European Journal of Operational Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Decision Support

Segmentation of scanning-transmission electron microscopy images using the ordered median problem

José J. Calvino[c], Miguel López-Haro[c], Juan M. Muñoz-Ocaña[a,*], Justo Puerto[b], Antonio M. Rodríguez-Chía[a]

[a] Departamento de Estadística e Investigación Operativa, Universidad de Cádiz, Cádiz, Spain
[b] IMUS, Instituto de Matemáticas de la Universidad de Sevilla, Sevilla, Spain
[c] Departamento de Ciencia de los Materiales e Ingeniería Metalúrgica y Química Inorgánica, Universidad de Cádiz, Cádiz, Spain

ARTICLE INFO

Article history:
Received 16 June 2021
Accepted 10 January 2022
Available online xxx

Keywords:
Location
Ordered median function
Segmentation
Clustering
Mixed integer linear programming

ABSTRACT

This paper presents new models for segmentation of 2D and 3D Scanning-Transmission Electron Microscope images based on the ordered median function. The main advantage of using this function is its good adaptability to the different types of images to be studied due to the wide range of weight vectors that can be cast. Classical segmentation models stand out for their ability to provide a segmentation of the original image very quickly and with low computational burden. However, they do not usually achieve high quality segmentations with a small number of clusters in order to classify the different elements which compose the structure represented in the image. The quality of the segmentation provided by our approach is analysed using different choices of the weight vector in some real instances. Moreover, improvements are proposed for the formulations to reduce the computational time needed to solve these problems by taking advantage of the weight vector structure.

1. Introduction

Nowadays, Nanoscience and Nanotechnology are of major relevance for the design of new nanomaterials with a wide range of applications in areas such as environmental protection, green energy sources and catalysis. These novel designs are based on the ability of controlling structure and morphology at nanometer scale. Scanning Transmission Electron Microscopy (STEM) has therefore become a powerful tool to rationalise the properties of nanomaterials.

In this technique, the nanomaterials are studied by recording images of their projected structure (2D-STEM images) either in a specific tilt or by acquiring a tilt-series around a single axis with constant increment of angle. In the latter, the reconstructions of the whole set of 2D-STEM images provide information about the 3D structure, most commonly the morphology of the object (3D-STEM images). The intensity displayed in the pixels of the 2D images must at least maintain a monotonic relationship with thickness.

By classifying the pixels in these images into groups of intensities it is possible to discriminate the different components which make up a material and quantify their morphological properties. For example, in a material composed of small particles dispersed on a surface, pixel classification (also known as segmentation) allows identification of the image areas or volumes corresponding to the particles as independent objects and further evaluation of their morphological features (e.g. size, shape or surface-to-volume ratio).

Different clustering methods focusing on pixel classification can be found in the field of STEM (Bai, Fan, & Dong, 2021; Gontar, Ozkaya, & Dunin-Borkowski, 2011). Otsu's method became popular for its simplicity, classifying pixels by minimising the intra-class variance of their intensities (Hindson, Saghi, Hernandez-Garrido, Midgley, & Greenham., 2011; Leary et al., 2012; Liu et al., 2020; Lopez-Haro et al., 2014). *K*-means clustering is also a widely used method to carry out segmentation of images, as it is one of the most effective methods to classify intensities (Belianinov et al., 2015). These clustering procedures play an important role in the analysis of electron microscopy images, since a high quality segmentation determines the success of microscopic characterisation (Yamamoto et al., 2014). The main advantage of these segmenta-

* Corresponding author.
E-mail addresses: jose.calvino@uca.es (J.J. Calvino), miguel.lopezharo@uca.es (M. López-Haro), juanmanuel.munoz@uca.es (J.M. Muñoz-Ocaña), puerto@us.es (J. Puerto), antonio.rodriguezchia@uca.es (A.M. Rodríguez-Chía).

https://doi.org/10.1016/j.ejor.2022.01.022
0377-2217/

tion methods is the ability to obtain solutions very quickly and with low computational burden. However, these models do not usually provide good classifications where there are few clusters or with noisy images, i.e., images where it is very difficult to distinguish between their different structures.

In recent decades, there has been increasing interest in the literature in solving clustering problems with mathematical programming methods (Hansen & Jaumard, 1997). Different objective functions are used for this purpose, such as minimising the maximum within-cluster distance (Saglam, Salman, Sayin, & Turkay, 2006), minimising the sum of within-cluster distances (Brusco, 2003) or minimising the sum of distances between each point and its cluster centre (Bradley, Fayyad, & Mangasarian, 1999). This objective function combined with feature selection was also studied in Benati & García (2014). See Benati, Puerto, & Rodríguez-Chía (2017), Benati, Ponce, Puerto, & Rodríguez-Chía (2021) for the latest advances in the use of mixed-integer linear programming formulations for clustering problems.

Following this line of research, in this paper we propose a new clustering method based on mathematical programming using the Discrete Ordered Median Problem (DOMP) as a criterion to segment STEM images. DOMP has attracted the attention of many studies in the area of discrete location, since it provides a unified framework for the most popular location problems used in discrete location literature (median, center, centdian, k-sum,...).

The idea behind the objective function of this problem consists of applying a penalisation to each distance between a client and its corresponding service facility depending on its position in the whole sequence of sorted distances (unlike classical models such as median or center, where this penalisation is assigned to each client regardless of the magnitude of the distance to its service facility). This adds a 'sorting' to the underlying facility location problem, making formulation and solution much more challenging. There are many studies based on this function such as Kalcsics, Nickel, & Puerto (2003), Ogryczak & Tamir (2003), Nickel & Puerto (2005), Boland, Domínguez-Marín, Nickel, & Puerto (2006), Puerto (2008), Marín, Nickel, Puerto, & Velten (2009) which introduce classical DOMP formulations, while Kalcsics, Nickel, Puerto, & Rodríguez-Chía (2010), Labbé, Ponce, & Puerto (2017), Aouad & Segev (2019), Olender & Ogryczak (2019), Blanco (2019), Deleplanque, Labbé, Ponce, & Puerto (2020), Espejo, Puerto, & Rodríguez-Chía (2021), Marín, Ponce, & Puerto (2020) offer an overview of recent advances in DOMP.

The goal of this paper is twofold. The first aim is to show that segmentations with the ordered median objective function provide high quality pixel classifications for certain choices of the weight vector. The second goal is to develop new formulations and improvements for DOMP with specific choices of the weight vector that allow us to solve the resulting optimisation problems in an efficient way.

This paper is structured as follows: Section 2 introduces the notation needed to formulate segmentation problems within the DOMP framework; Section 3 proposes different improvements to DOMP formulations to reduce the computational time needed to obtain a high quality segmentation by taking advantage of the weighted vector structure in the objective function to be minimised; Section 4 provides alternative improvements to the formulations based on the idea developed in Ogryczak & Tamir (2003); Section 5 offers an extensive computational analysis of the different formulations and the improvements developed in this paper; in Section 6 this model is validated with different images and a way of quantifying the segmentations obtained is proposed; Section 7 outlines the main conclusions of this paper; and finally, for the sake of completeness, the formulations used to compare with the ones proposed in this paper are set out in the Appendix.

2. Image segmentation and the ordered median problem

This section describes the elements that define image segmentation models and states the links to formulate these problems within the DOMP framework.

Let us suppose we have an image with $M \times N$ pixels to be segmented with the aim of identifying the elements which constitute the nanomaterial shown in the original image. Each pixel has a specific intensity which is an integer value in a range whose length will depend on the resolution (number of bits) of the image, for instance, between 0 and 2^8 or 0 and 2^{16}. The smallest intensities of this range correspond to the lowest densities of the object represented in the image (close to black colour) and the largest intensities correspond to the greatest densities (close to white colour). Let us also assume that we have n different intensities and $N := \{1, \dots, n\}$. The number of pixels having the same intensity is referred to as the frequency of that intensity and the set of frequencies of an image is denoted as $f := \{f_1, \dots, f_n\}$.

Segmenting an image consists of grouping its intensities into p ($\leq n$) different clusters. Each cluster is associated with an intensity which acts as its representative. Therefore, in terms of mathematical programming we can define the segmentation of an image as the choice/location of p cluster representatives and the allocation of each intensity to a cluster representative in such a way that some objective function is minimised. It is assumed that the set of candidate cluster representatives is the set of intensities. We define cluster j as the one having intensity j as its representative. Moreover, each intensity is allocated to only one cluster representative. Let $d = (d_{ij})_{i,j=1,\dots,n}$ be the $n \times n$ intensity weighted distance matrix where d_{ij} represents the intensity weighted distance for allocating intensity i to the cluster representative j. These intensity weighted distances are defined as the product of the frequency of the pixel intensity i (f_i) multiplied by the distance between intensities i and j, i.e. $d_{ij} = f_i|i-j|$. The distance between two intensities is obtained as the absolute value of the difference between both intensities, since they are on the real line ($|i-j|$). Let $J \subset N$ be the subset of p intensities selected as representatives of p different clusters. We define $d_i(J)$, $i \in N$ as the intensity-allocation weighted distance of intensity i to a cluster representative in J. It is assumed that each intensity i is allocated to a representative such that $j \in \arg\min_{k \in J} d_{ik}$ or in other words:

$$d_{ij} = d_i(J) := \min_{k \in J} d_{ik}.$$

2D and 3D-images are usually composed of different structures such as background, particles and support (materials which hold particles). Particles are usually characterised by intensities with low pixel frequencies in the original image because of their small size compared to the other structures of the image. Hence, if we consider the ordered vector of the intensity-allocation weighted distances, the first positions of that vector usually correspond to intensities associated with particles. We can therefore attempt to exploit this observation by applying a specific DOMP model with appropriate weights to achieve a good segmentation of images obtained with STEM. The intensity-allocation weighted distances are sorted to calculate the ordered median function where $d_{\leq}(J) := (d^1_{\leq}(J), \dots, d^n_{\leq}(J))$ will be this vector, such that:

$$d^1_{\leq}(J) \leq \dots \leq d^n_{\leq}(J).$$

The DOMP aims to minimise the ordered weighted average of vector $(d^k_{\leq}(J))$ with respect to a given set of λ-weights:

$$\min_{\substack{J \subset N \\ |J|=p}} \sum_{k \in N} \lambda^k d^k_{\leq}(J),$$

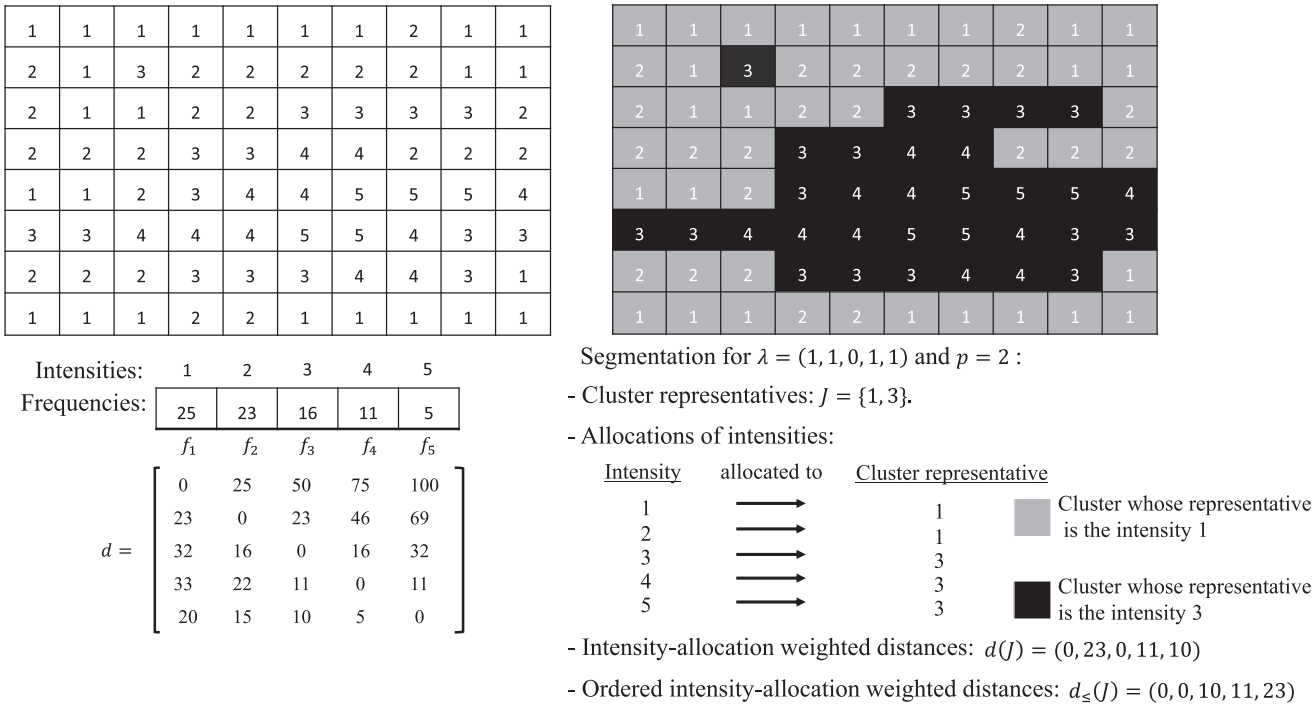

Fig. 1. Segmentation of an 8 × 10 image with 5 intensities and 2 clusters.

where $\lambda = (\lambda^1, \ldots, \lambda^n)$, with $\lambda^k \geq 0$, $\forall k \in N$. Figure 1 shows an example of an image segmentation with five intensities grouped into 2 clusters.

The above expression needs to be reformulated for its implementation in MIP solvers. There are several formulations involved in this task. The formulation F_{DOMP_G} proposed in Puerto, Ramos, & Rodríguez-Chía (2013), Puerto, Ramos, Rodríguez-Chía, & Sánchez-Gil (2016) and Espejo et al. (2021) and the formulation F_{OT_θ} introduced in Marín et al. (2020) are known to have the best performance in terms of solution times when λ-vectors have repetitions (components with equal values). For the sake of completeness, we have included these formulations in Appendix A.1 and A.2, respectively.

