their expenses, more profitable and more liquid. However, pronounced differences can be seen for environmental-related criteria, while a fuzzier picture appears for social and corporate governance-related differences.

To see the robustness of these results, we ran a sensitivity analysis, where we modify the stopping threshold (i.e. $\varepsilon = 0.35$ originally given by the DM) in the hierarchical clustering. Figure 8 shows how the distribution of the number of clusters that are found in the simulation environment changes when the threshold value, $\varepsilon$, changes accordingly. If the threshold is shifted to a stricter $\varepsilon = 0.40$, then the number of clusters peaks at $k = 2$ in just under 70% of
the 10,000 simulations. Meanwhile, if the threshold is decreased (i.e. become more ‘loose’), then the number of clusters increases accordingly. For a small decrease from 0.35 to 0.30, there are still four clusters in almost 60% of the simulations with only six banks out of the 208 (2.88%) moving to a different cluster when the ensemble clustering is done, and in particular moving from cluster 2 to cluster 3.

When the stopping threshold is decreased to 0.25, the distribution becomes flatter and the number of clusters peaks at a formation of six clusters. By decreasing the threshold even further to 0.20, the highest probabilistic number of clusters becomes 10. Understandably, this boils down to how strict or loose a DM’s preferences are. One way to handle this information follows. The DM could either start with an $\varepsilon$ value in mind, which they fine-tune to test the robustness of and then accept the clusters that are formed. For example, this involves adjusting the value to a reasonable extent, in the sense of finding an adequate % of simulations backing the number of formed clusters. Otherwise, the DM could start with the number of clusters that they would ideally like to form, and they then adjust the value until a desirable % of simulations supports this number. Of course, the latter option goes on a non-hierarchical multi-criteria clustering avenue that extends beyond the scope of this proposal. However, to illustrate this point, consider that the DM wants to form four clusters out of the 208 US banks in this sample. By looking at the distributions obtained through different $\varepsilon$ values, the DM could end up with any $\varepsilon$ value in the $[0.30, 0.35]$ range, with the highest % backing the $k = 4$ choice being that of $\varepsilon = 0.35$.

To complement these findings on classifying all-round performers, we compare our classification of banks’ inclusive performance with the classification of banks performance while taking into account only the financial aspects (CAMEL proxies) this time. This could permit us to display and compare the distributions of classifications overall, as well as between them in a confusion matrix. We conduct the exact same analysis that has been used up to this point and we use an $\varepsilon$ value of 0.35 to compare these results to the baseline results obtained previously.

Starting with a first overview of the two classifications, Figure 9 shows the % of banks
found to belong in the four classes in terms of their financial and inclusive (i.e. financial and non-financial) attributes. Seemingly, there is a distribution shift in favour of evaluating banks solely on financial criteria. In particular, the highest percent of banks in terms of inclusive evaluation belonged to Class 1 with just under 60% of banks belonging to that cluster. In contrast, just under 5% of banks are found to belong to Class 1 when one takes into account solely financial criteria, with the vast majority (just under 70%) of banks being evaluated in Class 2 in terms of their financial criteria. This pattern repeats itself with more banks being classified in a given class in terms of financial performance versus inclusive performance, which deteriorates as the number of class approaches the top performing banks (i.e. Class 4 banks). This may lead to a conjecture that well-performing (financially) banks have the capacity—and perhaps, the will—to cater for their inclusive aspects and satisfy a larger portion of shareholders compared to smaller and less performing banks, which might need to focus on their financial aspects first.

To disentangle these two classifications further, a comparison is done on the basis of a confusion matrix, where instead of a true and predicted classification, we have the classification of banks as C1 to C4 in terms of their inclusive performance and in terms of their financial performance. Theoretically speaking, if banks perform similarly in their financial aspects as
Figure 5: Typical bank performance between clusters (Environmental)
This figure illustrates the means (with 95% confidence intervals) of the banks belonging to each cluster. For a description of the criteria, see Table 1.

Figure 6: Typical bank performance between clusters (Social)
This figure illustrates the means (with 95% confidence intervals) of the banks belonging to each cluster. For a description of the criteria, see Table 1.

ey they do in their inclusive performances, then taking as a base of comparison the former, the diagonal of the confusion matrix should be close, or as close as possible to the row sums. A distribution shift underneath the diagonal means that the banks seem to perform better when the non-financial attributes are included in their evaluation, while the exact opposite pattern would be true if there was a distribution shift above the diagonal.

Table 2 shows the confusion matrix comparing the classification of banks' inclusive performance versus their classification in their financial performance. This shows the actual number of banks classified at a given position in terms of inclusive performance versus in terms of financial performance, as well as the % (by row sum) to facilitate comparison in relative terms having as a benchmark the inclusive performance. It appears that the vast majority of banks that are classified as $C_1$ in terms of their inclusive performance are actually classified as $C_2$ when taking into account only their financial performance. This is up to some extent reasonable because the banks classified in the lower classes in terms of their financial performance (i.e.
This figure illustrates the means (with 95% confidence intervals) of the banks belonging to each cluster. For a description of the criteria, see Table 1.

