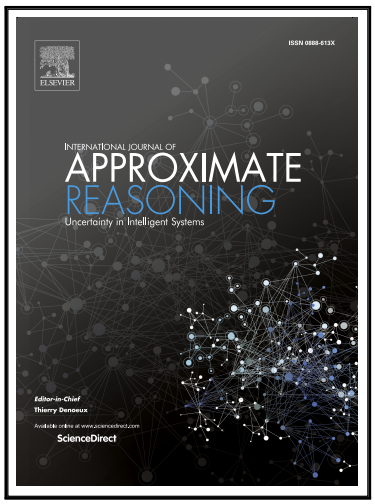
Journal Pre-proof



Fuzzy Clustering Methods for Compact Sets via Star-Shaped Formalization

Maria Brigida Ferraro, Gil González-Rodríguez,
Ana Belén Ramos-Guajardo

PII: S0888-613X(26)00084-8
DOI: https://doi.org/10.1016/j.ijar.2026.109708
Reference: IJA 109708

To appear in: *International Journal of Approximate Reasoning*

Received date: 14 January 2026
Revised date: 16 April 2026
Accepted date: 4 May 2026

Please cite this article as: Maria Brigida Ferraro, Gil González-Rodríguez, Ana Belén Ramos-Guajardo, Fuzzy Clustering Methods for Compact Sets via Star-Shaped Formalization, *International Journal of Approximate Reasoning* (2026), doi: https://doi.org/10.1016/j.ijar.2026.109708

Highlights

- A robust fuzzy k-means method is developed for clustering star-shaped sets.
- The approach is extended to compact sets integrating mixed geometric information.
- A tailored dissimilarity combines radial functions with numerical descriptors.
- A controlled toy example shows the benefits of mixed and robust formulations.
- The methods effectively cluster osteosarcoma cells and isolate atypical shapes.

Fuzzy Clustering Methods for Compact Sets via Star-Shaped Formalization

Maria Brigida Ferraro[a,1], Gil González-Rodríguez[b,c,1], Ana Belén Ramos-Guajardo[b,c,1,*]

[a] *Department of Statistical Sciences, Sapienza Università di Roma, Rome, Italy*
[b] *Department of Statistics, OR and MD, Universidad de Oviedo, Oviedo, Spain*
[c] *INDUROT, Universidad de Oviedo, Oviedo, Spain*

Abstract

Fuzzy clustering methods for compact sets are proposed by combining their star-shaped representations with numerical geometric descriptors. The classical fuzzy k-means algorithm, previously extended to the star-shaped setting through a distance that accounts for both the centers and their radial functions, is further enhanced here with a robust extension that incorporates a noise cluster to reduce the impact of atypical sets. The same ideas are then transferred to compact sets carrying mixed-type information by means of a dissimilarity measure that blends functional and real-valued components, giving rise to basic and robust mixed-data variants. The approach is illustrated with osteosarcoma cell morphology, where integrating shape-based and numerical features offers a more informative description of cellular variability and supports the identification of unusual cells.

Keywords: Star-shaped sets, fuzzy clustering method, distance-based approach

*Corresponding author
Email addresses: mariabrigida.ferraro@uniroma1.it (Maria Brigida Ferraro), gil@uniovi.es (Gil González-Rodríguez), ramosana@uniovi.es (Ana Belén Ramos-Guajardo)

[1] These authors contributed equally to this work.

1. Introduction

Since the foundational contributions of Bezdek (1981), Dunn (1974), Ruspini (1969), and Wee and Fu (1969), fuzzy clustering has attracted growing interest across a wide range of disciplines. A key development in this area is the well-known Fuzzy k-Means (FkM) clustering algorithm (Bezdek, 1981), which aims at identifying a limited number of homogeneous clusters by assigning objects to them based on degrees of membership ranging between 0 and 1. For a comprehensive overview of fuzzy clustering techniques, see, for example, Ferraro (2024) and Ruspini et al. (2019).

Fuzzy clustering has gained increasing attention in the scientific community, as shown by the growing number of studies on the topic (see, for example, the works of Doering et al. (2006), D'Urso and Giordani (2006), D'Urso and De Giovani (2015), Ferraro and Giordani (2013), Gustafson and Kessel Gustafson and Kessel (1979), Hwang et al. (2007), Masson and Denoeux (2008), and Ramos-Guajardo and Ferraro (2020)). These works explore both improvements to FkM and its practical uses in different areas.

In the context of classifying sets in $\mathbb{R}^p$, several clustering approaches have been proposed. For example, Joshi et *al.* introduced a density-based method for clustering polygons in Joshi et al. (2009), built upon the DBSCAN algorithm, which ensures spatial contiguity of the resulting clusters. Wang et al. (2014) developed a polygon clustering algorithm based on multilevel graph partitioning. In terms of spatial clustering, Chen et al. (2019) proposed an intuitionistic fuzzy similarity approach capable of quantifying similarities between polygons for clustering analysis. Additionally, in the field of image object segmentation, various shape-based fuzzy clustering techniques have been designed to explicitly segment geometrically regular objects by leveraging spatial information (see, for instance, Ameer et al. (2010), Shi et al. (2023), and references therein).

When multiple attribute types are collected, ignoring one or more of them during the clustering process can negatively impact the final results. In this context, the ability to incorporate mixed-type information becomes a valuable advantage for any proposed clustering methodology. In this regard, several recent fuzzy clustering methods designed for non-standard data types have been introduced in D'Urso and Massari (2019). Nevertheless, to the best of our knowledge, the problem of classifying general compact sets in $\mathbb{R}^p$ (including polygons in $\mathbb{R}^2$) while incorporating mixed-type information has not yet been addressed in the literature.

Compact sets can differ significantly in their structural complexity, ranging from basic geometric figures like polyhedra to highly irregular and elaborate shapes. Each of these sets can be described through a variety of geometric and morphological features, including volume, surface area, and centroid (or barycenter), as well as shape descriptors such as roundness, elongation, solidity, and eccentricity. These characteristics are typically represented using real-valued (numerical) variables.

Using star-shaped formalization to approach compact sets can be an advantage for extracting shape-related information. Star-shaped sets represent a versatile category of geometric objects, defined by their ability to retain their structure relative to a specified center point, a property that is effectively captured using the center-radial characterization (see Klain (1997) and González-Rodríguez et al. (2018) for further details). However, existing fuzzy clustering approaches do not address situations in which compact sets must be described simultaneously through a functional star-shaped representation and additional numerical geometric descriptors. Methods designed for star-shaped sets focus exclusively on their radial structure (Ferraro et al., 2023) and do not incorporate mixed-type information, while techniques for heterogeneous data (e.g., D'Urso and Massari (2019)) typically treat functional and numerical components separately. As a result, there is no unified fuzzy clustering framework capable of combining both sources of information or handling the irregularities often present in real geometric data, including those arising from atypical or highly irregular shapes.

This gap raises several methodological challenges. The first concerns the definition of a dissimilarity measure able to integrate the variability captured by the radial functions of star-shaped representations with the complementary information provided by numerical descriptors. A second challenge stems from the presence of atypical or highly irregular shapes, which require a mechanism that can mitigate their influence without distorting the resulting partition. These limitations motivate the development of the approach proposed in this work.

Ferraro et al. (2023) proposed a fuzzy clustering method for star-shaped sets based entirely on their center–radial representation and a tailored distance definition. Although effective for capturing functional shape variability, this approach does not incorporate additional geometric descriptors, nor does it address settings requiring a joint treatment of functional and numerical information or robustness to outliers. In compact-set applications, numerical features such as area, perimeter or circularity, as well as the loss incurred

when approximating the original shape by its star-shaped counterpart, provide complementary information that can be crucial for classification tasks. Such features have proven useful in related contexts, including polygon clustering and cellular morphology analysis, where geometric variability plays an essential role (see, for instance, Li et al. (2023)).

Building on these ideas, the methodology developed in this paper extends the framework introduced by Ferraro et al. (2023) to accommodate mixed-type information by introducing a fuzzy clustering approach for compact sets. The proposal relies on a dedicated dissimilarity measure that integrates the variations captured by the star-shaped representation together with a set of numerical geometric descriptors, allowing both components to contribute coherently to the clustering structure. In addition, a robust extension inspired by the noise-cluster mechanism of Davè (1991) is incorporated to isolate atypical elements and enhance the stability of the resulting partitions in the presence of irregular or extreme shapes.

The main contributions of the paper can be summarized as follows. First, building on existing extensions of fuzzy k-means to the space of star-shaped sets, where distances account for both the centers and the radial functions of the objects, we develop a robust formulation that incorporates a noise cluster, enabling atypical elements to be isolated without distorting the main clusters. Second, the general framework is transferred to compact sets endowed with mixed-type information by means of a dedicated dissimilarity measure that combines functional and numerical components, yielding both basic and robust mixed-data variants.

The rest of the paper is organized as follows. Section 2 introduces some preliminaries related to the space of star-shaped sets and defines a suitable distance between them. Section 3 presents the robust generalization of the fuzzy k-means method for star-shaped sets. Section 4 extends these methods by developing fuzzy clustering algorithms for compact sets, utilizing a generalized distance that accommodates differences in mixed-type information derived from the sets. To complement the methodological developments and to explicitly evaluate the effect of mixing star–shaped information with numerical descriptors, as well as the role of the noise cluster, a controlled toy example is presented in Section 5. In Section 6 the proposed methods are applied to the classification of various types of osteosarcoma cells. Finally, Section 7 outlines key conclusions and potential directions for future research.

2. Preliminaries

To properly address the fuzzy clustering proposals, it is essential to first establish some key concepts about the data being analyzed.

Consider $\mathbb{R}^p$, a p-dimensional Euclidean space equipped with the standard norm $\|\cdot\|$ and the inner product $\langle\cdot,\cdot\rangle$. Formally, the space of general star-shaped sets in $\mathbb{R}^p$ is denoted as $\mathbb{X}(\mathbb{R}^p) = (\mathbb{R}^p \times \mathbb{X}_0(\mathbb{R}^p))$, where $\mathbb{X}_0(\mathbb{R}^p)$ is the space of star-shaped sets in $\mathbb{R}^p$ with respect to 0, defined s. t.

$$\mathbb{X}_0(\mathbb{R}^p) = \left\{A \subset \mathbb{R}^p \,|\, \gamma a \in A \text{ for all } a \in A \text{ and all } \gamma \in [0,1]\right\}.$$

A star-shaped set $A \in \mathbb{X}(\mathbb{R}^p)$ is characterized by two elements: its center and the radial function of the associated set centered at 0, denoted by the pair (c_A, ρ_A). On one hand, the center $c_A \in \mathbb{R}^p$ is the location point such that for all $a \in A$, $\lambda c_A + (1-\lambda)a \in A$ for all $\lambda \in [0,1]$. For a discussion on selecting a center compatible with measure-preserving arithmetic, see González-Rodríguez (2026). On the other hand, the radial function of the centered version of A, $A^c = A - c_A$, is defined as $\rho_A : \mathbb{S}^{p-1} \to \mathbb{R}^+$ such that $\rho_A(v) = \sup\{\lambda \geq 0 : \lambda v \in A^c\}$, which relates to the imprecision of the set (see Schneider (1993)).