3. Specific formulations

The application of the state-of-the-art DOMP formulations to segmentation problems allows us to solve medium size instances. The most promising formulations for these types of problems are described in Appendix A.1 and A.2. However, still the CPU times required to solve these problems make them not competitive when compared with standard methods in the image segmentation area. For this reason, our goal is to improve their performance by exploiting specific aspects of the problem. Taking advantage of the particular λ-vector structure, we aim to provide new formulations that allow a reduction of the computational time required by the general formulations of the ordered median problem. Since the main goal of image segmentation techniques is the analysis of the particle characteristics contained in STEM experiments, these formulations will attempt to remove intensities that do not correspond to specific structures of the original image. These intensities could contain noise generated by the microscope when projections are recorded (intensities with the lowest frequencies). On the other hand, they could represent structures with larger sizes than particles. Therefore, to remove these intensities from the objective function, we attempt to assign the value 0 to the positions of the λ-vector where these intensities will hopefully be in the ordered intensity-allocation weighted distance vector and the value 1 to positions corresponding to relevant intensities.

In this section we propose different specific criteria depending on the type of sample to be segmented: i) anti-k-centrum ($\lambda = (1, \ldots, 1, 0, \ldots, 0)$), which minimises the sum of the k-smallest components of the ordered intensity-allocation weighted distance vector, will be selected to segment images with small relevant features (smallest frequencies) and very large unimportant structures (largest frequencies), since hopefully the first and last positions of the ordered intensity-allocation weighted distance vector will correspond to the intensities with the smallest and largest frequencies, respectively; ii) (k_1, k_2)-trimmed mean ($\lambda = (0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0)$), which minimises the sum of components between the $(k_1 + 1)$th and the $(n - k_2 - 1)$th positions of the ordered intensity-allocation weighted distance vector, will be applied to segment images where the k_1-smallest and the k_2-largest frequencies correspond to intensities that do not provide meaningful information about the image; and iii) (k_1, k_2)-anti-trimmed-mean ($\lambda = (1, \ldots, 1, 0, \ldots, 0, 1, \ldots, 1)$), which minimises the sum of the first k_1 plus the last k_2 components of the ordered intensity-allocation weighted distance vector, could be used for images whose k_1-smallest and k_2-largest frequencies correspond to intensities that represent the most important structures.

3.1. Anti-k-centrum

Let us suppose that the image to be segmented is made up of very small particles and the main aim is to detect only these particles regardless of the rest of the elements. Particles are usually in the first positions of the ordered intensity-allocation weighted distance vector, since there are fewer pixels with intensities associated with particles than pixels with intensities associated with the rest of elements of the image. That is, we are interested in assigning the intensities with the smallest intensity-allocation weighted distances in the best possible way. The anti-k-centrum problem may have a suitable structure for this situation, since the assignments with the smallest intensity-allocation weighted distances

penalise the objective function thanks to the non-zero values of λ-vector whereas the largest ones are hardly considered due to the zeros in the corresponding positions. This λ-vector assigns the value 1 to the first k-positions and zero otherwise, that is, $\lambda = (1, \ldots, 1, 0, \ldots, 0)$. To give a formulation for the anti-k-centrum model, we define the allocation r-variables, such that $r_{ij} = 1$ if intensity i is assigned to cluster representative j and $r_{ij} = 0$ otherwise. The locations of representatives are controlled by y-variables, defined as $y_j = 1$ if intensity j is selected as the representative of a cluster and $y_j = 0$ otherwise. In what follows, for the sake of brevity, we will say that cluster j exists if $y_j = 1$. Taking advantage of the particular structure of the λ-vector in anti-k-centrum, we propose the following formulation for this problem:

$$(\mathrm{F_{AkC}}) \quad \min \sum_{i,j\in N} d_{ij} r_{ij}$$

$$\text{s.t.} \quad \sum_{j\in N} y_j = p, \tag{1a}$$

$$r_{ij} \le y_j, \qquad \forall i, j \in N, \tag{1b}$$

$$\sum_{j\in N} r_{ij} \le 1, \qquad \forall i \in N, \tag{1c}$$

$$\sum_{i,j\in N} r_{ij} = k, \tag{1d}$$

$$r_{ij}, y_j \in \{0, 1\}, \qquad \forall i, j \in N. \tag{1e}$$

The objective function stands for the sum of the k-smallest components of the ordered intensity-allocation weighted distance vector. The equality constraint (1a) sets the number of cluster representatives to p. Constraints (1b) avoid allocating intensity i to j if j is not selected as cluster representative. Constraints (1c) ensure that each intensity is allocated to a maximum of one cluster. The equality constraint (1d) sets the number of allocations to k. Closest assignment constraints used in the formulation of Appendix A.2 (17d) are not included in the formulation, since the assignments with the $n-k$ largest intensity-allocation weighted distances may be post-processed to be allocated to their closest clusters. Although r-variables may be relaxed, we have considered them as binary variables since they provided us lower computational times. In the rest of the formulations included in this paper we will proceed in the same way.

We can adapt this formulation if there is more than one non-zero block in λ-vector (a block is defined as a set of consecutive non-null identical values in λ-vector). These blocks must be sorted in decreasing order, for instance $\lambda = (2, \ldots, 2, 1, \ldots, 1, 0, \ldots, 0)$. Consequently, we define as many sets of allocation variables as there are positive blocks. If there are two blocks with k_1 and k_2 elements which take the value of 2 and 1 respectively, r and s-variables will be defined to control each intensity allocation. The r-variables control the lowest k_1th allocation weights and the s-variables state the following k_2th ones.

$$(\mathrm{F_{AkC2}}) \quad \min \; 2\sum_{i,j\in N} d_{ij} r_{ij} + \sum_{i,j\in N} d_{ij} s_{ij}$$

$$\text{s.t.} \quad (1a),$$

$$r_{ij} + s_{ij} \le y_j, \qquad \forall i, j \in N, \tag{2a}$$

$$\sum_{j\in N} (r_{ij} + s_{ij}) \le 1, \qquad \forall i \in N, \tag{2b}$$

$$\sum_{i,j\in N} r_{ij} = k_1, \tag{2c}$$

$$\sum_{i,j\in N} s_{ij} = k_2, \tag{2d}$$

$$r_{ij}, s_{ij}, y_j \in \{0, 1\} \qquad \forall i, j \in N. \tag{2e}$$

Constraints (2a) avoid allocating intensity i to j if j is not selected as cluster representative. Constraints (2b) ensure that each intensity is assigned to at most one cluster. The number of allocations controlled by r and s-variables are set to k_1 and k_2 by constraints (2c) and (2d) respectively.

3.2. (k_1, k_2)-Anti-Trimmed mean

If one is interested in identifying only the smallest particles together with the largest ones, in general, these particles will correspond to the intensities with the smallest and largest frequencies, respectively. As mentioned above, the smallest elements of the image are usually in the first positions of the ordered intensity-allocation weighted distance vector and the largest ones are often in the last positions of the vector, which correspond to the background of the image. This situation fits to the structure of the (k_1, k_2)-anti-trimmed mean model, since it aims to minimise the k_1-smallest intensity-allocation weighted distances together with the k_2-largest ones, i.e., $\lambda = (1, \ldots, 1, 0, \ldots, 0, 1, \ldots, 1)$. We define three sets of allocation variables r_{ij}, s_{ij} and t_{ij}, with $i, j \in N$. These variables take the value 1 if intensity i is allocated to cluster j and the value 0 otherwise. r-variables are used for controlling the allocations with the smallest intensity-allocation weighted distances, t-variables manage the largest ones and s-variables are used to allocate intensities with intensity-allocation weighted distances corresponding to positions in which λ-vector takes the value 0. A formulation for the (k_1, k_2)-anti-trimmed mean model is as follows:

$$(\mathrm{F_{(k_1,k_2)ATM}}) \quad \min \sum_{i,j\in N} d_{ij} r_{ij} + \sum_{i,j\in N} d_{ij} t_{ij}$$

$$\text{s.t.} \quad (1a),$$

$$r_{ij} + s_{ij} + t_{ij} \le y_j, \qquad \forall i, j \in N, \tag{3a}$$

$$\sum_{j\in N} (r_{ij} + s_{ij} + t_{ij}) = 1, \qquad \forall i \in N, \tag{3b}$$

Ordering constraints (See Section 3.2.1), (3c)

$$\sum_{i,j\in N} t_{ij} = k_2, \tag{3d}$$

$$\sum_{i,j\in N} s_{ij} = n - (k_1 + k_2), \tag{3e}$$

$$r_{ij}, s_{ij}, t_{ij}, y_j \in \{0, 1\}, \qquad \forall i, j \in N. \tag{3f}$$

The objective function accounts for the k_1-smallest plus the k_2-largest intensity-allocation weighted distances. Constraints (3a) avoid allocating intensity i to j if j is not selected as cluster representative and (3b) ensure that each intensity is allocated to one cluster. Constraints (3c) ensure that any allocation using t-variables will have an intensity-allocation weighted distance larger than the ones associated with s-variables and will be described in Section 3.2.1. Constraints (3b) together with (3d) and (3e) determine the number of allocations controlled by each set of variables.

3.2.1. Alternative ordering constraints for $F_{(k_1,k_2)ATM}$ formulation

In this subsection we propose different ways of modelling ordering constraints, (3c). These constraints allow us to model the order of the intensity-allocation weighted distances. Hence, this family of constraints ensures that the intensity-allocation weighted distances of assignments controlled by s-variables must be smaller than or equal to the ones controlled by t-variables. Below, we introduce different sets of constraints with the same purpose of ordering the assignments controlled by s and t-variables. In Section 5 a preliminary computational study is carried out to choose the ordering constraints with the best performance to solve the $F_{(k_1,k_2)ATM}$ formulation.

First family: This allocation order may be controlled by a big M value to force that the intensity-allocation weighted distance of each s-variable assignment is smaller than or equal to the one of each allocation controlled by t-variables. Hence, these constraints may be written as follows:

$$\sum_{j\in N} d_{ij}s_{ij} \le (1-\sum_{j\in N} t_{i'j})M_i + \sum_{j\in N} d_{i'j}t_{i'j}, \quad \forall i,i' \in N. \tag{4a}$$

The value of M_i has been set to the maximum intensity-allocation weighted distance of intensity i, i.e., $M_i = \max_{j\in N} d_{ij}$. Constraints (4a) may be reinforced by including on the left side the sum of the intensity-allocation weighted distances of assignments controlled by r-variables, obtaining:

$$\sum_{j\in N}(d_{ij}s_{ij} + d_{ij}r_{ij}) \le (1-\sum_{j\in N} t_{i'j})M_i + \sum_{j\in N} d_{i'j}t_{i'j}, \quad \forall i,i' \in N. \tag{4b}$$

Second family: In what follows we consider a different way of modelling the ordering constraints without using big M values. To do so, we define the following constraints:

$$(n-k_1-k_2)t_{ij} \le \sum_{\substack{k,l\in N:\\ d_{kl}<d_{ij}}} s_{kl}, \qquad \forall i,j \in N. \tag{5a}$$

These constraints force that the intensity-allocation weighted distances of the $n-k_1-k_2$ assignments controlled by s-variables must be smaller than the intensity-allocation weighted distance of every assignment controlled by t-variables. The number of assignments controlled by r-variables may be included in (5a), since, if $t_{ij}=1$ then, the total number of assignments controlled by r and s-variables with intensity-allocation weighted distances smaller than d_{ij} must be larger than or equal to $n-k_2$:

$$(n-k_2)t_{ij} \le \sum_{\substack{k,l\in N:\\ d_{kl}<d_{ij}}} (s_{kl}+r_{kl}), \qquad \forall i,j \in N. \tag{5b}$$

Third family: An alternative way of modelling the ordering constraints is as follows:

$$k_2 s_{ij} \le \sum_{\substack{k,l\in N:\\ d_{kl}>d_{ij}}} t_{kl}, \qquad \forall i,j \in N. \tag{6a}$$

These inequalities ensure that the number of assignments controlled by t-variables must be equal to k_2 and they must have intensity-allocation weighted distances larger than the allocations controlled by s-variables.

Fourth family: The following constraints state the maximum number of assignments controlled by r and s-variables. If $t_{ij}=0$, then the number of assignments with intensity-allocation weighted distances larger than or equal to d_{ij} must be smaller than or equal to $n-k_2$. Moreover, if $t_{ij}=1$, there can be no assignments controlled by r and s-variables with intensity-allocation weighted distances larger than d_{ij}. These constraints may be written as follows:

$$\sum_{\substack{k,l\in N:\\ d_{kl}\ge d_{ij}}} (s_{kl}+r_{kl}) \le (1-t_{ij})(n-k_2), \qquad \forall i,j \in N. \tag{7a}$$

The values of t-variables may be added for intensity i if the intensity-allocation weighted distance is smaller than or equal to d_{ij}:

$$\sum_{\substack{k,l\in N:\\ d_{kl}\ge d_{ij}}} (s_{kl}+r_{kl}) \le (1-\sum_{\substack{l'\in N:\\ d_{il'}\le d_{ij}}} t_{il'})(n-k_2), \quad \forall i,j \in N. \tag{7b}$$

Fifth family: Another way to order each assignment is to define two variables ($u_{\max}$ and $U_{\min}$) which take the maximum intensity-allocation weighted distance value of the assignments controlled by r-variables and the minimum value of the ones controlled by t-variables, respectively. Four sets of ordering constraints are included in the $F_{(k_1,k_2)ATM}$ formulation:

$$d_{ij}(1-\sum_{k\in N} r_{ik} - \sum_{k\in N} t_{ik}) \le (1-y_j + \sum_{\substack{k\in N:\\ d_{ik}<d_{ij}}} y_k)M + U_{min}, \quad \forall i,j \in N, \tag{8a}$$

$$d_{ij}y_j + (1-y_j)M + (\sum_{k\in N} r_{ik} + \sum_{k\in N} t_{ik})M \ge u_{max}, \quad \forall i,j \in N, \tag{8b}$$

$$\sum_{k\in N} r_{ik}d_{ik} \le u_{max}, \qquad \forall i \in N, \tag{8c}$$

$$\left(1-\sum_{k\in N} t_{ik}\right)M + \sum_{k\in N} t_{ik}d_{ik} \ge U_{min}, \qquad \forall i \in N. \tag{8d}$$

Constraints (8a) and (8b) ensure that the maximum and minimum intensity-allocation weighted distance of each assignment controlled by s-variables must be smaller than or equal to $U_{\min}$ and larger than or equal to $u_{\max}$, respectively. Constraints (8c) and (8d) state the maximum intensity-allocation weighted distance of assignments managed by r-variables and the minimum intensity-allocation weighted distance of assignments controlled by t-variables, respectively.

