HIERARCHICAL CLUSTERING METHODOLOGIES FOR THE DETECTION OF SENESCENT CELLS

Master's thesis
University of Turku
Physics
2023
Giulia Parisi
Examiners:
Prof. Salvatore Miccichè
Dr. Johannes Niskanen

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using Turnitin Originality Check service.

UNIVERSITY OF TURKU
Department of Physics and Astronomy

Student, Giulia Parisi HIERARCHICAL CLUSTERING METHODOLOGIES FOR THE DETECTION OF SENESCENT CELLS

Master's thesis, 56 pp.
Physics
October 2023

This Master Thesis project is related to a data set of microscope images whose purpose is to analyze senescence. Senescence is a dynamic process whereby cells stop to duplicate and change their morphology.
The results here described are a first attempt to classify cells by their morphology, in order to find a method to detect the cluster of senescent cells.
The microscope images are first of all preprocessed and the objects are properly segmented. Then some features are extracted for each object, namely "area", "circularity", "eccentricity" and "convexity defects". With these features, an Agglomerative Hierarchical Clustering is applied with different methods. It results that, on the basis of the extracted features, the cells in 24 hours can be classified in 4 clusters with diffent morphological characteristics.

Contents

Introduction

The Master thesis project here presented is a preliminary analysis useful to develop a tool for the recognition of a cluster of senescent cells among a sample of cells in a cultured well plate.

Cellular senescence is a dynamic and heterogeneous process whereby cells stop to duplicate and change their morphology. An univocal method to detect senescent cells, analyzing just their shape and properties, does not exist yet and also it is not very clear how they are involved in processes such tumor arrest and aging[1]. For this reason, the only not invasive way to distinguish a senescent cell from a living one, is to follow its trajectory and to check if it duplicates or not; what could be useful instead is a method that would allow to detect the cluster of senescent cells collecting appropriate features, of morphological and kinematic type, in order to study possible collective behaviours and contagion effects.

The data set available is provided by *Fondazione Ri.MED*[2] and it consists of a time sequence of microscope images with which it is possible to follow the evolution of the cells over time. The main focus of this thesis work was that of recognizing cells in the different images and trying a tentative classification of them. Ideally such classification process would help in distinguishing senescent cells from normal ones. A first attempt of classification of the objects in the images is based on their morphological features. Classical clustering methods[3] therefore are used in order to find and characterize the objects observed in the images and to analyze how the behaviour of the cells inplanted in the culture changes with different levels of treatment that stimulates senescence. Other methodologies are in principle better suited for the classification purposes mentioned above. In fact, a machine learning approach could be used for cell recognition. However, such an approach would need the availability of data set much larger than the one considered. Larger datasets would allow to split it into a training and test dataset, and test the machine learning

protocol after having correctly classified cells in the training set with the help of biologists. Another approach would be that of using the support of fluorescence-based assays, that would allow to label the objects in the images, and recognize senescent cells based on their response to special biomarkers[4]. Due to data availability, such approaches are left for future studies.

The structure of the thesis can be summarized as follows: in the first chapter the data set is introduced and a description of the steps followed to detect the objects in the microscope images is given. This first section of pre-processing and image analysis was necessary to recognize, segment cells properly in each image and get an idea of the presence of systematic errors in the data set (movement of the lenses during the acquisition, light sources...); in the second chapter the first approach of classification is applied, using *Agglomerative Hierarchical Clustering* with different linkages[5], collecting the objects in the images all together and the clusters obtained are then interpreted, analyzing the evolution of the percentages of elements in each cluster over time; in the third chapter a characterization of the clusters is performed, looking if there are and which are the over-expressed features[6]. Using these information an alternative approach to find the groups of cells is carried out, performing the clustering algorithm separately image by image so that it's possible to make a comparison of the two different approaches.

The thesis work has been conducted during the *Double Degree Programme* with *Univerisity of Turku* and *Università degli Studi di Palermo* in close collaboration with the Advanced Data Analysis research group at *Fondazione Ri.MED*, led by Dr. Claudia Coronnello and with the *High-throughput Screening Laboratory*, led by Dr. Chiara Cipollina, where the experiments were carried out. They helped in giving the right biological interpretation of the results obtained during the investigations as well as in calibrating the use of the right method of analysis of the system.

1 Introduction to the Data Set and Cell Detection

The purpose of this first chapter is to describe the data set, which is composed by microscope images of cells in culture medium. The samples are analyzed in five different conditions, depending on the concentration of Hydrogen Peroxide (H_2O_2), which is added to induce senescence. A description of the steps used for cell detection in the images is also given and some ensemble trends are then analyzed.

1.1 Microscope Images Data Set

The work done in this thesis is based on the analysis of a data set of microscope images of cell culture; cell culture is an artificial technology which allows the growth of the cells in a controlled environment, with specific nutrients. The culture has a total duration of three days and the images have been acquired with a time step of one hour, whereby it is possible to follow the evolution of the cells for 72 hours.

The cells of the images are of HEK 293T type, this means that they are taken from the human embryonic kidney. The cells are plated in 96-well plates ($15X10^3$ cells/well) and incubated in the Incucyte S3[7] at 37°C and 5% of CO_2. The culture medium where to seed the cells is prepared with DMEM (Dulbecco's Modified Eagle Medium), FBS (Fetal Bovine Serum) and penicillin-strepromycin; furthermore H_2O_2 in the treated cases. The images dimensions are $1040X1408$ pixel and they where acquired by the instrument with an image resolution of 0.62 μm/pixel.

Only 15 wells of the plate have been used for the culture, which are five columns, namely 2-3-4-5-6 and three rows B, C, D. A schematic representation can be seen in *Fig.*1.

The dataset of the images is divided in non-treated cells, which are the samples in column 2, and treated-cells, where the environment was serial diluted with different

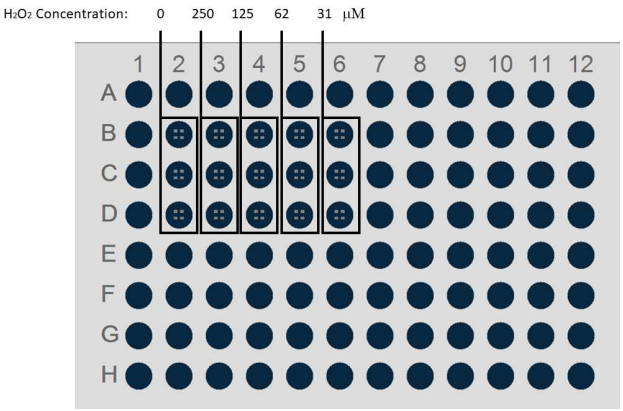


Figure 1: Scheme of the well plate used for the cell culture. Only 15 wells have been used, columns 2-3-4-5-6 and three rows B, C, D. Each column corresponds to a different value of H_2O_2 concentration.

concentration of H_2O_2 with these values:

- 250 μM (high level concentration) – column 3;
- 125 μM (medium level concentration) -column 4;
- 62 μM (medium level concentration) - column 5;
- 31 μM (low level concentration) - column 6.

At the beginning (time 00h00m), after that the cells are seeded in the wells, they have a spherical shape with a diameter typically between 11 and 15 μm and they are not attached at the bottom. *Fig.*2 is the image of a non-treated sample at time 00h00m.

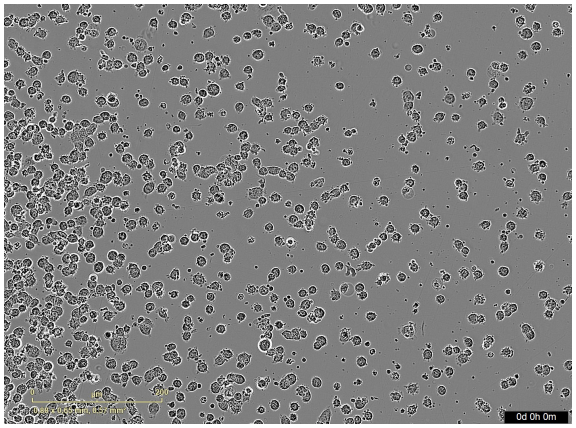


Figure 2: Image of a sample of cells in a non-treated environment at time 00h00m, it corresponds to the well plate B2. The cells are singular, well recognizable and have a spherical shape.

When the cells attach at the bottom, they change size and shape, becoming larger, and if they are in a non-treated environment after 15-20 hours they start to replicate exponentially, so that the initial number doubles in around 34-36 hours[8], and they keep splitting until they recover the entire space of the well. *Fig.*3 provides an example of how the same sample shown before looks like after 72 hours.

