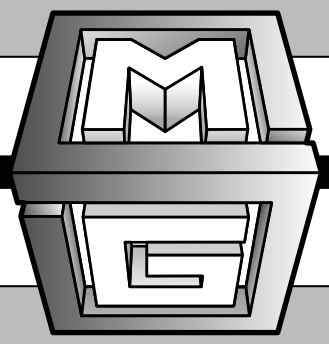

Université Libre de Bruxelles

CoDE - SMG

A Model for Spatial Multicriteria Hierarchical Clustering

CoDE-SMG – Technical Report Series

Karim Lidouh, Yves De Smet

CoDE-SMG – Technical Report Series

Technical Report No.
TR/SMG/2014-008

October 2014

CoDE-SMG – Technical Report Series
ISSN 2030-6296

Published by:
CoDE-SMG, CP 210/01
Université Libre de Bruxelles
Bvd du Triomphe
1050 Ixelles, Belgium

Technical report number TR/SMG/2014-008

A Model for Spatial Multicriteria Hierarchical Clustering

CoDE-SMG – Technical Report Series

Karim Lidouh klidouh@ulb.ac.be
Yves De Smet yvdesmet@ulb.ac.be
CoDE-SMG, Université Libre de Bruxelles, Brussels, Belgium

October 2014

A Model for Spatial Multicriteria Hierarchical Clustering

Karim Lidouh* and Yves De Smet

Department of Computer and Decision Engineering
École polytechnique de Bruxelles, Université libre de Bruxelles,
50 Avenue F.D. Roosevelt CP 210/01, 1050 Brussels, Belgium
E-mail: klidouh@ulb.ac.be
E-mail: yvdesmet@ulb.ac.be
*Corresponding author

Abstract: Research on the problem of multicriteria territory partitioning is at its begin. This is mainly due to the fact that it involves tools from fields that are to this day still young. To answer this shortage, we propose an adaptation of a multicriteria clustering method that takes spatial constraints into account. Two variants are described and tested on an illustrative case. This example deals with the partitioning of the Walloon region in Belgium into clusters with a similar level of well-being as perceived by its inhabitants. This gives us interesting results that illustrate the properties of the algorithm we use.

Keywords: Multiple criteria analysis, Hierarchical clustering, Preference modeling, Territory partitioning

Reference to this paper should be made as follows: Lidouh, K. and De Smet, Y. (xxxx) 'A Model for Spatial Multicriteria Hierarchical Clustering', *Int. J. Multicriteria Decision Making*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes:

Karim Lidouh has a degree as a Civil Engineer in Computer Science, a Master in Management and completed in 2014 a PhD thesis on the integration of multicriteria tools in geographical information systems. He works as a Teaching Assistant in the fields of statistics and quantitative methods at the Solvay Brussels School of Economics and Management (SBS-EM) of the Université libre de Bruxelles. He also gives occasional courses on operations research and optimisation methods at the IESEG School of Management.

Yves De Smet is Associate Professor at the Engineering Faculty of the Université libre de Bruxelles. He is both head of the Computer and Decision Engineering laboratory and of the SMG unit. Yves De Smet holds a degree in Mathematics (1998) and a PhD in Applied Sciences (2005). His research interests are focused on multicriteria decision aid and multi-objective optimization. Besides his academic activities he has been involved in different industrial projects. Since 2010, he has been co-founder of the Decision Sights spin-off.

1 Introduction

Some of the first cases to have been studied using multicriteria methods about 30 years ago involved spatial entities [1]. These have shown the advantages of working with the multicriteria paradigm when dealing with problems that involve spatial components [2]. However as can be seen in recent works [3], the integration between multicriteria decision aid and geographical information science has experienced a very slow evolution.