3.3. (k_1, k_2)-Trimmed mean

Let us now suppose that we are interested in segmenting an image where the intensities with the smallest and largest number of pixels do not provide useful information. These unimportant intensities usually represent noise (lowest frequencies) and the background (highest frequencies). Therefore, we need to find a model that does not consider the smallest intensity-allocation weighted distances together with the largest ones in the objective function. The (k_1, k_2)-trimmed mean model adapts well to this situation as it minimises intensity-allocation weighted distances excluding the k_1-smallest and k_2-largest ones, i.e., $\lambda = (0,\dots,0,1,\dots,1,0,\dots,0)$. To formulate this specific model, proceeding in a similar way to $F_{(k_1,k_2)ATM}$, we define two sets of allocation variables r and s. The r-variables which do not contribute to the objective function control the k_1-smallest intensity-allocation weighted distances and the s-variables control the $n-(k_1+k_2)$ following ones. This model may be written as follows:

$$(F_{(k_1,k_2)TM}) \quad \min \sum_{i,j\in N} d_{ij}s_{ij}$$

$$\text{s.t.} \quad (1a),$$

$$r_{ij} + s_{ij} \le y_j, \qquad \forall i,j \in N, \tag{9a}$$

$$\sum_{j\in N}(r_{ij}+s_{ij}) \le 1, \qquad \forall i \in N, \tag{9b}$$

Ordering constraints (See Section 3.3.1), (9c)

$$\sum_{i,j\in N} r_{ij} = k_1, \tag{9d}$$

$$\sum_{i,j\in N} s_{ij} = n - (k_1 + k_2), \tag{9e}$$

$$r_{ij}, s_{ij}, y_j \in \{0,1\}, \qquad \forall i,j \in N. \tag{9f}$$

The objective function provides the sum of assignments with intensity-allocation weighted distances larger than the k_1th position and smaller than the $n-k_2$th position in the ordered intensity-allocation weighted distance vector. Constraints (9a) ensure that each intensity is associated with a cluster that exists. Constraints (9b) force that each intensity is allocated to at most one cluster. Constraints (9c) ensure that the intensity-allocation weighted distances of assignments controlled by r-variables are smaller than the ones controlled by s-variables. The number of allocations made by r-variables and s-variables are given by (9d) and (9e), respectively.

3.3.1. Alternative ordering constraints for $F_{(k_1,k_2)TM}$ formulation

In this subsection, we analyse different ways of modelling ordering constraints (9c). These families of constraints aim to model the order of assignments controlled by r and s-variables in a similar way to the $F_{(k_1,k_2)ATM}$ formulation. We have proposed different ordering constraint families and a preliminary computational study is carried out to select the family that shows the best performance in the $F_{(k_1,k_2)TM}$ formulation (see Section 5).

First family: The first set of ordering constraints proposed to state the assignments controlled by r and s-variables such that intensity-allocation weighted distances of assignments managed by r-variables must be smaller than or equal to the ones controlled by s-variables. This can be written as follows:

$$\sum_{j\in N} d_{ij} r_{ij} \le (1 - \sum_{j\in N} s_{i'j}) M_i + \sum_{j\in N} d_{i'j} s_{i'j}, \quad \forall i,i' \in N, \tag{10a}$$

where $M_i = \max_{j\in N} d_{ij}$.

Second family: Similarly to the $F_{(k_1,k_2)ATM}$ formulation, constraints (10a) may be replaced by the following ones which also state the order of each assignment:

$$k_1 s_{ij} \le \sum_{\substack{k,l\in N:\\ d_{kl}<d_{ij}}} r_{kl}, \qquad \forall i,j \in N. \tag{11a}$$

These constraints ensure that the number of allocations with intensity-allocation weighted distances smaller than d_{ij} controlled by r-variables must be larger than or equal to k_1 if $s_{ij} = 1$. Constraints (9d) are removed from the $F_{(k_1,k_2)TM}$ formulation, since the number of allocations controlled by r-variables is stated in constraint (11a). Computational results have been obtained with this set of constraints. Valid inequalities to manage the order of allocations may also be included in the formulation:

$$(n - k_1 - k_2) r_{ij} \le \sum_{\substack{k,l\in N:\\ d_{kl}>d_{ij}}} s_{kl}, \qquad \forall i,j \in N. \tag{12a}$$

These constraints set as $n-k_1-k_2$ the number of assignments controlled by s-variables. Each one must have an intensity-allocation weighted distance larger than the intensity-allocation weighted distances of each assignment controlled by r-variables.

4. New formulations based on the dualisation of the k-sum problem

Ogryczak & Tamir (2003) developed a formulation for the ordered median problem using the dual of the problem that maximises the sum of k values of a set of n values (k-sum problem). Although this formulation has a good performance due to the reduced number of variables, it can only be extended to λ-vectors whose components are given in non-decreasing order, i.e., $0 = \lambda_0 \le \lambda_1 \le \cdots \le \lambda_n$. It requires a set of allocation variables x_{ij}, with $i,j \in N$, defined as $x_{ij} = 1$ if intensity i is assigned to cluster j and $x_{ij} = 0$ otherwise, $\forall i,j \in N$. In addition, the w_{ik}- and z_k-variables correspond to the dual variables of the problems that maximise the sums of k values for each $i \in N$. The formulation is given by:

$$(F_{OT}) \quad \min \quad \sum_{k\in N} (\lambda_{N-k+1} - \lambda_{N-k})(k z_k + \sum_{i\in N} w_{ik})$$

$$\text{s.t.} \quad (1a),$$

$$x_{ij} \le y_j, \qquad \forall i,j \in N, \tag{13a}$$

$$\sum_{j\in N} x_{ij} = 1, \qquad \forall i \in N, \tag{13b}$$

$$w_{ik} + z_k \ge \sum_{j\in N} d_{ij} x_{ij}, \qquad \forall i,k \in N, \tag{13c}$$

$$x_{ij}, y_j \in \{0,1\}, \qquad \forall i,j \in N, \tag{13d}$$

$$w_{ik} \ge 0, \qquad \forall i,k \in N, \tag{13e}$$

$$z_k \in \mathbb{R}, \qquad \forall k \in N. \tag{13f}$$

The kth addend of the objective function namely $k z_{ik} + \sum_{i\in N} w_{ik}$, represents the dual objective function of the problem that maximises the sum of k intensity-allocation weighted distances, and consequently, the telescopic sum of these addends when the weights are given in a non-decreasing way provides the ordered objective function. Constraints (13a) prevent each intensity being allocated to a cluster representative that does not exist. Moreover, constraints (13b) ensure that the allocation of each pixel is unique. Constraints (13c) are used to build the dual from the maximisation problem to calculate the sum of the k-largest weights.

The $F_{(k_1,k_2)ATM}$ and $F_{(k_1,k_2)TM}$ formulations require less computational time than the generic ordered median formulation. However, they still need a high computational time to solve large size instances. For this reason, we have exploited some aspects of the F_{OT} formulation to provide alternatives that permit to solve the aforementioned instances in smaller computation times.

4.1. OT-(k_1,k_2)-Anti-Trimmed mean

Let $\lambda = (1,\ldots,1,0,\ldots,0,1,\ldots,1)$ be the λ-vector corresponding to the (k_1,k_2)-anti-trimmed mean model, where the first and second blocks of ones have k_1 and k_2 elements, respectively. The k_1-smallest intensity-allocation weighted distances are minimised by using the anti-k-centrum criterion and the k_2-largest intensity-allocation weighted distances are minimised by exploiting the rationale behind the OT formulation to solve the k_2-centrum problem. Hence, r-variables are defined to control the assignments with the k_1-smallest intensity-allocation weighted distances and x-variables are used to apply the OT criterion to minimise the sum

of the k_2-largest intensity-allocation weighted distances:

$$(\mathrm{F}_{(k_1,k_2)\mathrm{ATM}_{\mathrm{OT}}}) \quad \min \quad \sum_{j,k\in N} d_{jk} r_{jk} + k_2 z + \sum_{i\in N} w_i$$

$$\text{s.t.} \quad (1a), (9d), (13a), (13b),$$

$$w_i + z \geq \sum_{j\in N} d_{ij} x_{ij}, \qquad \forall i \in N, \tag{14a}$$

$$r_{ij} \leq x_{ij}, \qquad \forall i, j \in N, \tag{14b}$$

$$x_{ij}, r_{ij}, y_j \in \{0,1\}, \qquad \forall i, j \in N, \tag{14c}$$

$$w_i \geq 0, \qquad \forall i \in N, \tag{14d}$$

$$z \in \mathbb{R}. \tag{14e}$$

The first term of the objective function computes the k_1-smallest intensity-allocation weighted distances and the sum of the k_2-largest ones are computed by the second term. Constraints (14a) relate primal and dual variables associated with the formulation that minimises the sum of the k_2-largest assignments. Constraints (14b) set $x_{jk} = 1$ if r_{jk} takes the value 1, since, in contrast to r-variables, which control k_1 assignments, x-variables are involved in each of the n allocations due to constraints (13b).

4.2. OT-(k_1, k_2)-Trimmed mean

We define $\lambda = (0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0)$ where the first block of zeros contains k_1 elements and the second block of zeros is composed of k_2 elements. Thus, Ogryczak and Tamir's formulation is applied to the first $n - k_2$ elements of λ-vector. To achieve this, the number of allocations controlled by x-variables will be set as $n - k_2$:

$$(\mathrm{F}_{(k_1,k_2)\mathrm{TM}_{\mathrm{OT}}}) \quad \min \quad (n - k_1 - k_2) z + \sum_{i\in N} w_i$$

$$\text{s.t.} \quad (1a), (13a), (14a),$$

$$\sum_{i,j\in N} x_{ij} = n - k_2, \tag{15a}$$

$$x_{ij}, y_j \in \{0,1\}, \qquad \forall i, j \in N, \tag{15b}$$

$$w_i \geq 0, \qquad \forall i \in N, \tag{15c}$$

$$z \in \mathbb{R}. \tag{15d}$$

The objective function implements Ogryczak and Tamir's formulation to the first $n - k_2$ components of λ. Constraints (15a) set the number of allocations controlled by x-variables, i.e., the number of assignments must be equal to $n - k_2$.

5. Computational results

This section provides a detailed computational analysis of the alternative formulations proposed for the different choices of the λ-vector in the discrete ordered median problem. Our focus is on the application of DOMP to segment electron microscopy images recorded during the structural characterisation of nanomaterials with potential applications in environmental catalysis. Therefore, we have restricted ourselves to meaningful choices of λ-vectors which are valid for this application. This analysis allows us to determine which formulations are the most efficient to solve the proposed segmentation problems. As is usual in the STEM field, the computational experience is carried out in simulated instances. Generating real experiments is highly time-demanding and consequently the time needed to obtain a large number of instances would not be affordable. The idea is to generate synthetic images (phantoms) with fully known features (e.g., sizes, shapes, intensities) that are close to real systems studied in STEM (Staniewicz & Midgley, 2015; Tovey et al., 2019).

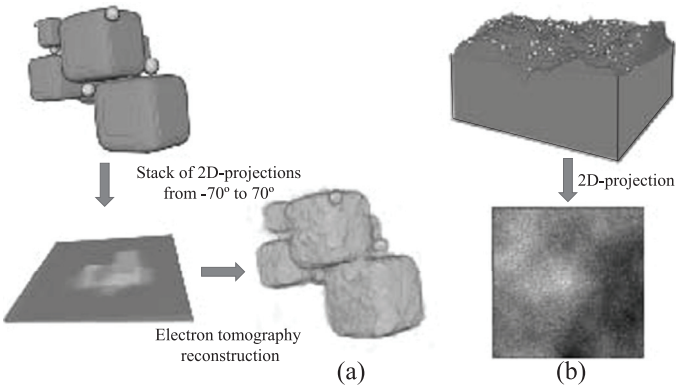

Fig. 2. (a) corresponds to a 3D-STEM image and (b) is a 2D-STEM image. Both have been used to assess the performance of the different formulations.

Figure 2 shows the procedures to generate the phantoms selected to implement the computational studies of the formulations and validation of the proposed models (see Section 6). We have decided to segment these images because of their importance in the field of nanoscience and nanotechnology. It is important to note that both phantoms represent complex nanomaterials which have potential applications in the field of heterogeneous and environmental catalysis (see Liu & Corma, 2018). Thus, high quality segmentations of the components that make up such systems are essential not only to obtain an accurate quantification but also to be able to link the structural properties at nm-scale to their chemical or physical behaviour (see López-Haro et al., 2018).

In particular, these datasets simulate 3D-STEM (Fig. 2(a)) and 2D-STEM (Fig. 2(b)) images. Both images are composed of a background (pixels close to black colour), the support (pixels with grey colours) and the particles (pixels close to white colour). Supports are necessary in STEM experiments to hold the nanometric objects, since these particles cannot be analysed individually due to their microscopic sizes. The goal for segmenting these images is to identify most of the particles to quantify their properties. To obtain Fig. 2(a), a 3D phantom was generated to simulate a nanocatalyst, and the shapes, sizes and intensities of these structures are related with the characteristics of a real nanocatalyst. Once the phantom had been generated, four different 3D-STEM images were reconstructed using a classical reconstruction algorithm by considering projections from -70 to 70 degrees obtained every 5, 10, 15, and 20 degrees (see Midgley, Ward, Hungría, & Thomas, 2007).