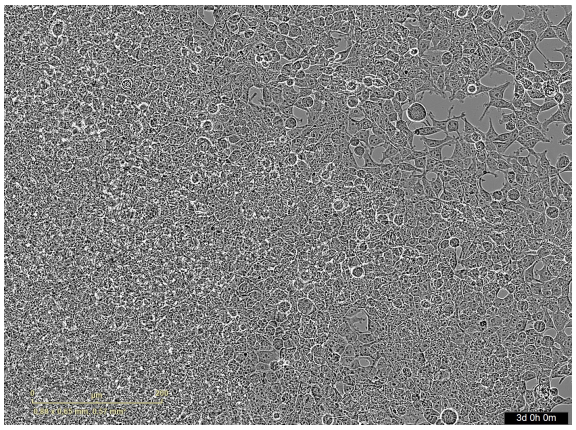


Figure 3: Image of a sample of cells in a non-treated environment at the time 72h00m, it corresponds to the well plate B2. The cells divided exponentially until they cover the whole plate after 72h00m. It's possible also to see that they overlap.

On the contrary, if cells are in an environment treated with H_2O_2 they do not replicate anymore, becoming senescent which impairs their ability to split and thus they are not able to cover all the space anymore. It is possible to see in *Fig.*4 how the sample treated with the maximum concentration of H_2O_2 looks like after 72 hours.

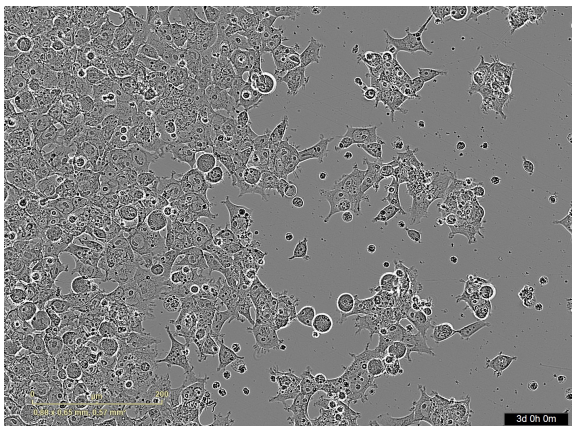


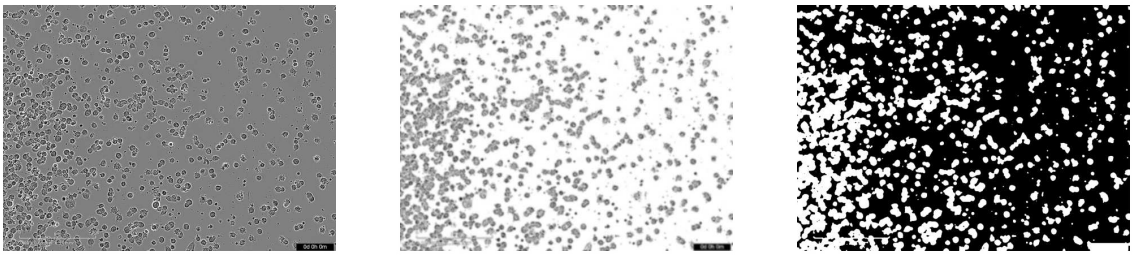
Figure 4: Image of a sample of cells in a treated environment with the maximum concentration of H_2O_2 at the time 72h00m, it corresponds to the well plate B3. The cells do not recover the whole space which indicates that most of them became senescent and thus could not divide.

1.2 Cell Detection

In order to find the properties of living and senescent cells, it is necessary first of all to detect the cells in the images. The problem consists in segmenting them[9], because when they start to split they also can overlap and form agglomerations, and distinguishing them from debris.

For this purpose the protocol of the software *ImageJ*[10] is automatize with the Python package *scipy.ndimage*[11]. The algorithm works in this way: since the images are in grayscale, first of all a value of gray as threshold is computed, and to do this the *Otsu Thresholding*[12] method included in *OpenCv*[13] Python package is used, where "a value of the threshold is not chosen but is determined automatically. A

bimodal image (two distinct image values) is considered. The histogram generated contains two peaks. So, a generic condition would be to choose a threshold value that lies in the middle of both the histogram peak values"[14]. Then a gaussian filter is applied and the threshold is again computed; at this point a mask which converts the values over the threshold to black and the values under it to white is applied and the inverted mask (black and white inverted) is used for the next step. The following images in *Fig.*5 illustrate the process.



(a) (b) (c)

Figure 5: Outputs of the algorithm steps for cells detection: (a) original image, (b) thresholded image with *Otsu thresholding* method and with gaussian filter applied, (c) masked image (inverted mask). The analyzed image is always that of the well plate B2 at time 00h00m.

At this point, it is necessary to separate the cells which in the masked image appear attached; to do this the *watershed transformation method*[15] is used for image segmentation from the Python package *skimage.segmentation*[16].

Selecting properly the parameters, the segmented image is obtained as shown in *Fig.*6.

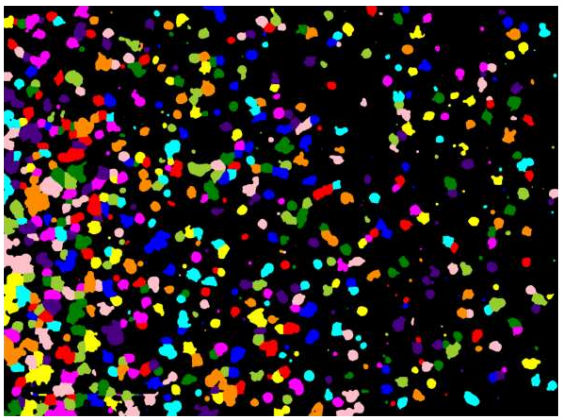


Figure 6: The result of watershed transformation method; most parts of the cells are now properly segmented.

To detect cells on the images it is also necessary to discard debris, small particles present on the background.

The area distribution of the objects detected until now is calculated; it results, as expected, that the areas follow a bimodal distribution where the first peak is referred to the mean size of the debris and the second one to the mean size of the cells. For this reason the particles whose area is under the threshold value of $A = 88.0\mu m^2$ are disregarded and only the particles with bigger size are considered as cells. At the end it is possible to detect the cells and the coordinates of their center of mass (Fig.8).

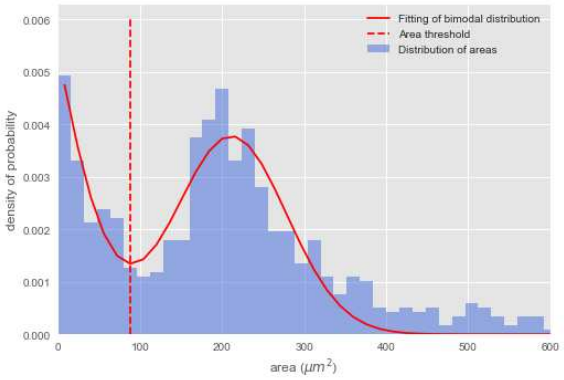


Figure 7: Fitting of the bimodal distribution of the areas. The areas under $A = 88.0\mu m^2$, which are referred to debris, are disregarded.

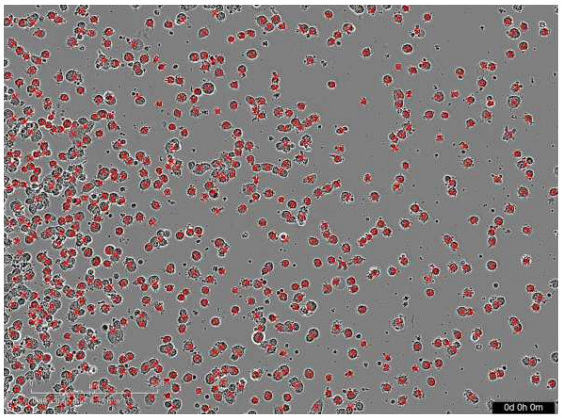


Figure 8: Final result of the cell detection; in the figure all the centers of mass detected are pointed out with a red cross

1.3 Ensemble Trends of the Identified Cell Samples

Some trends of the sets of cells are now analyzed over time. First of all, the trend of the number of cells over time is studied, for different cases: non-treated cells, cells treated with low level of H_2O_2, cells treated with medium level and with high level of H_2O_2; contextually also the trend of the median of areas of the whole cell sample over time is analyzed in order to find if there are some correlations. The following figures show the results.

