Visualizing proportions and dissimilarities by Space-filling maps: a Large Neighborhood Search approach*

Emilio Carrizosa[1], Vanesa Guerrero[1], and Dolores Romero Morales[2]

[1]Instituto de Matemáticas de la Universidad de Sevilla (IMUS), Seville, Spain
{ecarrizosa, vguerrero}@us.es
[2]Copenhagen Business School, Frederiksberg, Denmark
drm.eco@cbs.dk

Abstract

In this paper we address the problem of visualizing a set of individuals, which have attached a statistical value given as a proportion, and a dissimilarity measure. Each individual is represented as a region within the unit square, in such a way that the area of the regions represent the proportions and the distances between them represent the dissimilarities. To enhance the interpretability of the representation, the regions are required to satisfy two properties. First, they must form a partition of the unit square, namely, the portions in which it is divided must cover its area without overlapping. Second, the portions must be made of a connected union of rectangles which verify the so-called box-connectivity constraints, yielding a visualization map called Space-filling Box-connected Map (SBM). The construction of an SBM is formally stated as a mathematical optimization problem, which is solved heuristically by using the Large Neighborhood Search technique. The methodology proposed in this paper is applied to three real-world datasets: the first one concerning financial markets in Europe and Asia, the second one about the letters in the English alphabet, and finally the provinces of The Netherlands as a geographical application.
Keywords: Data Visualization, Box-connectivity, Proportions, Dissimilarities, Large Neighborhood Search

1 Introduction

Information Visualization has experienced a tremendous development in recent years as a result of the data storage growth, [38, 56]. Visualizing data with complex underlying structures has become a crucial task to help analysts to draw conclusions by detecting patterns and behaviours which remain hidden when classic techniques are applied, [9, 10, 23]. Mathematical Optimization plays an important role in the field of Information Visualization, both developing new tools to adapt existent visualization techniques and creating new ones, [7, 24, 28, 42]. In this paper, Mathematical Optimization is used to build a planar space-filling visualization map, in the sense that the whole region in which the data are visualized is covered. The framework developed

*This research is funded in part by Projects MTM2015-65915-R (Spain), P11-FQM-7603 and FQM-329 (Andalucía), all with EU ERD Funds.

in this work is able to properly represent statistical values (given as proportions) attached to individuals as well as proximity relationships (given by a dissimilarity matrix).

There exist several contexts in which proportions need to be properly visualized. Some straightforward examples are population rates or vote intention rates. Classic representations as pie or fan charts can handle this type of data, [53]. In addition to statistical values, there usually exists a relationship between the individuals in the form of a dissimilarity, such as geographical distances or affinity between political parties. Pie/fan charts and related ones, [19], are very limited to properly depict dissimilarities, since the set of possible locations for the different wedges is reduced to a choice of a permutation. Thus, more sophisticated visualization tools which simultaneously deal with both types of information are needed. For instance, there exist approaches which integrate Graph Drawing techniques, [2], with multivariate statistical approaches like Multidimensional Scaling (MDS), as in [17, 33], or Correspondence Analysis, as in [45]. For a comprehensive survey on recent visualization frameworks see [38].

Nevertheless, combining simultaneously the representation of proportions and dissimilarities in a planar visualization map is still a challenging problem to which very specific (ad-hoc) approaches exist. This is the case, for instance, of frameworks developed by Cartography. Cartograms try to reproduce the geographical distances between the regions while they are scaled, and thus deformed, according to a statistical variable, such as the population rates [57]. There are also approaches in which the regions are reshaped, for instance, to rectangles, [29, 36], piecewise rectangles, [1, 14], or circles [18]. Techniques to build cartograms take advantage of geographical information to properly depict dissimilarities, yielding very specific approaches that usually cannot be extended to more general examples, such as the ones handled in our computational section. Other disciplines which have dealt with the problem of representing simultaneously statistical values and dissimilarities are Facility Layout and Information Visualization. In Facility Layout, the aim is to locate the facilities with a given area according to the flow between facilities, [32, 34, 51], without necessarily producing space-filling layouts. A well-known example in Information visualization is the *treemap*, [4, 52], in which the unit square is sequentially subdivided to represent a hierarchy between the individuals satisfying that each region has a desired area. Treemaps are space-filling layouts which represent exactly the statistical values as the area of the rectangles but, in general, they fail to properly represent dissimilarities. Other attempts, such as [6, 17, 26, 27, 39, 55], either disregard dissimilarities or proportions, or they are not space-filling visualization maps. In [5, 20], space-filling rectangular maps are proposed, which take into consideration both dissimilarities and proportions. However, the geometrical shape used to represent each individual (a rectangle) seems to be too rigid to give a good fit in both dissimilarities and proportions. In [36], there is an example illustrating that it may be impossible to represent accurately both types of information in space-filling rectangular maps. In that paper, a dissimilarity matrix representing adjacencies is considered. A partition of a square into rectangles is sought such that the rectangles' areas depict the proportions and adjacencies between rectangles represent the given adjacency matrix. However, they show in a figure (Figure 2) how the exact representation of the adjacency matrix is not possible when an accurate visualization of the proportions is desired.

In this paper, we develop a general optimization-based framework to visualize a set of N individuals, $V = \{v_1, \ldots, v_N\}$, as N connected regions, $\boldsymbol{P} = (P_1, \ldots, P_N)$, on the unit square, $\Omega = [0,1]\times[0,1]$, with two types of information attached: a statistical value given as a proportion, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)$ such that $\sum_{r=1}^{N} \omega_r = 1$ and $\omega_r \geq 0$, $r = 1, \ldots, N$, as well as a dissimilarity measure $\boldsymbol{\delta} = (\delta_{rs})_{r,s=1,\ldots,N}$, such that $\delta_{rs} \geq 0$ and $\delta_{rr} = 0$. To enhance the interpretability of

the representation, the portions in $\boldsymbol{P}$ are required to satisfy two properties. First, they must form a partition of Ω, namely, the portions in which it is divided must cover its area without overlapping. Second, the portions must be made of a connected union of rectangles which verify the so-called *box-connectivity* property (to be formally stated later). We refer to such representation hereafter as a Space-filling Box-connected Map (SBM).

The remainder of this paper is structured as follows. In Section 2, we first establish the conditions that an SBM must satisfy. After defining the concept of box-connectivity, the problem of constructing an SBM is formally stated as a biobjective Mixed Integer Nonlinear Problem (MINLP) and this is then reformulated into two single-objective Mixed Integer Linear Programs (MILP). Finally, an equivalent formulation to these MILPs is provided, which, although less natural, is shown to be tighter in general. Due to the high computational effort needed to solve such optimization problems, heuristics are to be used. In Section 3, Large Neighborhood Search (LNS), [47, 50], is adapted to our problem. Section 4 contains the numerical experiments of our methodology on three datasets of different nature. The first one consists of visualizing as an SBM the world market portfolio of financial markets across Europe and Asia, [21], according to their correlation, [3]; the second dataset deals with the representation of the relative frequencies of letters in the English language, [49], and the confusion between acoustic Morse signals, [3]; the last one is a geographical example in which the provinces of The Netherlands are scaled by their population rates, [54], while preserving their geographical location. Section 5 includes some remarks and future work. Finally, the Appendix closes the paper with some technical details.

2 The model

In this section we first formally state the conditions that a Space-filling Box-connected Map (SBM) must satisfy, namely a visualization map in which the set of individuals V is represented as a partition into box-connected regions $\boldsymbol{P}$ (referred also as portions) on the unit square Ω. The construction of an SBM is modelled as a Mathematical Optimization problem, for which an equivalent reformulation, which turns out to be tighter in terms of its continuous relaxation, is also presented.

2.1 Problem statement

In order to construct an SBM, we consider Ω partitioned into K rows and L columns, called in what follows (K, L)-grid. Pairs of cells forming a (K, L)-grid are considered adjacent if they share one full side, which implies that a cell can have at most four adjacent cells. Grid structures are commonly used in Reserve Network design to arrange a connected union of sites to protect ecosystems and species [31, 43, 44]. Since the grid layout provides a well-structured and compact representation, it has been also used in visualization map designs as in [26, 39, 55], and in Facility Layout as a tool to easily measure the area of the facilities [34].

Definition 1. *Given two cells (i, j) and (i', j') in a (K, L)-grid, we define the box generated by (i, j) and (i', j') as*

$$B\left((i,j),(i',j')\right) = \{(i'',j'') \in (K,L)\text{-}grid: \quad \min\{i,i'\} \leq i'' \leq \max\{i,i'\}, \\ \min\{j,j'\} \leq j'' \leq \max\{j,j'\}\}.$$

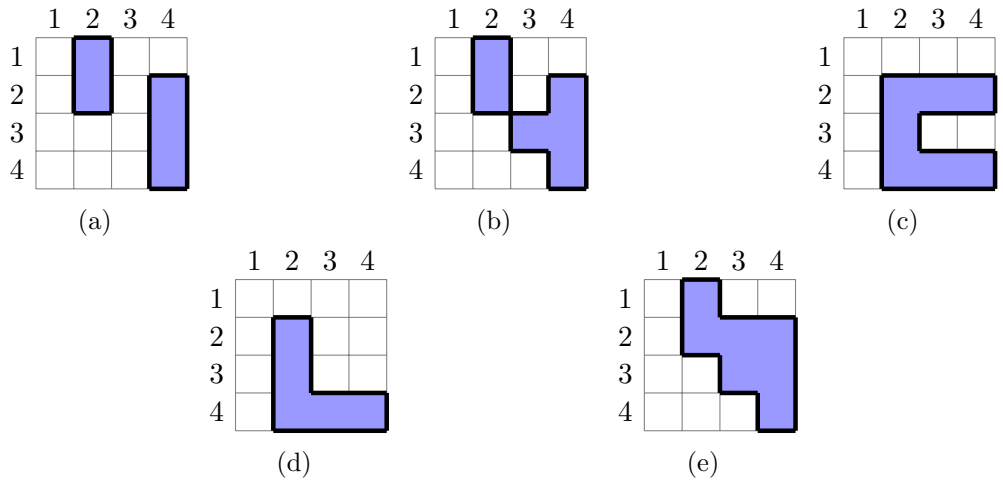


Figure 1: Shaded cells in (a), (b) and (c) are not box-connected regions, while shaded cells in (d) and (e) are box-connected regions.