Figure 2 (b) simulates a 2D projection provided by electron microscopes before applying a reconstruction algorithm (2D-STEM image). Four instances were created by modifying the support structure and the location, number and sizes of every particle within the image. Supports were obtained by generating 3D surfaces very similar to the structures that hold the particles in these types of experiments. Particles were represented by spheres whose centres (x, y, z) were obtained by generating uniformly distributed values between 0 and 512 pixels for each dimension. Values related to the radius of these spheres were calculated with a uni-

Table 1
Different ordering constraints included in the $F_{(k_1,k_2)ATM}$ formulation.

		(4a)				(4b)				(5a)				(5b)			
n	p	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes
128	2	3201.88(1)	1.14	99.94	22550.75	1524.01	0	99.94	11612.5	7208.06(4)	99.99	99.88	8079.75	7311.02(4)	99.99	99.94	4560.75
128	3	4423.04(1)	4.58	99.93	30106.75	1837.45	0	99.93	10774.75	7207.14(1)	99.98	99.85	6210.50	-	-	-	-
128	4	5445.62(2)	28.66	99.90	36150.25	2111.39	0	99.90	19921	7205.35(4)	99.98	99.77	9085.75	-	-	-	-
128	5	7200.97(4)	70.06	99.15	42022.75	6208.16(3)	46.34	99.91	45656.75	7206.14(4)	99.99	99.76	8425.25	-	-	-	-
256	2	7200.90(4)	97.61	99.86	1214.25	7200.57(4)	99.94	99.76	0	-	-	-	-	-	-	-	-
256	3	7202.06(4)	97.54	99.88	1658	7200.56(4)	99.63	99.73	1670.75	-	-	-	-	-	-	-	-
256	4	7201.61(4)	98.30	99.89	648.75	7201.23(4)	99.74	99.73	2218.5	-	-	-	-	-	-	-	-
256	5	7201.12(4)	98.78	99.89	428	7201.44(4)	99.76	99.67	86.33	-	-	-	-	-	-	-	-
		(6a)				(7a)				(7b)				(8a), (8b), (8c), (8d)			
128	2	1127.49	0	34.05	1165.25	7227.16(4)	99.76	99.94	1546.75	7207.51(4)	99.75	99.60	3.5	7201.19(4)	0.7815	99.95	859252.25
128	3	1250.88	0	40.50	2371.25	7208.01(4)	99.84	99.92	1926.50	7213.70(4)	99.74	99.43	503.5	7202.15(4)	99.90	99.94	170940.25
128	4	1284.98	0	64.41	833	7207.35(4)	99.87	99.89	2349.50	7205.43(4)	99.77	99.13	16.5	7201.99(4)	99.67	99.90	399771
128	5	6735.35(3)	46.84	74.35	12269.5	7247.24(4)	99.84	99.89	2456.50	7204.15(4)	99.68	99.05	229.75	7204.99(4)	99.88	99.89	223942.33
256	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
256	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
256	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
256	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

form distribution between 3 and 8 pixels. After creating these 3D structures, they were projected on the horizontal plane to obtain 2D-STEM images.

The instances to be segmented were generated with different numbers of intensities to study the performance of every formulation with different instance sizes (128, 256, 512, and 1024 intensities) and four values of p for each instance were selected (2, 3, 4, and 5 clusters). Moreover, k, k_1 and k_2 parameters were set depending on the size of the instances: k was set to $\frac{n}{2}$, where n is the number of intensities of the image under study, i.e., k was set to 64, 128, 256, and 512 to segment an image with 128, 256, 512, and 1024 intensities, respectively; meanwhile, k_1 and k_2 were both set to approximately $\frac{n}{10}$, i.e., $k_1 = k_2 = 10, 25, 50$, and 100 for 128, 256, 512, and 1024 intensities, respectively. These choices correspond to a preliminary computational analysis over different experiments, where the values of these parameters always reported very good performance.

All the formulations were implemented in MATLAB R2020b and solved with CPLEX 12.10 thanks to the API that links both codes. All the experiments were performed on an Intel Xeon W-2245 workstation, 256 Gb RAM, NVIDIA Quadro RTX 4000. The time for solving each instance was limited to 7200 CPU seconds. All the tables report the average of 4 instances and the number of instances for which the optimal solution was not obtained within the time limit is denoted using a superscript. Moreover, '-' means that more than 2 h were required to obtain a feasible solution for each of the 4 instances.

5.1. Finding the best ordering constraints for $F_{(k_1,k_2)ATM}$ and $F_{(k_1,k_2)TM}$

A preliminary computational study was carried out to select the most efficient ordering constraint family described in Sections 3.2.1 and 3.3.1 using the instances introduced above with 128 and 256 intensities (see Tables 1 and 2). The first two columns contain the number of intensities and clusters, respectively. The rest of the table is divided into different blocks. Each one of them reports the time needed to obtain the optimal solution in seconds, the gap between the best solution and the best bound (Gap BB), the gap between the optimal solution of the integer problem and the optimal solution of the linear relaxation (Gap LR), and finally, the number of nodes explored in the branching tree by each formulation. If the optimal solution is not obtained with any formulation, the best solution found among all the formulations is chosen to compute Gap LR.

In Table 1, we can see that constraints (6a) provide the best computational times to solve the $F_{(k_1,k_2)ATM}$ formulation for 128 intensities, but instances with 256 intensities are not solved before 7200 s. Alternatively, constraints (4a) and (4b) provided feasible solutions for 256 intensity instances. Therefore, constraints (4b) will be included in our segmentation model, since these constraints solve instances with 128 intensities in less time than constraints (4a).

As in the $F_{(k_1,k_2)ATM}$ formulation, we also analysed the results provided by the different ordering constraints included in the $F_{(k_1,k_2)TM}$ formulation. These computational results are shown in Table 2, which shows that the best computational times to solve the $F_{(k_1,k_2)TM}$ formulation for 128 intensities are given by constraints (12a). However, this family of constraints does not even provide feasible solutions for instances with 256 intensities. For this reason, constraints (10a) were chosen to carry out the computational study for $F_{(k_1,k_2)TM}$.

5.2. Comparing formulations

Table 3 shows the computational results obtained applying the two best formulations existing in the literature to solve the anti-k-centrum problem, F_{DOMP_G} and F_{OT_θ} (see Appendix A.1 and A.2, respectively, for more details about these formulations), compared with the one proposed in Section 3.1 using 3D-STEM images.

The F_{DOMP_G} formulation only solved instances with 128 intensities. The F_{OT_θ} formulation has provided segmentation up to 512 intensities and the computational time was significantly reduced. Nevertheless, this formulation could not provide solutions of instances with 1024 intensities in less than 7200 s. Finally, the F_{AkC} formulation reported the best performance, reducing the computing time substantially to obtain the optimal solution of instances with 128, 256 and 512 intensities. Moreover, instances with 1024 intensities were only solved in less than ten minutes with the F_{AkC} formulation. All the formulations provided very good linear relaxation values with gaps smaller than 0.3% and the Nodes columns in the table show that most of the instances were solved in the root node.

Table 4 provides the computational results obtained by applying F_{DOMP_G}, F_{OT_θ} and F_{AkC} formulations to the anti-k-centrum problem to segment 2D-STEM images. This table is organised in the same way as Table 3 and the results obtained show the performance of the formulations is similar to what was observed for 3D-STEM images.

Table 2
Different ordering constraints included in the $F_{(k_1,k_2)TM}$ formulation.

		(10a)				(11a)				(12a)			
n	p	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes
128	2	600.24	0	32.97	8055.25	899.69	0	33.76	1882.5	931.17	0	24.75	3475.50
128	3	3333.32	0	37.70	115815.75	3181.24	0	34.73	14403.25	1899.87	0	27.39	10398.75
128	4	5653.98(2)	6.45	39.46	94996.5	4862.83(1)	1.68	32.65	24,553	1794.43	0	25.55	13302.25
128	5	4207.94	0	40.71	52518.25	5219.34(1)	4.11	31.32	26,188	1296.91	0	24.77	11,954
256	2	7204.33(4)	79.24	65.83	2737.75	-	-	-	-	-	-	-	-
256	3	7206.98(4)	89.67	63.78	0	-	-	-	-	-	-	-	-
256	4	7204.99(4)	76.06	63.89	1.25	-	-	-	-	-	-	-	-
256	5	7206.05(4)	68.66	58.76	480.75	-	-	-	-	-	-	-	-

Table 3
Computational results of the F_{DOMP_G}, F_{OT_θ} and F_{AkC} formulations to solve the anti-k-centrum problem using 3D-STEM images.

		F_{DOMP_G}				F_{OT_θ}				F_{AkC}			
n	p	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes
128	2	315.57	0	0.11	0	13.71	0	0.11	0	0.43	0	0.11	0
128	3	264.52	0	0	0	13.21	0	0	0	0.31	0	0	0
128	4	268.09	0	0	0	15.03	0	0.06	0	0.33	0	0.06	0
128	5	240.85	0	0	0	13.01	0	0	0	0.29	0	0	0
256	2	-	-	-	-	121.01	0	0.10	0	16.16	0	0.10	0
256	3	-	-	-	-	132.79	0	0.01	0	17.39	0	0.01	0
256	4	-	-	-	-	114.15	0	0.01	0	15.81	0	0.01	0
256	5	-	-	-	-	117.93	0	0	0	13.41	0	0	0
512	2	-	-	-	-	1014.45	0	0.05	5.5	110.77	0	0.11	0
512	3	-	-	-	-	1830.42	0	0.01	0	109.64	0	0.01	0
512	4	-	-	-	-	1643.14	0	0	0	99.02	0	0	0
512	5	-	-	-	-	1470.52	0	0	0	101.52	0	0	0
1024	2	-	-	-	-	-	-	-	-	661.56	0	0.28	10.25
1024	3	-	-	-	-	-	-	-	-	669.84	0	0	0
1024	4	-	-	-	-	-	-	-	-	583.23	0	0	0
1024	5	-	-	-	-	-	-	-	-	564.76	0	0	0

Table 4
Computational results of the F_{DOMP_G}, F_{OT_θ} and F_{AkC} formulations to solve the anti-k-centrum problem using 2D-STEM images.

		F_{DOMP_G}				F_{OT_θ}				F_{AkC}			
n	p	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes
128	2	217.39	0	0	0	20.20	0	0	0	0.58	0	0	0
128	3	207.42	0	0	0	19.85	0	0	0	0.58	0	0	0
128	4	178.39	0	0	0	20.7	0	0	0	0.55	0	0	0
128	5	169.50	0	0	0	20.94	0	0	0	0.53	0	0	0
256	2	5985.12(2)	48.30	0	0	217.69	0	0	0	26.14	0	0	0
256	3	5824.79(2)	47.43	0	0	202.04	0	0	0	29.20	0	0	0
256	4	5915.04(2)	48.02	0	0	210.63	0	0	0	17.86	0	0	0
256	5	5125.26(1)	24.15	0	0	236.43	0	0	0	22.05	0	0	0
512	2	-	-	-	-	-	-	-	-	138.77	0	0.01	0
512	3	-	-	-	-	-	-	-	-	156.35	0	0.02	0
512	4	-	-	-	-	-	-	-	-	155.42	0	0.02	0
512	5	-	-	-	-	-	-	-	-	133.40	0	0.08	190.25
1024	2	-	-	-	-	-	-	-	-	617.91	0	0.02	0
1024	3	-	-	-	-	-	-	-	-	642.72	0	0.02	0
1024	4	-	-	-	-	-	-	-	-	634.04	0	0.04	38
1024	5	-	-	-	-	-	-	-	-	497.77	0	0.04	284

Tables 5 (3D-STEM images) and 6 (2D-STEM images) show the computational times, gap and nodes needed to solve the (k_1, k_2)-trimmed mean problem using F_{DOMP_G}, $F_{(k_1,k_2)TM}$, F_{OT_θ}, and $F_{(k_1,k_2)TM_{OT}}$ (see Appendix A.1 and A.2 for more details of the F_{DOMP_G} and F_{OT_θ} formulations, respectively). The F_{OT_θ} and $F_{(k_1,k_2)TM_{OT}}$ formulations provided optimal solutions for larger instances than those provided by F_{DOMP_G} and $F_{(k_1,k_2)TM}$. Instances with 256 and 512 intensities were solved applying F_{OT_θ} and $F_{(k_1,k_2)TM_{OT}}$ formulations within the time limit. It is worth highlighting that the computational times to obtain optimal solutions provided by $F_{(k_1,k_2)TM_{OT}}$ are one order of magnitude lower than the time needed by F_{OT_θ}. For the case of instances with 1024 intensities, although $F_{(k_1,k_2)TM_{OT}}$ did not provide optimal solutions for most instances within 7200 s, it reports feasible solutions with a small gap. The number of instances without any solution obtained within 7200 s is shown in parenthesis. In this case, the Gap LR values are larger than those obtained for the F_{AkC} formulation, and very few instances are solved in the root node (see the Nodes column).

Tables 7 and 8 report the computational results with the (k_1, k_2)-anti-trimmed mean problem for 3D-STEM and 2D-STEM images, respectively. In contrast to the F_{DOMP_G} formulation, which

Table 5
Computational results of the F_{DOMP_G}, $F_{(k_1,k_2)TM}$, F_{OT_θ}, and $F_{(k_1,k_2)TM_{OT}}$ formulations to solve the (k_1, k_2)-trimmed mean problem using 3D-STEM images.