Then, by considering the so-called *center-radial characterization* defined above, A can be expressed as

$$A = \{c_A + \gamma\rho_A(v) \mid \gamma \in [0,1], v \in \mathbb{S}^{p-1}\}.$$

The center of A can be any point within the kernel of A, where the kernel, denoted as $ker(A)$, consists of those points in A that satisfy

$$\lambda x + (1-\lambda)a \subset A \;\forall\lambda \in [0,1] \text{ and } \forall a \in A\;. \tag{1}$$

The centroid (or barycenter) is typically chosen as the representative center. Moreover, from an operational point of view, it is convenient to consider radial functions in the space $\mathcal{L}^2(\mathbb{S}^{p-1})$, and thus the following class of star-shaped sets will be considered:

$$\mathbb{X}^*(\mathbb{R}^p) = \{A \in \mathbb{X}(\mathbb{R}^p) \mid \rho_A \in \mathcal{L}^2(\mathbb{S}^{p-1})\}.$$

The distance between two star-shaped sets $A, B \in \mathbb{X}^*(\mathbb{R}^p)$ is defined using an $\mathcal{L}_2$-type distance metric introduced in González-Rodríguez et al. (2018) as follows:

$$d_\tau(A,B) = \sqrt{\tau\|c_A - c_B\|^2 + (1-\tau)\|\rho_A - \rho_B\|}^2_{\mathcal{L}^2}, \tag{2}$$

where $\tau \in (0,1)$ is a weighting parameter that balances the importance of the spatial location (center) versus the directional imprecision (radial function). Besides, $\|\cdot\|_{\mathcal{L}^2}$ represents the usual L_2-type norm in $\mathcal{L}^2(\mathbb{S}^{p-1})$.

3. Robust fuzzy clustering method for star-shaped sets

This section presents a robust variant of the FkM generalization tailored for star-shaped sets, developed in Ferraro et al. (2023). The proposal is designed to enhance clustering performance and account for data irregularities and outliers.

Suppose there are n star-shaped sets, denoted by $\mathbf{x}_i = (c_{\mathbf{x}_i}, \rho_{\mathbf{x}_i})$, where each $\mathbf{x}_i$ is characterized by a center $c_{\mathbf{x}_i} \in \mathbb{R}^p$ and a radial function $\rho_{\mathbf{x}_i} \in \mathcal{L}^2(\mathbb{S}^{p-1})$. The objective is to partition these n sets into k clusters such that sets with similar features are grouped into the same cluster to maximize homogeneity, while those with distinct features are assigned to different clusters to ensure adequate separation. To achieve this, it is crucial to adopt an appropriate distance or dissimilarity measure specifically designed for the object space of star-shaped sets. This measure must account for both the location of the centers $c_{\mathbf{x}_i}$ and the directional imprecision captured by the radial functions $\rho_{\mathbf{x}_i}$.

The fuzzy k-means algorithm, widely used for clustering, is extended here to accommodate star-shaped sets by employing several steps. In a first stage, k cluster prototypes are randomly initialized, each represented by a tuple, a cluster center in $\mathbb{R}^p$ and a radial function in $\mathcal{L}^2(\mathbb{S}^{p-1})$. A suitable dissimilarity measure combining location and radial imprecision is then used, such as the $\mathcal{L}_2$-type distance:

$$d_\tau(\mathbf{x}_i, \mathbf{h}_g) = \sqrt{\tau\|c_{\mathbf{x}_i} - c_{\mathbf{h}_g}\|^2 + (1-\tau)\|\rho_{\mathbf{x}_i} - \rho_{\mathbf{h}_g}\|^2_{\mathcal{L}^2}},$$

where $\mathbf{h}_g = (c_{\mathbf{h}_g}, \rho_{\mathbf{h}_g})$ represents the g-th cluster prototype, and τ balances the weight of center-based and radial function-based dissimilarity.

The selection of a dissimilarity measure is critical for effectively capturing both the location and shape features of the star-shaped sets. A generalization of the FkM method to the framework of star-shaped sets is addressed in Ferraro et al. (2023). This method will now be extended to handle noisy data during the clustering process.

The proposal, inspired by the work of Davè (1991), integrates a noise cluster mechanism, which significantly improves its robustness against variability and imprecision within the data. By isolating data points with high uncertainty or inconsistency into a specific noise cluster, the algorithm ensures that the primary clusters remain well-defined and meaningful. This separation not only preserves the integrity of the main k clusters but also reduces the influence of noise on the overall clustering process. The noise cluster does not represent a proper cluster, as it lacks homogeneity (compactness); moreover, the dissimilarity δ^2 of each unit from the noise prototype $\mathbf{h}_{k+1}$ is fixed in advance. When applied to star-shaped sets, the robust version, which will be called RFkM-SSS, considers a customized distance measure that accounts for their unique structural properties, which ensures that the clustering process effectively captures the characteristics of these complex data sets.

Given n star-shaped sets $\mathbf{x}_i = (c_{\mathbf{x}_i}, \rho_{\mathbf{x}_i})$, we define the RF$k$M-SSS optimization problem with a noise cluster as:

$$\begin{aligned}
\min_{\mathbf{U},\mathbf{H}} J_{\text{RF}k\text{M-SSS}} &= \min_{\mathbf{U},\mathbf{H}} \left(\sum_{i=1}^{n} \sum_{g=1}^{k} u_{ig}^m d_\tau^2 \left((c_{\mathbf{x}_i}, \rho_{\mathbf{x}_i}), (c_{\mathbf{h}_g}, \rho_{\mathbf{h}_g}) \right) + \sum_{i=1}^{n} u_{i,\text{noise}}^m \delta^2 \right) \\
&= \min_{\mathbf{U},\mathbf{H}} \left(\sum_{i=1}^{n} \sum_{g=1}^{k} u_{ig}^m \left(\tau \left\| c_{\mathbf{x}_i} - c_{\mathbf{h}_g} \right\|^2 + (1-\tau) \| \rho_{\mathbf{x}_i} - \rho_{\mathbf{h}_g} \|_{\mathcal{L}^2}^2 \right) \right. \\
&\qquad \left. + \sum_{i=1}^{n} u_{i,\text{noise}}^m \delta^2 \right), \\
&\text{s.t.} \quad u_{ig} \in [0,1], \quad i = 1, \cdots, n, \quad g = 1, \cdots, (k+1), \\
&\qquad \sum_{g=1}^{k+1} u_{ig} = 1, \quad i = 1, \cdots, n
\end{aligned} \tag{3}$$

where $\mathbf{U} = [u_{ig}]$ is the membership degree matrix of order $(n \times (k+1))$, $u_{i,\text{noise}} = u_{i,k+1}$ is the membership value of $\mathbf{x}_i$ to the noise cluster, for $i = 1, \cdots, n$. These conditions imply that outliers can have low membership degrees to the proper k clusters, and high degree to the noise cluster, thereby reducing their influence on the proper ones. $\mathbf{H} = [\mathbf{h}_1, \cdots \mathbf{h}_g, \cdots, \mathbf{h}_k]$, with $\mathbf{h}_g = (c_{\mathbf{h}_g}, \rho_{\mathbf{h}_g})$, is the matrix of k prototypes, and m is the parameter of fuzziness, which is typically greater than 1 (commonly $m = 2$). As m approaches 1, the membership degrees become closer to 0 and 1, causing FkM to converge to the standard k-means algorithm. Conversely, as m increases and moves away from 1, the membership degrees diverge from 0 and 1, resulting in a more fuzzy partition.

In order to obtain the optimal fuzzy membership degree matrix $\mathbf{U}$ of the RFkM-SSS method, we consider the following Lagrangian function defined

as

$$L(\mathbf{U},\mathbf{H},\lambda)=\sum_{i=1}^{n}\sum_{g=1}^{k+1}u_{ig}^{m}d_{\tau}^{2}\left((c_{\mathbf{x}_i},\rho_{\mathbf{x}_i}),(c_{\mathbf{h}_g},\rho_{\mathbf{h}_g})\right)\quad+\sum_{i=1}^{n}\lambda_i\left(\sum_{g=1}^{k+1}u_{ig}-1\right) \tag{4}$$

To find the optimal value of $\mathbf{U}$ we compute the partial derivatives of (4) with respect to u_{ig} and λ_i and we set them equal to 0:

$$\frac{\vartheta L}{\vartheta u_{ig}}=0\Leftrightarrow mu_{ig}^{m-1}d_{\tau}^{2}\left((c_{\mathbf{x}_i},\rho_{\mathbf{x}_i}),(c_{\mathbf{h}_g},\rho_{\mathbf{h}_g})\right)-\lambda_i=0, \tag{5}$$

$$\frac{\vartheta L}{\vartheta \lambda_i}=0\Leftrightarrow\sum_{g=1}^{k+1}u_{ig}-1=0. \tag{6}$$

Then, by carrying out the usual calculations we get

$$u_{ig}=\frac{1}{\sum\limits_{g'=1}^{k}\left(\frac{d_{\tau}^{2}\left((c_{\mathbf{x}_i},\rho_{\mathbf{x}_i}),(c_{\mathbf{h}_g},\rho_{\mathbf{h}_g})\right)}{d_{\tau}^{2}\left((c_{\mathbf{x}_i},\rho_{\mathbf{x}_i}),(c_{\mathbf{h}_{g'}},\rho_{\mathbf{h}_{g'}})\right)}\right)^{\frac{1}{m-1}}+\left(\frac{d_{\tau}^{2}\left((c_{\mathbf{x}_i},\rho_{\mathbf{x}_i}),(c_{\mathbf{h}_g},\rho_{\mathbf{h}_g})\right)}{\delta^{2}}\right)^{\frac{1}{m-1}}}, \tag{7}$$

for $i=1,\ldots,n$ and $g=1,\ldots,k$, and

$$u_{i,\text{noise}}=u_{i,k+1}=\left(1-\sum_{g=1}^{k}u_{ig}\right). \tag{8}$$

On the other hand, by considering the partial derivatives of (4) with respect to $c_{\mathbf{h}_g}$ and $\rho_{\mathbf{h}_g}$, for $g=1,\cdots,k$, and setting them equal to 0, the centroid components are given by

$$c_{\mathbf{h}_g}=\frac{\sum\limits_{i=1}^{n}u_{ig}^{m}c_{\mathbf{x}_i}}{\sum\limits_{i=1}^{n}u_{ig}^{m}}\quad\text{and}\quad\rho_{\mathbf{h}_g}(v)=\frac{\sum\limits_{i=1}^{n}u_{ig}^{m}\rho_{\mathbf{x}_i}(v)}{\sum\limits_{i=1}^{n}u_{ig}^{m}},\qquad g=1,\cdots,k, \tag{9}$$

where $v\in\mathbb{S}^{p-1}$ denotes a direction on the unit sphere. As the noise cluster does not constitute a proper cluster, its prototype is merely an abstract

construct, and computing it as one would for the other k prototypes is not meaningful.

Special attention must be paid to the selection of the noise dissimilarity threshold δ^2. Following the ideas of Davè (1991), δ^2 should be a proportion of the mean of the squared distances between points and cluster prototypes. Thus, it is set to

$$\delta^2 = \lambda \frac{1}{n \cdot k} \left(\sum_{i=1}^{n} \sum_{g=1}^{k} d_\tau^2 \left((c_{\mathbf{x}_i}, \rho_{\mathbf{x}_i}), (c_{\mathbf{h}_g}, \rho_{\mathbf{h}_g}) \right) \right), \tag{10}$$

with λ being a user-defined parameter determining the proportion: the smaller the value of λ, the higher the proportion of points that are considered as outliers.