Definition 2. *Let S be a subset of cells of a (K, L)-grid and let $|S|$ denote its cardinality. S is said to be box-connected if one of these conditions holds:*

1. $|S| = 1$.
2. $|S| = 2$ *and its two cells are adjacent.*
3. $|S| \geq 3$ *and for all non-adjacent* $(i, j), (i', j') \in S$, *there exists* $(i'', j'') \in B\left((i, j), (i', j')\right) \cap S$ *such that* $(i'', j'') \neq (i, j)$ *and* $(i'', j'') \neq (i', j')$.

Each individual v_r must be depicted in the SBM as a unique box-connected portion P_r made of cells in the (K, L)-grid. Shaded cells in Figures 1 (a) - 1 (c) represent portions that are not allowed in our representation, since they do not verify the box-connectivity stated in Definition 2. Regarding Figure 1 (a), there are no shaded cells in the intersection of the set containing the shaded cells and the box $B((2, 2), (2, 4))$, except for $(2, 2)$ and $(2, 4)$. Analogously, in 1 (b) considering $B((2, 2), (3, 3))$. Thus, we observe that disconnected portions are obviously not box-connected. Finally, Figure 1 (c) violates box-connectivity when considering, for instance, the box $B((2, 4), (4, 4))$. Possible shapes allowed to represent the individuals in V are depicted in Figures 1 (d) and 1 (e).

The construction of an SBM regards the partition of Ω into N box-connected portions, which are made of cells in the (K, L)-grid in which Ω is subdivided, and in such a way that the areas of the portions depict the statistical values, $\boldsymbol{\omega}$, and the distances between the portions are proportional to the dissimilarities, $\boldsymbol{\delta}$. These statements can be summed up in the following conditions:

(D1) The portions in $\boldsymbol{P} = (P_1, \ldots, P_N)$ form a partition of $\Omega = [0, 1] \times [0, 1]$,

(D2) P_r is a box-connected region, made up of a collection of cells of the (K, L)-grid in which Ω is divided, $r = 1, \ldots, N$,

(D3) $\text{distance}(P_r, P_s) \propto \delta_{rs}$, $r, s = 1, \ldots, N$, $r \neq s$,

(D4) $\text{area}(P_r) = \omega_r$, namely $\dfrac{1}{K \times L}|P_r| = \omega_r,\ \forall r = 1, \ldots, N$, where $|P_r|$ denotes the number of cells in P_r, $r = 1, \ldots, N$.

Since the portions in $\boldsymbol{P}$ are made of connected unions of cells and they form a partition of Ω, the grid structure helps to easily measure the closeness and the area of the portions. In order to measure the distance between portions P_r and P_s in (D3), those usually used in Cluster Analysis like the single, complete and average linkage, [28], seem suitable for an SBM.

2.2 The Mathematical Optimization model

An SBM which satisfies conditions (D1) and (D2) can be obtained straightforwardly by allocating cells belonging to the same portion sequentially until filling Ω. However, including conditions (D3) and (D4) as hard requirements might make the problem unfeasible, [36]. Thus, our model consists of building SBMs in which the errors made by approximating the scaled dissimilarities, through a real positive variable κ, by the distances between the portions, and the statistical values by their areas, both measured in absolute value, are minimized. Our model considers then the violation of conditions (D3) and (D4) as objectives to be minimized, yielding a Biobjective Space-filling Box-Connected Map (BSBM) model, which reads as follows

$$
\begin{array}{ll}
\min & \displaystyle\sum_{\substack{r,s=1,\ldots,N \\ r \neq s}} |distance(P_r, P_s) - \kappa\delta_{rs}| \\
\min & \displaystyle\sum_{r=1,\ldots,N} |area(P_r) - \omega_r| \qquad (BSBM) \\
\text{s.t.} & \mathbf{P} = (P_1, \ldots, P_N) \text{ satisfying (D1) and (D2)} \\
& \kappa \geq 0.
\end{array}
$$

The $(BSBM)$ problem can be reformulated as a parametric problem, parametrized by a real positive number α, in which the error between the distances in the SBM and the scaled dissimilarities is minimized among those maps whose area error is less or equal than α, yielding the single objective problem $(\alpha - SBM)$. $(BSBM)$ can be also parametrized by a real positive number β, in which the error between the areas depicted in the SBM is minimized among those maps whose error between the scaled dissimilarities and the distances in the map is less or equal than β, yielding the single objective problem $(\beta - SBM)$:

$$
\begin{array}{ll}
\min & \displaystyle\sum_{\substack{r,s=1,\ldots,N \\ r \neq s}} |distance(P_r, P_s) - \kappa\delta_{rs}| \\
\text{s.t.} & \mathbf{P} = (P_1, \ldots, P_N) \text{ satisfying (D1) and (D2)} \\
& \displaystyle\sum_{r=1,\ldots,N} |\omega_r - area(P_r)| \leq \alpha \\
& \kappa \geq 0.
\end{array}
$$

$(\alpha - SBM)$

$$
\begin{array}{ll}
\min & \displaystyle\sum_{r=1,\ldots,N} |area(P_r) - \omega_r| \\
\text{s.t.} & \mathbf{P} = (P_1, \ldots, P_N) \text{ satisfying (D1) and (D2)} \\
& \displaystyle\sum_{\substack{r,s=1,\ldots,N \\ r \neq s}} |distance(P_r, P_s) - \kappa\delta_{rs}| \leq \beta \\
& \kappa \geq 0.
\end{array}
$$

$(\beta - SBM)$

In what follows, we formally state the problem $(\alpha - SBM)$ as a Mixed Integer Nonlinear Program (MINLP), and a similar approach can be used for $(\beta - SBM)$.

Let x_{rij}, $r = 1, \ldots, N$, $i = 1, \ldots, K$ and $j = 1, \ldots, L$, be binary variables which determine if cell (i, j) belongs to portion P_r or not, namely

$$x_{rij} = \begin{cases} 1 & \text{if cell } (i,j) \text{ belongs to portion } P_r \\ 0 & \text{otherwise.} \end{cases}$$

Thanks to these variables, we express portion P_r as $P_r(x) = \{(i,j) : x_{rij} = 1, i = 1,\dots K, j = 1,\dots,L\}$. Therefore, $(\alpha - SBM)$ can be stated as follows:

$$\min \sum_{\substack{r,s=1,\dots,N \\ r\neq s}} |distance(P_r(x), P_s(x)) - \kappa\delta_{rs}| \tag{1}$$

s.t.

$$\sum_{r=1,\dots,N} x_{rij} = 1 \qquad i = 1,\dots,K,\ j = 1,\dots,L \tag{2}$$

$$\sum_{\substack{i=1,\dots,K \\ j=1,\dots,L}} x_{rij} \geq 1 \qquad r = 1,\dots,N \tag{3}$$

$$x_{rij} \in \{0,1\} \qquad r = 1,\dots,N,\ i = 1,\dots,K,\ j = 1,\dots,L \tag{4}$$

$$\kappa \geq 0 \tag{5}$$

$$\sum_{\substack{(i'',j'')\in B\left((i,j),(i',j')\right) \\ (i'',j'')\neq(i,j) \\ (i'',j'')\neq(i',j')}} x_{ri''j''} \geq x_{rij} + x_{ri'j'} - 1 \qquad r = 1,\dots,N,\ i,i' = 1,\dots,K,\ j,j' = 1,\dots,L, \text{ such that cells } (i,j) \text{ and } (i',j') \text{ are non-adjacent} \tag{6}$$

$$\sum_{r=1,\dots,N} \left| \left(\frac{1}{KL} \sum_{\substack{i=1,\dots,K \\ j=1,\dots,L}} x_{rij} \right) - \omega_r \right| \leq \alpha. \tag{7}$$

Constraint (2) models condition (D1), since it forces that every cell must belong to exactly one portion. In order to have at least one cell assigned to every portion, constraint (3) is considered. Constraints (4) and (5) establish the type of the variables. The box-connectivity in Definition 2, and required in (D2), is enforced through constraint (6). The box-connectivity of $P_r(x)$ is enforced by imposing that the box generated by each pair of non-adjacent cells belonging to $P_r(x)$ (two cells that do not share a common boundary) must contain also cells of $P_r(x)$, namely the intersection between such box (excluding its two generator cells) and the portion must be nonempty. Finally, the error incurred by approximating the statistical values $\boldsymbol{\omega}$ by the area of the portions is modeled through constraint (7).