		F_{DOMP_G}				$F_{(k_1,k_2)TM}$				F_{OT_θ}				$F_{(k_1,k_2)TM_{OT}}$			
n	p	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes
128	2	7203.12(4)	8.33	10.54	71701.25	600.24	0	32.97	8055.25	18.20	0	0.70	13.25	1.22	0	0.91	0
128	3	7202.09(4)	5.90	10.25	77007.25	3333.33	0	37.70	115815.75	17.19	0	0.42	0	1.50	0	0.76	1.5
128	4	7202.50(4)	4.74	10.01	128,057	5653.98	6.45	39.47	94996.5	15.40	0	0.14	0	1.14	0	0.45	2.5
128	5	7201.44(4)	2.85	10.02	153064.75	4207.94	0	40.73	52518.25	16.06	0	0.23	0	0.96	0	0.59	0
256	2	-	-	-	-	7204.33(4)	79.24	41.37	2737.75	154.74	0	0.24	42.75	13.86	0	1.11	0
256	3	-	-	-	-	7206.98(4)	89.67	44.49	0	182.61	0	0.16	49.25	15.39	0	1.36	605.75
256	4	-	-	-	-	7204.99(4)	76.06	45.27	1.25	187.47	0	0.16	0	12.90	0	0.93	158.5
256	5	-	-	-	-	7206.05(4)	68.66	45.91	480.75	208.75	0	0.28	290.5	12.46	0	1.14	587.75
512	2	-	-	-	-	-	-	-	-	2819.58	0	0.19	151	262.89	0	1.32	1991
512	3	-	-	-	-	-	-	-	-	4133.13	0	0.16	351.5	695.10	0	1.62	7973.5
512	4	-	-	-	-	-	-	-	-	3999.28	0	0.16	587	234.02	0	1.21	3969.25
512	5	-	-	-	-	-	-	-	-	5138.95	0	0.24	2073.25	381.47	0	1.40	7934.75
1024	2	-	-	-	-	-	-	-	-	-	-	-	-	7200.87(4)	2.81	2.82	1081
1024	3	-	-	-	-	-	-	-	-	-	-	-	-	7200.43(1)(4)	11.10	11.12	0
1024	4	-	-	-	-	-	-	-	-	-	-	-	-	7201.15(4)	22.79	22.81	710.25
1024	5	-	-	-	-	-	-	-	-	-	-	-	-	6932.82(3)	1.66	1.80	531.75

Table 6
Computational results of the F_{DOMP_G}, $F_{(k_1,k_2)TM}$, F_{OT_θ}, and $F_{(k_1,k_2)TM_{OT}}$ formulations to solve the (k_1, k_2)-trimmed mean problem using 2D-STEM images.

		F_{DOMP_G}				$F_{(k_1,k_2)TM}$				F_{OT_θ}				$F_{(k_1,k_2)TM_{OT}}$			
n	p	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes
128	2	7202.00(4)	2.02	8.16	462050.25	1282.68	0	49.26	8031.75	27.64	0	0.12	0	1.2	0	0.61	0
128	3	7202.06(4)	1.75	8.17	267664.75	4041.77	0	51.48	19264.75	27.78	0	0.21	0	1.38	0	0.79	0
128	4	7203.58(4)	2.29	26.46	983072.75	4834.27	0	52.31	26931.75	29.46	0	0.34	0	1.88	0	0.98	0
128	5	7201.62(4)	1.76	8.28	370,114	5604.49(1)	1.72	54.86	23,347	32.52	0	0.45	0	2.09	0	1.17	0
256	2	-	-	-	-	7209.91(4)	79.29	56.71	60.25	293.08	0	0.14	0	10.90	0	0.98	0
256	3	-	-	-	-	7204.93(4)	76.43	57.79	0	325.94	0	0.27	0	16.26	0	1.19	0
256	4	-	-	-	-	7205.27(4)	72.91	59.02	659.5	376.12	0	0.38	0	19.15	0	1.43	162.25
256	5	-	-	-	-	7203.63(4)	66.37	60.32	655.75	379.6	0	0.45	145	29.96	0	1.59	250.5
512	2	-	-	-	-	-	-	-	-	6503.94(2)(2)	0	0.14	0	194.04	0	1.17	460.5
512	3	-	-	-	-	-	-	-	-	6208.88(2)(2)	0	0.27	213.5	260.40	0	1.44	884
512	4	-	-	-	-	-	-	-	-	6038.99(2)(2)	0	0.35	389.5	708.23	0	1.65	4198.75
512	5	-	-	-	-	-	-	-	-	7002.03(2)(3)	0.54	0.76	202.5	832.08	0	1.84	5037.75
1024	2	-	-	-	-	-	-	-	-	-	-	-	-	6657.31(3)	1.67	2.20	2168.75
1024	3	-	-	-	-	-	-	-	-	-	-	-	-	7201.70(4)	2.46	2.54	1851.25
1024	4	-	-	-	-	-	-	-	-	-	-	-	-	7201.12(4)	21.74	21.74	267.5
1024	25	-	-	-	-	-	-	-	-	-	-	-	-	7201.34(4)	3.61	3.64	196.33

Table 7
Computational results of the F_{DOMP_G}, $F_{(k_1,k_2)ATM}$, F_{OT_θ}, and $F_{(k_1,k_2)ATM_{OT}}$ formulations to solve the (k_1, k_2)-anti-trimmed mean problem using 3D-STEM images.

		F_{DOMP_G}				$F_{(k_1,k_2)ATM}$				F_{OT_θ}				$F_{(k_1,k_2)ATM_{OT}}$			
n	p	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes
128	2	7203.52(4)	48.49	77.86	4365.75	1524.00	0	99.94	11612.5	15.55	0	1.68	0	6.12	0	1.68	0
128	3	7203.07(4)	48.30	66.78	7842.25	1837.45	0	99.93	10774.75	17.53	0	4.87	0	7.95	0	4.87	0
128	4	7203.05(4)	63.71	57.91	9502	2111.39	0	99.90	19,921	16.65	0	7.72	0	6.28	0	7.72	0
128	5	7202.51(4)	87.48	63.65	10098.75	6208.16	46.34	99.91	45656.75	22.11	0	21.30	0	7.75	0	21.30	0
256	2	-	-	-	-	7200.57(4)	99.59	99.76	0	157.78	0	0.57	0	42.49	0	0.57	0
256	3	-	-	-	-	7200.56(4)	99.63	99.73	1670.75	189.00	0	2.14	0	58.65	0	2.14	0
256	4	-	-	-	-	7201.23(4)	99.74	99.73	2218.5	210.66	0	4.97	0	47.05	0	4.97	0
256	5	-	-	-	-	7201.44(4)	99.76	99.67	86.33	287.41	0	10.18	0	71.77	0	10.18	56.25
512	2	-	-	-	-	-	-	-	-	2837.15	0	0.52	0	408.74	0	0.52	0
512	3	-	-	-	-	-	-	-	-	3076.41	0	1.17	0	493.77	0	1.17	0
512	4	-	-	-	-	-	-	-	-	4948.72(1)(4)	28.50	1.91	0	463.45	0	2.69	0
512	5	-	-	-	-	-	-	-	-	7201.38(4)	16.15	16.77	11.75	1243.08	0	8.21	790.25
1024	2	-	-	-	-	-	-	-	-	-	-	-	-	2780.90	0	1.96	528.75
1024	3	-	-	-	-	-	-	-	-	-	-	-	-	4591.78(2)	3.12	3.29	1202.25
1024	4	-	-	-	-	-	-	-	-	-	-	-	-	4045.44(3)	4.85	5.36	251
1024	5	-	-	-	-	-	-	-	-	-	-	-	-	5401.93(4)	31.97	32.01	4.75

Table 8
Computational results of the F_{DOMP_G}, $F_{(k_1,k_2)ATM}$, F_{OT_θ}, and $F_{(k_1,k_2)ATM_{OT}}$ formulations to solve the (k_1, k_2)-anti-trimmed mean problem using 2D-STEM images.

		F_{DOMP_G}				$F_{(k_1,k_2)ATM}$				F_{OT_θ}				$F_{(k_1,k_2)ATM_{OT}}$			
n	p	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes	Time	Gap BB	Gap LR	Nodes
128	2	7203.63(4)	41.69	75.01	30038.75	6901.19(3)	61.24	98.68	17622.25	33.32	0	16.24	0	16.97	0	16.24	0
128	3	7203.20(4)	43.46	74.73	37,649	7200.66(4)	97.68	99.94	10262.5	45.98	0	25.95	88.75	19.97	0	25.95	0
128	4	7202.87(4)	42.55	74.61	38892.75	7200.39(4)	98.85	99.96	11056.25	66.34	0	28.91	202.75	15.57	0	28.91	296.25
128	5	7202.74(4)	35.97	75.03	28,942	7201.16(4)	98.74	99.95	15,342	91.97	0	29.75	546.75	18.37	0	29.75	454.5
256	2	-	-	-	-	-	-	-	-	366.13	0	11.82	49.25	98.02	0	11.82	42
256	3	-	-	-	-	-	-	-	-	541.22	0	18.24	528.25	116.20	0	18.24	333.25
256	4	-	-	-	-	-	-	-	-	2056.11	0	21.37	1972	184.06	0	21.37	1678.75
256	5	-	-	-	-	-	-	-	-	3548.77(1)	1.5	23.65	5665.25	620.93	0	23.65	10403.75
512	2	-	-	-	-	-	-	-	-	6506.18(2)(2)	0	9.68	298.5	769.88	0	11.38	389
512	3	-	-	-	-	-	-	-	-	7202.25(2)(4)	19.98	19.98	0	866.34	0	16.33	2043.75
512	4	-	-	-	-	-	-	-	-	7204.49(2)(4)	25.42	25.44	1.5	4273.16(2)	7.06	20.13	5347.25
512	5	-	-	-	-	-	-	-	-	5401.40(3)(4)	18.26	18.85	0	7201.18(2)(4)	13.66	21.59	10182.33
1024	2	-	-	-	-	-	-	-	-	-	-	-	-	6060.20(3)	9.66	12.89	411
1024	3	-	-	-	-	-	-	-	-	-	-	-	-	7201.96(4)	75.52	75.60	2
1024	4	-	-	-	-	-	-	-	-	-	-	-	-	7202.06(2)(4)	50.99	51.00	3.5
1024	5	-	-	-	-	-	-	-	-	-	-	-	-	7202.09(3)(4)	31.23	31.24	4

does not provide the optimal solution within the time limit with 128 intensities, $F_{(k_1,k_2)ATM}$ achieves the optimal solution for most instances of this size. The F_{OT_θ} and $F_{(k_1,k_2)ATM_{OT}}$ formulations make considerable improvements on the computational times of the others and thus allow us to solve larger instances. In particular, $F_{(k_1,k_2)ATM_{OT}}$ provided the optimal solution for every instance with 512 intensities. Moreover, this formulation achieved solutions for 1024 intensities, with acceptable gaps in most instances.

The general conclusion is that, to apply and solve large size instances of image segmentation with the above-mentioned λ-weights for DOMP, it is advisable to use our new formulations that exploit the structure of these problems, resulting in improvements in CPU time of one order of magnitude.

6. Validation of the model

In this section, we assess the performance of the discrete ordered median model in 2D and 3D-STEM images. We have used the phantoms introduced in the previous section to validate the proposed models and evaluate the quality of the segmentations. The particles in Fig. 2(a) are the smallest structures, whereas the background corresponds to the largest one. Therefore, it is expected that the first positions of the ordered intensity-allocation weighted distance vector will correspond to the intensities with the smallest frequencies. This justifies the use of the anti-k-centrum model to obtain a segmentation of this image using the F_{AkC} formulation. However, particles are not the smallest structures in Fig. 2(b), since there is a wide range of intensities that correspond to the support and intensities representing the particles are not found in the first positions of the ordered intensity-allocation weighted distance vector. Hence, the (k_1, k_2)-trimmed mean model seems to be suitable to perform segmentation of the image in Fig. 2(b) using the $F_{(k_1,k_2)TM_{OT}}$ formulation. After a preliminary analysis with different models, the aforementioned criteria provided the best results to segment these two images. In addition, a real experiment recorded by an electron microscope will be segmented to verify the performance in a real instance.

These analyses are carried out to compare the efficiency provided by DOMP formulations and classical segmentation models such as Otsu's method and p-means in the field of STEM (Belianinov et al., 2015; Hindson et al., 2011; Leary et al., 2012; Liu et al., 2020; Lopez-Haro et al., 2014). To perform a comparison between segmentations obtained by using classical models and DOMP with the formulations introduced in this work, the number of clusters (p) has been set to 4 and 5 for 3D-STEM images and to 2 and 3 for 2D-STEM images. These choices aim to identify the different elements that constitute the original image with as few clusters as possible.

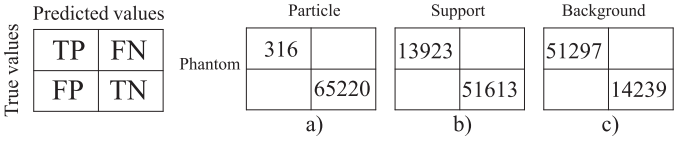

Fig. 3. TP represents the number of pixels that constitute each structure of the original image (a) particle, b) support and c) background). TN is the number of pixels that form the remaining parts (a) support and background, b) particle and background and c) particle and support).

In order to quantify the quality of the different segmentations we have divided the elements which constitute the image into three confusion matrices, because visually the image is composed of three elements (background, support and particles). Since the main goal of the experiment is to detect true particles, to compare the quality of the segmented particles we define the null hypothesis H_0 and the alternative hypothesis H_1 as:

$$\begin{cases} H_0: \text{ pixel that does not represent a particle,} \\ H_1: \text{ pixel that represents a particle.} \end{cases}$$

Each element of the confusion matrix referring to particles is defined as follows:

- True Positive (TP): Pixels successfully detected as particle. H_0 is false and it is rejected.
- True Negative (TN): Pixels successfully detected as different from particles. H_0 is true and it is not rejected.
- False Positive (FP): Pixels wrongly detected as particles. H_0 is true and it is rejected (Type I error).
- False Negative (FN): Pixels not detected as particles, in a wrong way. H_0 is false and it is not rejected (Type II error).

The same applies to the confusion matrices for the support and background. Figure 3 shows the number of pixels that correspond to particles, support and background in the phantom used to obtain the reconstruction shown in Fig. 2(a). This figure has 65536 pixels, of which 316 correspond to particles, 13,923 are generated by the support and 51,297 are the background of the image.

As is to be expected, the main goal is to avoid FP for particles, i.e., to avoid classifying as particles pixels that actually do not cor-

respond to particles. Indeed, these are the errors that produce the largest distortion in the analysis of characteristics of a nanomaterial. Therefore, although the goal is to have a small number of FN and FP, it would be better to have more FN than FP.