This paper will focus on a particular type of spatial problem which is territory partitioning or districting. From a methodological point of view, this problem can be seen as a clustering problem with additional constraints [4], namely that only connected areas can be grouped in the same class or cluster. And even though the literature abounds with cases of territory partitioning (see [5]), very few of them actually take explicitly multicriteria information into account. One of the sole examples is demonstrated by the work of Tavares-Pereira et al. [6], where the multicriteria profiles of all the areas to be clustered were taken into account and partitions where generated using a genetic algorithm. To our knowledge, aside from that contribution, very few others proposed significant works that tackle this particular problem.

Our work proposes an extension of the multicriteria nominal clustering method initially developed by De Smet and Montano Guzmán [7]. This method was based on the k-means algorithm. Let us note that this approach was later extended by De Smet and Eppe in the context of partial ordered clustering [8]. Our extension adds support for the spatial connectivity of the areas to be clustered. We also changed the approach to a hierarchical one. This means that the process can be stopped at any iteration if one wishes to study the progress on a particular setting. Two variants of the proposed extension are also presented with a description of their respective properties.

Section 2 of this paper presents the model we use and explains its differences with the initial clustering method. In section 3, we put this model to use on an illustrative case which studies the well-being in the Walloon region of Belgium. Two maps are proposed with the two variants we have developed. Finally, Section 4, concludes this paper and presents some perspectives.

2 Model

As previously stated, the model we use is based on an earlier article by De Smet and Montano Guzmán [7]. In that first version, the authors proposed an extension of the well-known k-means algorithm to multicriteria clustering problems. The result was a nominal clustering that was based on four preference relations. Those allow to characterize the profile of each action, i.e. its relative "preferential position" with respect to the whole dataset. The model was based on the idea that all the actions in a cluster should have similar behaviours in terms of preference, indifference, and incomparability relations. The algorithm thus adopts an approach similar to the k-means method [9, 10] and stops when the cluster memberships no longer change.

Our approach differs from the previous one in that it is based on a hierarchical procedure. At each step of the process all actions are compared and the pair of actions with the smallest distance between profiles gets merged into a single cluster with a resulting profile. Furthermore, since we are applying this to territory partitioning problems, comparisons are only made between neighbouring actions. This ensures that

only connected actions or clusters are merged. After a number of steps equal to the number of actions, all actions are merged together into a single cluster.

In this section, the set of actions to be clustered will be denoted $A = \{a_1, a_2, \ldots, a_n\}$ (where n denotes the number of actions). The actions merged together at each given step of the method will be referred to as $B = \{b_1, b_2, \ldots, b_{n^*}\}$ where n^* indicates the current number of clusters (n at the start of the algorithm, and 1 at the end). The information on neighbouring actions will be stored in an $n \times n$ adjacency matrix denoted $C = (c_{ij})$ with $i, j = 1, 2, ..., n$ and $c_{ij} = 1$ if a_i and a_j are spatially adjacent.

2.1 Profiles

As in [7], the profiles of each action will be built using the traditional $\langle P, I, J\rangle$ (Preference, Indifference, and Incomparability) relations [11]. Each action a_i's profile will be a 4-uple $\langle J(a_i), P^-(a_i), I(a_i), P^+(a_i)\rangle$ where:

- $J(a_i) = \{a_j \in A | a_i J a_j\} = P_1(a_i)$
- $P^-(a_i) = \{a_j \in A | a_j P a_i\} = P_2(a_i)$
- $I(a_i) = \{a_j \in A | a_i I a_j\} = P_3(a_i)$
- $P^+(a_i) = \{a_j \in A | a_i P a_j\} = P_4(a_i)$

We refer the interested reader to the article by De Smet and Montano Guzmán [7] for a detailed definition of the preference structure we use.

For practical reasons, these profiles will be stored in a $n \times 4n$ binary matrix defined as follows:

$$P = \begin{pmatrix} P_1(a_1), P_2(a_1), P_3(a_1), P_4(a_1) \\ \cdots \\ P_1(a_n), P_2(a_n), P_3(a_n), P_4(a_n) \end{pmatrix} = \begin{pmatrix} J(a_1), P^-(a_1), I(a_1), P^+(a_1) \\ \cdots \\ J(a_n), P^-(a_n), I(a_n), P^+(a_n) \end{pmatrix}$$

where each element is equal to 1 if the corresponding relation between the two actions is true, and equal to 0 otherwise.