The steps of the RFkM-SSS algorithm are described in Algorithm 1.

Algorithm 1 RFkM-SSS

Step 0 (Initialization) Generate randomly a feasible membership degree matrix $\mathbf{U^{(0)}}$ of order $(n \times (k+1))$ subject to (3).
Step 1 Compute the centroid matrix $\mathbf{H^{(t)}}$ according to (9) using $\mathbf{U^{(t-1)}}$.
Step 2 Update the fuzzy membership degree matrix $\mathbf{U^{(t)}}$ according to (7) and (8), keeping fixed $\mathbf{H^{(t)}}$.
Step 3 Check convergence. If the convergence condition is not satisfied, go to **Step 1**.

Note that the method introduced in Ferraro et al. (2023), FkM-SSS, can be obtained as a particular case of RFkM-SSS, when the noise component is not considered, that is, when the second term in the optimization function (3) is equal to 0. This implies that we have only k proper clusters and the updating equation of the membership degrees will be

$$u_{ig} = \frac{1}{\sum_{g'=1}^{k} \left(\frac{d_\tau^2\left((c_{\mathbf{x}_i}, \rho_{\mathbf{x}_i}), (c_{\mathbf{h}_g}, \rho_{\mathbf{h}_g})\right)}{d_\tau^2\left((c_{\mathbf{x}_i}, \rho_{\mathbf{x}_i}), (c_{\mathbf{h}_{g'}}, \rho_{\mathbf{h}_{g'}})\right)} \right)^{\frac{1}{m-1}}}, \quad i = 1, \ldots, n, \quad g = 1, \ldots, k. \tag{11}$$

The updating equation for the k prototypes remains unchanged.

4. Fuzzy clustering methods for compact sets

Consider now the space of compact subsets of $\mathbb{R}^p$, denoted by $\mathcal{K}(\mathbb{R}^p)$. These compact sets can vary in complexity, ranging from simple shapes like polyhedrons to irregular and intricate forms. Each compact set possesses geometric properties such as volume, surface area, centroid (barycenter), as well as other characteristics like roundness, elongation, solidity or eccentricity, among others. All these properties can be modeled using numerical/real variables.

On the other hand, a compact set $A \in \mathcal{K}(\mathbb{R}^p)$ may have a shape that is not necessarily star-shaped. In such cases, it is possible to approximate its original shape using the shape of the largest star-shaped set visible from the centroid of A and contained within A, denoted as A^{SS}, along with a measure of the volume lost when using such approximation, which is represented by a real number, and other real geometric measures as the ones provided above.

Therefore, it would be beneficial to extend the methods proposed in Section 3 to the case of compact sets using mixed data. This approach would not only consider a star-shaped set, but also incorporate some real variables that would help us to describe other important geometric features of the sets under consideration.

Consider n compact sets, $\{\mathbf{x}_i\}_{i=1}^n$, to be classified in k clusters. Each compact set $\mathbf{x}_i$ is characterized by the largest star-shaped set visible from the centroid and contained within the original set, denoted as $\mathbf{x}_i^{SS} = \left(c_{\mathbf{x}_i^{SS}}, \rho_{\mathbf{x}_i^{SS}}\right)$, along with a vector of q real-valued variables, $\mathbf{x}_i^q$, that describe certain geometric properties of the compact set or other numerical attributes of the represented entity.

In this case, a suitable dissimilarity measure between compact sets should combine the distance between the star-shaped part of the sets and the distance between the real-valued part. Thus, if $\mathbf{x}_i$ is a compact set and $\mathbf{h}_g$ represents the g-th cluster prototype, with star-shaped part $\mathbf{h}_g^{SS}$ and corresponding real-valued part $\mathbf{h}_g^q$, the distance between both elements is defined by

$$d_{mix}\left(\mathbf{x}_i, \mathbf{h}_g\right) = \sqrt{d_\tau^2(\mathbf{x}_i^{SS}, \mathbf{h}_g^{SS}) + d^2(\mathbf{x}_i^q, \mathbf{h}_g^q)}, \tag{12}$$

where d_τ is the distance metric for star-shaped sets and d is the Euclidean distance.

In this section two fuzzy clustering methods for compact sets are introduced: FkM-Mix and its robust extension, RFkM-Mix. The first one is the FkM

method for compact sets of $\mathbb{R}^p$ that incorporates mixed-type information, while the second one is its variant with the additional noise cluster.

The FkM-Mix minimization problem can be formulated as follows:

$$\begin{aligned}\min_{\mathbf{U},\mathbf{H}} J_{\text{F}k\text{M-Mix}} &= \min_{\mathbf{U},\mathbf{H}} \left(\sum_{i=1}^{n} \sum_{g=1}^{k} u_{ig}^m d_{mix}^2 \left(\mathbf{x}_i, \mathbf{h}_g\right) \right) \\ &= \min_{\mathbf{U},\mathbf{H}} \left(\sum_{i=1}^{n} \sum_{g=1}^{k} u_{ig}^m \left(d_\tau^2(\mathbf{x}_i^{SS}, \mathbf{h}_g^{SS}) + d^2(\mathbf{x}_i^q, \mathbf{h}_g^q) \right) \right) \\ \text{s.t.} \quad & u_{ig} \in [0,1], \quad i = 1, \cdots, n, \quad g = 1, \cdots, k, \\ & \sum_{g=1}^{k} u_{ig} = 1, \quad i = 1, \cdots, n \end{aligned} \tag{13}$$

where $\mathbf{x}_i = (\mathbf{x}_i^{SS}, \mathbf{x}_i^q)$, with $\mathbf{x}_i^{SS} \in \mathbb{X}^*(\mathbb{R}^p)$ and $\mathbf{x}_i^q \in \mathbb{R}^q$, for $i = 1, \cdots, n$, is a compact set from the sample, while $\mathbf{H} = [\mathbf{h}_1, \cdots \mathbf{h}_g, \cdots, \mathbf{h}_k]$, with $\mathbf{h}_g = (\mathbf{h}_g^{SS}, \mathbf{h}_g^q)$ ($\mathbf{h}_g^{SS} \in \mathbb{X}^*(\mathbb{R}^p)$ and $\mathbf{h}_i^q \in \mathbb{R}^q$), is the matrix of k prototypes.

By introducing the noise cluster, we obtain the RFkM-Mix, whose optimization function is

$$J_{\text{RF}k\text{M-Mix}} = \sum_{i=1}^{n} \sum_{g=1}^{k} u_{ig}^m d_{mix}^2 \left(\mathbf{x}_i, \mathbf{h}_g\right) + \sum_{i=1}^{n} u_{i,\text{noise}}^m \delta^2 \tag{14}$$

where δ^2 is defined, as in Section 3.2. as a proportion of the mean squared dissimilarity based on d_{mix}.

In this case, the minimization of $J_{\text{RF}k\text{M-Mix}}$ with respect to $\mathbf{U}$ subject to the constraints provided in (13) leads to

$$u_{ig} = \frac{1}{\sum_{g'=1}^{k} \left(\frac{d_{mix}^2(\mathbf{x}_i, \mathbf{h}_g)}{d_{mix}^2(\mathbf{x}_i, \mathbf{h}'_g)} \right)^{\frac{1}{m-1}} + \left(\frac{d_{mix}^2(\mathbf{x}_i, \mathbf{h}_g)}{\delta^2} \right)^{\frac{1}{m-1}}}, \quad i = 1, \ldots, n, \quad g = 1, \ldots, k, \tag{15}$$

$$u_{i,\text{noise}} = u_{i,k+1} = \left(1 - \sum_{g=1}^{k} u_{ig} \right), \quad i = 1, \ldots, n, \tag{16}$$

For $g = 1, \ldots, k$ (k proper clusters), the prototypes are updated by combining the star-shaped ($\mathbf{h}_g^{SS} = (c_{\mathbf{h}_g^{SS}},\ \rho_{\mathbf{h}_g^{SS}}(v))$) and numerical components

$(\mathbf{h}_g^q)$:

$$c_{\mathbf{h}_g^{SS}} = \frac{\sum_{i=1}^{n} u_{ig}^m c_{\mathbf{x}_i^{SS}}}{\sum_{i=1}^{n} u_{ig}^m}, \quad \rho_{\mathbf{h}_g^{SS}}(v) = \frac{\sum_{i=1}^{n} u_{ig}^m \rho_{\mathbf{x}_i^{SS}}(v)}{\sum_{i=1}^{n} u_{ig}^m}, \quad \text{and} \quad \mathbf{h}_g^q = \frac{\sum_{i=1}^{n} u_{ig}^m \mathbf{x}_i^q}{\sum_{i=1}^{n} u_{ig}^m}. \quad (17)$$

The steps describing the RFkM-Mix approach can be found in Algorithm 2.

Algorithm 2 RFkM-Mix

Step 0 (Initialization) Generate randomly a feasible membership degree matrix $\mathbf{U^{(0)}}$ of order $(n \times (k+1))$ subject to (14).
Step 1 Compute the prototype components $\mathbf{h_g^{SS(t)}}$ and $\mathbf{h_g^{q(t)}}$, for $g = 1, \cdots, k$ according to (17) using $\mathbf{U^{(t-1)}}$.
Step 2 Update the fuzzy membership degree matrix $\mathbf{U^{(t)}}$ according to (15) and (16), keeping fixed $\mathbf{h_g^{SS(t)}}$ and $\mathbf{h_g^{q(t)}}$.
Step 3 Check convergence. If the convergence condition is not satisfied, go to **Step 1**.

5. Comparative analysis through a toy example

To illustrate the behaviour of the proposed fuzzy clustering methods and to evaluate the contribution of the mixed-type geometric information, we consider a controlled toy example consisting of planar compact sets constructed through a prescribed center–radial representation. The main dataset is composed of 100 compact sets arranged in two groups, G1 and G2, each containing 50 elements. Although G1 and G2 share the same concavity locations (at fixed angular positions across all objects), several key characteristics distinguish the two families. On the one hand, the concavities in G2 are substantially deeper than those in G1, resulting in a more pronounced inward deformation of the boundary; on the other hand, G2 exhibits systematically larger high-frequency roughness, capturing more irregular fluctuations along the boundary. Taken together, these features yield radial functions in G2 that are both more oscillatory and more sharply concave compared to G1, thus ensuring identifiable differences between the two groups. Representative examples are shown in Figure 1.

Each compact set is represented through three descriptions. The first is its convex hull (CH), a purely geometric simplification. Figure 2 displays the

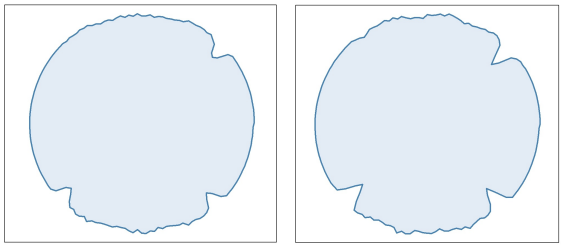


Figure 1: Representative compact sets from groups G1 (left) and G2 (right).

corresponding convex hulls, highlighting the geometric information that is lost when convexification is applied.

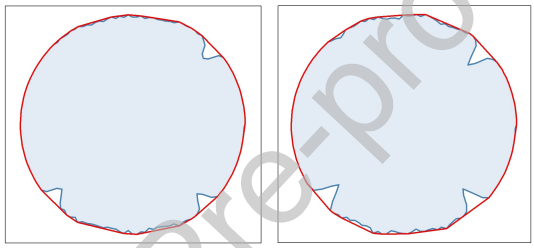


Figure 2: Convex hulls of the examples shown in Figure 1.