The objective in (1) remains to be modeled through the variables stated. Indeed, its expression will depend on the choice of the distance function. With the choices mentioned above related to Cluster Analysis, namely the single linkage (SL), the complete linkage (CL) and the average linkage (AvL), these distances can be easily expressed through the binary variables of our mathematical optimization problem, namely x_{rij}. Firstly, we need to define how to measure the distance between two single cells. Since the grid structure naturally calls for the use of the ℓ_1-norm, given two cells (i,j) and (i',j'), the distance between them is equal to $|i-i'|+|j-j'|$, where $i,i' = 1,\dots,K$ and $j,j' = 1,\dots,L$. Thus, the distance between two portions $P_r(x)$ and $P_s(x)$ on an SBM, $distance(P_r(x), P_s(x))$, can be expressed respectively as $SL(P_r(x), P_s(x))$, or

$CL(P_r(x), P_s(x))$, or $AvL(P_r(x), P_s(x))$, defined as

$$SL(P_r(x), P_s(x)) = \min_{\substack{i,i'=1,\dots,K \\ j,j'=1,\dots,L}} \left\{ |i-i'| + |j-j'| : x_{rij} = x_{si'j'} = 1 \right\}, \tag{8}$$

$$CL(P_r(x), P_s(x)) = \max_{\substack{i,i'=1,\dots,K \\ j,j'=1,\dots,L}} \left\{ |i-i'| + |j-j'| : x_{rij} = x_{si'j'} = 1 \right\}, \tag{9}$$

$$AvL(P_r(x), P_s(x)) = \frac{1}{|P_r(x)||P_s(x)|} \sum_{\substack{i,i'=1,\dots,K \\ j,j'=1,\dots,L}} (|i-i'| + |j-j'|) \cdot x_{rij} \cdot x_{si'j'}, \tag{10}$$

where $|P_r(x)|$ (resp. $|P_s(x)|$) denotes the number of cells of the portion $P_r(x)$, i.e., $|P_r(x)| = \sum_{\substack{i=1,\dots,K \\ j=1,\dots,L}} x_{rij}$.

Summing up, the problem $(\alpha - SBM)$ is stated as an MINLP as

$$\min \{(1) : \text{s.t. } (2) - (7)\}, \tag{$\alpha - SBM$}$$

where $distance(P_r(x), P_s(x))$ in (1) is replaced by either $SL(P_r(x), P_s(x))$, $CL(P_r(x), P_s(x))$ or $AvL(P_r(x), P_s(x))$, as in (8)-(10) respectively.

One has that, for the SL and CL distances, $(\alpha - SBM)$ can be easily reformulated as a MILP, and as an approximation to an MILP in the case of AvL. These three reformulations of $(\alpha - SBM)$, one for each distance function, are included in the Appendix, and they are called $(\alpha - SBM)^L_{SL}$, $(\alpha - SBM)^L_{CL}$ and $(\alpha - SBM)^L_{AvL}$, respectively. Note that the statement of the problem $(\beta - SBM)$ as an MINLP and its pertinent reformulation as an MILP or an approximation to an MILP, yielding $(\beta - SBM)^L_{SL}$, $(\beta - SBM)^L_{CL}$ and $(\beta - SBM)^L_{AvL}$, are straightforward from the work done for $(\alpha - SBM)$.

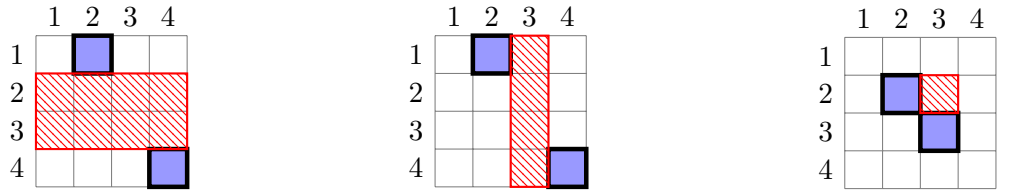
2.3 A tighter model

Problems $(\alpha - SBM)$ and $(\beta - SBM)$ enforce the box-connectivity of the portions $\boldsymbol{P}$ in the SBM through constraint (6). There exist several attempts in the literature which deal with the problem of modeling connectivity with integer programming, for instance using graph theory, [31, 43], designing the connected portions according to fixed locations, [44], or considering *node-cut sets*, [8, 58].

Definition 3. *Let $G = (U, E)$ be a graph, whose nodes U are the cells of a (K, L)-grid. If two nodes $u, u' \in U$ are adjacent on the grid, then there exists an edge $(u, u') \in E$ linking the two nodes.*

Given two non-adjacent nodes $u, v \in U$, a set of nodes $C \subseteq U \setminus \{u, v\}$ is a node-cut separating u and v if there is no path between u and v in the subgraph $G' = (U \setminus C, E \setminus E')$, where $E' = \{(w, w') : w, w' \in C \cup \{u, v\}\}$.

Figure 2 illustrates some possible node-cuts on a $(4, 4)$-grid: a horizontal cut for cells $(1, 2)$ and $(4, 4)$ in Figure 2 (a), a vertical cut for the same cells in Figure 2 (b) and a diagonal cut in Figure 2 (b) for cells $(2, 2)$ and $(3, 3)$.



(a) Horizontal node-cut (b) Vertical node-cut (c) Diagonal node-cut

Figure 2: Dashed cells illustrate node-cuts of shaded cells on a $(4, 4)$-grid

Thanks to Definition 3, the box-connectivity constraint, namely (6), can be modeled through intersections of node-cuts yielding an equivalent formulation for problems $(\alpha - SBM)$ and $(\beta - SBM)$. In what follows, we prove that box-connectivity can be equivalently modeled through the following set of constraints:

$$\sum_{\substack{i \le i'' < i' \\ \min\{j,j'\} \le j'' \le \max\{j,j'\}}} x_{ri''j''} \ge x_{rij} + x_{ri'j'} - 1 \qquad r = 1, \dots, N, \quad i = 1, \dots, K-2, \; i' = i+2, \dots, K, \; j, j' = 1, \dots, L, \tag{6a}$$

$$\sum_{\substack{\min\{i,i'\} \le i'' \le \max\{i,i'\} \\ j < j'' < j'}} x_{ri''j''} \ge x_{rij} + x_{ri'j'} - 1 \qquad r = 1, \dots, N, \quad i, i' = 1, \dots, K, \; j = 1, \dots, L-2, \; j' = j+2, \dots, L \tag{6b}$$

$$x_{r,i-1,j} + x_{r,i,j-1} \ge x_{rij} + x_{r,i-1,j-1} - 1 \qquad r = 1, \dots, N, \quad i = 2, \dots, K, \; j = 2, \dots, L \tag{6c}$$

$$x_{r,i-1,j} + x_{r,i,j+1} \ge x_{rij} + x_{r,i-1,j+1} - 1 \qquad r = 1, \dots, N, \quad i = 2, \dots, K, \; j = 1, \dots, L-1 \tag{6d}$$

$$x_{r,i+1,j} + x_{r,i,j-1} \ge x_{rij} + x_{r,i+1,j-1} - 1 \qquad r = 1, \dots, N, \quad i = 1, \dots, K-1, \; j = 2, \dots, L \tag{6e}$$

$$x_{r,i+1,j} + x_{r,i,j+1} \ge x_{rij} + x_{r,i+1,j+1} - 1 \qquad r = 1, \dots, N, \quad i = 1, \dots, K-1, \; j = 1, \dots, L-1 \tag{6f}$$

Constraint (6a) considers bounded horizontal node-cuts, in the sense that it considers that if two cells belong to the same portion and are placed in different and non-contiguous rows, namely there is at least a row in between, then there must exist cells belonging also to that portion in the rows strictly in between them. On the other hand, constraint (6b) addresses bounded vertical node-cuts, which analogous to the horizontal ones but considering columns instead of rows. Finally, constraints (6c)-(6f) model diagonal node-cuts, as this in Figure 2 (c), namely if two cells belong to the same portion and they touch in one corner (shaded cells), then one of the two cells in the diagonal crossing that corner must also belong to such portion (striped cells).

Let us consider the problem $(\alpha - SBM)$ stated as

$$(\alpha - SBM)_{box} = \min \{(1) : \text{s.t. } (2) - (7)\}$$

and

$$(\alpha - SBM)_{cut} = \min\{(1) : \text{s.t. } (2) - (5), (6a) - (6f), (7)\}.$$

Proposition 1. *One has that problems* $(\alpha - SBM)_{box}$ *and* $(\alpha - SBM)_{cut}$ *are equivalent.*

Proof. Observe that constraints (2)-(5) and (7) appear in both problems, and thus it remains to prove that constraint (6) is equivalent to constraints (6a)-(6f). Let us prove such statement for a portion $P_r(x)$, denoted S to simplify.

Let us prove that, if S is box-connected, then (6a)-(6f) hold.

Suppose that (6a) fails. Then, the summation on the left-hand-side is equal to zero, which means that there exist two cells $(i,j),(i',j') \in S$, such that $i < i' - 1$ and

$$\left\{(i'',j'') : i < i'' < i', \min\{j,j'\} \leq j'' \leq \max\{j,j'\}\right\} \cap S = \emptyset. \tag{11}$$

In particular, (11) is also true when $j' = j$. This means

$$\left\{(i'',j) : i < i'' < i'\right\} \cap S = \emptyset. \tag{12}$$

On the other hand, S is box-connected and then,

$$\exists (i'',j) \in B\left((i,j),(i',j)\right) \cap S, \text{ such that } i'' \neq i \text{ and } i'' \neq i',$$

which is clearly in contradiction with (12).