As in other references (see Murugan et al., 2020), different quantifiers have been used to compare the quality of the segmentations obtained using each formulation. In the following, we consider different types of metrics to measure the performance of segmentation models depending on the pixels classified (correctly or wrongly).

6.1. Rates of wrong classification (RWC)

These coefficients report the percentages of wrongly classified pixels. Accordingly, the lower the values of these coefficients (perfect segmentations correspond to 0), the better the quality of the segmentation obtained. The false negative rate (FNR) means the probability of classifying pixels as features different from particles when they are in fact particles:

$$FNR = \frac{FN}{FN + TP}.$$

The false positive rate (FPR) means the probability of wrongly classifying pixels as particles when they correspond to other features:

$$FPR = \frac{FP}{FP + TN}.$$

This rate plays an important role assessing the quality of the segmentations, since detecting false particles negatively affects on the nano-object analysis. Therefore, high quality segmentations are obtained if false positives are detected with a low probability.

6.2. Rates of correct classification (RCC)

We have named the coefficients which report the percentages of correctly clustered pixels 'RCC'. The higher the values (the value of 1 represents a perfect segmentation), the better the pixel classification. The true positive rate (TPR), also known as *recall* or *sensitivity*, measures the percentage of pixels correctly classified as particles (TP) among all pixels that form the particles in the original image (TP+FN):

$$TPR = \frac{TP}{TP + FN} = 1 - FNR.$$

Similarly, the true negative rate (TNR) or *specificity* computes the proportion of pixels correctly classified as different to particles (TN) among all the pixels that do not correspond to particles in the original image (TN+FP):

$$TNR = \frac{TN}{TN + FP} = 1 - FPR.$$

The efficiency of each classification model can be measured by plotting the false positive rate against the true positive rate in the ROC space. The perfect classification is represented by the point in the ROC space corresponding to $FPR = 0$ and $TPR = 1$. The Area Under the Curve (AUC) is computed as:

$$AUC = \frac{1 - FPR + TPR}{2} = \frac{specificity + sensitivity}{2}.$$

The F1 score is used to obtain a balance between *sensitivity* and *precision*, where *precision* measures the percentage of pixels correctly classified as particles (TP) among all the pixels identified as particle (TP+FP), i.e.,

$$precision = \frac{TP}{TP + FP}.$$

Thus, the F1 score is the harmonic mean between both metrics,

$$F1 = \frac{2TP}{2TP + FP + FN}.$$

The accuracy (ACC) measures the percentage of pixels correctly classified (TP+TN) among all the pixels that constitute the segmented image:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}.$$

Finally, the Rand Index (RI) and the Adjusted Rand Index (ARI) have been computed to obtain the similarity between the correct classification and the segmentation provided by the different formulations introduced in this paper (Hubert & Arabie, 1985). The RI takes a value between 0 and 1, while the ARI takes values between -1 and 1. The closer the value is to 1, the higher is the similarity between both classifications. The following matrix reports the overlaps between both segmentations (original and approximate):

Original \ Segmentation					sums
	c_{11}	c_{12}	...	c_{1p}	s_{x1}
	c_{21}	c_{22}	...	c_{2p}	s_{x2}
	⋮	⋮	...	⋮	⋮
	c_{p1}	c_{p2}	...	c_{pp}	s_{xp}
sums	s_{y1}	s_{y2}	...	s_{yp}	

c_{ij} represents the number of pixels which must belong to the cluster i but which are classified in cluster j. These coefficients are obtained as follows:

$$RI = \frac{2\sum_{ij}\binom{c_{ij}}{2} - \sum_i \binom{s_{x_i}}{2} - \sum_i \binom{s_{y_j}}{2} + \binom{n}{2}}{\binom{n}{2}}$$

$$ARI = \frac{\sum_{ij}\binom{c_{ij}}{2} - \left[\sum_i \binom{s_{x_i}}{2}\sum_j \binom{s_{y_j}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{s_{x_i}}{2} + \sum_j \binom{s_{y_j}}{2}\right] - \left[\sum_i \binom{s_{x_i}}{2}\sum_j \binom{s_{y_j}}{2}\right]/\binom{n}{2}}.$$

6.3. Anti-k-centrum

Figure 4 compares the solutions of 3D-STEM segmentations provided by the p-means, Otsu and anti-k-centrum models with $n = 128$ intensities and two numbers of clusters ($p = 4$, $p = 5$). The value of k in anti-k-centrum is set to $\frac{n}{2}$. One slice has been selected from the whole segmented volume to analyse the segmentation quality in an efficient way. The performance of segmentations for the remaining slices is similar. Ideally, intensities which contain information about particles should be grouped in the same cluster and it should be different to the clusters containing intensities representing the rest of structures. In Fig. 4, cluster 4 for the case $p = 4$ and cluster 5 for $p = 5$ will represent particles. It can be observed with $p = 4$ that p-means and Otsu models assign to the same cluster the particles and a large number of pixels corresponding to support. Hence, these models provide a large false positive rate (0.1538 and 0.1415, respectively) due to the large number of false positives obtained (10029 and 9226). However, the anti-k-centrum model identified pixels corresponding to particles in a highly effective way. This model had a very small false positive rate (0.0001), since very few pixels were identified as particles in a wrong way (55 pixels). Similar results were obtained by applying p-means for $p = 5$. However, the Otsu's method with 5 clusters provided 0 pixels classified as particles wrongly but at the expense of identifying a small number of pixels as particles successfully (18 pixels). In addition, the RI and ARI coefficients took the highest values for segmentations obtained using the anti-k-centrum model (the RI reported 0.9791 and 0.9738 for $p = 4$ and $p = 5$ respectively and 0.9536 and 0.9420 for the ARI coefficient).

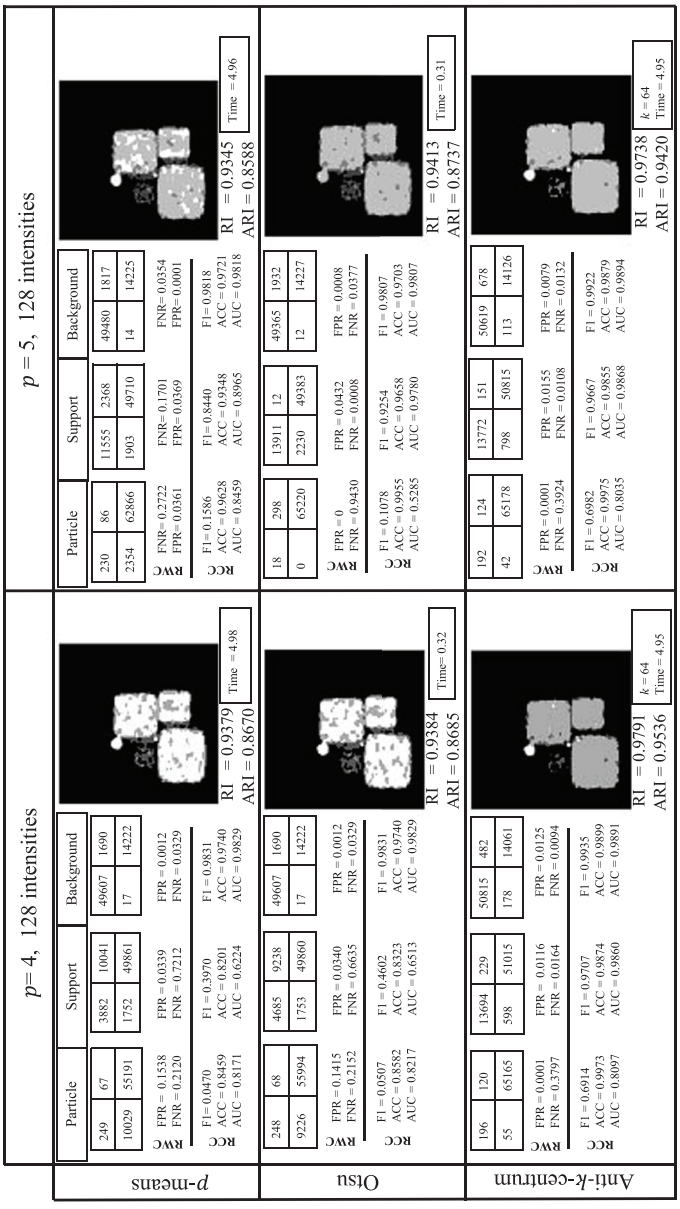

Fig. 4. Quantification of the results obtained by applying p-means, Otsu and anti-k-centrum models to segment a 3D-STEM image with $p = 4$ and $p = 5$.

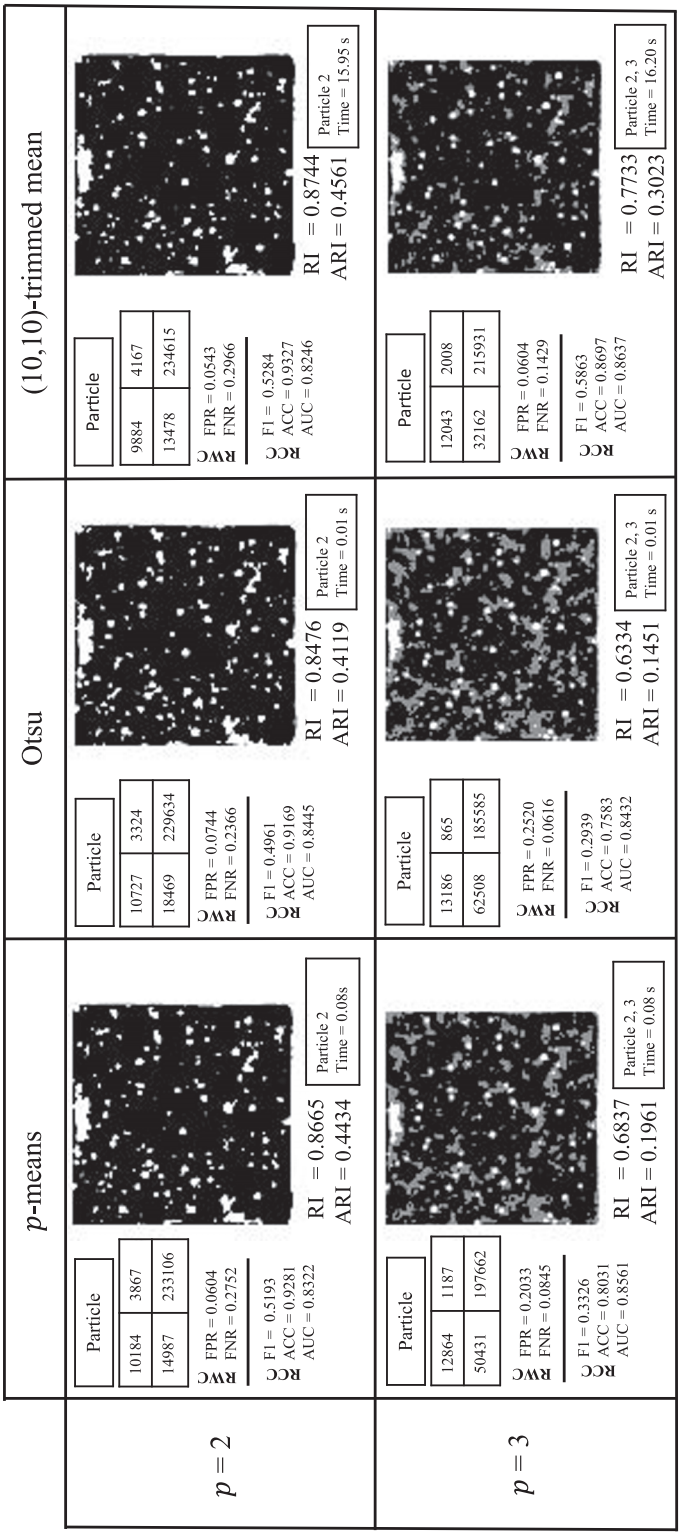

Fig. 5. Quantification of the results obtained by applying p-means, Otsu and (k_1, k_2)-trimmed mean models to segment a 2D-STEM image.

The remaining rates of RCC coefficients have a similar performance, since the highest values correspond to the anti-k-centrum model. Therefore, we can ensure that this formulation provided the most accurate segmentation. The quality obtained with the different coefficients may be visually confirmed with the images in Fig. 4. Segmentations provided by the anti-k-centrum correctly identify most pixels belonging to particles (white colour). However, p-means and Otsu segmentations contain a large number of pixels classified as particles, although they actually belong to support.

In the STEM segmentation field, clustering is carried out on the vector of intensities in such a way that the clusters are delimited by certain intensities, which are called threshold values. Therefore, some intensities are identified as particles (associated with the cluster of particles) if their intensities are larger than the lower threshold value defining this cluster. Figure 6 shows the compromise between TPR and FPR depending on the threshold value from which an intensity is considered as a particle for segmentations shown in Fig. 4 with $p = 4$ and $p = 5$. The red dot represents the anti-k-centrum model, the green dot is the solution provided by the Otsu's method and the p-means is represented by a yellow dot. Moreover, the results provided by the (k_1, k_2)-trimmed mean and (k_1, k_2)-anti-trimmed mean are represented by black and orange dots, respectively. The p-means, Otsu, (k_1, k_2)-trimmed mean, and (k_1, k_2)-anti-trimmed mean provide a higher TPR, but at the cost of a high FPR. However, the anti-k-centrum model achieves a

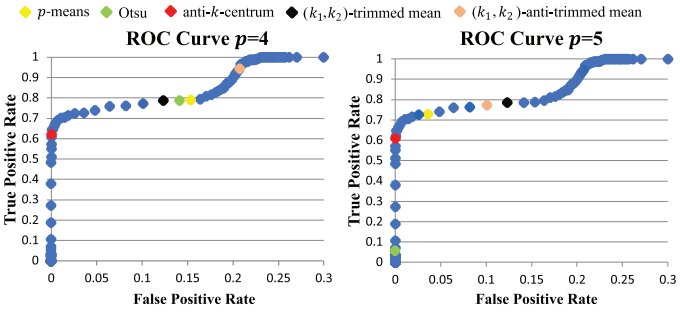

Fig. 6. ROC curves obtained by 3D-STEM image segmentations which represent a comparison between TPR and FPR depending on the minimum intensity value considered as a particle.