The second representation is the star-shaped formalization $\mathbf{X_1}$, which fully preserves concavities and oscillatory behaviour encoded in the radial functions. The third is a mixed representation $(\mathbf{X_1}, \mathbf{X_2})$ in which $\mathbf{X_2}$ includes two real-valued geometric descriptors: the perimeter and solidity (ratio between the area of the shape and the area of its convex hull) of the original polygon. Specifically, lower values of solidity correspond to shapes displaying stronger concavities or irregular boundaries.

To assess the robustness of the proposed methods, four additional compact sets are included in a group of outliers denoted by O. Unlike G1 and G2, whose concavities always occur at the same angular locations, these outliers exhibit concavities or local oscillations at different angular positions. The representation of the four outliers is displayed in Figure 3.

On the basis of these representations, the proposed fuzzy clustering methods are applied. First, the fuzzy k–means method is applied to the convex–hull representation (FkM–CH) and to the star–shaped sets (FkM–SSS). To evaluate the contribution of numerical descriptors, mixed fuzzy k–means models are examined, combining the star–shaped representation with so-

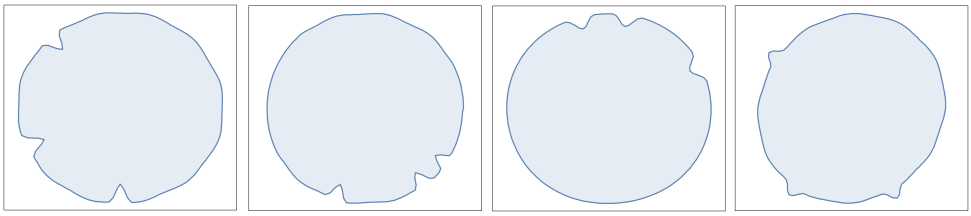


Figure 3: Outliers incorporated into the toy dataset.

lidity, the perimeter, or with both descriptors (FkM–Mix(S), FkM–Mix(P), and FkM–Mix(S+P)). Finally, the corresponding robust extensions are also included, namely RFkM–CH, RFkM–SSS, RFkM–Mix(S), RFkM–Mix(P), and RFkM–Mix(S+P), which incorporate a noise cluster to reduce the influence of atypical shapes.

To ensure comparability across models, all distances are computed over the same angular discretization. For internal validation, we adapt the fuzzy silhouette index of Campello and Hruschka (2006) to compact sets, including the mixed-type setting introduced in this paper. The modified fuzzy silhouette (MFS) index is based on d_τ for star-shaped sets and on d_{mix} when mixed-type data are used, providing a more appropriate assessment of compact sets' dissimilarities. The resulting partitions for the toy dataset are summarized in Table 1.

Table 1: Partitions obtained for the toy dataset under all non–robust and robust fuzzy k–means variants.

Method	**Cluster 1**			**Cluster 2**			**Noise**		
	G1	**G2**	**O**	**G1**	**G2**	**O**	**G1**	**G2**	**O**
FkM–CH	21	31	2	29	19	2	0	0	0
FkM–SSS	37	9	5	13	41	0	0	0	0
FkM–Mix(S)	47	0	4	3	50	0	0	0	0
FkM–Mix(P)	0	50	0	50	0	4	0	0	0
FkM–Mix(S+P)	0	50	0	50	0	4	0	0	0
RFkM–CH	20	20	0	19	19	0	11	11	4
RFkM–SSS	35	2	0	5	29	0	10	19	4
RFkM–Mix(S)	48	0	0	2	50	0	0	0	4
RFkM–Mix(P)	50	0	0	0	0	50	0	0	4
RFkM–Mix(S+P)	50	0	0	0	0	50	0	0	4

Table 1 summarizes the partitions obtained for the 104 compact sets. As expected, FkM–CH struggles to recover the two groups due to the loss of concavity information, whereas FkM–SSS improves the separation but still exhibits some overlap. In contrast, the mixed-type methods achieve nearly perfect group recovery: both perimeter and solidity incorporate complementary geometric variation that enhances discrimination. In all robust methods based on mixed information, the four outliers are consistently assigned to the noise cluster, confirming that the proposed noise mechanism effectively isolates atypical shapes.

These observations are fully consistent with the corresponding MFS indices. Among the non–robust methods, FkM–CH and FkM–SSS yield relatively low silhouette values (0.197 and 0.287, respectively), reflecting the substantial mixing observed in their partitions. The incorporation of numerical descriptors leads to a marked improvement: FkM–Mix(S) attains an MFS of 0.592, FkM–Mix(P) reaches 0.628, and the combined model FkM–Mix(S+P) achieves the highest value (0.708), corroborating its excellent cluster separation. The robust variants exhibit MFS values that closely mirror those of their non–robust counterparts (RFkM–CH = 0.072, RFkM–SSS = 0.265, RFkM–Mix(S) = 0.592, RFkM–Mix(P) = 0.629, RFkM–Mix(S+P) = 0.751), indicating that the introduction of the noise cluster enhances stability while preserving the overall clustering structure. Taken together, these results illustrate the benefits of combining star–shaped information with complementary geometric descriptors.

6. Application on the classification of osteosarcoma cells

The methods proposed in this manuscript are suitable methods for clustering geometric and functional data that contain mixed-type information. They could also be applied in different fields such as image analysis, shape modeling, and morphological characterization. In particular, the manuscript addresses a classification problem involving different types of cancer cells.

The dataset under consideration, known as AXCFP2019, contains images of murine osteosarcoma (bone cancer) cells and was originally introduced in Alizadeh et al. (2019) to investigate whether variations in cell morphology are associated with metastatic potential in both murine and human osteosarcomas. Among the cell lines included in this dataset are two distinct types: DLM8 and DUNN. The DLM8 cell line is derived from the DUNN cell line with selection for metastatic capability (see Asai et al. (1998)). Therefore,

DLM8 is closely related to DUNN except for the level of invasiveness: the DLM8 cell line exhibits higher aggressiveness, whereas the DUNN cell line represents a less aggressive cancer phenotype.

The dataset also includes cells exposed to treatments with specific single drugs that target the cellular cytoskeleton, an essential component responsible for maintaining cell shape, facilitating division, and enabling intracellular transport (see, for instance, Davis and Gruebele (2020)). These treatments allow for the investigation of the effects of cytoskeletal disruption on cellular behavior and properties. Both DLM8 and DUNN cell lines were either left untreated (control) or treated with cytochalasin D (*cytD*) or jasplakinolide (*jasp*), two cancer drugs that differently affect cellular morphology. Specifically, *cytD* induces actin depolymerization, whereas *jasp* enhances it (see Kováčiková et al. (2018) for further details).

Each cell in the dataset originates from a raw image containing multiple cells. These raw images were processed using a thresholding technique to produce binarized images, which differentiate cell regions from the background based on intensity levels (see Alizadeh et al. (2019) for further details). After the binarization process, contouring algorithms were applied to identify and isolate individual cells. This step involved the extraction of cell boundaries, which were subsequently represented as a discrete curve. A example of a binarized actin cell (actin being the most dynamic component of the cytoskeleton) is shown in Figure 4.

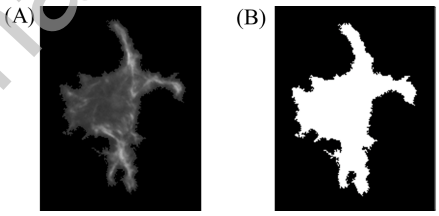


Figure 4: (A) Gray scale images using labeled actin. Intensity of each pixel is recorded and represented as a grayscale intensity plot. (B) Binary image of actin. Source: `https://doi.org/10.1371/journal.pone.0217346.g002`

The dataset was also utilized in Li et al. (2023) to explore how an elastic metric can be applied to study and compare cellular morphology based on the contours formed by cells on planar surfaces. The preprocessing steps, data analysis and code used in that study are thoroughly documented at `https://`

geomstats.github.io/notebooks/11_real_world_applications__cell_shapes_analysis.html. After the data preprocessing, the boundary of each cell is represented as a counter-clockwise ordered sequence of 2D coordinates, allowing for precise geometric characterization of cell shapes. The resulting 2D dataset is publicly accessible at https://github.com/wxli0/dyn/tree/main%4092c7a58/dyn/datasets/osteosarcoma/aligned/full.

The dataset comprises a total of 650 cells, systematically organized into treatment groups. Each cell is labeled according to two parameters: its cell line of origin, classified as either DLM8 or DUNN, and its treatment condition, categorized as control (no treatment), cytochalasin D (*cytD*), or jasplakinolide (*jasp*). During an initial visual inspection, six instances were identified as double cells. To avoid potential bias in the analysis, these were excluded from the study, resulting in a final dataset of 644 cells. The distribution of cells across cell lines and treatment conditions is summarized in Table 2.

Table 2: Number of cells belonging to each origin cell line and treatment condition

Treatment	**DLM8**	**DUNN**
control	114	202
cytD	80	93
jasp	60	95
Total	254	390

This classification provides a structured framework for analyzing the cellular responses to different treatments, taking into account different levels of cancer aggressiveness. In addition, a visualization of the mean of each group is provided in Figure 5.

As illustrated in Figure 5, the mean shapes of the control groups for both cell lines exhibit a regular and consistent morphology. In contrast, distinct variations are observed in the mean shapes of the treated groups. Specifically, the mean shape associated with *cytD* treatment displays the highest degree of irregularity across both cell lines. Conversely, the mean shape resulting from *jasp* treatment varies between cell lines: for the DLM8 cell line, it appears more elongated, whereas for the DUNN cell line, it presents a more irregular

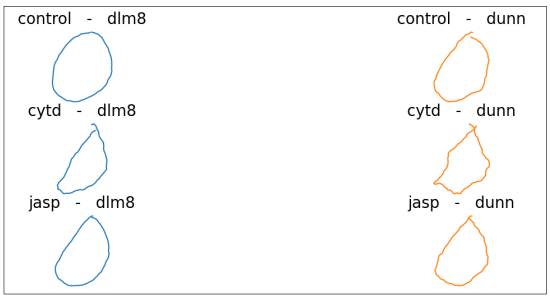


Figure 5: Mean shape of the cells belonging to each group according to their cell line of origin and their treatment condition. Source: `https://geomstats.github.io/notebooks/11_real_world_applications__cell_shapes_analysis.html`

shape. These observations highlight the differential impact of the treatments on cellular morphology, depending on both the treatment type and the cell line.

It should be noticed that the 2D representation of each cell can be viewed as a compact subset of $\mathbb{R}^2$. The objective is to apply the methods described in Section 4 to partition the cells of each cell line separately into k clusters. The goal is to group cells with similar features into the same cluster to maximize homogeneity while assigning those with distinct features to different clusters. To determine the optimal value of k, we introduce and use a modified version of the fuzzy silhouette index proposed by Campello and Hruschka (2006), adapted to the case of compact sets. Additionally, before applying the clustering methods, it is necessary to consider mixed-type information from each cell representation, combining its star-shaped formalization with other real geometric measurements of the cell.

6.1. Data loading and preprocessing

The aforementioned dataset was preprocessed using the open-source software R (R Core Team, 2020) in order to apply the FkM-Mix and RFkM-Mix clustering algorithms. Data preprocessing was performed using the R libraries `pliman`, `pracma`, and `starshapedsets`.