Using a similar procedure, we can prove that (6b)-(6f) also hold.

Conversely, let us prove that, if (6a)-(6f) are satisfied, then S is box-connected.

Let $(i,j),(i',j') \in S$ such that $i < i' - 1$. Since (6a) is satisfied, there exists $(i'',j'') \in S$ such that $i < i'' < i'$ and $\min\{j,j'\} \leq j'' \leq \max\{j,j'\}$, which implies that $(i'',j'') \in B\left((i,j),(i',j')\right) \cap S$, and it is different from (i,j) and (i',j'). This proves the desired result.

The case in which $j < j' - 1$, constraint (6b) applies, showing that S is box-connected. Finally, for pairs of cells belonging to S such that $\{(i,j),(i-1,j-1)\}$, $\{(i,j),(i-1,j+1)\}$, $\{(i,j),(i+1,j-1)\}$ and $\{(i,j),(i+1,j+1)\}$, constraint (6c)-(6f) yield, respectively, the box-connectivity of S.

□

In order to choose between the $(\alpha - SBM)_{box}$ or the $(\alpha - SBM)_{cut}$ formulations, we study their continuous relaxation tightness. The formulation with the tightest continuous relaxation will be the preferred. Thus, let us consider the continuous relaxations of the MINLP problems $(\alpha - SBM)_{box}$ and $(\alpha - SBM)_{cut}$, obtained by substituting constraint (4) by $0 \leq x_{rij} \leq 1$, $r = 1,\ldots,N$, $i = 1,\ldots,K$, $j = 1,\ldots,L$, and denoted as $R((\alpha - SBM)_{box})$ and $R((\alpha - SBM)_{cut})$ respectively. We show in what follows that, while the MINLP problems $(\alpha - SBM)_{box}$ and $(\alpha - SBM)_{cut}$ are equivalent, as stated in Proposition 1, it turns out that $R((\alpha - SBM)_{cut})$, is, in general, tighter than $R((\alpha - SBM)_{box})$.

Let $v(R((\alpha - SBM)_{box}))$ and $v(R((\alpha - SBM)_{cut}))$ be the optimal objective values of problems $R((\alpha - SBM)_{box})$ and $R((\alpha - SBM)_{cut})$, respectively.

Proposition 2. *One has that*

$$v(R((\alpha - SBM)_{box})) \leq v(R((\alpha - SBM)_{cut})). \tag{13}$$

Proof. In order to show inequality (13), we prove that each solution x to $R(\alpha-(SBM)_{cut})$ is also feasible to $R((\alpha-SBM)_{box})$. Let x be a feasible solution to $R(\alpha-(SBM)_{cut})$, we need to prove that x satisfies constraints (2)-(7). First, constraints (2)-(5) and (7) are satisfied, since they appear in both problems. Then, let us check that (6) also holds. Since x verifies (6a), one has

$$x_{rij}+x_{ri'j'}-1 \leq \sum_{\substack{i<i''<i' \\ \min\{j,j'\}\leq j''\leq \max\{j,j'\}}} x_{ri''j''} \leq \sum_{\substack{(i'',j'')\in B((i,j),(i',j')) \\ (i'',j'')\neq(i,j) \\ (i'',j'')\neq(i',j')}} x_{ri''j''} \quad r=1,\ldots,N$$

$$i=1,\ldots,K-2,\ i'=i+2,\ldots,L$$
$$j,j'=1,\ldots,L.$$

In other words, x is shown to satisfy (6) because it also satisfies (6b)-(6f), and then the result holds. □

It turns out that the feasible region of $R(\alpha-(SBM)_{cut})$ is smaller than the feasible region of $R((\alpha-SBM)_{box})$, due to problem instances for which feasible solutions to $R(\alpha-(SBM)_{box})$ are unfeasible for $R(\alpha-(SBM)_{cut})$. For instance, let us consider $N=2$, $K=3$ and $L=2$. One has that the following solution is feasible for $R(\alpha-(SBM)_{box})$ but unfeasible for $R(\alpha-(SBM)_{cut})$, taking $\alpha=\infty$ and $\kappa\geq 0$:

$$x_{111}=\frac{3}{4} \quad x_{112}=\frac{1}{5} \quad x_{121}=\frac{1}{5} \quad x_{122}=\frac{1}{5} \quad x_{131}=\frac{1}{5} \quad x_{132}=\frac{3}{4}$$

$$x_{211}=\frac{1}{4} \quad x_{212}=\frac{4}{5} \quad x_{221}=\frac{4}{5} \quad x_{222}=\frac{4}{5} \quad x_{231}=\frac{4}{5} \quad x_{232}=\frac{1}{4}$$

The reasoning applied for finding a tighter reformulation to problem $(\alpha-SBM)$ also holds for the problem $(\beta-SBM)$. MILPs reformulations described in the Appendix also hold in case that problems $(\alpha-SBM)$ and $(\beta-SBM)$ are modeled through connectivity constraints (6a)-(6f) instead of (6). These MILPs reformulations let us solve to optimality small instances of the optimization problems, i.e., when the number of individuals is not very large as well as the (K,L)-grid is coarse enough. However, big instances remain hard to solve by standard MILPs optimizers, and heuristic techniques seem to be necessary to handle real-world examples. In Section 3 we describe how the LNS metaheuristic can be integrated in our methodology to construct SBMs.

3 A Large Neighborhood Search approach

In this section, we adapt the Large Neighborhood Search (LNS) methodology to the problem of building an SBM. Since small instances of the MILPs reformulations of problems $(\alpha-SBM)$ and $(\beta-SBM)$ can be quickly solved by standard MILP optimizers and thanks to the grid structure considered, LNS seems to be a good candidate to take advantage from such facts. The LNS metaheuristic was first proposed in [50]. It has been successfully applied in recent years to problems of different nature, for instance Vehicle Routing Problems, [16, 40, 48], Scheduling, [35, 46] and Visualization [59]. Roughly speaking, LNS performs a search for good solutions in a neighborhood of a starting point. A neighborhood of a certain point contains all the solutions that can be reached from the starting one by a *destroy* procedure, which erases part of

the solution, and a *repair* method, which rebuilds the previously destroyed solution in order to obtain a new and better one. The good performance of LNS does not only depend on the quality of the starting point, but also on the destroy and repair methods. A tradeoff between the degree of destruction and the rebuilding process needs to be established: the degree of destruction should be such that a large part of the search space can be explored, and the rebuilding process should be efficient. Algorithm 1 contains the pseudocode of the LNS metaheuristic. For further details on LNS see [47] and references therein.

Algorithm 1 LNS pseudocode, [47]

Input: A feasible solution x, an objective function f, *destroy* and *repair* procedures

$x^* \leftarrow x$
repeat
  $x^t \leftarrow repair(destroy(x^*))$;
  if $f(x^t) < f(x^*)$ **then**
    $x^* \leftarrow x^t$;
  end if
until stop condition is met

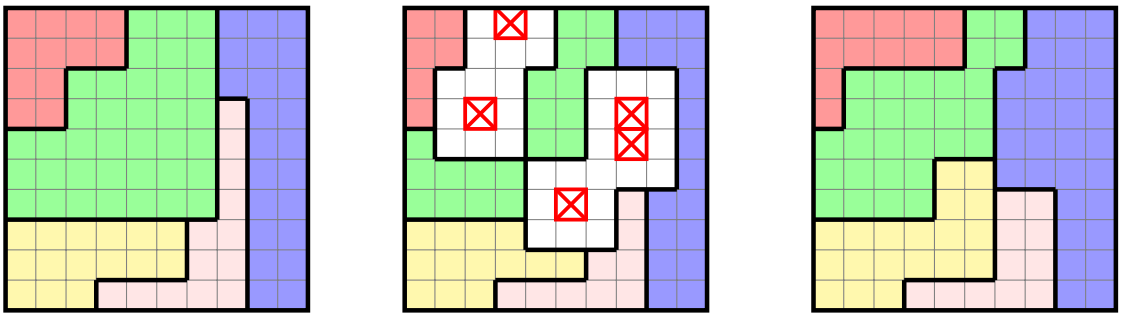
Output: x^*

In order to explain the destroy and repair procedures for our model, we introduce some necessary concepts. We define the *incidence degree* γ of a cell (i,j) in portion P_r as the number of connected cells surrounding (i,j) that also belong to P_r, namely $(i-1,j)$, $(i,j+1)$, $(i+1,j)$ and $(i,j-1)$. We say that a cell (i,j) is *redundant* if its removal keeps the portion which it belongs still connected. Observe that a cell is redundant if $\gamma = 0, 1, 2$ or 3, while $\gamma = 4$ implies that it is non-redundant.

Given an SBM, $\boldsymbol{P}$, the destroy phase consists of selecting randomly a number μ of redundant cells in $\boldsymbol{P}$, and removing them as well as their eight surrounding neighbors, yielding an incomplete SBM, $destroy(\boldsymbol{P})$. Figure 3 (b) contains $\mu = 5$ selected redundant cells (crossed bold) from the initial solution in Figure 3 (a), and in white their surrounded neighbors, which are also deleted. The repair step consists of rebuilding the destroyed solution $destroy(\boldsymbol{P})$ by assigning the free cells to portions satisfying (D1)-(D4), yielding a new SBM, $repair(destroy(\boldsymbol{P}))$, see Figure 3 (c).