2.2 Distance

In order to determine which actions or clusters are to be merged together at each step, it is necessary to compute a distance between their profiles. As these are only composed of 0 and 1 values in the P matrix, this is easily done using the following equation:

$$\min d(b_i, b_j) = 1 - \frac{1}{n^*} \sum_{l=1}^{4} |P_l(b_i) \cap P_l(b_j)| = 1 - \frac{1}{n^*} \sum_{m=1}^{4n} (p_{im} \cdot p_{jm})$$

In doing so, two clusters will be considered close the more their profiles are alike. The distance between them would then be closer to 0 than 1.

2.3 Construction of resulting profiles

There are several ways to construct a resulting profile for a cluster made of a set of actions. We propose two versions for this algorithm.

Voting procedure

The initial version of the multicriteria clustering algorithm [7] featured only one way of constructing a resulting profile when actions are merged into a single cluster. This profile $P(c_i)$ was determined using a voting procedure:

$$a_j \in P_k(a_i) \Leftrightarrow \arg\max \sum_{a_{i_l}} \mathbb{1}_{\{a_j \in P_k(a_{i_l})\}}$$

where a_{i_l} are the actions belonging to the considered cluster.

This voting procedure allows to obtain a resulting profile that matches the profiles within the cluster as closely as possible. When several profiles can be used, one of them is selected randomly. A direct consequence of this is that the results expected from this procedure will not always be consistent as is shown in Section 3.

Intersection

The second procedure we propose consists in replacing the set of profiles in a cluster with their intersection. The idea behind this approach is that we only keep the common part between all the profiles. Using the same notations as before, we have:

$$a_j \in P_k(a_i) \Leftrightarrow a_j \in \bigcap_{a_{i_l}} P_k(a_{i_l})$$

This approach ensures that the results are always the same. However this generates incomplete profiles which lead us to distances that become greater after less steps. Therefore, if the number of actions is great, at some point the associations might no longer make any sense. The smallest distance found could therefore be used as a termination criterion. Indeed one could for example stop the algorithm as soon as the smallest distance becomes equal to 1, meaning that all the cluster profiles left no longer present any similarities.

2.4 Algorithm

Algorithm 1 shows all the steps of the method as we implemented it. The only part that can be adapted is the construction of the resulting profiles at each iteration that we described at the end of the previous section.

When using a voting procedure, we encounter the same drawbacks as with the k-means method. This is common for such approaches for which there is no uniqueness of results due to the starting conditions and the construction of prototypes for the clusters. Profiles built using a voting procedure will indeed be set randomly when there is an equal number of votes for two types of relations. However, since in our case the approach is hierarchical, this effect is limited. It is also further limited due to the fact that we only merge actions that are adjacent.

3 Illustrative Case: The Index of Conditions of Well-Being (ICWB) in Wallonia

In order to illustrate this method, we chose to use a recently published study from the Walloon Institute for Evaluation, Prospective, and Statistics (IWEPS): the Index of

Algorithm 1 Spatial Multicriteria Hierarchical Clustering

for all numbers of clusters n^* **do**
  for all clusters b_i in B **do**
    Compute the smallest distance to its adjacent neighbours
  end for
  Select the cluster b_i with the smallest distance $d(b_i, b_j)$ and assign its neighbour b_j to it
  Update the set B by removing cluster b_j
  Update matrix C by removing row j and replacing row i with the max of both rows (apply the same to the columns i and j)
  Update matrix P by removing row j and replacing row i by the resulting profile for the new cluster (apply the same to the columns for each relation P_k)
end for

Conditions of Well-Being in Wallonia (ICWB) [12]. This index was constructed out of 58 indicators that were grouped in 19 dimensions and then 8 families. All of them were evaluated on the 262 municipalities that constitute the Walloon region in Belgium. Figure 1 shows the hierarchical structure of the ICWB. The data for the 19 dimensions was extracted from the report that was published by the IWEPS institute. As the table is quite big (262×19) we decided not to include it in this paper, but the original report is freely available on the IWEPS website.