In the original dataset, the 2D contour of a cell is represented in a discretized form by a matrix with 1999 rows, where each row corresponds to a point in the counter-clockwise ordered sequence of 2D coordinates defining the cell's boundary. Consequently, the contour can be interpreted as a polygon in $\mathbb{R}^2$ with 1999 vertices.

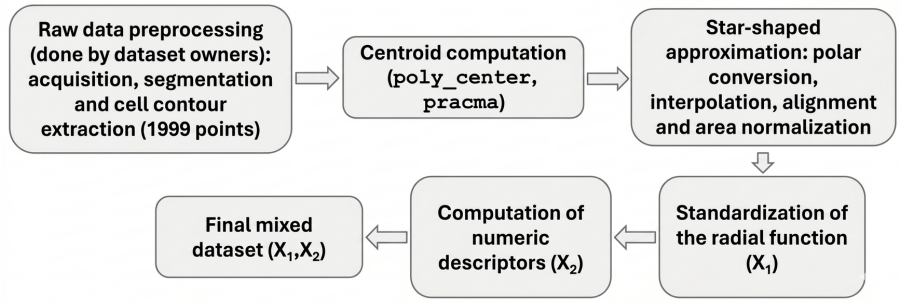


Figure 6: Overview of the preprocessing pipeline.

Figure 6 provides an overview of the preprocessing pipeline. The first block (raw data preprocessing) was performed by the dataset owners, whereas the other blocks correspond to the procedures implemented in this work.

Since the original contour of a cell is not always star-shaped, the first step involves approximating this shape by taking into account the shape of the largest star-shaped set visible from the centroid and contained within the original cell. Additionally, several numerical variables are also computed to quantify the loss in transitioning from the original contour to its star-shaped approximation.

To obtain the largest star-shaped set visible from the centroid and contained within the original contour of a cell, the data is first centered by computing the centroid of each polygon using the `poly_center` function from the `pracma` library, which in this case coincides with the barycenter. It is important to note that the absolute location is not relevant to the analysis, as the positions of the sets provided in the study are arbitrary. Next, a custom function named `polygon2D.polarinterpolation()`, built upon several functions from the `starshapedsets` library, is applied in order to convert the contour points into polar coordinates and interpolate them across a set of evenly spaced angles. The goal is to derive the largest star-shaped set visible from the centroid and contained within the original shape as output. The resulting star-shaped set is aligned such that its longest radius lies along the positive X-axis, and its area is normalized to 1 using the function `polyarea` included in the library `pracma`. This normalization ensures that all resulting star-shaped representations of the cells have the same area, so that the radial function ρ of each cell reflects only its shape, since the area will be treated separately as an additional attribute. Finally, the resulting data is

standardized by dividing by the corresponding standard deviation, in order to assign them a weight comparable to that of the other numerical variables considered for the optimization problem, yielding a collection of star-shaped sets denoted as $\mathbf{X}_1$.

To illustrate the morphological variability within each group, Figure 7 shows the mean radial function and its pointwise standard deviation for each cell line–treatment combination. The curves are obtained after aligning the star-shaped representations by their maximal radius and enforcing a consistent orientation.

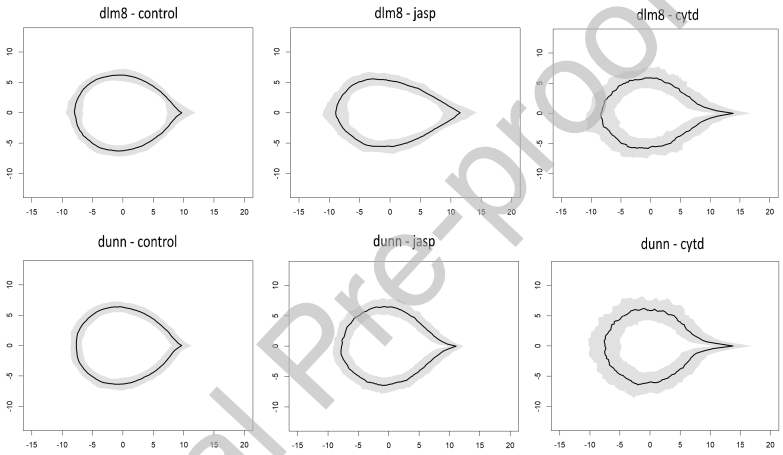


Figure 7: Mean radial shape (solid line) and pointwise standard deviation (shaded band) for each combination of cell line (DLM8/DUNN) and treatment (control, jasp, cytd).

To quantify the loss incurred when transitioning from the original cell contour to its star-shaped formalization, four key real-valued measures have been considered: the elongation, the roundness, the area of the resulting star-shaped set, and the difference in area between the original polygon and its star-shaped approximation.

On one hand, elongation represents how stretched or extended the shapes are. Clusters with higher elongation values may correspond to data points distributed in more linear or elliptical patterns, while lower values may indicate more compact or circular shapes. This metric is calculated using the `poly_elongation` function from the `pliman` library and is represented as X_2^E.

On the other hand, roundness quantifies how closely a shape approximates

a perfect circle by evaluating the uniformity of its boundary point distribution. Higher roundness values indicate that the shape is more circular, with smoother and more uniform boundaries, whereas lower values correspond to shapes that deviate significantly from circularity, often exhibiting angular features. Mathematically, this metric, denoted as X_2^R, is computed using the perimeter and the area of the polygon (via the `poly_length` and `polyarea` functions from the `pracma` library) as $X_2^R = \frac{4\pi \cdot \text{area}}{(\text{perimeter})^2}$, where the parameter reflects the deviation from ideal circularity.

The area of the resulting star-shaped set provides insight into the size of the star-shaped regions associated with each cluster. If a group shows a larger centroid area than another, it suggests that the elements in that group tend to be larger than those in the other. This measure is calculated using the `polyarea` function included in the `pracma` library and is denoted as $X_2^{A^{SS}}$.

Finally, the difference between the original area and the star-shaped area captures the discrepancy between the area of the polygon determining the cell contour and that of the corresponding star-shaped set. A larger difference may indicate that the cluster's distribution deviates significantly from a star-shaped configuration, whereas a smaller difference suggests closer conformity. The area difference, denoted as X_2^{diffA}, is computed using the `polyarea` function from the `pracma` library.

Consequently, the final dataset considered in the following sections comprises two main components. The first, denoted as $\mathbf{X_1}$, contains the star-shaped approximations of the cell contours. The second component, represented by the matrix $\mathbf{X_2}$, includes a set of quantitative descriptors described above, conveniently standardized to ensure comparability with the other variables involved in the optimization problem, with the configuration depending on the specific scenario under consideration. Together, these elements provide a comprehensive representation of each cell's geometric characteristics, enabling a more detailed and structured classification in the following sections.

6.2. FkM-Mix and RFkM-Mix results

The FkM-Mix and RFkM-Mix clustering methods for compact sets, introduced in Section 4, are applied to the preprocessed osteosarcoma dataset, which combines both the star-shaped contour representations ($\mathbf{X_1}$) and a set of associated geometric descriptors ($\mathbf{X_2}$). To ease comparison, the results

of both algorithms are presented jointly for each dataset in the following sections, first for DLM8 and then for DUNN.

6.2.1. FkM-Mix and RFkM-Mix applied to DLM8 cells

The FkM-Mix and RFkM-Mix clustering algorithms were applied to the dataset derived from the DLM8 cell line, which comprises 254 cells. To determine the optimal number of clusters (k), the fuzzy silhouette was adapted to accommodate the mixed-type nature of the data, integrating both the star-shaped representations ($\mathbf{X_1}$) and the associated geometric descriptors ($\mathbf{X_2}$). In this case, the numerical features that maximize the modified fuzzy silhouette index—thereby optimizing the quality of the partition—are elongation, the area of the resulting star-shaped set, and the area difference between the original polygon and its star-shaped approximation.

Table 3 summarizes the index values obtained for k ranging from 2 to 6 for both approaches, using a fuzziness parameter $m = 1.5$, which is the most commonly adopted value in fuzzy clustering. The robust method was computed using a threshold δ^2, defined in Equation (10), and set to 6.04959.

Table 3: Modified Fuzzy Silhouette (MFS) index for different values of k (FkM-Mix and RFkM-Mix methods, DLM8 cells)

k	2	3	4	5	6
MFS (FkM-Mix)	.5826	.6366	.5389	.5546	.5522
MFS (RFkM-Mix)	.5732	.6222	.5183	.5584	.4513

Based on the results presented in Table 3, the optimal number of clusters in both cases (excluding the noise cluster in the RFkM-Mix approach) corresponds to the maximum MFS value, which is attained at $k = 3$.

The star-shaped components of the prototypes obtained with both FkM-Mix and RFkM-Mix for the DLM8 dataset are summarized in Table 4. For each cluster, the center–radial characterization is expressed in terms of the centroid (which in this case is $(0, 0)$ for all groups after alignment) and the mean radial value.

Table 4: Star-shaped prototypes for the DLM8 cells when using the FkM-Mix and RFkM-Mix methods

Cluster	1	2	3
Star-shaped prototypes (FkM-Mix)	$(0,0) \pm 7.1$	$(0,0) \pm 6.8$	$(0,0) \pm 6.8$
Star-shaped prototypes (RFkM-Mix)	$(0,0) \pm 7.1$	$(0,0) \pm 6.9$	$(0,0) \pm 7$

Figure 8 displays the star-shaped prototypes obtained with FkM-Mix for the DLM8 dataset. These shapes correspond to the centroid–radial representation of each cluster and complement the numerical prototype information shown in Table 4. Similarly, Figure 9 shows the prototypes obtained with the robust RFkM-Mix method.

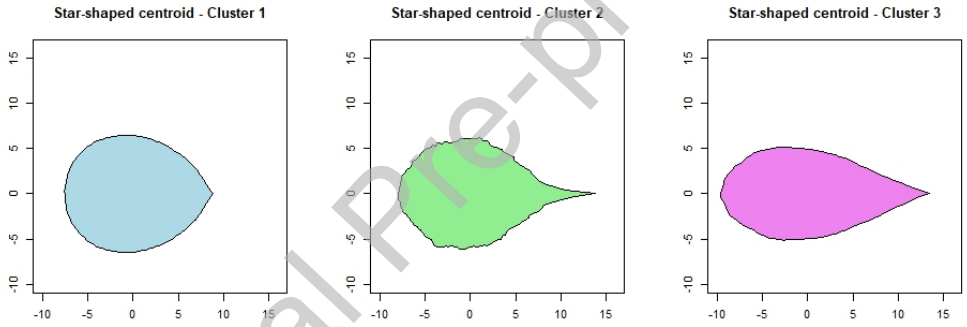


Figure 8: Star-shaped prototypes of the three clusters (FkM-Mix method, DLM8 cells)

The numerical descriptors of the prototypes are provided in matrices (18) and (19). Each row corresponds to a cluster. The first column represents the elongation, the second the area of the associated star-shaped set, and the third the difference between the original cell area and the star-shaped area.

$$\begin{pmatrix} -0.4883745 & 1.0507582 & -0.7653004 \\ 0.3267751 & -0.9450836 & 0.8073102 \\ 1.4698102 & -0.1581266 & -0.2930397 \end{pmatrix} \tag{18}$$

$$\begin{pmatrix} -0.5272168 & 1.16017706 & -0.8072122 \\ 0.1384659 & -0.58754004 & 0.1025047 \\ 1.5188650 & -0.09157736 & -0.3664565 \end{pmatrix} \tag{19}$$

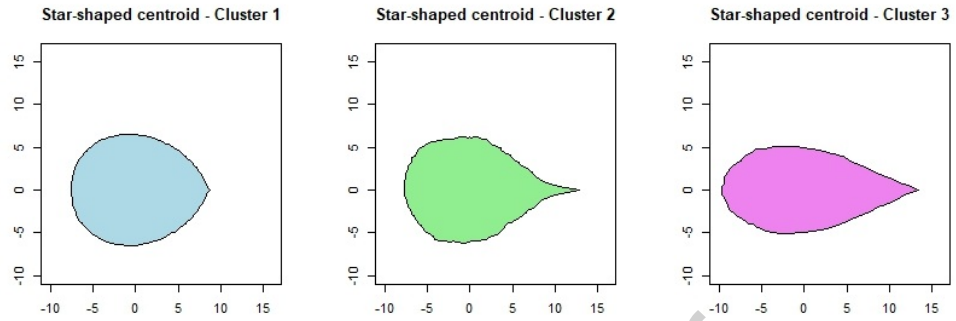


Figure 9: Star-shaped prototypes of the three clusters (RFkM-Mix method, DLM8 cells)

The prototypes of both the star-shaped part and the numerical descriptors, expressed as standardized values (with zero mean and unit standard deviation), offer the following insights.