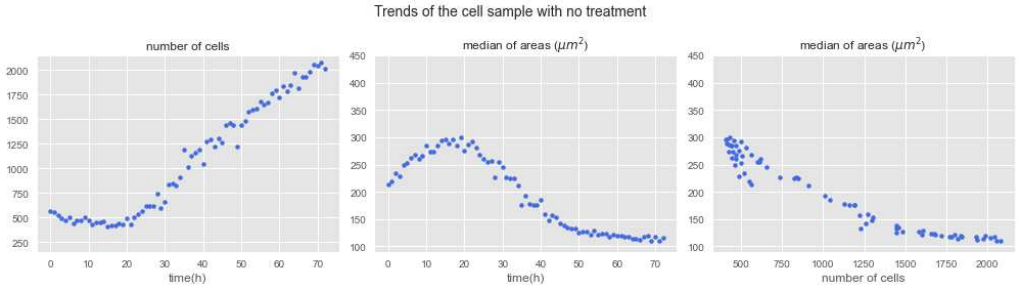
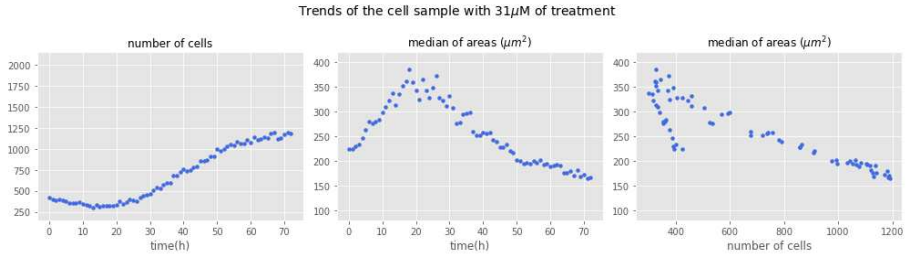
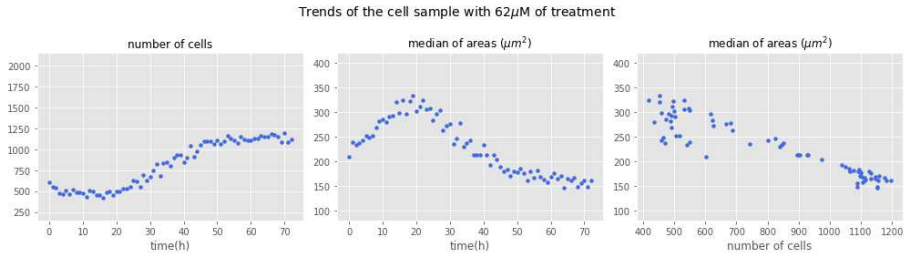


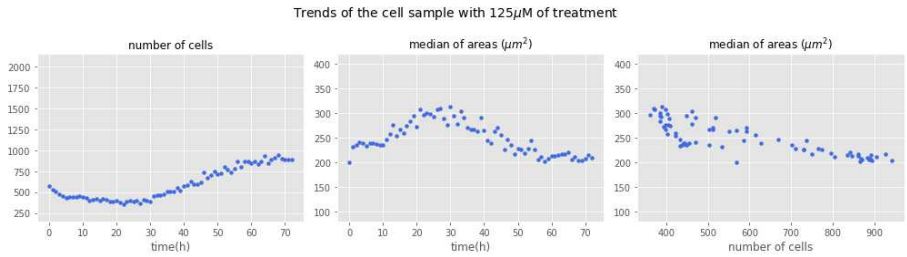
Figure 9: Trends of sample of cells in culture with no treatment (well plate B2).



(a)



(b)



(c)

Figure 10: Trends of cells in well plates (a) B6-treated with 31μM concentration of H_2O_2, (b) B5-treated with 62μM concentration of H_2O_2, (c) B3-treated with 125μM concentration of H_2O_2.

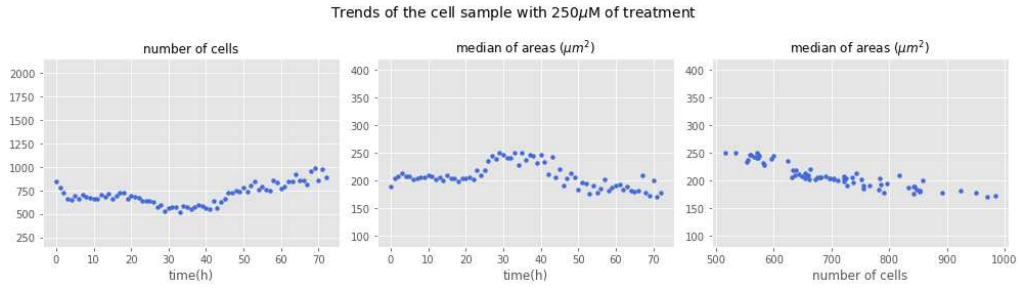


Figure 11: Trends of the sample of cells with maximum value of concentration of H_2O_2.

In *Fig.*9, i.e. when the sample is not treated at all, the cells around 10-20 hours enlarge (the median of areas arises indeed) and then they start to divide, as it is possible to see after 20h the number of cells increases. After 36 hours, as expected, the number doubles. After 72 hours the final number has almost quadrupled in respect to the beginning. The cells would keep dividing until the whole space is covered, but since in this case after 72h they still have space, the number simply increases until the end. In the third figure (median of areas as a function of number of cells) it is possible to see that the highest values of area median is recorded when the number of cells is still small, while when the number increases the medians become lower since the cells overlap and do not allow to the near ones to enlarge further.

The situation is different when they are treated; in *Fig.*10 the samples are treated with low and medium levels of H_2O_2 and the number of cells starts to increase later, after around 30 hours. The cells divide up to a certain point from which onwards the number is constant. This happens after around 50-60 hours probably induced by the treatment. An explanation can be that the number of new cells is compensated by the dying ones. In any case, the total number of cells at the end is lower than in the non-treated sample.

In *Fig.*11, when the cells are treated with the maximum level of H_2O_2, the number

keeps decreasing since a lot of cells die, their size reduces and they are detected by the algorithm as debris and discarded from the detection. After 40 hours the still living ones are able to divide and the number increases.

At this point it is possible to compare the growth rate of the number of cells in the growing-regime for all the configurations. To do this, each set of data is divided by the number of cells at the first hour.
Since the trend should grow as 2^n, the data are rescaled in log_2-scale and the slopes of the lines are compared. As it is observed, the highest slope is recorded for the sample without treatment, since the cells are more vital and they keep splitting very fast, while the lower the value of the slope, the higher the level of treatment and the lowest value of slope corresponds to the highest level of treatment.

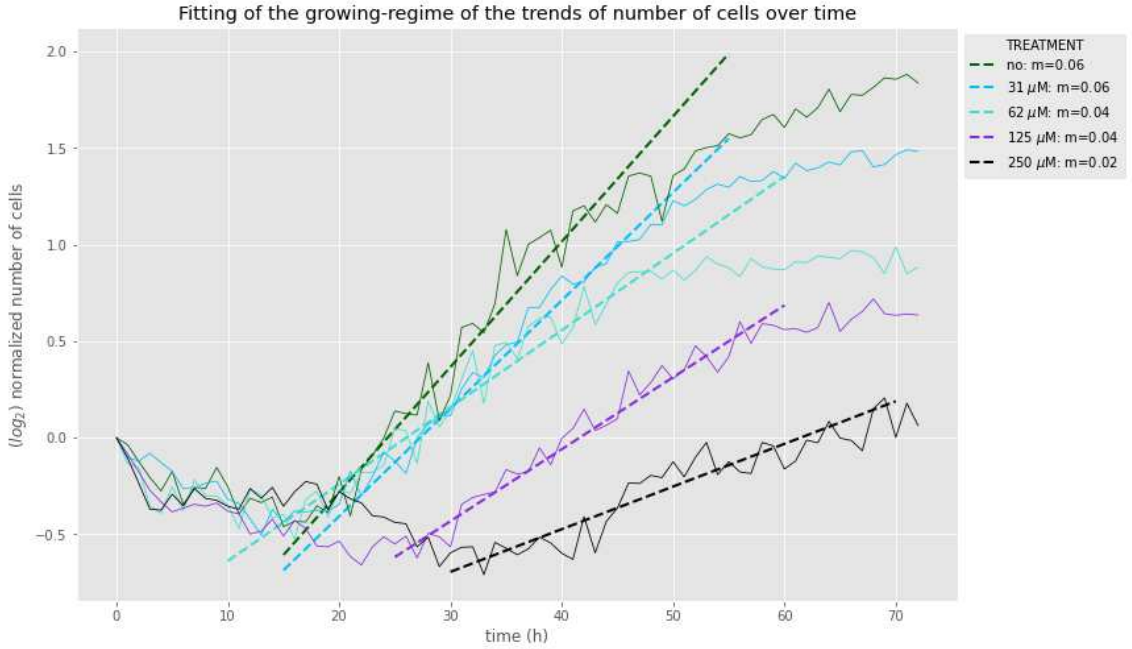
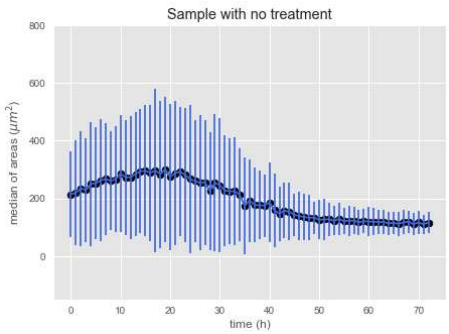
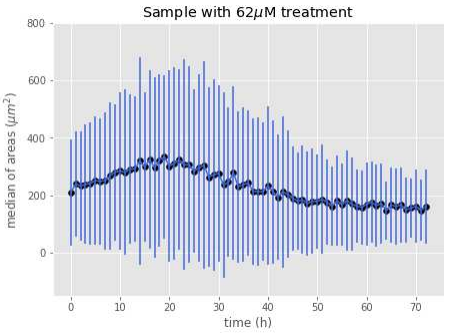


Figure 12: Comparison of the growing-regime trend for all the configurations; as expected, the highest slope is recorded for the configuration without treatment (green line) and the lowest for that one with highest level of treatment (black line).