Figure 1: Hierarchical structure for the ICWB (based on [12])

3.1 Preference structure

Instead of applying a standardisation similar to the one used by the IWEPS researchers, we chose to use preference functions inspired from the PROMETHEE methodology [13]. For each criterion, we defined a preference function where both preference and indifference thresholds were both equal to half of the greatest difference between evaluations (see

Figure 2). This ensured that we would obtain varied profiles with some common parts. Using these functions for each pair of actions, we counted the number of criteria on which an action a_i is preferred to another a_j and the number of criteria were the opposite relation was observed:

- If both numbers were different from zero, the actions were considered incomparable ($a_j J a_j$)
- If both numbers were equal to zero, the actions were considered indifferent ($a_i I a_j$)
- If only one of the numbers was equal to zero, the better action was preferred to the other ($a_i P a_j$ or $a_j P a_i$)

By doing this, we considered equal weights for all the criteria. Let us point out that a description of detailed and justified preferences goes beyond the scope of this paper. This is why we rely on a simplified model.

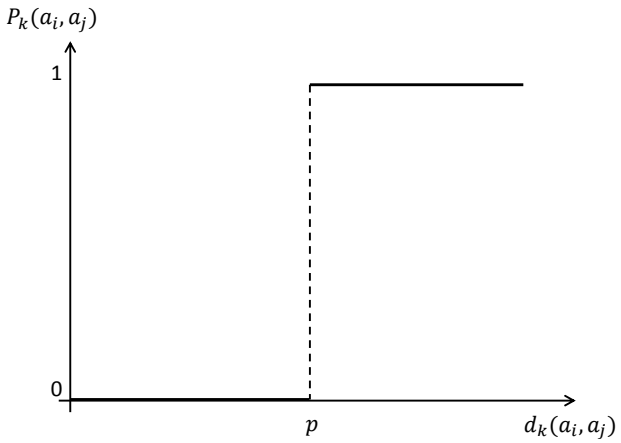

Figure 2: Type of preference function used

We also considered two separate sets of criteria: on the one hand the 19 dimensions and on the other hand the 8 aggregated families. This allowed us to see how the algorithm behaved when applied on datasets that significantly differ in size. Finally, in order to compare the results, we stopped the algorithm after 200 steps, leaving us with 62 clusters of variable sizes. In the next subsections, we describe two of the results we obtained using the different methods to construct resulting profiles for the clusters.

3.2 Voting procedure

When applying the voting procedure on such a large problem and for a great number of steps, we immediately see that the voting procedure will tend to easily merge actions and obtain clusters with heterogeneous contents. This can be seen in Figure 3 where we see that a single cluster (i.e. cluster 5) covers almost the entire map. Since the procedure replaces the actions of a cluster by a single profile, small differences do not matter so much.

The numbers indicate the cluster to which each municipality belongs. The colors represent the ICWB score of each cluster computed by using the average of all the municipalities' scores. A high score is displayed in red while low scores are displayed in yellow.