FkM-Mix. Regarding the star-shaped component, the prototypes provided in Figure 8 suggest a certain degree of symmetry or balance around the origin, with distinct spatial deviations characterizing each cluster. Specifically, the star-shaped prototype associated with Cluster 1 appears less elongated and exhibits a smoother contour, with fewer irregularities compared to the other two clusters. Cluster 3 is characterized by a highly elongated star-shaped prototype and a relatively smooth boundary. In contrast, Cluster 2 presents a moderately elongated form (less rounded than Cluster 1) but with a boundary that displays considerable irregularities. With respect to the numerical descriptors of Cluster 1, the elongation value (-0.4883745) suggests a slightly below-average elongation compared to the dataset's mean. The star-shaped area (1.0507582) is above average (note that the representation in Figure 8 has been standardized) indicating that cells in this cluster tend to have larger star-shaped areas. The difference between the original cell area and the star-shaped area (-0.7653004) is below average, indicating the smallest discrepancy between the two. As far Cluster 2 is concerned, the elongation value (0.3267751) is slightly above average, reflecting moderate elongation. The star-shaped area (-0.9450836) is considerably below average, representing the smallest star-shaped areas among the clusters. The difference between the cell area and the star-shaped area (0.8073102) is above average, showing that the cell area surpasses the star-shaped area more significantly here. Finally, Cluster 3 presents an elongation value (1.4698102) significantly

above average, indicating that it contains the most elongated objects. The star-shaped area (-0.1581266) is slightly below average, while the area difference (-0.2930397) is moderately below average, reflecting a closer alignment between the original and star-shaped areas compared to Cluster 2.

RFkM-Mix. The prototypes obtained with the robust method (Figure 9 and matrix (19)) display similar relative tendencies. Cluster 1 remains the smoothest group, Cluster 2 retains its more irregular and compact morphology, and Cluster 3 continues to show the most elongated shapes. Therefore, the three clusters identified by RFkM-Mix correspond closely to those obtained with FkM-Mix, although the robust method prevents the most atypical shapes from influencing the prototype estimates by assigning them to the noise cluster.

Assuming that the assignment of a cell $\mathbf{x}_i$ to a cluster g is determined by the maximum membership value u_{ig} exceeding 0.5 (for $g = 1, 2, 3$), the resulting partitions for the DLM8 dataset are summarized in Table 5.

Table 5: Partition of DLM8 cells by considering the FkM-Mix and RFkM-Mix algorithms

Method	**Cluster**	**control**	***cytD***	***jasp***
FkM-Mix	1	90	5	25
	2	4	48	3
	3	20	27	32
RFkM-Mix	1	84	1	19
	2	7	33	7
	3	17	20	26
	noise	6	26	8

As shown in Table 5, when using the FkM-Mix method, Cluster 1 is predominantly composed of control cells (90), with a smaller presence of *cytD* (5) and *jasp* (25) cells. Cluster 3 exhibits a more balanced composition, with comparable numbers of *cytD* and *jasp* cells (27 and 32, respectively), and a smaller contribution from control cells (20). Cluster 2 is strongly associated with *cytD* cells (48), with minimal participation from the other groups. These distributions highlight distinct clustering patterns, with *cytD* playing

a significant role in both clusters 2 and 3. Figure 10 illustrates three representative DLM8 cells and their corresponding star-shaped approximations, each selected for having the highest membership degree (close to 1) within their respective clusters.

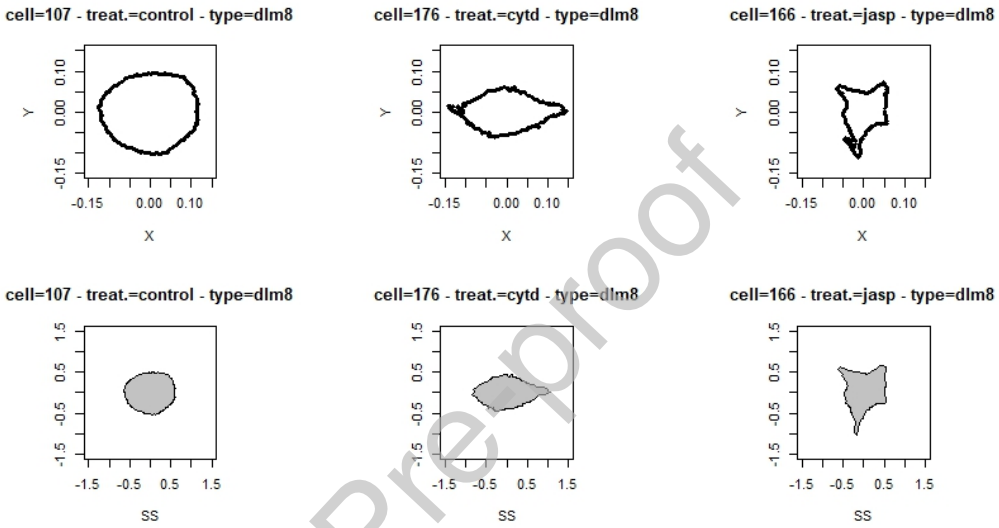


Figure 10: Three original DLM8 cells exhibiting the highest membership degrees to Cluster 1, Cluster 2, and Cluster 3, respectively, when the FkM-Mix algorithm is applied (top panel of the graphic). For each cell, the corresponding star-shaped representation is shown (bottom panel of the graphic).

For the RFkM-Mix method, several observations arise. Cluster 1 is primarily composed of control cells (84), with smaller contributions from *jasp* (19) and *cytD* (1), suggesting a group of cells that retain characteristics close to baseline morphology. Cluster 2 contains a majority of *cytD* cells (33), while control and *jasp* cells are less represented (7 each), reflecting the specific effects of *cytD* on DLM8 cells. Finally, Cluster 3 presents a balanced distribution, with control (17), *cytD* (20), and *jasp* (26) cells, indicating the presence of shared morphological traits across treatment conditions.

In this case, the three cells and their respective star-shaped representations that exhibit the highest membership degree (close to 1) in each of the three clusters correspond to those reported in Figure 10.

It is worth noting that one of the clusters contains a mixture of control, cytD and jasp cells. This behaviour is consistent with the substan-

tial morphological overlap observed across treatments, as illustrated by the mean–shape analysis in Figure 3 and by the cluster prototypes shown above. A subset of cells exhibits intermediate or treatment-independent morphological traits, which naturally leads to the formation of a mixed cluster. This does not indicate a limitation of the method, but rather reflects the intrinsic variability of the cell shapes: the effects of cytD and jasp do not always produce fully separated morphological phenotypes, especially in the presence of high within-treatment dispersion. This observation also suggests the possible existence of finer morphological subgroups within each treatment, which could be explored in future work by incorporating additional geometric descriptors or expert knowledge to refine the interpretation of these mixed groups.

When the RFkM-Mix method was applied to the DLM8 cells, 40 outliers were assigned to the noise cluster, each with a membership degree greater than 0.5. Among these, the three with the highest membership degrees (all corresponding to the *cytD* treatment, cells 243, 187, and 189) are shown in Figure 11.

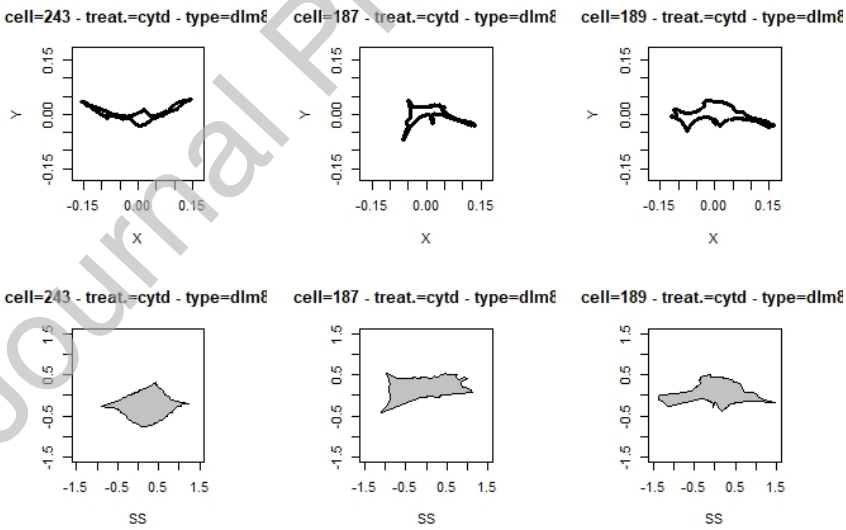


Figure 11: Three DLM8 cell outliers exhibiting the highest membership degrees to the noise cluster when the RFkM-Mix algorithm is applied (top panel of the graphic). For each outlier, the corresponding star-shaped representation is also shown (bottom panel of the graphic).

6.2.2. FkM-Mix and RFkM-Mix applied to DUNN cells

The FkM-Mix and RFkM-Mix clustering algorithms are now applied to the dataset obtained from the DUNN cell line, consisting of 390 cells. To determine the optimal number of clusters (k), the MFS index is once again utilized. In this case, the numerical descriptors for which the index is optimized include roundness, the area of the resultant star-shaped set, and the area difference between the original cell and its star-shaped approximation. Note that roundness is chosen instead of elongation in this case, as it has been confirmed that elongation does not effectively differentiate clusters. This is due to the fact that the boundary of DUNN-type cells generally exhibits a much more rounded shape across all treatments compared to DLM8-type cells, whose elongation varies depending on the specific treatment.

The MFS values obtained for k ranging from 2 to 6 for both clustering methods are summarized in Table 6. As in the previous case, the fuzziness parameter is set to $m = 1.5$, and the threshold δ^2 used in the robust method is set to 6.731584.