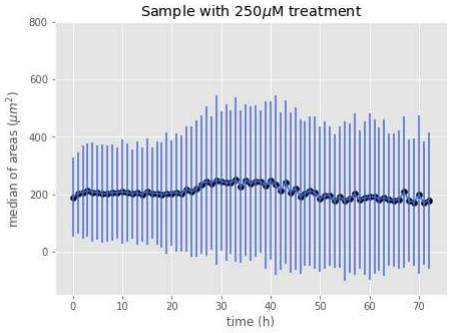
Looking again at the *Fig.*9-11 in all the five different cases it is possible to observe how the values of median decrease as the number of cells increases. This probably means that the cells in culture are smaller when they are a lot because they occupy more space and cannot enlarge further. When then number of cells is small instead, the values of median of areas seem to have two different trends; the smaller values probably refer to the first hours, when the cells are just seeded and they are few and medium-sized, the larger values instead refer to the range of hours 10-20, when the cells are still few but have a lot of space at disposition so their size becomes larger. Even if the median trends of areas appear quite smooth, taking a look at the standard deviations *Fig.*13, it is possible to notice that all the samples at every hour are very heterogeneous.



(a)



(b)



(c)

Figure 13: Plots of area median over time with the standard deviations, for three different configurations: (a) no treatment, (b) 62 μM-medium level of treatment and (c) 250 μM-medium level of treatment.

These high values of standard deviations are probably due to the fact that the segmentation does not work properly image by image and agglomeration are then always present and they influence the distribution.

By further analyzing the images hour by hour it would be expected to find different types of cells with different sizes. Cell sizes, indeed, will be used in the next section as a feature to perform clustering and classify the objects in the images. It is possible to see that for the sample with no and medium level of treatment, the standard deviation reduces over time, and this is because the cells recover almost all the space of the well plate and they all become smaller, while in the case of high level of treatment, the standard deviation values increase when the first phase is overcome and the still living cells start to split. In any case, the values stay high and this means that the sample is heterogenous and it would be expected to find objects of different size almost at every hour.

2 Agglomerative Hierarchical Clustering for Objects Classification

Different *unsupervised* approaches of clustering are now performed in order to classify the objects in each image, according to some morphological features.

In the first method, all the objects in all the configurations (treatments) from time 00h00m to time 23h00m on the well plate B are grouped; this means that the clustering algorithm is performed on the objects present on 120 images, which are in total 77702. For all the objects some features are extracted, which will be used to compute the euclidean distances among them and an *Agglomerative Hierarchical Clustering* (AHC)[5] is applied, with different linkage methods, namely *Ward*[17] and *Average Linkage Cluster Analysis*[18]. The aim is to find a classification of the cells based on the selected set of features.

Three score evaluation criteria are used to find the best number of clusters and then the model which returns the possible classification of the objects is applied. In this way it is possible to visualize the objects of each cluster image by image and study how the percentage of elements in each cluster evolves during the period of time considered.

2.1 Agglomerative Hierarchical Clustering

The objects expected to find in the images are of different types: debris, living cells which are able to split, dead and senescent cells. Since there is no an univocal method to recognize which type the cell belongs to, just looking at its morphology[1], it is not possible to perform any *supervised* approach. Furthermore, the number of objects available in the images is not sufficient to split the data set in a training and test set for machine learning techniques.

For these reasons the decision was to apply an *unsupervised* classification approach on the unlabeled data and to perform *Agglomerative Hierarchical Clustering* (AHC), in order to detect the possible clusters of objects.

The algorithm used from the *Scikit-Learn*[19] python package can be resumed in these following steps[20]:

1. at the beginning each point is considered as an independent entity, so the total number of clusters is equal to *N*, which is the total number of objects;

2. the two closest data points are merged in one cluster, so the total number of resulting clusters is N-1;

3. the two closest clusters are merged in an unique cluster, in this way the total number of clusters is N-2;

4. the steps are repeated until the set of data is divided in the number of clusters chosen as an input.

The way to define two points and clusters as the "closest" depends on the choose of the *metric* and the *linkage*, respectively.
The decision was to implement the algorithm selecting the *Euclidean* metric, which works well for low dimension[21]. The euclidean distances are indeed computed after that some proper features of the objects are extracted which, as it will be explained in the next section, are four for each of them; this means that each object is characterized by a vector of four features, that are used to define objects as closest. The AHC with *Ward* linkage and *Average* linkage is then performed; the first one is also known as MISSQ (Minimal Increase of Sum-of-Squares), which tries to minimize the increase of the sum of squares errors at each step, therefore minimizing the error[17], the second one instead is defined as the average of distances between all pairs of objects. The distance matrix T is defined as:

$$\begin{cases} t_{hj} = \frac{N_h t_{hj} + N_k t_{kj}}{N_h + N_k}, & \text{if } j \neq h \text{ and } j \neq k \\ t_{ij} = t_{ij}, & \text{otherwise} \end{cases}$$

where N_h and N_k are the number of elements inside respectively the cluster h and the cluster k and at each step the clusters h and k, for which T is minimized, are merged[18].

2.2 Features Extraction and Data Preparation Steps

In order to perform the clustering, it is necessary to characterize the objects by some appropriate features. The watershed procedure was already described, which segments the objects in the image and at this point it is possible to find the contours of each of them and thus extract information; for this purpose, *FindContours* command and the methods of the python package *OpenCV*[13] are used. The extracted features then are all of morphological type, in particular the following:

- area;
- perimeter;
- radius, it was computed after finding the minimum enclosing circle;
- circularity as: $C = \frac{4 \cdot \pi \cdot area}{perimeter^2}$;
- eccentricity as $E = \frac{\sqrt{a^2 - b^2}}{a}$ it was computed after performing an elliptical fit, so that $a = \frac{axis_+}{2}$ and $b = \frac{axis_-}{2}$;
- number of convexity defects.

The last feature is a measure of how much the contour of a cell is irregular. To compute such a feature, it is necessary to consider first the contour of the masked object and the convex hull (i.e. a convex curve around the object); any deviation of the object from this hull can be considered as convexity defect.

As an example of how the algorithm works, *Fig.*14 and *Fig.*15 provide the masked images of two cells. The first one with a jagged contour and the second one with a circular contour. As it is possible to see, the number of convexity defects is larger in the first case, while only one defect is present in the second case.

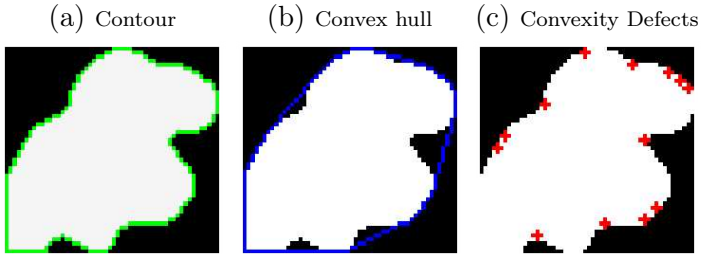


Figure 14: Steps of the algorithm to find the contour irregularities of an object. In this case the cell is jagged and presents a lot of irregularity points.

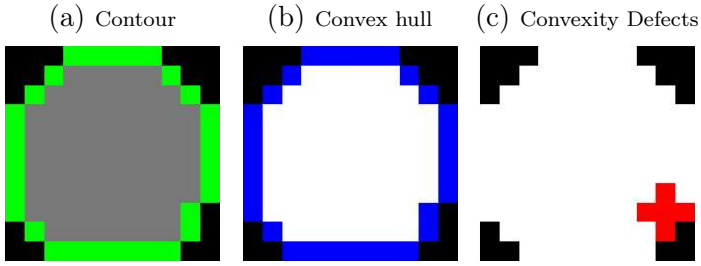


Figure 15: In this case the contour of the object is more regular and circular and indeed the convexity defect point detected is just one.

These objects features are computed for all five image configurations (samples without and with treatments) in the well plate B from the time 00h00m to time 23h00m - thus the clustering will be performed considering all the objects present in 120 images, which are in total 77702.

At this stage it is decided to restrict the period of the analysis to 24 hours, since the extracting features are only morphological; indeed, the first 24 hours, as it is

possible to see in the previous chapter, is a "stationary" state of the samples, i.e. cells do not split and do not merge yet and the number of objects stays quite the same; in order to consider a longer period, some kinematic and dynamical features should also be extracted and taken into account to make the clustering more accurate. This is done as after 24 hours cells can start to duplicate or can disappear (they can become debris).