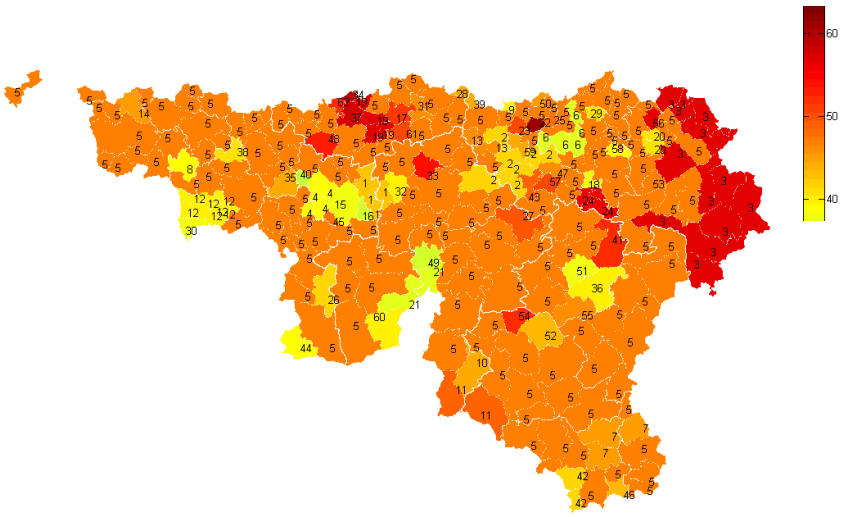

Figure 3: Spatial clustering in 62 areas based on the ICWB (Voting procedure)

As mentioned earlier, each attempt at applying this approach leads to different results due to the randomisation aspect of the procedure. Furthermore the great number of steps increases the likelihood that small differences at the beginning might have a great impact on the rest of the computations.

However, an interesting aspect of this approach is that it identifies outliers quite well. Indeed, those municipalities will usually stay isolated in their own cluster until the last steps of the algorithm. The smallest clusters indicate characteristics of Wallonia that match the study done by IWEPS. For instance, the set of yellow clusters that form a belt in the upper part of the region correspond to the industrial and urban areas where the ICWB values are at their lowest. These outliers seem to be the only constant characteristic of the varied results we obtain when applying the method several times.

3.3 Intersection

When applying the intersection procedure the results are fundamentally different. Indeed, as can be seen in Figure 4, each difference counts as the resulting profiles at each step become more and more different from each other, leading us to several clusters of roughly the same size.

Once again the outliers are present, but this time there is no single cluster that occupies the majority of the map. The variety in profiles is indeed preserved and is shown with the different scores in a map that matches the individual ICWB scores more accurately.

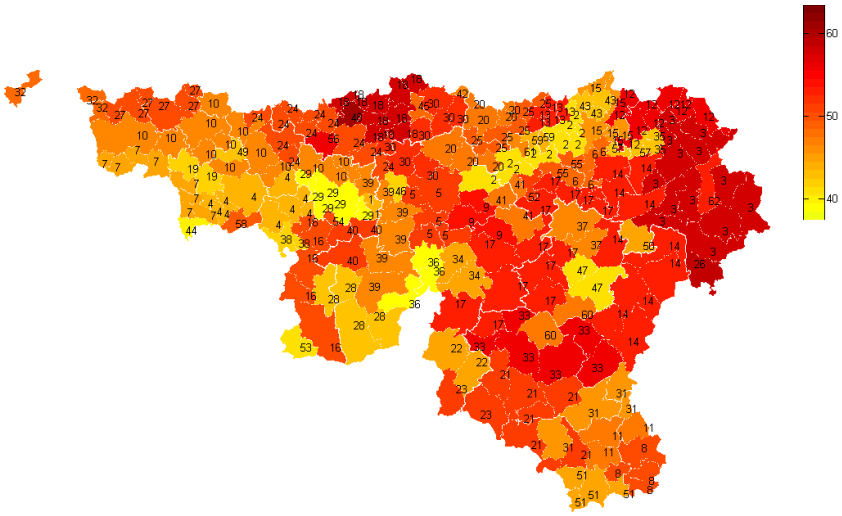

Figure 4: Spatial clustering in 62 areas based on the ICWB (Intersection)

3.4 Influence of the size of the problem

The number of criteria used does have a significant influence on the results. As expected, we noticed indeed that with 19 criteria we frequently find pairs of alternatives that are incomparable. This of course can lead to several problems as the profiles are almost completely heterogeneous from the start. This means that an approach based on the intersection of profiles will no longer work. Indeed, only a few steps are enough to start having empty profiles that appear in our set of clusters. From that point, the distances evaluated are always maximal and the next steps of the algorithm have no real meaning.