The decision of selecting redundant cells in the destroy stage instead of non-redundant ones is not arbitrary. If only non-redundant cells were chosen, the risk of getting stuck on the initial SBM, $\mathbf{P}$, increases. The reason comes from imposing box-connectivity in the portions forming the $repair(destroy(\boldsymbol{P}))$ SBM, which may force to reconstruct $\mathbf{P}$ time after time because any other partition is feasible. On the other hand, selecting only redundant cells yields to a $destroy(\boldsymbol{P})$ configuration with more freedom (regarding to box-connectivity) to obtain a new allocation of the free cells distinct from $\boldsymbol{P}$.

As we highlighted in the previous section, small instances of the MILP reformulations of $(\alpha - SBM)$ and $(\beta - SBM)$ can be effortlessly solved by standard MILPs optimization routines, and that is the key of our repairing procedure. To obtain the repaired solution, cells which have not been removed in the destroy stage become constants in the MILP problem, in the sense that their values are known and they are thus fixed. Those cells that have been removed in the destroy step remain decision variables when solving the MILP problem. Thanks to this



(a) Initial solution: $\boldsymbol{P}$ (b) $destroy(\boldsymbol{P})$ (c) $repair(destroy(\boldsymbol{P}))$

Figure 3: LNS algorithm for SBM

variable-fixing procedure, the number of constraints is also substantially reduced, since there are already cells assigned to portions (reducing the number of constraints in (2) and (3)), there are also connected parts in the portions (reducing then the number of constraints in (6a)-(6f)), and even the number of extra constraints included in the linear reformulations. Therefore, the size of the MILPs to solve in the repairing procedure is considerably reduced compared to the full-size problems. Thus, depending on which criteria is to be improved in the current iteration of the LNS algorithm, either the MILP version of $(\alpha - SBM)$ or $(\beta - SBM)$ is solved when repairing the previously destroyed SBM to find a better one.

4 Numerical illustrations

The performance of the approach described in previous sections is tested in three data sets of different sizes and nature on a $(40, 40)$-grid. The LNS algorithm has been coded in AMPL, [25], and all the MILPs problems have been solved with CPLEX v12.6, [12], with a time limit of five minutes, on a PC Intel® Core™ i7-2600K, 16GB of RAM.

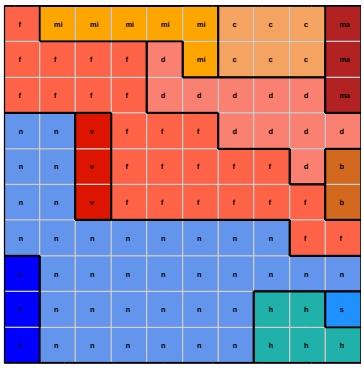
The first dataset, `Markets`, consists of $N = 11$ financial markets across Europe and Asia. The statistical value ω_r relates to the importance of market r relative to the world market portfolio, [21], and the dissimilarity δ_{rs} is based on the correlation between markets r and s, [3]. The second dataset, `Morse`, comes from a study regarding the confusion between the acoustic Morse signals used for the $N = 26$ letters of the English alphabet, [49]. The dissimilarity δ_{rs} between a signal r and a signal s is computed as the average percentage with which the answer 'Same!' was given in each combination of those signals, [3]. The statistical values $\boldsymbol{\omega}$ come from the relative frequencies of letters in the English language, [37]. In the last dataset, `Netherlands`, the individuals are the $N = 12$ provinces of The Netherlands, and the statistical values $\boldsymbol{\omega}$ are their population rates, see [54]. The dissimilarity between two provinces is related to their geographical location, and this is measured as the length of the shortest path in the graph obtained from considering two nodes adjacent if the corresponding provinces are adjacent in the geographical map.

In order to compute an SBM with the methodology described in this paper, some decisions should be made in advance. Firstly, we need to choose between any of the distances proposed in Section 2.1, namely single, complete or average linkage. Due to the characteristics of the problem we are dealing with, we consider that the average linkage reflects properly the visualization aim

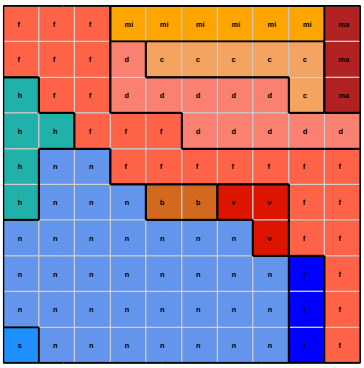
of the SBM, since it summarizes the distance information between every pair of cells belonging to the portions and thus, they become something global instead of something intrinsic of a single cell as the single or complete linkage would do. Thus, the mathematical optimization models considered throughout this section are $(\alpha - SBM)^L_{AvL}$ and $(\beta - SBM)^L_{AvL}$, in which box-connectivity has been modeled through constraints (6a)-(6f). In order to perform the LNS procedure described in Section 3, we need to determine how the redundant cells are selected in the destroy stage and which mathematical optimization problem, either $(\alpha - SBM)^L_{AvL}$ and $(\beta - SBM)^L_{AvL}$, is solved in the repair phase. On one hand, in the destroy step, μ redundant cells are selected with nonuniform probabilities, depending on their incidence degree: a redundant cell (i, j) with incidence degree equal to $\gamma = 0, \ldots, 3$, will be selected with probability proportional to $2^{3-\gamma}$. This way, cells with a low incidence degree, and thus those which have more chances to be allocated to different portions, are selected with higher probability. On the other hand, when repairing the destroyed solution, we assume that we have an SBM for which the error when approximating $\boldsymbol{\omega}$ by the portions' areas is pretty low, and therefore we can use a small value of parameter α. Then, the problem $(\alpha - SBM)^L_{AvL}$ is solved in this phase to improve the dissimilarities fit. We made this decision because we considered that it is crucial that areas are very well fitted, whereas the dissimilarities admit more flexibility when interpreting the SBM. As a stopping condition for LNS we establish a maximum number of iterations.

In order to construct an SBM on a $(40, 40)$-grid for the three previously described datasets, we firstly obtain an initial SBM on a $(10, 10)$-grid by solving a surrogate of problem $(\beta - SBM)^L_{AvL}$, which has an accurate representation of the statistical values as the area of the portions and also some information about the dissimilarities. To do that, the problem $(\beta - SBM)^L_{AvL}$ is solved, setting $\beta = \infty$, including an extra set of constraints that imposes that some cells are already assigned to some portions. These cells are found accordingly to the dissimilarities between individuals by using the Multidimensional Scaling for Rectangular Maps described in [5]. Then, 100 iterations of the LNS algorithm are performed, with $\mu = 4$ and α equal to the objective value obtained when solving the surrogate $(\beta - SBM)^L_{AvL}$. The so-obtained SBM is embedded into a $(20, 20)$-grid, and it is considered as an initial SBM for 50 iterations of the LNS algorithm with $\mu = 8$ and α a 15% smaller than the objective value obtained when solving the surrogate $(\beta - SBM)^L_{AvL}$ in the first stage. Finally, the so-obtained SBM is embedded into a $(40, 40)$-grid, and it is considered as an initial SBM for 25 iterations of the LNS algorithm with $\mu = 16$ and α a 5% smaller than the objective value obtained when solving the surrogate $(\beta - SBM)^L_{AvL}$ in the previous stage.

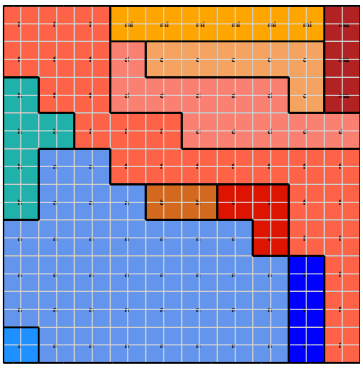
The process to construct an SBM for `Markets` dataset is illustrated in Figure 4. The initial SBM on a $(10, 10)$-grid obtained by solving the surrogate $(\beta - SBM)^L_{AvL}$ is depicted in Figure 4 (a). After 100 iterations of LNS the SBM in Figure 4 (b) is obtained. Then, this SBM is embedded into a $(20, 20)$-grid, see Figure 4 (c), and it is set as the initial solution for the LNS algorithm. After 50 iterations and a reduction of a 15% of the parameter α, the solution depicted in Figure 4 (d) is obtained. This SBM is embedded into a $(40, 40)$-grid, see Figure 4 (e), and it is set as the initial solution for the LNS algorithm. After 25 iterations and imposing a 5% reduction of the parameter α, the solution in Figure 4 (f) is obtained. Figures 5-6 show the SBMs for `Morse` and `Netherlands` datasets, which are also obtained with the procedure described above.