Table I shows the first rows of the dataframe on which are collected all the information about the ~ 80000 objects which are used to perform AHC.
In addition to the morphological features described before, are collected also the identification number ("cell id"), provided by the watershed algorithm after performing the segmentation in the image, the treatment, the hour and the coordinates ("x position" and "y position", referred to the position of the centers of mass in pixel in the images).

cell id	treatment	hour	x position	y position	area (μm^2)	perimeter (μm)	radius (μm)	eccentricity	circularity	convexity defects
1	B2	0	933	11	203.54	58.31	9.92	0.80	0.75	5
2	B2	0	1019	8	111.86	39.86	6.82	0.54	0.88	6
3	B2	0	681	11	26.14	21.17	3.72	0.62	0.73	5
4	B2	0	198	15	328.66	81.17	13.02	0.48	0.63	8
5	B2	0	865	13	2.88	6.35	0.62	0.58	0.90	0
6	B2	0	1263	16	64.58	31.90	5.58	0.74	0.80	5

Table I: First rows of the dataframe which collects the extracted features of the objects. The clustering is then performed from these results.

After collecting all the data, it is necessary to preprocess them in order to get the best performance of the model; this includes to check for *outliers* (i.e. records that are very different from all others) or missing values and to avoid dependencies among features, in order to make the data less redundant as possible[22]. What it should be obtained is indeed an unbiased and representative data set, i.e. a data set

that contains all information about the inherent patterns and rules[23].

First of all the distribution of the numerical features collected before are shown as *boxplots*, which are a compact way to summarize the main characteristics of the samples and to find outliers[22]. The boxplots are built in this way for each of the respective feature: *(i)* the median value of the distribution is computed, *(ii)* the 25%- and the 75%- quartiles q_1 and q_3 are computed and a box limited by these two quartiles is drawn, *(iii)* the interquartile range (iqr) $q_3 - q_1$ is computed and the inner fence is defined as the two values $f_1 = q_1 - 1.5 \cdot iqr$ and $f_3 = q_3 + 1.5 \cdot iqr$, *(iv)* the smallest data point greater than f_1 is found and the largest data point smaller than f_3 and "whiskers" are added to the box extending a line to these data points[23]. The data points outside the whiskers will be considered as outliers. The distributions of the numerical features can be visualized in *Fig.*16.

As it is possible to see, there are a lot of outliers in the "area" distribution (a lot of records have an area value greater than 603.5 μm^2). A possible explanation could be that sometimes the segmentation procedure fails, and glued cells are detected as a single big-size agglomeration. These points are discarded as the values of cell areas are not plausible and they could affect the clustering afterwards.

It is also necessary to check the correlations among the features, using *Pearson correlation coefficient*[24,25]. The results are provided in *Fig.*17, showing the highest coefficients between the features "area", "perimeter" and "radius".

At this point, the preparation of the data set is ready; the outliers are removed and the features "perimeter" and "radius" are discarded from the analysis, since they are strongly correlated with the feature "area" and thus are redundant.
The last step of pre processing is to standardize the data, which is important in order

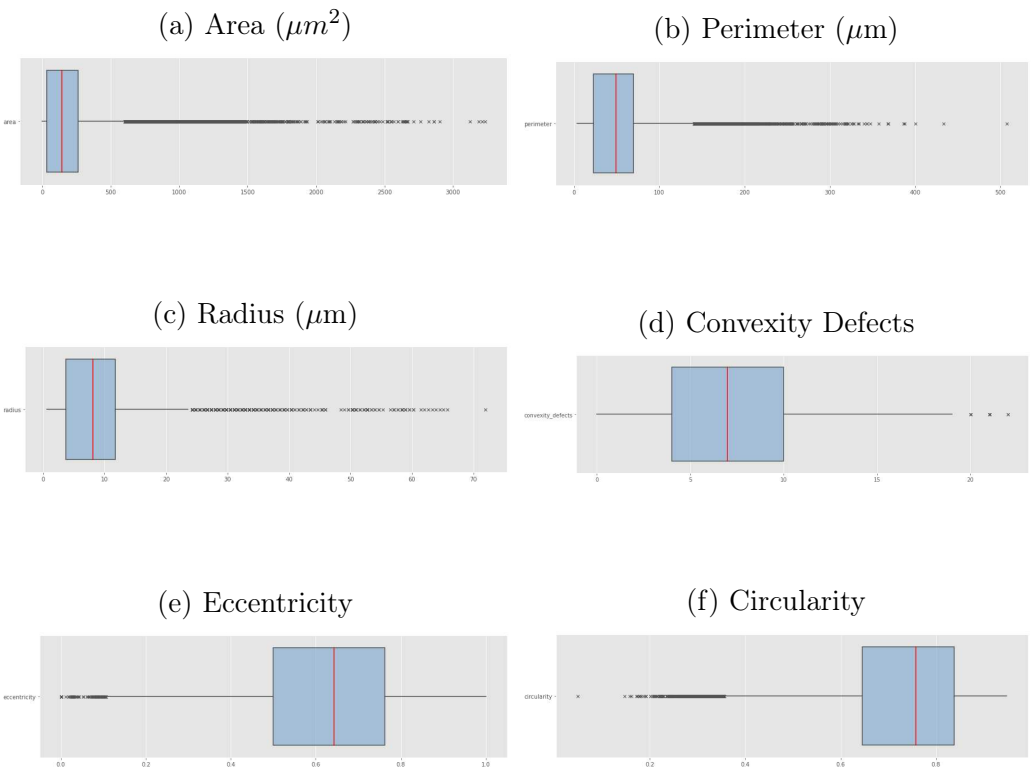


Figure 16: Boxplots of the extracted features. The blue box contains all the values among the quartiles q_1 and q_3. The red line inside the box refers to the median value of the distribution. The whiskers are added to the box and they are represented by a segment. The greatest amount of outliers is recorded for the "area". This is probably due to segmentation issues and it is necessary to discard them for the analysis.

to have all the features at the same scale. The class *sklearn.preprocessing.StandardScaler()* is used in order to implement *z-score standardization*.

Z-score standardization: given a feature X with mean sample μ_X and standard deviation σ_X, the z-score standardization is defined as:

$$s : \text{domX} \to \mathbb{R}, \qquad x \to \frac{x - \mu_X}{\sigma_X}$$

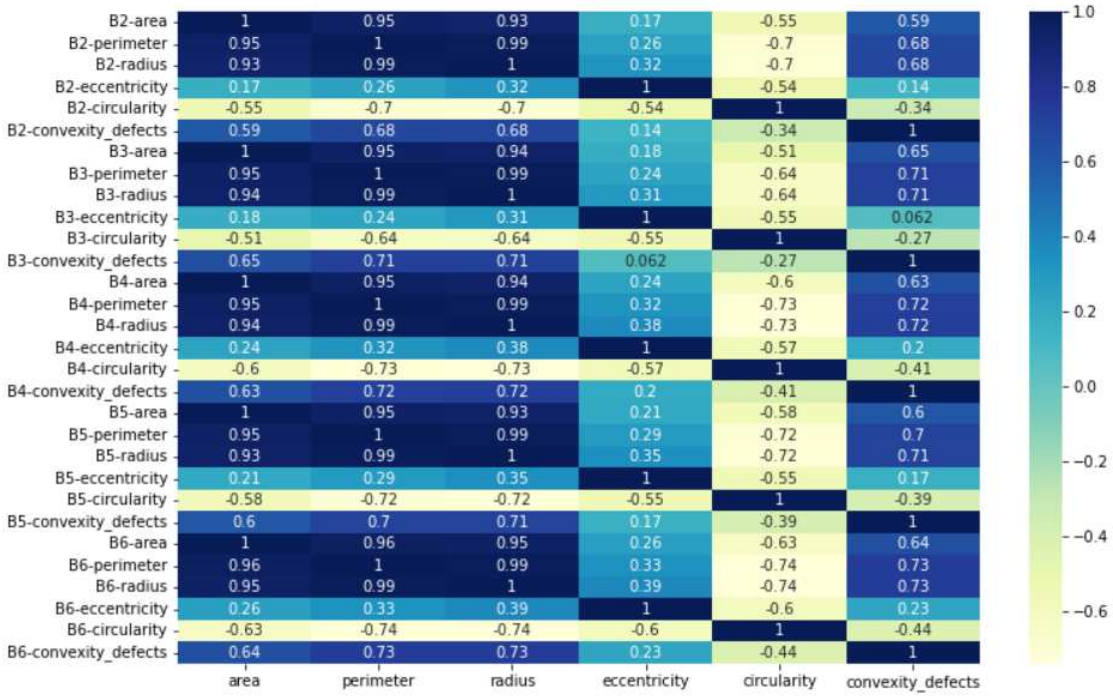


Figure 17: Correlation matrix of the features for each treatment. The highest values of the Pearson correlation coefficients are recorded for pairwise of the features "area", "perimeter" and "radius".