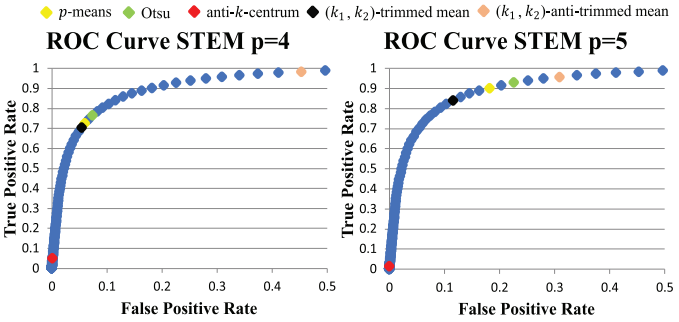

Fig. 7. ROC curves obtained by 2D-STEM image segmentations which represent a comparison between TPR and FPR depending on the minimum intensity value considered as a particle.

very good compromise, with a slightly lower TPR value but an FPR close to zero. Accordingly, the anti-k-centrum model results in a better segmentation, since electron tomography segmentations try to avoid false particles, i.e., the value of FP must be close to zero.

6.4. (k_1, k_2)-Trimmed mean

Following the comparison performed for 3D-STEM images, Fig. 5 shows the evaluation of the quality of the segmentations provided by the p-means, Otsu and (k_1, k_2)-trimmed mean models using formulations for a 2D-STEM image with 128 intensities. In this case, only one confusion matrix for particle and background was obtained, since this image has only two different structures (particles and background). Two different numbers of clusters are considered, $p = 2$ and $p = 3$, and $k_1 = k_2 = 10$. It may be observed that the FPR takes the lowest values for the (k_1, k_2)-trimmed mean model (0.0543 and 0.0604 for $p = 2$ and $p = 3$, respectively) while the p-means (0.0604 and 0.2033) and Otsu (0.0744 and 0.2520) result in higher FPR values. Regarding the RCC coefficients, they take the highest values for the (k_1, k_2)-trimmed mean model. This performance resulting from the RWC and RCC coefficients is confirmed in Fig. 7. It may be observed with $p = 2$ that the compromise between false positive rate and true positive rate is very similar for the p-means, Otsu and (k_1, k_2)-trimmed mean. Nevertheless, if 3 clusters are considered, there are differences in the ROC curve among these formulations. Clearly, the (k_1, k_2)-trimmed mean achieved the best compromise between false positive and true positive rates. The images in Fig. 5 report the quality of the segmentations in a visual way. It may be observed that the p-means, Otsu and (k_1, k_2)-trimmed mean provide similar results for $p = 2$. However, the (k_1, k_2)-trimmed mean identifies particles with a better quality than the p-means and Otsu for $p = 3$.

In Fig. 9 of Appendix we have carried out a similar analysis for the (k_1, k_2)-anti-trimmed mean with a 2D-STEM image that represents a small number of particles and low noise. The results show that the (k_1, k_2)-anti-trimmed mean provides better quality segmentation than the OTSU and p-means.

6.5. Segmentation of an experimental 2D-STEM image

In order to extend the analysis beyond phantom images, the p-means and (k_1, k_2)-trimmed mean models were used to solve an experimental 2D-STEM image (see Fig. 8). Note how the image shows small bright areas corresponding to the nanoparticles on

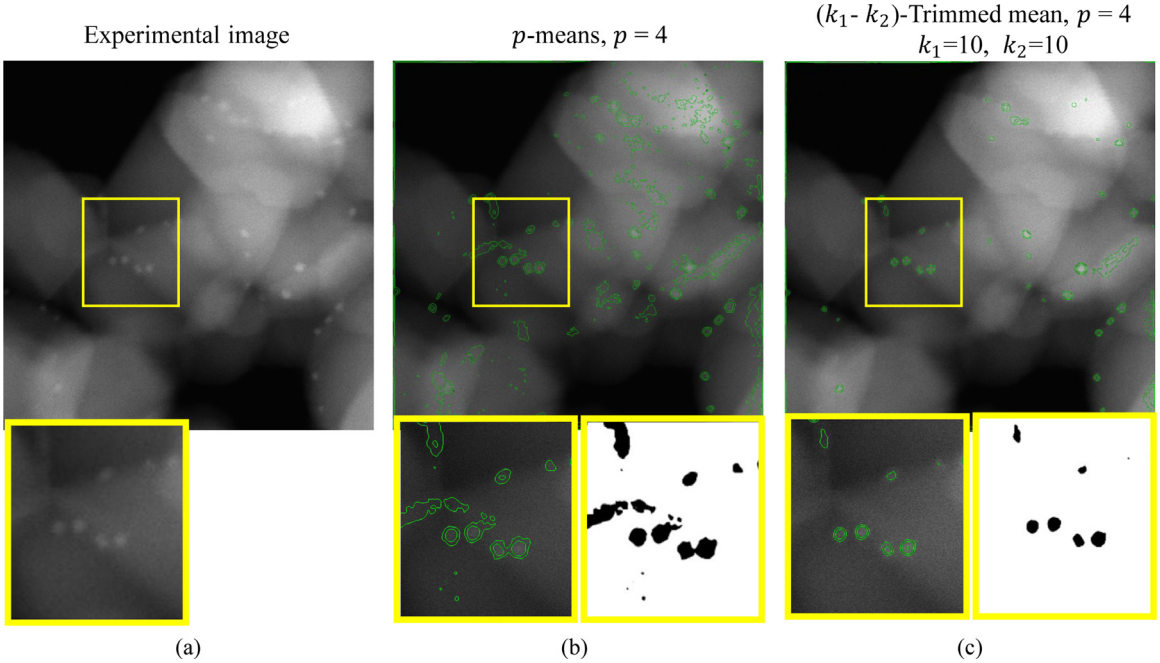

Fig. 8. Segmentation of an experimental 2D-STEM image: (a) 2D-STEM image. (b) p-means segmentation and (c) (k_1, k_2)-trimmed mean segmentation. Yellow rectangles correspond to enlargements of the areas marked by the squares.

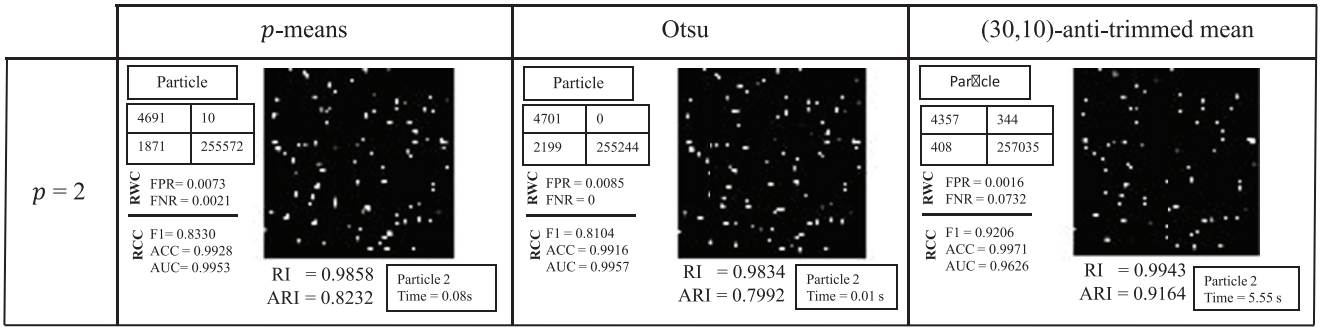

Fig. 9. Quantification of the results obtained by applying the p-means, Otsu and (k_1, k_2)-trimmed mean models to segment a 2D-STEM image.

top of the non-homogenous larger particles (support), which display areas with similar intensities to those corresponding to the nanoparticles. These features make it difficult to distinguish between the intensities belonging to the particles and the support. See Liu & Corma (2018) for further details about the structure of this type of images.

To perform the segmentation of the 2D-STEM image, the following parameters were considered to classify the intensities of the pixels for both the p-means and (k_1, k_2)-trimmed mean model: 128 intensities, $p = 4$, $k_1 = 10$ and $k_2 = 10$. The results obtained for each model were superimposed on the original image as a transparent green contour (see Fig. 8(b) and (c)). The pixels belonging to clusters 2, 3 and 4 are included in these contour images. Generally speaking, in both cases there is a good correlation between the particles and the segmented areas. Nevertheless, it can be seen how the p-means (Fig. 8(b)) has provided a much noisier image segmentation than the (k_1, k_2)-trimmed mean model. This is clearly shown in the enlargement marked with a yellow square in Fig. 8(b) and (c). This specific area shows the transparent green contour and black and white images. The latter shows the pixels classified as background (cluster corresponding to $p = 1$) in white and the clusters $p > 1$ as black. In this specific area, the p-means model not only shows a higher number of false positives but also the particles identified are not well segmented, showing large artifacts, e.g. they are not well separated and some tails appear in the surrounding areas of the particles. These artifacts appear because in the selected clusters not only the pixels belonging to the particles are classified but also pixels corresponding to the background are included. The (k_1, k_2)-trimmed mean model achieves a high level of accuracy, identifying every particle of the original image and providing better results than the p-means. In addition, this result is very closely aligned with those obtained from the phantom images. Therefore, our proposed models are very promising tools to obtain high quality segmentations that allow us to quantify the structural properties of nanomaterials and obtain a better insight into their chemical or physical properties.

7. Conclusions

This paper has studied an application of the ordered median problem to segment 3D-STEM and 2D-STEM images. Classical models do not provide good quality segmentations of small particles. However, the ordered median operator allows us to select in the objective function the frequencies which usually represent small particles thanks to the λ-vector values. Different formulations have been proposed depending on the λ-vector structure (anti-k-centrum, (k_1, k_2)-trimmed mean, (k_1, k_2)-anti-trimmed mean) to improve the computational times needed to obtain the optimal solution for each segmentation model. The formulations introduced in this paper have substantially reduced the computing time to obtain the optimal solution, making it possible to solve larger size instances. The importance of solving large size STEM instances must be emphasised, since the larger the instances solved the more accurate the segmentations are with respect to the original images. Finally, this paper also proposes an efficient way of quantifying the segmentations obtained by analysing their confusion matrices.

Acknowledgment

The authors have been partially supported by projects PID2020-114594GB-{C21, C22}, PID2020-113006RB-I00, PID2019-110018GA-I00, MAT2017-87579-R funded by MCIN/AEI/10.13039/501100011033 and project NetmeetData Fundacion BBVA convocatoria 2019. This work has been co-funded by the 2014–2020 ERDF Operational Programme and by the Department of Economy, Knowledge, Business and University of the Regional Government of Andalusia; projects FEDER-UCA18-106895, FEDER-UCA18-107139, FEDER-US-1256951 and P18-FR-1422.

Appendix A.

In the computational analysis described in Section 5, the formulations proposed in this manuscript have been compared with the two most promising formulations in the state of the art of DOMP. These formulations, are the block formulation (Espejo et al., 2021; Puerto et al., 2013; Puerto et al., 2016) and OT_θ formulation (Marín et al., 2020). For the sake of completeness, we have described both formulations in this appendix.

A1. Ordered median problem with blocks

The DOMP formulation with blocks is developed in Puerto et al. (2013), Puerto et al. (2016), Espejo et al. (2021). A block is defined as a set of consecutive non-null identical values in the λ-vector. The structure of this formulation takes advantage of sequences of repetitions in some classical λ-vectors by defining new vectors which provide the information about the length of every block.

We consider $\lambda = (1, 1, 0, 1, 1)$ a vector with 2 blocks as an example. Let I be the number of non-null blocks in λ-vector and $I := \{1, \ldots, I\}$. Let us define the vector $\gamma = (\gamma_1, \ldots, \gamma_I)$, being γ_i, $i \in I$ the value of the elements in the ith block of repeated elements in λ, i.e., in our example γ-vector is $\gamma = (1, 1)$. We define the vector $\alpha = (\alpha_1, \ldots, \alpha_I, \alpha_{I+1})$ where α_i, with $i \in I$, is the number of elements taking the value of zero between the $(i-1)$th and ith blocks of positive elements in λ-vector and α_{I+1} the number of zeros after the Ith positive block in λ, i.e., in our example $\alpha = (0, 1, 0)$. In addition, we set the vector $\beta = (\beta_1, \ldots, \beta_I)$ where β_i, with $i \in I$, is the number of elements in the ith block of positive elements in λ-vector, i.e., $\beta = (2, 2)$ in the example. This formulation considers an ordered intensity-allocation weighted distance vector created from the matrix d by removing the duplicated

Table 9
Dimensions of the F_{DOMP_G} (Appendix A.1), F_{OT} (Section 4) and F_{OT_θ} (Appendix A.2) formulations, where $K^+ = \{k \in N : \Delta_k > 0\}$ and $K^- = \{k \in N : \Delta_k < 0\}$.