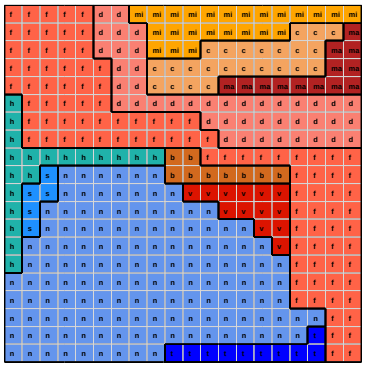
(a) Initial SBM on a $(10, 10)$-grid



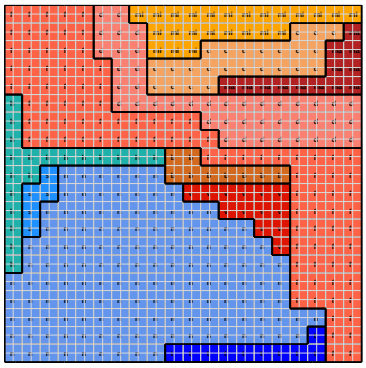
(b) SBM on a $(10, 10)$-grid after LNS



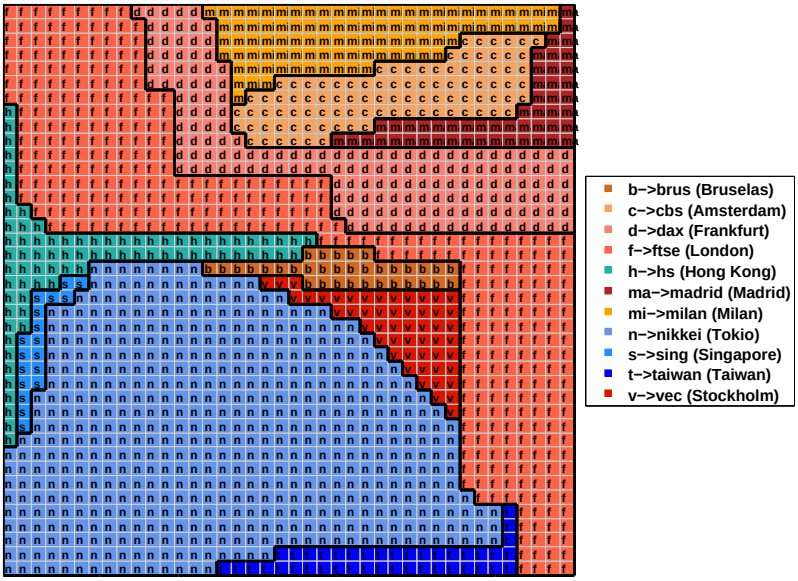
(c) Initial SBM on a $(20, 20)$-grid



(d) SBM on a $(20, 20)$-grid after LNS



(e) Initial SBM on a $(40, 40)$-grid



(f) SBM on a $(40, 40)$-grid after LNS

Figure 4: Illustration of the methodology proposed to construct an SBM for `Markets` dataset using a $(40, 40)$-grid.

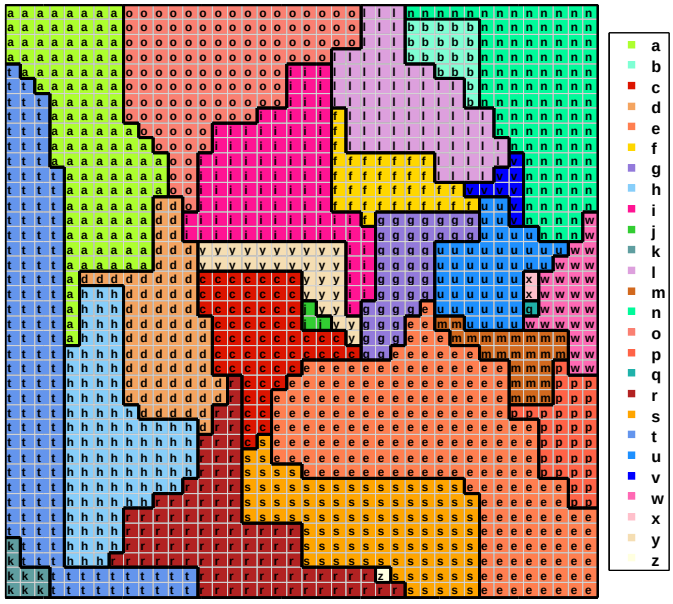


Figure 5: SBM for `Morse` using a $(40, 40)$-grid.

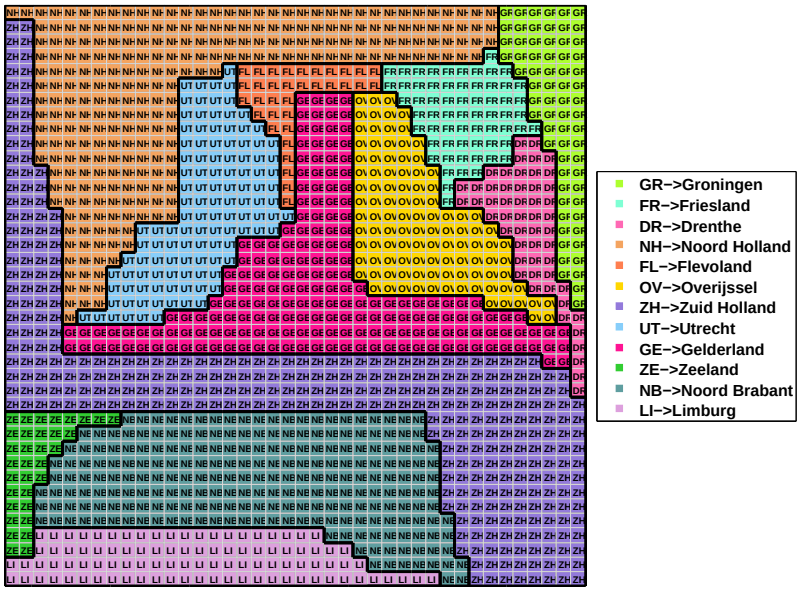


Figure 6: SBM for `Netherlands` using a $(40, 40)$-grid.

5 Conclusions and future research

In this paper, we have formally introduced a mathematical optimization formulation for the problem of visualizing as a space-filling box-connected map a set of individuals which have attached a statistical variable, represented as a proportion, and a dissimilarity measure. This visualization problem is addressed by simultaneously optimizing the error incurred when approximating the statistical values by the area of the portions and the dissimilarities by the distances among them, satisfying the condition that the portions are box-connected and they must form a partition of the unit square. Such approach yields a biobjective optimization problem, which is solved as a single-objective problem that optimizes the fit in distances ensuring a very accurate fit in the sizes of the portions. The Large Neighborhood Search metaheuristic has been proposed to deal with big instances of such problems, due to the fact that the combinatorial structure underlying in the optimization problems makes solving them to optimality too time-demanding. The usefulness of our approach has been illustrated in a variety of data sets, related to market indices, the Morse code and a geographical map.

There are several interesting extensions to the methodology proposed in this paper which deserve further study. For instance, considering visualization regions different from the unit square, in which a regular grid cannot be obtained in a straightforward manner, is a challenging problem when modeling portions' connectivity, [8, 58]. Representing simultaneously dissimilarities not only between individuals but also between groups of individuals is a demanding task which deserves further study, [11, 30]. This extension can be applied when designing the arrangement of the members of a parliament, since all the members belonging to the same party should be seated close to each other, in the sense that they must form a connected portion and everybody must be as close as possible to everyone in his/her party. In addition, the different parties should be ideally placed in the parliament accordingly to their ideology. There are many contexts in which dissimilarities and the statistical values vary over time, for instance stock markets, or in which individual appear and disappear over time, for instance most used words related to a trending topic in Twitter. Thus, modeling dynamical SBMs which can handle these temporal changes seems to be also a very interesting problem, [13, 15, 41].

References

[1] M.J. Alam, T. Biedl, S. Felsner, M. Kaufmann, S.G. Kobourov, and T. Ueckerdt. Computing cartograms with optimal complexity. *Discrete & Computational Geometry*, 50(3):784–810, 2013.

[2] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis. *Graph drawing: algorithms for the visualization of graphs*. Prentice Hall PTR, 1998.

[3] I. Borg and P.J.F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2005.

[4] K. Buchin, D. Eppstein, M. Löffler, M. Nöllenburg, and R. I. Silveira. Adjacency-preserving spatial treemaps. In *Algorithms and Data Structures*, pages 159–170. Springer, 2011.

[5] E. Carrizosa, V. Guerrero, and D. Romero Morales. A multi-objective approach to visualize adjacencies in weighted graphs by rectangular maps. Technical report, Optimization Online, 2015. http://www.optimization-online.org/DB_HTML/2015/12/5226.html.

[6] E. Carrizosa, V. Guerrero, and D. Romero Morales. Visualizing data as objects by DC (difference of convex) optimization. Technical report, Optimization Online, 2015. http://www.optimization-online.org/DB_HTML/2015/12/5227.html.

[7] E. Carrizosa and D. Romero Morales. Supervised classification and mathematical optimization. *Computers & Operations Research*, 40(1):150–165, 2013.

[8] R. Carvajal, M. Constantino, M. Goycoolea, J. P. Vielma, and A. Weintraub. Imposing connectivity constraints in forest planning models. *Operations Research*, 61(4):824–836, 2013.

[9] C. P. Chen and C.-Y. Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314–347, 2014.

[10] J. Choo and H. Park. Customizing computational methods for visual analytics with big data. *IEEE Computer Graphics and Applications*, 33(4):22–28, 2013.

[11] S. Clémençon, H. De Arazoza, F. Rossi, and V.C. Tran. Hierarchical clustering for graph visualization. In *19th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2011)*, pages 227–232, 2011.

[12] IBM ILOG CPLEX. http://www.ilog.com/products/cplex/, 2014.

[13] W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu. Context preserving dynamic word cloud visualization. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 121–128. IEEE, 2010.

[14] M. de Berg, E. Mumford, and B. Speckmann. Optimal BSPs and rectilinear cartograms. *International Journal of Computational Geometry & Applications*, 20(02):203–222, 2010.

[15] R. Dantas de Pinho, M. C. Ferreira de Oliveira, and A. de Andrade Lopes. An incremental space to visualize dynamic data sets. *Multimedia Tools and Applications*, 50(3):533–562, 2010.