The preprocessed and standardized distributions of the selected features are shown in the boxplots in the figure below *Fig.*18.

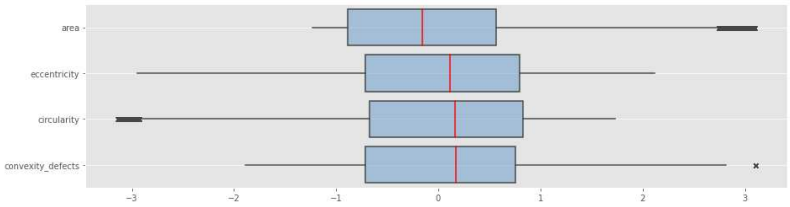


Figure 18: Boxplots of the standardized features. These are the results of the data preparation steps, which include removal of outliers and redundant features. Standardization is carried out with z-score method.

2.3 Number of Clusters Decision Making

The algorithm described in the *Agglomerative Hierarchical Clustering* Section is implemented selecting the linkages methods *Ward* and *Average*. The *Euclidean* distance metric is then selected, which is suitable for low dimension data set.
To choose the best number of clusters (classes of objects), the AHC with the selected parameters is performed, letting it vary from 2 to 10 and evaluating the results of three different scores, namely *(i)* Silhouette score[26], *(ii)* Calinski-Harabasz index[27] and *(iii)* Davies-Bouldin score[28]. A short definition is given for each of these indexes scores, according to the annotation of the article *"Analysis of Clustering Evaluation Considering Features of Item Response Data Using Data Mining Technique for Setting Cut-Off Scores"*. Symmetry (2017)[29]:

Silhouette Score: "the Silhouette Score, represented by the Silhouette index (SI), calculates a measure for each point k. This measure is based on the membership of the point in any cluster and is then averaged over all observations:

$$SI_k = \frac{1}{n}\sum_{i=1}^{n}\frac{b_i - a_i}{max(a_i, b_i)}$$

where n is the total number of points, a_i is the average distance between point i and all other points in its own cluster (*mean intra-cluster distance*), and b_i is the minimum of the average dissimilarities between i and points in other clusters(*mean nearest-cluster distance*)".

The optimal cluster partition is the one with the highest *SI*.

Calinski-Harabasz Index: "the Calinski-Harabasz index (CH index) assesses cluster partition performance by considering the average between-cluster (or inter-clusters) sum of squares and within-cluster (or intra-clusters) sum of squares:

$$CH(k) = \frac{[traceB/K - 1]}{[traceW/N - K]} for K \in \mathbb{N}$$

where K is the number of cluster, N is the total number of data points, B represents the error sum of squares between inter-cluster and W indicates the squared

differences of all objects in a cluster from their respective cluster centers".

Again the higher value corresponds to the better cluster partition.

Davies-Bouldin Index: "the Davies-Bouldin index (DB index) is expressed as follows:

$$DB_k = \frac{1}{k}\sum_{i=1}^{k} \max_{\substack{j=1,\dots,k \\ i\neq j}} \left\{ \frac{diam(c_i) + diam(c_j)}{d(c_i, c_j)} \right\}$$

where k denotes the number of clusters, i and j are cluster labels, and $d(c_i, c_j)$ denotes the distance between centers of cluster c_i and c_j. The diameter of a cluster is defined as:

$$diam(c_i) = \left(\frac{1}{n_i} \sum_{x \in c_i} \|x - z_i\|^2 \right)^{1/2}$$

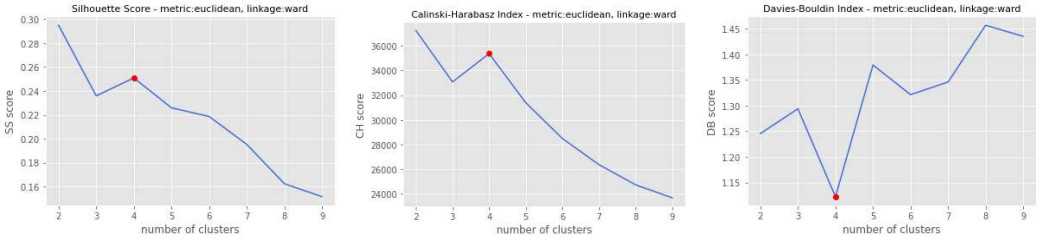
where n_i is the number of points and z_i is the center of cluster c_i".

In this case lower values indicate better separated clusters and the minimum score is zero.

*Fig.*19 and *Fig.*20 show what is obtained for *Ward* and *Average* linkage respectively. In the first case, although the highest values of *SI* (*Fig.*19a) and *CH* (*Fig.*19b) score are observed for a number of clusters equal to 2, it is also recorded a local maximum value for 4 clusters in both cases. Two clusters of objects are plausible if just the first hours are considered, when the objects in the images are recognizable as round-shaped cells just seeded in the well plate and debris. After the first few hours, as it is also possible to see in the previous section, cells start to change their size and their morphology, so what it should be expected is a more heterogeneous ensemble of objects which can be classified in more than 2 clusters. Since the lowest value of DB score is also recorded for 4 clusters (*Fig.*19c), for these reasons, the number of clusters is set equal to 4 for the AHC implementation with the linkage *Ward*.

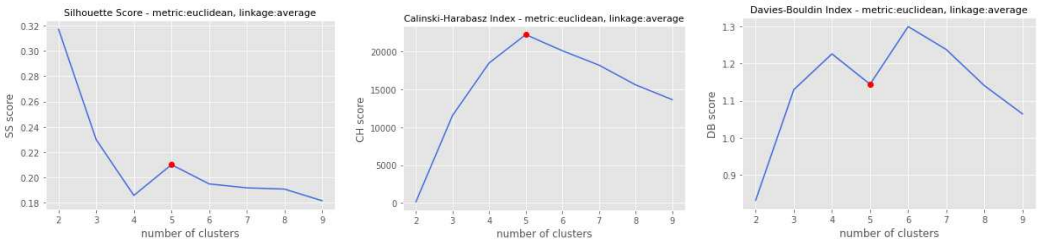
Similar considerations are made for the results referred to *Average* linkage (*Fig.*20). In this case SI score has the maximum value for 2 clusters, but a local maximum is also recorded at 5 clusters, while CH and DB scores respectively have a maximum

and a minimum value for 5 clusters; this value is chosen as the optimal to perform AHC with *Average* linkage.



(a) SI score (b) CH index (c) DB index

Figure 19: Results of the three different scores described to evaluate the best number of clusters to perform AHC with linkage *Ward*. As it is possible to see SI and CH index present the highest value for 2 clusters, but they also both show a local maximum for 4 clusters. Also DB index show the lowest value (which is considered the best for the partitioning) at 4 clusters. Since from preliminary considerations it should be expected to classify objects until 24 hours in more than two clusters, the number of clusters is set equal to 4 for the AHC with *Ward* linkage.



(a) SI score (b) CH index (c) DB index

Figure 20: Results for the linkage *Average*. Also in this case SI presents the highest value for 2 clusters, but it shows a local maximum for 5 clusters. Since both CH and DB scores present respectively a maximum and a minimum value for 5 clusters, the number of clusters is set equal to 5 for the AHC with *Average* linkage.

2.4 AHC Implementation

2.4.1 Ward Linkage

The results of the Agglomerative Hierarchical Clustering performed with *Euclidean* distance, *Ward* linkage and number of clusters equal to 4 are now shown in *Fig.*21 as a pairplot of the features. The mean values of the features for each cluster are provided in *Table* II.

CLUSTER	area (μm^2)	perimeter (μm)	radius (μm)	eccentricity	circularity	convexity defects	fraction of elements
0	314.31	79.14	13.37	0.73	0.61	9.51	0.32
1	88.69	36.93	6.37	0.75	0.71	5.67	0.23
2	176.02	51.58	8.39	0.50	0.80	9.58	0.25
3	28.63	18.72	2.89	0.48	0.85	3.27	0.20

Table II: The table shows the mean values of the features which characterize each cluster, obtained performing AHC with *euclidean* metric, *ward* linkage and 4 clusters. The fraction of objects over the whole data collected in 24 hours is also included.

As it is possible to observe, looking both at the pairplot and at the table, the objects in the clusters can be characterized in the following way:

- *Cluster 0* is composed of objects with the greatest size, therefore it is possible since now to classify them as cells (the mean value of the area is widely above the threshold used to identify debris), with an irregular shape-both eccentricity and circularity have high values-and jagged contours;
- *Cluster 1* is composed of small objects with irregular shape;
- *Cluster 2* is composed of objects which can be classified as cells and they are also round shaped, according to the values of circularity and eccentricity;
- *Cluster 3* is composed of very small objects with a circular shape and a regular contour: this is most likely the cluster of the debris.