The voting approach however seems to deal with this difficulty a bit better as it will always produce complete profiles. Nonetheless the results obtained with this method should be treated with care as the clusters themselves might contain actions very different from their resulting profile.

Another point to be mentioned is the impact of weights of lack thereof. Indeed, as this method does not make use of any differentiated weights (e.g. we simply counted the criteria for which an alternative is preferred to another, see Section 3.1), it considers all criteria equally to generate the starting profiles of all actions. This does not constitute a problem as such but we need to be aware of this aspect when giving interpretations on the results. Depending on the values, this feature's influence might have an even greater impact when combined with a large number of criteria.

4 Conclusion

In this paper we developed an extension of an existing multicriteria clustering method. We adapted it to the spatial problem of territory partitioning and proposed two variants to the algorithm. The main advantage of this method compared to other districting techniques is

that it take into account the multicriteria information of all the areas before grouping them in clusters.

The first difference we introduced was changing the method into a hierarchical clustering technique. This eliminated the problem of initial conditions that influence the results and lead us to a method that can more easily produce stable results.

Out of the two variants we proposed, the first one uses a voting procedure similar to the one in the existing clustering method. The effect of this approach is that all the areas that are similar are more easily grouped while only outliers and particular cases are kept isolated until the last iterations of the algorithm. This is due to the fact that a new artificial profile is built for each cluster of areas which ignores differences that can exist within the cluster.

The other approach we proposed is more strict when it comes to differences between profiles and will therefore only group areas when these are all similar within the group. This second approach therefore gives us results that are reproducible and which present clusters of roughly the same size.

Further testing could help us imagine new variants of this clustering method. Moreover, applying this technique to other cases or comparing the obtained results to actual studies of territories could help us understand the characteristics that are highlighted by these variants. Finally, the characterisation of the geographical partition quality has still to be further investigated.

References and Notes

1 P Bertier and J de Montgolfier. Approche multicritère des problèmes de décision. *Edition Hommes et Techniques, Paris*, 1978.

2 Bernard Roy. *Multicriteria methodology for decision aiding*, volume 12. Springer, 1996.

3 Karim Lidouh. On the motivation behind mcda and gis integration. *International journal of multicriteria decision making*, 3(2):101–113, 2013.

4 Constantin Zopounidis and Michael Doumpos. Multicriteria classification and sorting methods: A literature review. *European Journal of Operational Research*, 138(2):229–246, 2002.

5 Federica Ricca, Andrea Scozzari, and Bruno Simeone. Political districting: from classical models to recent approaches. *Annals of Operations Research*, 204(1):271–299, 2013.

6 Fernando Tavares-Pereira, José Rui Figueira, Vincent Mousseau, and Bernard Roy. Multiple criteria districting problems. *Annals of Operations Research*, 154(1):69–92, 2007.

7 Yves De Smet and Linett Montano Guzmán. Towards multicriteria clustering: An extension of the k-means algorithm. *European Journal of Operational Research*, 158(2):390–398, 2004.

8 Yves De Smet and Stefan Eppe. Multicriteria relational clustering: The case of binary outranking matrices. In *Evolutionary Multi-Criterion Optimization*, pages 380–392. Springer, 2009.

9 James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. California, USA, 1967.

10 Michael R Anderberg. Cluster analysis for applications. Technical report, DTIC Document, 1973.

11 Philippe Vincke. *Multicriteria decision-aid*. John Wiley & Sons, 1992.

12 Julien Charlier, Isabelle Reginster, Christine Ruyters, and Laurence Vanden Dooren. Index of conditions of well-being in Wallonia, 1st exercice, april 2014. Technical report, Institut Wallon de l'Evaluation, de la Prospective et de la Statistique (IWEPS), 04 2014.

13 Jean-Pierre Brans and Ph Vincke. A preference ranking organisation method: (the promethee method for multiple criteria decision-making). *Management science*, 31(6):647–656, 1985.