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>Row Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Financial</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>3</td>
<td>90</td>
<td>19</td>
<td>1</td>
<td>113</td>
</tr>
<tr>
<td></td>
<td>2.65%</td>
<td>79.65%</td>
<td>16.81%</td>
<td>0.88%</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>6</td>
<td>45</td>
<td>23</td>
<td>0</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>8.11%</td>
<td>60.81%</td>
<td>31.08%</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td><strong>Inclusive</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>9.09%</td>
<td>54.55%</td>
<td>27.27%</td>
<td>9.09%</td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>0.00%</td>
<td>0.00%</td>
<td>20.00%</td>
<td>80.00%</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Confusion matrix: Financial vs Inclusive

C1 and C2) might not have the resources or capacity to upscale their inclusive attributes and thus are clustered in a lower performing class when everything is taken into account. Looking at C2 banks in terms of their inclusive performance, the consistency rises to about a 60% classification accuracy, with just under 10% of banks being classified as C1 in terms of their financial performance and about a third of banks being classified as C3 instead. Interestingly, when looking at C3 banks in terms of their inclusive performance, only a third of these banks perform adequately to be classified as C3 in both types of attributes (i.e. financial and non-financial). In particular, more than half of the banks (54.55%) are classified as C2 in terms of their financial performance. This shows that a good portion of financially fair-performing (C2) banks tend to outperform their peers in the ESG arena and actually improve their overall score to be classified as C3 banks when all attributes are taken into account. Finally, the vast majority of high performing banks in terms of their inclusive performance (i.e. C4 banks) are also found to be classified in the C4 class in terms of their financial performance (80%), which
shows that larger and more established banks do tend to have the capacity (and seemingly, the will) to extend their good performance to other ESG attributes. Interestingly, 20% of banks were classified as \( C_3 \) in terms of their financial performance but \( C_4 \) in terms of their inclusive performance. This shows again that a few banks of lower calibre in terms of financial aspects tend to outperform their peers in ESG aspects and improve their inclusive score.

### 4.4 Comparison to other MCDA clustering proposals

To illustrate how the proposed algorithm fares overall, in this section we compare the results abridged in the previous section to the equivalent set of results obtained using the state of the art MCDA clustering algorithms. In particular, the most closely related algorithm to our proposed algorithm is that of De Smet (2014), which offers a clustering solution based on PROMETHEE’s net flow scores. An example of how that algorithm works is offered in Section 2, although we refer the interested reader to the authors’ original study.

To render our results comparable, we make use of De Smet’s (2014) algorithm (hereafter...
referred to as DSA) in a SMAA setting, where we use the exact same vectors of preferences that have been used in our analysis so far (i.e. same uniformly generated weights, indifference thresholds set to zero and preference thresholds set to the maximum difference between alternatives for each criterion). We also implement the algorithm in such a way to obtain four clusters in each simulation. Notably, this gives the edge to the DSA because the number of clusters varies in our proposed algorithm between two and six (see Fig.2) but is always fixed to four clusters (desired number for overall classification) using their algorithm. A final clustering solution is then obtained through the same ensemble clustering method applied to the abridged set of results to highlight the final clustering set of banks in our sample. We followed the same procedure for the overall results (financial and inclusive criteria combined) or individual sets of results (only financial and only inclusive criteria) with very similar distinguishable differences obtained between the two algorithms. To conserve space, we only report the overall results abridged at the beginning of Section 4, although the full set is available in an online supplementary appendix.

Starting with an overall comparison of the two clustering algorithms (proposed algorithm and DSA), the correlation is rather strong with Spearman’s correlation coefficient amounting to 0.84 (0.85 for Pearson’s and 0.79 for Kendall’s tau). A more detailed break-down illustration of the discrepancy between the two classifications is illustrated in Figure 10. It appears that all banks classified as C4 with the proposed method are also classified as C4 using the DSA. However, it also appears that the proposed algorithm is more selective because there is a discrepancy for a large portion of banks that are classified as C3 or C2 using our proposed
algorithm and are instead classified as C3 and C4 using the DSA. In the following, we will try to disentangle which algorithm seems more appropriate based on the set of results obtained.