Table 6: Modified Fuzzy Silhouette (MFS) index for different values of k (FkM-Mix and RFkM-Mix methods, DUNN cells)

k	2	3	4	5	6
MFS (FkM-Mix)	.5607	.5916	.5825	.5279	.5208
MFS (RFkM-Mix)	.5106	.5788	.5432	.5109	.5244

From the results shown in Table 6, the optimal number of clusters (corresponding to the maximum value of the fuzzy silhouette index) is determined to be $k = 3$ in both cases.

The star-shaped components of the prototypes derived from the FkM-Mix and RFkM-Mix clustering procedures applied to DUNN cells are summarized in Table 7. As in the DLM8 case, the center–radial representation of each prototype is expressed by the centroid (which is $(0, 0)$ after alignment) and its mean radial value.

To better visualize the star-shaped components, their graphical representations are provided in Figures 12 and 13.

Table 7: Star-shaped prototypes for the DUNN cells when using the FkM-Mix and RFkM-Mix methods

Cluster	1	2	3
Star-shaped prototypes (FkM-Mix)	$(0,0) \pm 7.1$	$(0,0) \pm 6.8$	$(0,0) \pm 7$
Star-shaped prototypes (RFkM-Mix)	$(0,0) \pm 7.1$	$(0,0) \pm 6.9$	$(0,0) \pm 7$

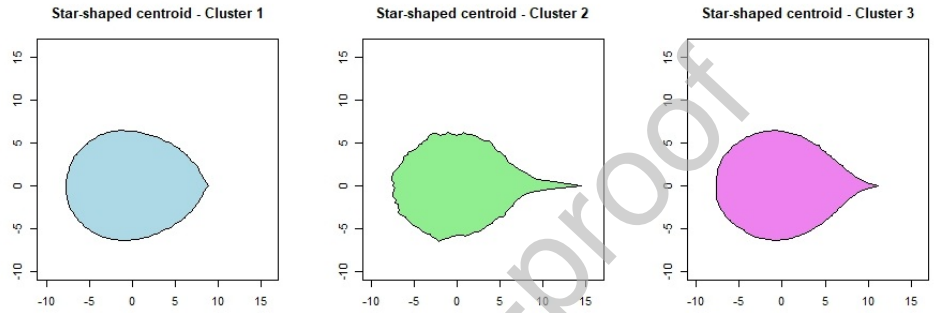


Figure 12: Star-shaped prototypes of the three clusters (FkM-Mix method, DUNN cells)

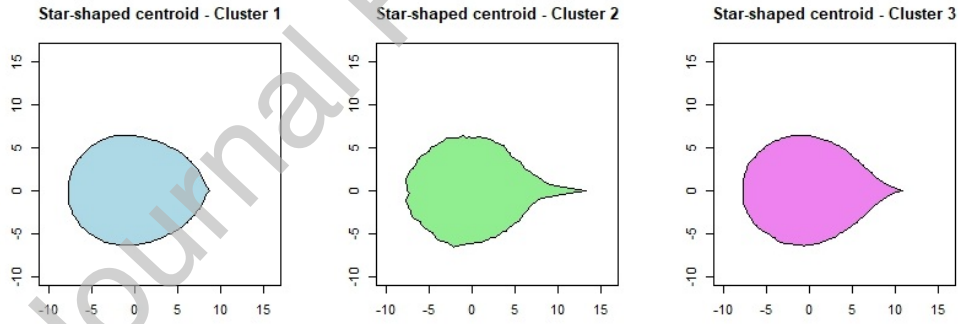


Figure 13: Star-shaped prototypes of the three clusters (RFkM-Mix method, DUNN cells)

In addition, the numerical descriptors characterizing the prototypes are defined as the ones in matrices (20) and (21). In this case, the first column corresponds to the roundness of the cells, the second to the area of the associated star-shaped set, and the third to the difference between the original cell and its star-shaped approximation.

$$\begin{pmatrix} -0.54531009 & 1.0512494 & -0.7771703 \\ 1.62150002 & -1.2699668 & 1.3746292 \\ -0.01855696 & -0.5826998 & 0.3850152 \end{pmatrix} \tag{20}$$

$$\begin{pmatrix} -0.5693007 & 1.1374912 & -0.8159494 \\ 0.9590012 & -1.1242116 & 1.2145805 \\ -0.1032009 & -0.4882132 & 0.1882336 \end{pmatrix} \tag{21}$$

Both the star-shaped component and the numerical descriptors, standardized as before, lead to the following observations.

FkM-Mix. As previously discussed, the star-shaped prototypes shown in Figure 12 exhibit similar elongation. However, the smoothness of the cell contours varies across clusters. The prototype corresponding to Cluster 1 features a notably smoother boundary with fewer irregularities compared to the other two clusters. While Cluster 3 is distinguished by an intermediate level of boundary smoothness, Cluster 2 displays the highest degree of irregularity. Regarding the numerical descriptors, the following key observations can be highlighted. Cluster 1 is characterized by a roundness value (-0.54531009) slightly below average, reflecting that these cells are more circular than those in the other clusters. The star-shaped area (1.0512494) is the largest, and the area difference (-0.7771703) is the smallest, indicating that the star-shaped representation closely preserves the original shape with minimal structural modifications. Cluster 2 presents the highest roundness value (1.62150002), hence it consists of the most irregularly shaped cells. The star-shaped area (-1.2699668) is the smallest, while the area difference (1.3746292) is the largest, indicating that the star-shaped transformation leads to significant structural changes in these cells. Concerning Cluster 3, all numerical descriptors in this cluster are relatively balanced. The near-zero roundness value (-0.01855696) that these cells closely resemble the dataset's average roundness. The star-shaped area (-0.5826998) is moderately smaller, indicating that the star-shaped area of the cells is somewhat smaller compared to the dataset's average. The difference between the original cell area and the star-shaped area is positive (0.3850152), which implies a moderate structural change following the star-shaped transformation.

RFkM-Mix. The prototypes obtained with the robust method (Figure 13 and matrix (21)) convey similar qualitative differences among clusters. Cluster 1 again shows the smoothest morphology, Cluster 2 retains its greater irregularity and smaller star-shaped area, and Cluster 3 remains intermediate.

Thus, the structure of the three main clusters under RFkM-Mix corresponds closely to that identified with FkM-Mix, with the additional advantage that the robust method isolates atypical shapes by assigning them to the noise cluster rather than allowing them to influence the estimated prototypes.

In this case, the obtained partition is the one presented in Table 8.

Table 8: Partition of DUNN cells by considering the FkM-Mix and RFkM-Mix algorithms

Method	**Cluster**	**control**	***cytD***	***jasp***
FkM-Mix	1	122	1	8
	2	6	63	7
	3	74	29	80
RFkM-Mix	1	115	1	5
	2	8	50	14
	3	70	19	69
	noise	9	23	7

Table 8 shows that Cluster 1 exhibits a balanced distribution across treatments, with a relatively high presence of *jasp* (80) and control (74) cells, along with a moderate number of *cytD* cells (29). The dominant presence of control cells (122) suggests that Cluster 2 may represent a baseline or untreated population, with minimal influence from *cytD* (1) or *jasp* (8) treatments. Finally, Cluster 3 is highly enriched with *cytD* cells, indicating that *cytD* treatment strongly differentiates this group from the others. Figure 14 presents three representative DUNN cells alongside their respective star-shaped approximations, selected for exhibiting the highest membership degree within their assigned clusters.

On the other hand, Table 8 also provides key insights into the clustering of cells under different treatment conditions when method RFkM-Mix is applied. Cluster 3 consists mainly of control cells (115), with minimal representation of *cytD* (1) and *jasp* cells (5). This suggests that this cluster largely consists of untreated cells, maintaining characteristics closest to the initial state without intervention. Overall, the partition reveals that the treatments influence cell clustering, with *cytD* forming a more distinct group,

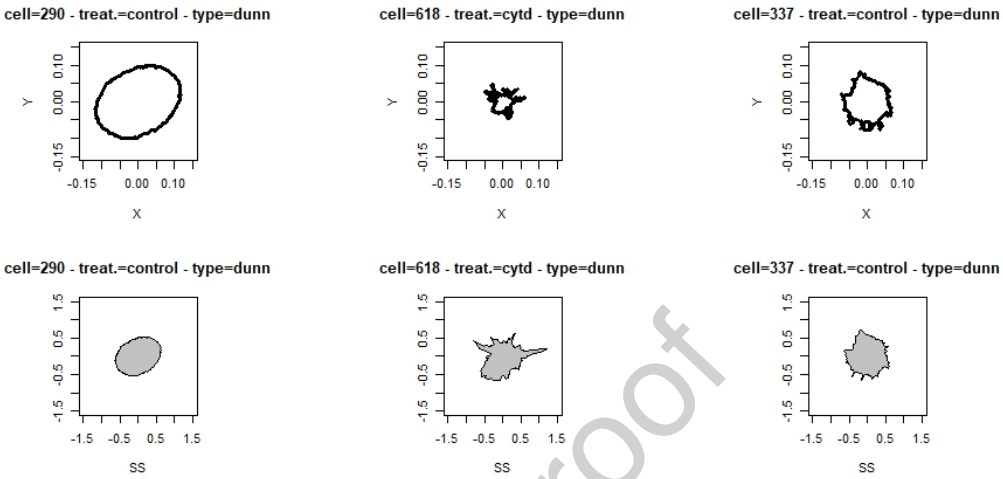


Figure 14: Three original DUNN cells exhibiting the highest membership degrees to Cluster 1, Cluster 2, and Cluster 3, respectively, when the FkM-Mix algorithm is applied (top panel of the graphic). For each cell, the corresponding star-shaped representation is shown (bottom panel of the graphic).

while *jasp* shows some overlap with the control group. In contrast, Cluster 2 consists primarily of *cytD* cells (50), with fewer cells from the control group (8) and *jasp* treatment (14), indicating that this cluster may capture distinctive traits induced by the *cytD* treatment. Lastly, Cluster 3 shows a relatively balanced distribution between the control group (70 cells) and the *jasp* treatment group (69 cells), while *cytD* is less represented (19 cells). Thus, this cluster groups cells with characteristics similar to those in the control group and *jasp* treatment but with less affinity for *cytD*.

Figure 15 presents three representative DUNN cells alongside their respective star-shaped approximations, selected for exhibiting the highest membership degree within their assigned clusters. Note that in this case such cells do not coincide with the ones provided in Figure 14 when the FkM-Mix method was applied to DUNN cells.

As in the DLM8 case, one of the clusters includes cells from all three treatments. This simply reflects that some DUNN cells share very similar shapes regardless of treatment, so they naturally cluster together. This overlap is a feature of the data rather than a limitation of the method.