[16] E. Demir, T. Bektaş, and G. Laporte. An adaptive large neighborhood search heuristic for the pollution-routing problem. *European Journal of Operational Research*, 223(2):346–359, 2012.

[17] M. Dörk, S. Carpendale, and C. Williamson. Visualizing explicit and implicit relations of complex information spaces. *Information Visualization*, 11(1):5–21, 2012.

[18] D. Dorling. Area cartograms: their use and creation. In *Concepts and Techniques in Modern Geography series no. 59*. University of East Anglia: Environmental Publications, 1996.

[19] G. M. Draper, Y. Livnat, and R. F. Riesenfeld. A survey of radial methods for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 15(5):759–776, 2009.

[20] F. S. Duarte, F. Sikansi, F. M. Fatore, S. G. Fadel, and F. V. Paulovich. Nmap: A novel neighborhood preservation space-filling algorithm. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2063–2071, 2014.

[21] T. Flavin, M. Hurley, and F. Rousseau. Explaining stock market correlation: A gravity model approach. *The Manchester School*, 70:87–106, 2002.

[22] R. Fortet. Applications de l'algèbre de Boole en recherche opérationelle. *Revue Française de Recherche Opérationelle*, 4:17–26, 1960.

[23] K. Fountoulakis and J. Gondzio. Performance of first- and second-order methods for ℓ_1-regularized least square problems. Technical Report ERGO-15-005, 2015. arXiv:1503.03520.

[24] K. Fountoulakis and J. Gondzio. A second-order method for strongly convex ℓ_1-regularization problems. *Mathematical Programming*, 156(1):189–219, 2016.

[25] R. Fourer, D. M. Gay, and B. W. Kernighan. *AMPL: A Modeling Language for Mathematical Programming*. Thomson/Brooks/Cole, 2003.

[26] O. Fried, S. DiVerdi, M. Halber, E. Sizikova, and A. Finkelstein. Isomatch: Creating informative grid layouts. In *Computer Graphics Forum*, volume 34, pages 155–166. Wiley Online Library, 2015.

[27] E. Gómez-Nieto, F. San Roman, P. Pagliosa, W. Casaca, E. S. Helou, M. C. F. de Oliveira, and L. G. Nonato. Similarity preserving snippet-based visualization of web search results. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):457–470, 2014.

[28] P. Hansen and B. Jaumard. Cluster analysis and mathematical programming. *Mathematical Programming*, 79(1-3):191–215, 1997.

[29] R. Heilmann, D. A. Keim, C. Panse, and M. Sips. Recmap: Rectangular map approximations. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 33–40. IEEE Computer Society, 2004.

[30] I. Herman, G. Melançon, and M.S. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000.

[31] N. Jafari and J. Hearne. A new method to solve the fully connected reserve network design problem. *European Journal of Operational Research*, 231(1):202–209, 2013.

[32] I. Jankovits, C. Luo, M. F. Anjos, and A. Vannelli. A convex optimisation framework for the unequal-areas facility layout problem. *European Journal of Operational Research*, 214(2):199–215, 2011.

[33] M. Klimenta and U. Brandes. Graph drawing by classical multidimensional scaling: new perspectives. In *Graph Drawing*, volume 7704, pages 55–66. Springer, 2013.

[34] J. S. Kochhar, B. T. Foster, and S. S. Heragu. HOPE: a genetic algorithm for the unequal area facility layout problem. *Computers & Operations Research*, 25(7):583–594, 1998.

[35] J. E. Korsvik, K. Fagerholt, and G. Laporte. A large neighbourhood search heuristic for ship routing and scheduling with split loads. *Computers & Operations Research*, 38(2):474–483, 2011.

[36] M. van Kreveld and B. Speckmann. On rectangular cartograms. *Computational Geometry*, 37(3):175–187, 2007.

[37] R. E. Lewand. *Cryptological Mathematics*. The Mathematical Association of America, Washington, D. C., 2000.

[38] S. Liu, W. Cui, Y. Wu, and M. Liu. A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12):1373–1393, 2014.

[39] X. Liu, Y. Hu, S. North, and H. W. Shen. Correlatedmultiples: Spatially coherent small multiples with constrained multi-dimensional scaling. *Computer Graphics Forum*, pages 1–12, 2015.

[40] A. Lodi, E. Malaguti, N. E. Stier-Moses, and T. Bonino. Design and control of public-service contracts and an application to public transportation systems. *Management Science*, 62(4):1165–1187, 2015.

[41] D. Mashima, S. G. Kobourov, and Y. Hu. Visualizing dynamic data with maps. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1424–1437, 2012.

[42] S. Olafsson, X. Li, and S. Wu. Operations research and data mining. *European Journal of Operational Research*, 187(3):1429–1448, 2008.

[43] H. Önal and R. A. Briers. Optimal selection of a connected reserve network. *Operations Research*, 54(2):379–388, 2006.

[44] H. Önal, Y. Wang, S. T. M. Dissanayake, and J. D. Westervelt. Optimal design of compact and functionally contiguous conservation management areas. *European Journal of Operational Research*, 251(3):957–968, 2016.

[45] J. Owen-Smith, M. Riccaboni, F. Pammolli, and W.W. Powell. A comparison of U.S. and European university-industry relations in the life sciences. *Management Science*, 48(1):24–43, 2002.

[46] D. Pacino and P. Van Hentenryck. Large neighborhood search and adaptive randomized decompositions for flexible jobshop scheduling. In *International Joint Conference on Artificial Intelligence*, 2011.

[47] D. Pisinger and S. Ropke. Large neighborhood search. In M. Gendreau and J. Y. Potvin, editors, *Handbook of Metaheuristics*, volume 146, chapter 13, pages 399–419. Springer US, 2010.

[48] G. M. Ribeiro and G. Laporte. An adaptive large neighborhood search heuristic for the cumulative capacitated vehicle routing problem. *Computers & Operations Research*, 39(3):728–735, 2012.

[49] E. Z. Rothkopf. A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology*, 53(2):94, 1957.

[50] P. Shaw. Using constraint programming and local search methods to solve vehicle routing problems. In M. Maher and J.F. Puget, editors, *Principles and Practice of Constraint Programming - CP98*, volume 1520, pages 417–431. Springer, Berlin Heidelberg, 1998.

[51] H. D. Sherali, B.M.P. Fraticelli, and R.D. Meller. Enhanced model formulations for optimal facility layout. *Operations Research*, 51(4):629–644, 2003.

[52] B. Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, 11(1):92–99, 1992.

[53] I. Spence and S. Lewandowsky. Displaying proportions and percentages. *Applied Cognitive Psychology*, 5(1):61–77, 1991.

[54] Statistics Netherlands. Population; gender, age, marital status and region, January 1. www.cbs.nl, 2013. Retrieved on: 2013-10-31.

[55] G. Strong and M. Gong. Self-sorting map: An efficient algorithm for presenting multimedia data in structured layouts. *IEEE Transactions on Multimedia*, 16(4):1045–1058, 2014.

[56] J. Thomas and P.C. Wong. Visual analytics. *IEEE Computer Graphics and Applications*, 24(5):20–21, 2004.

[57] W. Tobler. Thirty five years of computer cartograms. *Annals of the Association of American Geographers*, 94(1):58–73, 2004.

[58] Y. Wang, A. Buchanan, and S. Butenko. On imposing connectivity constraints in integer programs. Technical report, Optimization Online, 2015. http://www.optimization-online.org/DB_HTML/2015/02/4768.html.

[59] V. Yoghourdjian, T. Dwyer, G. Gange, S. Kieffer, K. Klein, and K. Marriott. High-quality ultra-compact grid layout of grouped networks. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):339–348, 2016.

Appendix

In this appendix, the MINLP problem $(\alpha - SBM)$ problem is formally stated as an MILP, when using the Single and the Complete Linkage (SL and CL, respectively), defined in (8) and (9), and as an approximation to an MILP when considering the Average Linkage (AvL), stated in (10). In all the cases, we use the usual techniques to linearize the product of binary variables proposed in [22] and the absolute values both in the objective function and constraint (7). Thus, we consider the following reformulations **R1**, **R2** and **R3** stated as follows:

R1: let $u_{rsiji'j'}$ be defined as $u_{rsiji'j'} = x_{rij} \cdot x_{si'j'}$, which implies that

$$u_{rsiji'j'} = \begin{cases} 1 & \text{if cell } (i,j) \text{ belongs to portion } P_r \text{ and the } (i',j') \text{ belongs to portion } P_s \\ 0 & \text{otherwise.} \end{cases}$$

The following set of constraints is also needed to be included in the linear reformulation to properly linearize the product:

$$u_{rsiji'j'} \leq x_{rij} \qquad r,s = 1,\dots N,\ i,i' = 1,\dots,K,\ j,j' = 1,\dots,L \tag{14}$$
$$u_{rsiji'j'} \leq x_{si'j'} \qquad r,s = 1,\dots N,\ i,i' = 1,\dots,K,\ j,j' = 1,\dots,L \tag{15}$$
$$x_{rij} + x_{si'j'} \leq u_{rsiji'j'} + 1 \qquad r,s = 1,\dots N,\ i,i' = 1,\dots,K,\ j,j' = 1,\dots,L \tag{16}$$
$$u_{rsiji'j'} \in \{0,1\} \qquad r,s = 1,\dots N,\ i,i' = 1,\dots,K,\ j,j' = 1,\dots,L. \tag{17}$$