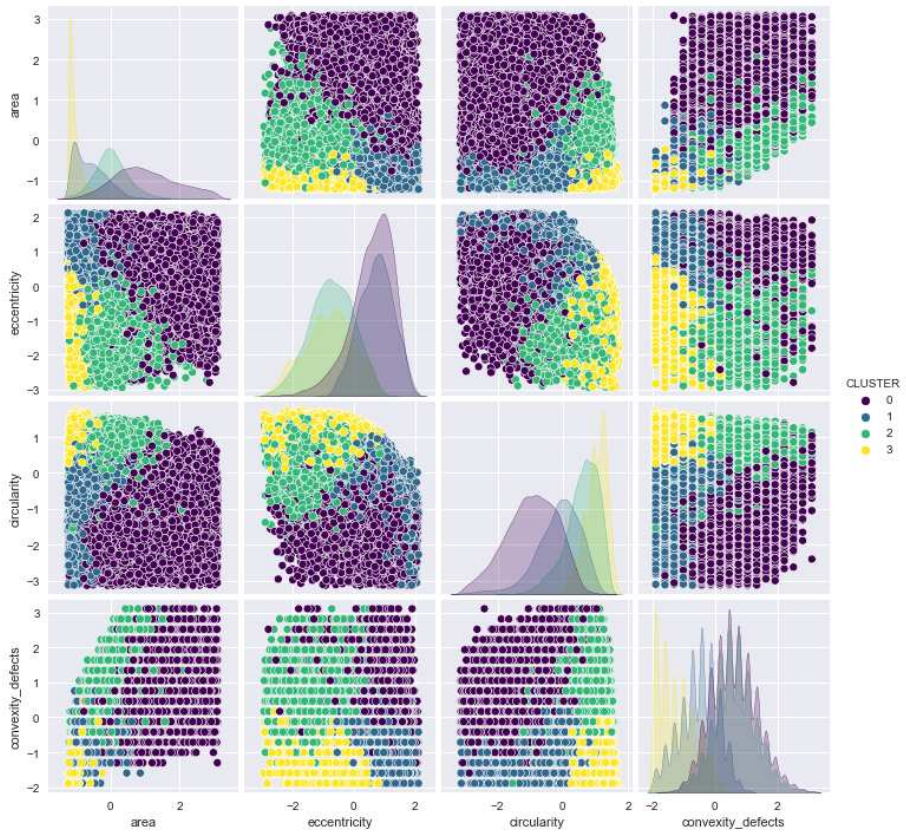


Figure 21: Pairplot of the standardized features, obtained with AHC (metric: euclidean, linkage:ward, number of clusters:4). Coloring indicates the clusters and in the diagonal the *pdf*s (probability density function) of the respective features are shown.

To visualize how the clusters of objects found look like, *Fig.*22 shows as an example these clusters for the control configuration (sample without treatment) at time 12h00m.

Once that the clustering is performed, it is possible to analyze in an image what object belongs to what cluster and so compute how many objects of a given cluster there are in that image (i.e. at the corresponding hour). By doing this it is also possible to study how the percentages of elements in each cluster evolve during 24 hours, in order to make some conclusions about the classification of the objects.

For this purpose barplots are used; a specific configuration (treatment) is set and the bar at each hour is drawn, where the different colours stand for the percentage of cell in a cluster. The results can be seen in the figures *Fig.*23.

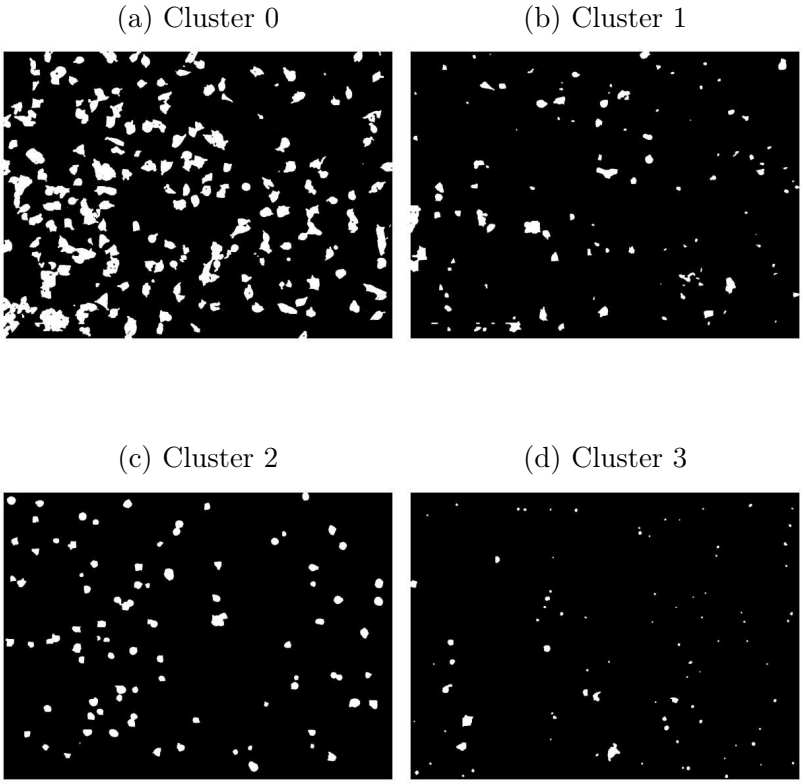


Figure 22: Clusters of objects obtained performing the AHC with *Ward* linkage; it is possible to recognize (a) cluster 0: big and jagged cells, (b) cluster 1: small and irregular objects, (c) cluster 2: medium and circular cells, (d) cluster 3: small and circular objects that can be classified as debris.

In order to understand how the amount of treatment influences the evolution of the objects in the clusters, also the linear plots of the percentages over time for each cluster are shown in *Fig.* 24. To make the results comparable for all the configurations, since the number of cells of the samples is different in the well plates with different concentration of treatment, the percentages of objects of each cluster are normalized, dividing the values of all the hours by the value of percentage of objects in that cluster computed at the first hour.

Looking at the results it is now possible to make an interpretation of the classified objects:

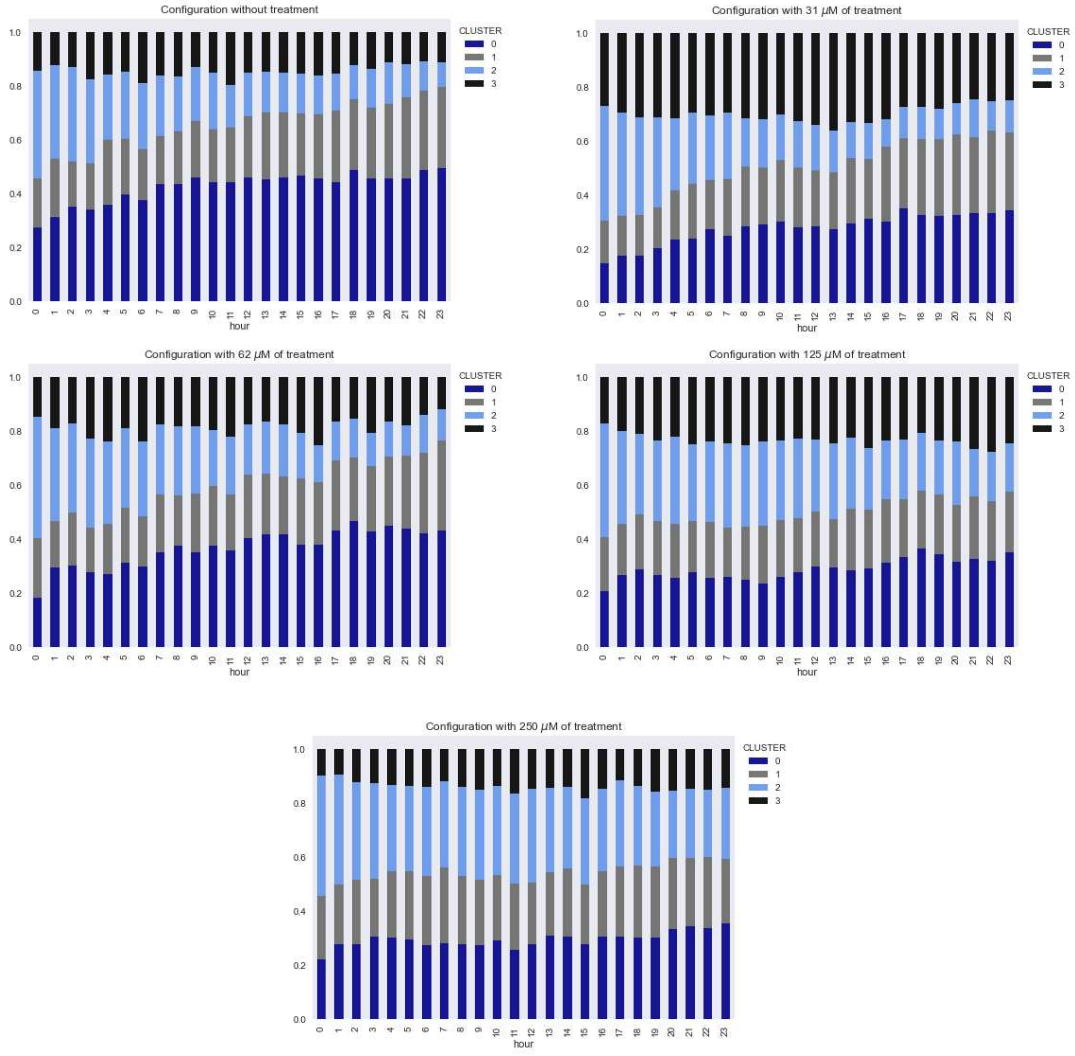
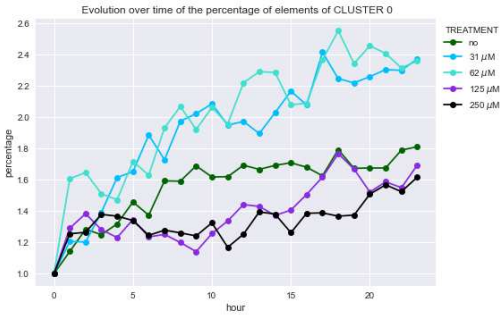