	Variables							Constraints										
	x	y	u	v	w	z	θ	(1a)	(13a)	(13b)	(13c)	(16a)	(16b)	(16c)	(16d)	(17a)	(17b)	(17d)
F_{DOMP_G}	n^2	n	$\lvert I\rvert\cdot\lvert H\rvert$	$\lvert I\rvert\cdot\lvert H\rvert$	-	-	-	1	n^2	n	-	$\lvert H\rvert$	$(\lvert I\rvert-1)\cdot\lvert H\rvert$	$(\lvert I\rvert-1)\cdot\lvert H\rvert$	$\lvert I\rvert\cdot\lvert H\rvert$	-	-	-
F_{OT}	n^2	n	-	-	n^2	n	-	1	n^2	n	n^2	-	-	-	-	-	-	-
F_{OT_θ}	n^2	n	-	-	$n\cdot\lvert K^+\rvert$	$\lvert K^+\rvert$	$n^2\cdot\lvert K^-\rvert$	1	n^2	n	n^2	-	-	-	-	$\lvert K^-\rvert$	$n^2\cdot\lvert K^-\rvert$	n^2

values and sorting its components in increasing sequence. Let H be the number of different nonzero elements of the intensity-allocation weighted distance matrix and $H := \{1, \ldots, H\}$. The ordered intensity-allocation weighted distance vector is built as follows:

$$d_{(0)} = 0 < d_{(1)} < \ldots < d_{(H)} := \max\{d_{ij} : i, j \in N\}.$$

Let $h \in H$ and $k \in I$, we define the set of binary variables u_{kh} which take the value of one if the $(\sum_{j=1}^{k}\alpha_j + \sum_{j=1}^{k-1}\beta_j + 1)$th intensity-allocation weighted distance is at least $d_{(h)}$ and the value of zero otherwise. In addition, variables v_{kh} give the number of allocations in the kth block between positions $\sum_{j=1}^{k}\alpha_j + \sum_{j=1}^{k-1}\beta_j + 1$ and $\sum_{j=1}^{k}(\alpha_j + \beta_j)$ with a weight higher than $d_{(h)}$. Thus, the ordered median problem with blocks is formulated as follows:

$$(F_{DOMP_G}) \quad \min \quad \sum_{k\in I}\sum_{h\in H}\gamma_k(d_{(h)} - d_{(h-1)})v_{kh}$$

$$\text{s.t.} \quad (1a), (13a), (13b)$$

$$\sum_{k\in I}\alpha_k u_{kh} + \sum_{k\in I} v_{kh} + \alpha_{I+1} \geq \sum_{i\in N}\sum_{\substack{j\in N:\\ d_{ij}\geq d_{(h)}}} x_{ij}, \quad \forall h \in H, \tag{16a}$$

$$u_{kh} \geq u_{k-1,h}, \quad \forall k = 2\cdots, I,\ h \in H, \tag{16b}$$

$$\beta_{k-1}u_{kh} \geq v_{k-1,h}, \quad \forall k = 2\cdots, I,\ h \in H, \tag{16c}$$

$$v_{kh} \geq \beta_k u_{kh}, \quad \forall k \in I,\ h \in H, \tag{16d}$$

$$x_{ij}, y_j \in \{0,1\}, \quad \forall i, j \in N, \tag{16e}$$

$$u_{kh} \in \{0,1\}, \quad \forall k \in I, h \in H. \tag{16f}$$

$$v_{kh} \in \mathbb{Z} \cap [0, \beta_k], \quad \forall k \in I, h \in H. \tag{16g}$$

The objective function is the ordered sum of the intensities-allocation weighted distances. Constraints (16a) guarantee that the number of intensities with an intensity-allocation weighted distance greater than or equal to $d_{(h)}$ is either equal to the number of allocations with an intensity-allocation weighted distance at least $d_{(h)}$ whenever $v_{lh} > 0$ or less than or equal to α_{l+1} otherwise. Constraints (16b) control that u_{kh} must be greater than or equal to $u_{k-1,h}$. Upper and lower bounds of variables v_{kh} are given by constraints (16c) and (16d) respectively.

A2. OT_θ Formulation

Marín et al. (2020) introduce a formulation of the ordered median problem based on the rationale that explains the original OT formulation (Ogryczak & Tamir, 2003) (see Section 4). Differently from the original OT formulation, this one is valid for any λ-vector structure, non necessarily monotone. Let $\Delta_k = \lambda_k - \lambda_{k-1}$. It requires new variables θ_{ij}^k that take the value of 1 if the intensity-allocation weighted distance of assigning intensity i to cluster j is sorted in position k and 0 otherwise. Then, the formulation that results is:

$$(F_{OT_\theta}) \quad \min \quad \sum_{\substack{k\in N:\\ \Delta_k>0}} \Delta_k\left((n-k+1)z_k + \sum_{i\in N} w_{ik}\right) + \sum_{\substack{k\in N:\\ \Delta_k<0}} \Delta_k \sum_{i\in N}\sum_{j\in N} d_{ij}\theta_{ij}^k$$

$$\text{s.t.} \quad (1a), (13a), (13b), (13c),$$

$$\sum_{i\in N}\sum_{j\in N}\theta_{ij}^k = n-k+1, \quad \forall k \in N : \Delta_k < 0, \tag{17a}$$

$$\theta_{ij}^k \leq x_{ij}, \quad \forall i, j, k \in N : \Delta_k < 0, \tag{17b}$$

$$\theta_{ij}^k \geq 0, \quad \forall i, j, k \in N : \Delta_k < 0, \tag{17c}$$

$$\sum_{a:d_{ia}>d_{ij}} x_{ia} + y_j \leq 1, \quad \forall i, j \in N, \tag{17d}$$

$$w_{ik}, z_k \geq 0, \quad \forall i, k \in N : \Delta_k > 0, \tag{17e}$$

$$x_{ij}, y_j \in \{0,1\}, \quad \forall i, j \in N. \tag{17f}$$

The first term of the objective function contains the sum of the $(n-k+1)$th greatest intensity-allocation weighted distances multiplied by Δ_k for $k \in N$ such that $\Delta_k > 0$. The second term contains the sum of the remaining $(n-k+1)$th greatest intensity-allocation weighted distances multiplied by Δ_k for $k \in N$ such that $\Delta_k < 0$. Constraints (17a) state the number of allocations controlled by variables θ_{ij}^k and (17b) force that every allocation managed by θ_{ij}^k is also controlled by x_{ij}. Finally, the family of constraints (17d) ensures the allocation of each intensity to the closest cluster.

To have an idea of the complexity of the different formulations, we show in Table 9 a comparative of the number of variables and constraints needed to formulate the F_{DOMP_G}, F_{OT} (see Section 4) and F_{OT_θ}.

A3. Validation of the (k_1, k_2)-anti-trimmed mean model.

In this Appendix, a 2D-STEM image with a low density of particles supported on a continuous thin layer is segmented by using the (k_1, k_2)-anti-trimmed mean model. This 2D-STEM phantom image represents an example of nanomaterials used in devices for alternative energy, see Fig. 9. This new instance contains a small amount of particles and low noise, then we are interested in the smallest and largest intensity-allocation weighted distances. Ideally, the smallest intensity-allocation weighted distance will correspond to the ones of particles and the largest ones to those of the background. Therefore, taking $\lambda = (1, \ldots, 1, 0, \ldots, 0, 1, \ldots, 1)$ the resulting segmentation model attemps to identify two different structures: particles and background with a low level of noise. Hence, we only need 2 clusters to classify the intensities (see Fig. 9).

References

Aouad, A., & Segev, D. (2019). The ordered k-median problem: Surrogate models and approximation algorithms. *Mathematical Programming*, 1–29.
Bai, F., Fan, M., & Dong, L. (2021). Image segmentation method for coal particle size distribution analysis. *Particuology, 56*, 163–170.
Belianinov, A., Vasudevan, R., Strelcov, E., Steed, C., Yang, S. M., Tselev, A., … Kalinin, S. (2015). Big data and deep data in scanning and electron microscopies: Deriving functionality from multidimensional data sets. *Advanced Structural and Chemical Imaging, 1*(6).
Benati, S., & García, S. (2014). A mixed integer linear model for clustering with variable selection. *Computers & Operations Research, 43*, 280–285.
Benati, S., Ponce, D., Puerto, J., & Rodríguez-Chía, A. M. (2022). A branch-and-price procedure for clustering data that are graph connected. *European Journal of Operational Research, 297*(3), 817–830.
Benati, S., Puerto, J., & Rodríguez-Chía, A. M. (2017). Clustering data that are graph connected. *European Journal of Operational Research, 261*, 43–53.
Blanco, V. (2019). Ordered p-median problems with neighbourhoods. *Computational Optimization and Applications, 73*(2), 603–645.
Boland, N., Domínguez-Marín, P., Nickel, S., & Puerto, J. (2006). Exact procedures for solving the discrete ordered median problem. *Computers & Operations Research, 33*(11), 3270–3300.
Bradley, P. S., Fayyad, U. M., & Mangasarian, O. L. (1999). Mathematical programming for data mining: Formulations and challenges. *Journal on Computing, 11*, 217–238.
Brusco, M. J. (2003). An enhanced branch-and-bound algorithm for a partitioning problem. *British Journal of Mathematical and Statistical Psychology, 56*, 83–92.
Deleplanque, S., Labbé, M., Ponce, D., & Puerto, J. (2020). A branch-price-and-cut procedure for the discrete ordered median problem. *INFORMS Journal on Computing, 32*(3), 582–599.
Espejo, I., Puerto, J., & Rodríguez-Chía, A. M. (2021). A comparative study of different formulations for the capacitated discrete ordered median problem. *Computers & Operations Research, 125*, 105067.
Gontar, L. C., Ozkaya, D., & Dunin-Borkowski, R. E. (2011). A simple algorithm for measuring particle size distributions on an uneven background from TEM images. *Ultramicroscopy, 111*(2), 101–106.
Hansen, P., & Jaumard, B. (1997). Cluster analysis and mathematical programming. *Mathematical Programming, 79*, 191–215.
Hindson, J. C., Saghi, Z., Hernandez-Garrido, J. C., Midgley, P. A., & Greenham, N. C. (2011). Morphological study of nanoparticle-polymer solar cells using high-angle annular dark-field electron tomography. *Nano Letters, 11*(2), 904–909.
Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.
Kalcsics, J., Nickel, S., & Puerto, J. (2003). Multifacility ordered median problems on networks: A further analysis. *Networks, 41*(1), 1–12.
Kalcsics, J., Nickel, S., Puerto, J., & Rodríguez-Chía, A. M. (2010). Distribution systems design with role dependent objectives. *European Journal of Operational Research, 202*, 491–501.
Labbé, M., Ponce, D., & Puerto, J. (2017). A comparative study of formulations and solution methods for the discrete ordered p-median problem. *Computers & Operations Research, 78*, 230–242.
Leary, R., Saghi, Z., Armbrüster, M., Wowsnick, G., Schlögl, Robert, T., … Midgley, P. (2012). Quantitative high-angle annular dark-field scanning transmission electron microscope (HAADF-STEM) tomography and high-resolution electron microscopy of unsupported intermetallic $gapd_2$ catalysts. *The Journal of Physical Chemistry, 116*(24), 13343–13352.
Liu, K., & Corma, A. (2018). Metal catalysts for heterogeneous catalysis: From single atoms to nanoclusters and nanoparticles. *Chemical Reviews, 118*, 4981–5079.
Liu, L., Lopez-Haro, M., Lopes, C. W., Rojas-Buzo, S., Concepcion, P., Manzorro, R., … Corma, A. (2020). Structural modulation and direct measurement of subnanometric bimetallic PtSn clusters confined in zeolites. *Nature Catalysis, 3*(8), 628–638.
Lopez-Haro, M., Guetaz, L., Printemps, T., Morin, A., Escribano, S., Jouneau, P. H., … Gebel, G. (2014). Three-dimensional analysis of nafion layers in fuel cell electrodes. *Nature Communications, 5*(1).
López-Haro, M., Tinoco, M., Fernández-Garcia, S., Chen, X., Hungria, A. B., Cauqui, M. A., & Calvino, J. J. (2018). A macroscopically relevant 3D-metrology approach for nanocatalysis research. *Particle & Particle Systems Characterization, 35*, 1700343.
Marín, A., Nickel, S., Puerto, J., & Velten, S. (2009). A flexible model and efficient solution strategies for discrete location problems. *Discrete Applied Mathematics, 157*(5), 1128–1145.
Marín, A., Ponce, D., & Puerto, J. (2020). A fresh view on the discrete ordered median problem based on partial monotonicity. *European Journal of Operational Research, 286*(3), 839–848.
Midgley, P. A., Ward, E. P. W., Hungría, A. B., & Thomas, J. M. (2007). Nanotomography in the chemical, biological and materials sciences. *Chemical Society Reviews, 36*(9), 1477–1494.
Murugan, K., Daniel, F., Shanmugaraja, T., Venkatesh, T., Siddarthraju, K., Dhivya Devi, R., & Supriya, M. (2020). A critical review on medical image processing techniques. *Journal of Critical Reviews, 7*(5), 576–580.
Nickel, S., & Puerto, J. (2005). *Location theory - a unified approach*. Springer Verlag.
Ogryczak, W., & Tamir, A. (2003). Minimizing the sum of the k largest functions in linear time. *Information Processing Letters, 85*(3), 117–122.
Olender, P., & Ogryczak, W. (2019). A revised variable neighborhood search for the discrete ordered median problem. *European Journal of Operational Research, 274*(2), 445–465.
Puerto, J. (2008). A new formulation of the capacitated discrete ordered median problems with {0, 1}-assignment. *Operations research proceedings 2007*. Springer Berlin Heidelberg.
Puerto, J., Ramos, A. B., & Rodríguez-Chía, A. M. (2013). A specialized branch & bound & cut for single-allocation ordered median hub location problems. *Discrete Applied Mathematics, 161*(16), 2624–2646.
Puerto, J., Ramos, A. B., Rodríguez-Chía, A. M., & Sánchez-Gil, M. C. (2016). Ordered median hub location problems with capacity constraints. *Transportation Research Part C: Emerging Technologies, 70*, 142–156.
Saglam, B., Salman, F. S., Sayin, S., & Turkay, M. (2006). A mixed-integer programming approach to the clustering problem with an application in customer segmentation. *European Journal of Operational Research, 173*(3), 866–879.
Staniewicz, L., & Midgley, P. A. (2015). Machine learning as a tool for classifying electron tomographic reconstructions. *Advanced Structural and Chemical Imaging, 1*, 9.
Tovey, R., Benning, M., Brune, C., Lagerwerf, M. J., Collins, S. M., Leary, R. K., … Schönlieb, C. B. (2019). Directional sinogram inpainting for limited angle tomography. *Inverse Problems, 35*(02), 024004.
Yamamoto, Y., Arai, S., Esaki, A., Ohyama, J., Satsuma, A., & Tanaka, N. (2014). Statistical distribution of single atoms and clusters of supported Au catalyst analyzed by global high-resolution HAADF-STEM observation with morphological image-processing operation. *Microscopy, 63*(3), 209–218.