Applying the RFkM-Mix method to the DUNN cells resulted in the clas-

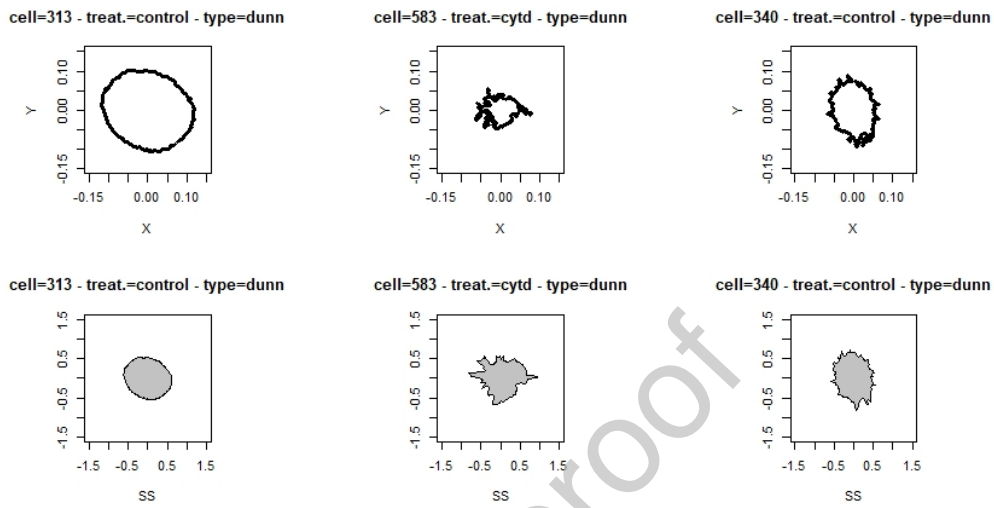


Figure 15: Three original DUNN cells exhibiting the highest membership degrees to Cluster 1, Cluster 2, and Cluster 3, respectively, when the RFkM-Mix algorithm is applied (top panel of the graphic). For each cell, the corresponding star-shaped representation is also shown (bottom panel of the graphic).

sification of 39 outliers into the noise cluster, each with a membership degree exceeding 0.5. Among these, the three outliers with the highest membership degrees (cells 613, 572, and 638, all associated with the *cytD* treatment) are depicted in Figure 16.

7. Conclusions and open problems

This study introduces FkM-type methods for compact sets in $(\mathbb{R}^p)$, leveraging mixed data that integrates geometric features with shape approximations to enhance clustering and analysis. To address variations in the mixed-type information extracted from the sets, a generalized distance measure has been developed. Furthermore, to account for potential outliers, a robust variant incorporating a noise cluster mechanism has been introduced. The comparative analysis conducted in the toy example (Section 5) confirms the advantages of the proposed mixed and robust formulations, showing improved separation and enhanced stability even in a controlled setting. The effectiveness of the proposed methodologies is also demonstrated through a real-world application in cancer cell analysis, highlighting their capability to process complex

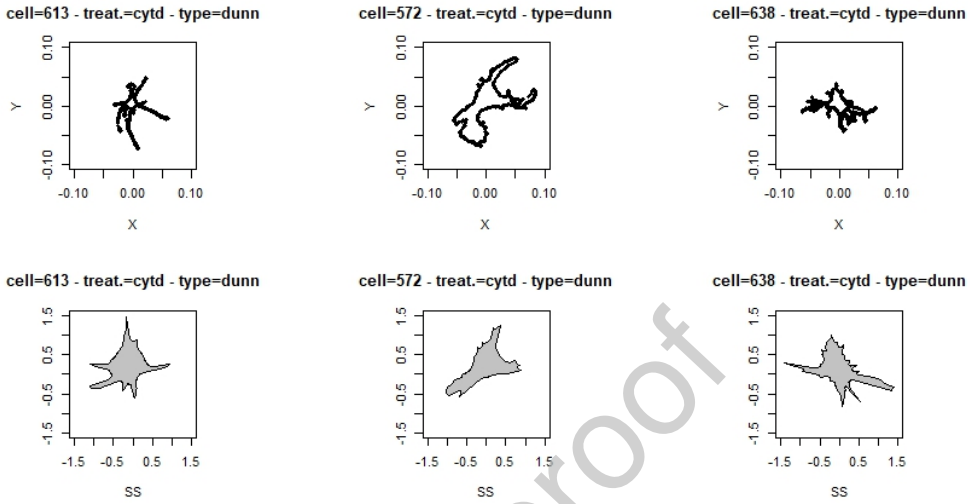


Figure 16: Three DUNN cell outliers exhibiting the highest membership degrees to the noise cluster when the RFkM-Mix algorithm is applied (top panel of the graphic). For each outlier, the corresponding star-shaped representation is also shown (bottom panel of the graphic).

biological data and extract meaningful insights.

As potential directions for future research, alternative measures (such as kernel-based or anisotropic metrics) could be explored depending on the application. The methods may also be extended to integrate additional constraints or regularization terms, incorporating domain-specific requirements or prior knowledge about cluster structures while accounting for noise. Furthermore, other clustering techniques, including hierarchical clustering or spectral clustering, could be adapted for compact sets, providing complementary solutions for diverse scenarios. Finally, the presence of mixed clusters in the application highlights a natural overlap in cell shapes and opens the possibility of using additional geometric or medical descriptors to better capture subtle differences.

Funding The research in this paper has been partially supported by the project Pilot Plan on Sediment Management in the lower Nalón River (2021-2023), funded by TRAGSA to the Cantabrian Water Authority (CHCantábrico), within the Plan PIMA Adapta Agua. The work was also benefited from the methods developed by the Cost HiTEc, CA21163, supported by

COST (European Cooperation in Science and Technology).

Data availability The dataset that support the findings of this study is openly available at the web page `https://github.com/wxli0/dyn/tree/main%4092c7a58/dyn/datasets/osteosarcoma/aligned/full`

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alizadeh, E., Xu, W., Castle, J., Foss, J., Prasad, A. (2019). Tismorph: a tool to quantify texture, irregularity and spreading of single cells. PLoS ONE, 14(6), 0217346.

Ameer Ali, M., Karmakar, G.C., Dooley, L.S. (2010) Detection and separation of generic-shaped objects by fuzzy clustering. International Journal of Intelligent Computing and Cybernetics, 3(3), 365–390.

Asai, T., Ueda, T., Itoh, K., Yoshioka, K., Aoki, Y., Mori, S., et al. (1998). Establishment and characterization of a murine osteosarcoma cell line (LM8) with high metastatic potential to the lung. International Journal of Cancer, 76(3), 418–422.

Bezdek, J.C. (1981). Pattern recognition with fuzzy objective function algorithm. Plenum Press, New York.

Campello, R.J.G.B., Hruschka, E.R. (2006) A fuzzy extension of the silhouette width criterion for cluster analysis. Fuzzy Sets and Systems, 157, 2858– 2875.

Chen, Z., Ma, X., Wu, L. Xie, Z. (2019) An intuitionistic fuzzy similarity approach for clustering analysis of polygons. ISPRS International Journal of Geo-Information, 8(2), 98.

Davè, R. (1991) Characterization and detection of noise in clustering. Pattern Recognition Letters, 12, 657–664.

Davis, C.M., Gruebele, M. (2020) Cytoskeletal drugs modulate off-target protein folding landscapes inside cells. Biochemistry, 59(28), 2650–2659.

Doering, C., Lesot, M.J., Kruse, R. (2006). Data analysis with fuzzy clustering methods. Computational Statistics and Data Analysis, 51, 192–214.

Dunn, J.C. (1974) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern., 3, 32–57.

D'Urso, P., De Giovanni, L. (2015) Trimmed fuzzy clustering for interval-valued data. Advances in Data Analysis and Classification, 9, 21–40.

D'Urso, P., Giordani, P. (2006) A weighted fuzzy c-means clustering model for fuzzy data. Computational Statistics and Data Analysis, 50, 1496–1523.

D'Urso, P., Massari, R. (2019) Fuzzy clustering of mixed data. Information Sciences, 505, 513–534.

Ferraro, M.B., Giordani, P. (2013) On possibilistic clustering with repulsion constraints for imprecise data, Inf. Sci., 245, 63–75.

Ferraro, M.B., Fernández Iglesias, E., Ramos-Guajardo, A.B., González-Rodríguez, G. (2023). On Clustering of Star-Shaped Sets with a Fuzzy Approach: An Application to the Clasts in the Cantabrian Coast. In: García-Escudero, L.A., et al. Building Bridges between Soft and Statistical Methodologies for Data Science. Advances in Intelligent Systems and Computing, vol 1433.

Ferraro, M.B. (2024) Fuzzy k-Means: history and applications. Econometrics and Statistics, 30, 110–123.

González-Rodríguez, G. Ramos-Guajardo, A.B., Colubi, A., Blanco-Fernández, Á. (2018) A new framework for the statistical analysis of set-valued random elements. International Journal of Approximate Reasoning 92, 279–294.

González-Rodríguez, G. (2026) Choosing the center of star-shaped set-valued data compatible with measure-preserving arithmetic. International Journal of Approximate Reasoning 188.

Gustafson, E., Kessel, W. (1979) Fuzzy clustering with a fuzzy covariance matrix. In: Proc. of IEEE CDC, San Diego, CA, USA, 761–766.

Hwang, C., Rhee, F. C. (2007). Uncertain fuzzy clustering: Intervals with examples. International Journal of Approximate Reasoning, 44(2), 114–135.

Klain, D. (1997). Invariant valuations on star-shaped sets. Advances in Mathematics 125(1), 95–113.

Kováčiková, M., Vaškovicová, N., Nebesářová, J., Valigurová, A. (2018). Effect of jasplakinolide and cytochalasin D on cortical elements involved in the gliding motility of the eugregarine Gregarina garnhami (Apicomplexa). European Journal of Protistology, 66, 97–114.

Joshi, D., Samal, A., Soh, L.K. (2009) Density-Based Clustering of Polygons. In: IEEE Symposium on Computational Intelligence and Data Mining, 2009.

Li, W., Prasad, A. Miolane, N., Dao Duc, K. (2023). Using a Riemannian elastic metric for statistical analysis of tumor cell shape heterogeneity. In: Nielsen, F. and Barbaresco, F. (eds.) Geometric Science of Information, pp. 583–592. Springer Nature Switzerland.

Masson, M.H., Denoeux, T. (2008) ECM: an evidential version of the fuzzy C-means algorithm. Pattern Recognition 41, 1384–1397.

Matheron, G. (2018). Random Sets and Integral Geometry. Wiley & Sons, New York.

Molchanov, I., Molinari, F. (2014). Applications of random set theory in econometrics. Annual Review of Economics 6, 229–251.

R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Ramos-Guajardo, A.B., Ferraro, M.B. (2020) A fuzzy clustering approach for fuzzy data based on a generalized distance. Fuzzy Sets and Systems, 389, 29–50.

Ramos-Guajardo, A.B., González-Rodríguez, G., Colubi, A., Ferraro, M.B., Blanco-Fernández, Á. (2018). On Some Concepts Related to Star-Shaped Sets. Studies in Systems, Decision and Control 142, 699–70.

Ruspini, E.A. (1969) A new approach to clustering, Inf. Control, 15, 22–32.

Ruspini, E.A., Bezdek, J.C., Keller, J.M. (2019). Fuzzy Clustering: A Historical Perspective. IEEE Computational Intelligence Magazine, 14(1), 45–55.

Schneider, R. (1993). Convex bodies: the Brunn-Minkowski theory. Cambridge University Press, Cambridge.

Shi, P., Guo, L., Cui, H., Chen, L. (2023) Geometric consistent fuzzy cluster ensemble with membership reconstruction for image segmentation. Digital Signal Processing, 134, 130901.

Simó, A., De Ves, E., Ayala, G. (2004). Resuming shapes with applications. Journal of Mathematical Imaging and Vision 20(3), 209–222.

Wang, W., Du, S., Guo, Z., Luo, L. (2014) Polygonal Clustering Analysis Using Multilevel Graph-Partition. Transations in GIS, 19(5), 716–736.

Wee, W.G., Fu, K.S. (1969) A formulation of fuzzy automata and its application as a model of learning systems, IEEE Trans. Syst. Sci. Cybern., 5, 215–223.