R2: The objective function of $(\alpha - SBM)$ is written as a linear objective subject to additional constraints by considering the positive continuous variables $\varphi_{rs}, \psi_{rs} \geq 0$, $r,s = 1,\dots,N$, $r \neq s$:

$$\min \left\{ \sum_{\substack{r,s=1,\dots,N \\ r \neq s}} |distance(P_r(x), P_s(x)) - \kappa\delta_{rs}| : \text{ s.t. } (2)-(7) \right\} =$$
$$= \min \sum_{\substack{r,s=1,\dots,N \\ r \neq s}} (\varphi_{rs} + \psi_{rs})$$
$$\text{s.t.} \begin{cases} (2)-(7) & \\ distance(P_r(x), P_s(x)) - \kappa\delta_{rs} = \varphi_{rs} - \psi_{rs} & r,s = 1,\dots,N,\ r \neq s \\ \varphi_{rs} \geq 0 & r,s = 1,\dots,N,\ r \neq s \\ \psi_{rs} \geq 0 & r,s = 1,\dots,N,\ r \neq s \end{cases}$$

R3: Constraint (7) is written as the following set of linear constraints by adding the positive continuous variables y_r, $r = 1,\dots,N$:

$$\left(\frac{1}{KL} \sum_{\substack{i=1,\dots,K \\ j=1,\dots,L}} x_{rij} \right) - \omega_r \leq y_r \qquad r = 1,\dots,N \tag{7a}$$
$$\left(\frac{1}{KL} \sum_{\substack{i=1,\dots,K \\ j=1,\dots,L}} x_{rij} \right) - \omega_r \geq -y_r \qquad r = 1,\dots,N \tag{7b}$$

$$\sum_{r=1,\dots,N} y_r = \alpha \tag{7c}$$

$$y_r \geq 0 \qquad r = 1,\dots,N \tag{7d}$$

Single Linkage

Let z_{rs} and $\eta_{rsiji'j'}$ be defined as follows:

$$z_{rs} \;=\; \min_{\substack{i,i'=1,\dots,K\\ j,j'=1,\dots,L}} \left\{ |i-i'| + |j-j'| :\; x_{rij} = x_{si'j'} = 1 \right\}$$

$$\eta_{rsiji'j'} = \begin{cases} 1 & \text{if the minimum in } SL(P_r(x), P_s(x)) \text{ is attained at pair } (i,j),\ (i',j') \\ 0 & \text{otherwise.} \end{cases}$$

Thus, $(\alpha - SBM)_{SL}$ is stated by using these set of variables plus the convenient constraints to ensure that the distances between portions correspond with the Single Linkage.

$$\begin{aligned}
&(\alpha - SBM)_{SL} = \\
&= \min \left\{ \sum_{\substack{r,s=1,\dots,N\\ r\neq s}} |SL(P_r(x), P_s(x)) - \kappa\delta_{rs}| :\ \text{s.t. } (2)-(7) \right\} \\
&= \min \sum_{\substack{r,s=1,\dots,N\\ r\neq s}} |z_{rs} - \kappa\delta_{rs}| \\
&\quad \text{s.t.} \begin{cases}
(2)-(7) & \\
z_{rs} \geq (|i-i'| + |j-j'|)\cdot \eta_{rsiji'j'} & r,s = 1,\dots,N,\ r\neq s \\
& i,i' = 1,\dots,K,\ j,j' = 1,\dots,L \\
z_{rs} \leq (|i-i'| + |j-j'|) + (K+L-2)(1 - x_{rij}\cdot x_{si'j'}) & r,s = 1,\dots,N,\ r\neq s \\
& i,i' = 1,\dots,K,\ j,j' = 1,\dots,L \\
\displaystyle\sum_{\substack{r,s=1,\dots,N\\ r\neq s\\ i,i'=1,\dots,K\\ j,j'=1,\dots,L}} \eta_{rsiji'j'} \geq 1 & \\
\eta_{rsiji'j'} \leq x_{rij}\cdot x_{si'j'} & r,s = 1,\dots,N,\ r\neq s \\
& i,i' = 1,\dots,K,\ j,j' = 1,\dots,L \\
\eta_{rsiji'j'} \in \{0,1\} & r,s = 1,\dots,N,\ r\neq s \\
& i,i' = 1,\dots,K,\ j,j' = 1,\dots,L.
\end{cases}
\end{aligned}$$

First and second added constraints ensure that variables z_{rs} are well-defined, i.e., if the minimum is achieved in cells (i,j) and (i',j'), then z_{rs} is equal to the minimum distance between portions P_r and P_s. Therefore, these constraints are inactive either if $\eta_{rsiji'j'} = 0$ or if (i,j) or (i',j') do not belong to P_r and P_s, respectively. Third constraint is to impose that there must exist at least one pair of cells which give the minimum distance, and fourth that each pair must belong to the corresponding portion. Finally, we impose the binary requirement for $\eta_{rsiji'j'}$. Observe that, by applying **R1**-**R3** reformulations, the formulation of $(\alpha - SBM)_{SL}$ as an MILP, namely $(\alpha - SBM)^L_{SL}$, is straightforward.

Complete Linkage

Similar reasoning applied to Single Linkage leads to the statement of $(\alpha - SBM)_{CL}$ as a MILP. Let z_{rs} and $\eta_{rsiji'j'}$ be defined as follows:

$$z_{rs} = \max_{\substack{i,i'=1,\dots,K \\ j,j'=1,\dots,L}} \left\{ |i-i'| + |j-j'| : x_{rij} = x_{si'j'} = 1 \right\}$$

$$\eta_{rsiji'j'} = \begin{cases} 1 & \text{if the maximum in } CL(P_r(x), P_s(x)) \text{ is attained at pair } (i,j),\ (i',j') \\ 0 & \text{otherwise.} \end{cases}$$

Thus, $(\alpha - SBM)_{CL}$ is stated by using these set of variables plus the convenient constraints to ensure that the distances between portions correspond with the Complete Linkage.

$$\begin{aligned}
&(\alpha - SBM)_{CL} = \\
&= \min \left\{ \sum_{\substack{r,s=1,\dots,N \\ r \neq s}} |CL(P_r(x), P_s(x)) - \kappa\delta_{rs}| : \text{ s.t. } (2)-(7) \right\} \\
&= \min \sum_{\substack{r,s=1,\dots,N \\ r \neq s}} |z_{rs} - \kappa\delta_{rs}| \\
&\quad \text{s.t.} \begin{cases}
(2)-(7) & \\
z_{rs} \geq (|i-i'| + |j-j'|) \cdot x_{rij} \cdot x_{si'j'} & r,s = 1,\dots,N,\ r \neq s \\
& i,i' = 1,\dots,K,\ j,j' = 1,\dots,L \\
z_{rs} \leq (|i-i'| + |j-j'|) \cdot \eta_{rsiji'j'} + (K+L-2)(1-\eta_{rsiji'j'}) & r,s = 1,\dots,N,\ r \neq s \\
& i,i' = 1,\dots,K,\ j,j' = 1,\dots,L \\
\displaystyle\sum_{\substack{r,s=1,\dots,N \\ i,i'=1,\dots,K \\ j,j'=1,\dots,L}} \eta_{rsiji'j'} \geq 1 & \\
\eta_{rsiji'j'} \leq x_{rij} \cdot x_{si'j'} & r,s = 1,\dots,N,\ r \neq s \\
& i,i' = 1,\dots,K,\ j,j' = 1,\dots,L \\
\eta_{rsiji'j'} \in \{0,1\} & r,s = 1,\dots,N,\ r \neq s \\
& i,i' = 1,\dots,K,\ j,j' = 1,\dots,L
\end{cases}
\end{aligned}$$

First added constraint ensures that for any pair of cells belonging to two different portions, the distance between those portions is greater or equal than the distance between such cells. Second one is to ensure that the distance between two portions is exactly the maximum distance between all possible pairs of cells, since $\eta_{rsiji'j'}$ are precisely forcing that. Third constraint is to impose that there must exist at least one pair of cells which give the maximum distance, and fourth that each pair must belong to the corresponding portion. Finally, we impose the binary requirement for $\eta_{rsiji'j'}$. Observe that, by applying **R1**-**R3** reformulations, the formulation of $(\alpha - SBM)_{CL}$ as an MILP, namely $(\alpha - SBM)^L_{CL}$, is straightforward.

Average Linkage

Since the statistical value associated to portion P_r is represented through its area, we consider a surrogate of the expression of the average distance, (10), by approximating the number of cells

of P_r by $|P_r| = \omega_r KL$, yielding the following expression for the Approximated Average Linkage (AvL_{app})

$$AvL_{app}(P_r(x), P_s(x)) = \frac{1}{\omega_r \omega_s K^2 L^2} \sum_{\substack{i,i'=1,\dots,K \\ j,j'=1,\dots,L}} (|i - i'| + |j - j'|) \cdot x_{rij} \cdot x_{si'j'}. \tag{18}$$

$$(\alpha - SBM)_{AvL} =$$

$$= \min \left\{ \sum_{\substack{r,s=1,\dots,N \\ r \neq s}} |AvL_{app}(P_r(x), P_s(x)) - \kappa \delta_{rs}| : \text{ s.t. } (2) - (7) \right\}$$

By applying **R1**-**R3** reformulations, the formulation of $(\alpha - SBM)_{AvL}$ as an MILP, namely $(\alpha - SBM)^L_{AvL}$, is obtained.