The two clustering algorithms can be compared by going back to the simulation results and looking at the pairwise winning indices (PWI) computed through the SMAA-PROMETHEE method (Corrente et al., 2014). The PWIs show how much a bank is inferior or preferred to its counterparts in the space of preferences simulated in the Monte Carlo environment. We compare the PWIs of the 18 banks classified by the DSA as C4 in order to check which of the two algorithms results in more reasonable clustering. Out of these 18 banks, 10 were indeed classified as C4 based on our algorithm, while the remaining 8 were classified in the C3 cluster. For each and every of these 18 banks, we computed their PWIs to show the % of simulations in which each bank is inferior to the remaining 17 in the C4 cluster. Figure 11 shows the PWIs of the 18 banks. While all those are clustered together according to the DSA, those highlighted in blue are actually clustered together as C4 in our proposed algorithm, and those highlighted in yellow are found to be clustered with other banks as C3. The red-dotted line shows the average inferiority of each set of banks (highlighted in blue or yellow) with respect to the remaining \( n - 1 \) banks in the C4 according to DSA. As one may observe from the plot, the average inferiority of each bank actually clustered as C4 based on our proposed algorithm is just under 30%, which jumps to a stark 80% for those that are clustered as C3 based on our algorithm. This highlights that the latter (highlighted in yellow) banks are very inferior to their remaining counterparts and thus are relatively very distant in terms of preference in the whole space of preferences simulated within SMAA. Therefore, it could be more reasonable to separate those banks by assigning them into different clusters, which is the case with the proposed algorithm.

An additional way to delve into the two algorithms’ discrepancies is through calculation of the average net flow in the 10,000 simulations. Figure 12 shows (left-hand sub-plot) the
Figure 11: Pairwise Winning Indices (PWIs) of C4 banks according to De Smet (2014)

This figure shows the % of simulations according to which each C4 bank (according to the DSA) is inferior to its counterparts in the same cluster. The banks that are also found to be C4 using the proposed algorithm are highlighted in blue, whereas the banks found to belong in the C3 based on the proposed algorithm are highlighted in yellow. Red-dotted lines are averages of percentages of the banks labeled on the horizontal axis.

The typical net flow score of each of the 208 banks grouped based on the ensemble clustering result of each algorithm, as well as a more abridged picture of the same results, summed via their means and 95% confidence intervals (right-hand sub-plot). As can be seen from the figure, the average net flow of C4 banks according to the DSA could reasonably be split into two clusters, with some of the banks (classified as C4, with an average net flow in the $(0.012, 0.028)$ range) that are relatively distant from their counterparts (also classified as C4, though with an average net flow in the $(0.037, 0.047)$ range) actually joining the C3 cluster, which implies that some of the lower-performing C3 banks join the C2 cluster and so on. This justifies the discrepancy observed in the confusion matrix (Fig. 10). The same picture may be illustrated rather more clearly in the right-hand sub-plot of the same figure portraying the cluster means with their confidence intervals. Our proposed clustering algorithm appears to cluster bank performance more smoothly when compared to the other algorithm, which shows an abrupt difference between C4 and C3. This may be due to the fact that the compared algorithm (by its very nature) tries to form clusters based on positive and negative net flows, which under certain circumstances may amplify the differences between clusters.

5 Conclusions

In this paper, we have developed a divisive hierarchical clustering algorithm, SMAA-MDHC, which groups alternatives into homogeneous groups with similar preferences. Unlike existing studies in this strand, an ensemble clustering method is embedded into the SMAA-PROMETHEE framework to obtain clustering solutions that are more robust and stable. While
SMAA-PROMETHEE is used to generate different clustering solutions by changing different values of thresholds and weight parameters, the ensemble clustering method combines them into a consensus solution. Because a large number of clustering solutions are obtained through the algorithm, our approach also helps analysts to identify the most representative number of clusters.

We have illustrated the performance of SMAA-MDHC on a real-world dataset. A sample of 208 US banks are partitioned into clusters according to a set of financial, environmental, social and governance criteria. The results show that SMAA-MDHC yields reasonable and robust solutions across different parameter settings. Moreover, and in line with expectations, larger and more established banks tend to outperform their peers in terms of a holistic, inclusive performance, while there is a set of banks that prove the exception to the rule and (despite being lower calibre in terms of financial aspects) tend to outperform their peers in the ESG criteria and improve their final score. Additionally, comparing US banks in terms of financial, and financial and ESG criteria combined, we infer that there is a cohesive pattern of classification for C4 banks and partially for C2 banks. Interestingly, we find that half of C3 banks in terms of inclusive performance were actually classified as C2 in terms of financial performance, which shows that even modestly performing (financially) banks may venture further and focus on ESG aspects. However, this is not the norm given that a third of banks classified as C2 in terms of inclusive performance are actually classified as C3 in terms of financial performance.

For further research, an interactive approach that incorporates the preferences of the DM throughout the algorithm can be developed. The DM can guide the process to achieve a more
consolidated consensus solution, not only in parameter selection but also in the ensemble clustering. In addition, further research may develop a pre-processing algorithm that studies the characteristics of the multiple clustering solutions and eliminates some of them before applying the consensus function. This may help to improve the quality of the consensus solution because some poor clustering solutions in the initial cluster ensemble may have mislead the process. Finally, and from a more general perspective, one could further study the link between the results of a clustering procedure with $k$ fixed clusters and a sorting procedure with $k$ ordered clusters.

References


MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate