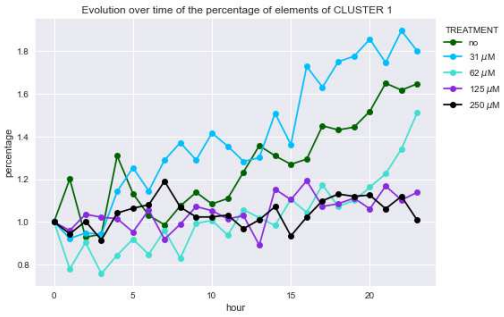
Figure 23: Barplots for the 5 different configurations (samples without and with different level of treatment), each bar represents with different colours the percentages of cells in the clusters.

- *Cluster 0* (*Fig.*24a) is the cluster composed of cells with the biggest mean size, irregular shape and jagged contours. Because of these characteristics they could be identified either as cells that are going to split or agglomeration of cells that are merging. The number of cells which belong to this cluster keeps increasing over time in all the configurations, even if the increase during 24 hours is higher for those ones with low and medium level of treatment and lower for the configurations without treatment and high level of treatment.

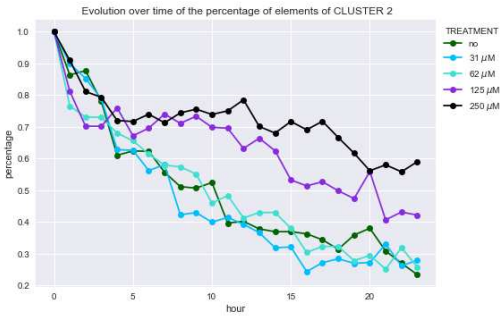
(a) Cluster 0



(b) Cluster 1



(c) Cluster 2



(d) Cluster 3

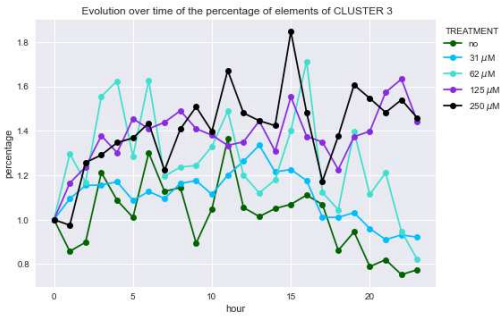


Figure 24: Linear plot of the normalized percentages of cells in the clusters over time. The number of objects in each cluster is normalized with the number of objects at time 00h00m and the values at all the hours are divided for the same number. In this way it is possible to make a comparison of how the percentages of elements of each cluster vary according to the treatment of the sample and to compare how the amount of cells varies respect to different level of treatment.

The possible explanation in this last case is that, because of the high level of H_2O_2, cells lose the possibility of duplicate as they are less vital and therefore their morphology does not change over time more than when they were seeded; this is also compatible with the preliminary result obtained in the *Chapter 1* for the median of the areas, which does not increase during the first 20 hours. In the case of the configuration without treatment instead the explanation could be that, since they are left to split without any poisoning, they quickly keep

dividing and changing morphology, so that during the 24 hours the sample is always very heterogeneous and a net increase of one type of cells is not recorded. Finally, for the configurations with 31 μM and 62 μM the maximum increase of cells is observed in cluster 0. The explanation could be that, since they are not poisoned as in the cases of higher amount of H_2O_2, they are able to enlarge and to split, but they are slower in comparison to the case of no treatment and so during 24 hours a net increase of these cells is recorded;

- *Cluster 1* (*Fig.*24b) is composed of small and irregular objects, it can be interpreted as a class of small cells which are changing from their rounded shaped and/or debris. The number of elements increases for the configurations without and with the lowest level of treatment, where a lot of irregular and jagged cells would be expected, while it stays constant in the other configurations;

- *Cluster 2* (*Fig.*24c) is made of round shaped cells, which is the morphology of cells just seeded in the well plate with culture. As expected, the number of cells in this cluster keeps decreasing over time, although with different slope for different treatments. Indeed, as it can be observed also looking at the barplots in *Fig.*23, the decrease is evident in the samples with low level of treatment, where the most part of cells change their morphology, while it is less rapid for configuration with high level (125 μM and 250 μM) of H_2O_2. This in agreement with the interpretation given before, that cells in these samples are poisoned and they are forced to not change and move;

- *Cluster 3* (*Fig.*24d), was already identified as the cluster of the debris, because of the mean characteristics of its elements which are very small, round shaped and with regular contours. The trend of the relative fraction is noisy, and a net increase or decrease of debris is not recorded during 24 hours and the initial amount stays quite constant in each sample.

2.4.2 Average Linkage

Now the results of AHC with *Average* linkage and number of clusters equal to 5 are provided in *Fig.*25 and *Table* III.

CLUSTER	area (μm^2)	perimeter (μm)	radius (μm)	eccentricity	circularity	convexity defects	fraction of elements
0	365.14	88.30	14.67	0.70	0.58	9.56	0.1800
1	424.11	79.58	12.60	0.34	0.83	15.40	0.0007
2	186.77	54.24	9.09	0.62	0.76	9.07	0.4800
3	31.01	19.44	3.00	0.46	0.84	3.29	0.2000
4	69.31	34.25	6.01	0.82	0.65	4.85	0.1400

Table III: The table shows the mean values of the features which characterize each cluster, obtained performing AHC with *euclidean* metric, *average* linkage and 5 clusters as well as the fraction of objects over the whole data collected in 24 hours.

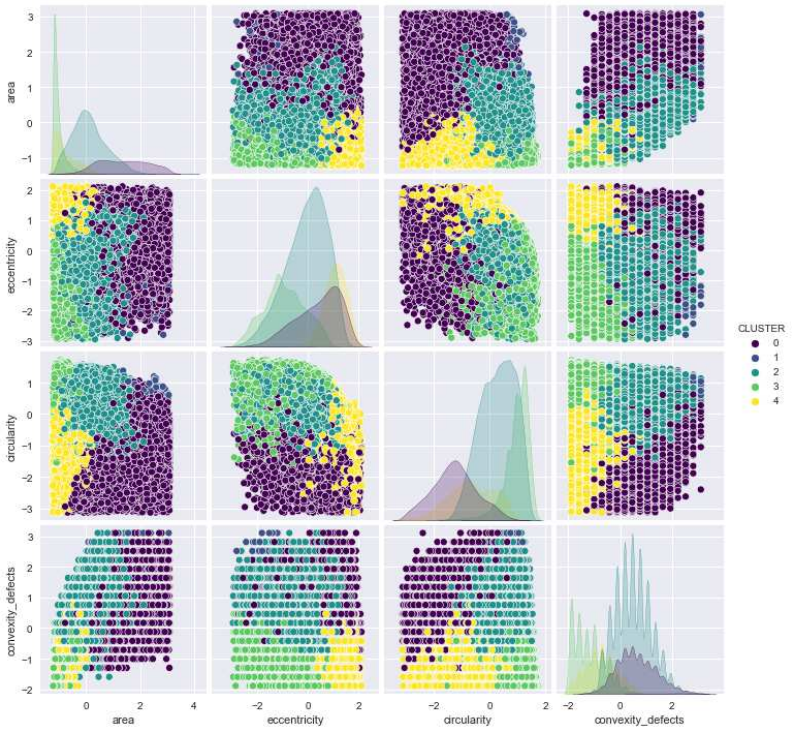


Figure 25: Pairplot of the features, obtained with AHC (metric: euclidean, linkage:average, number of clusters:5). Coloring indicates the clusters and in the diagonal the *pdf*s (probability density function) of the respective features are shown.

Again, looking at the plot and at the mean values it is possible to give a first characterization of the clusters:

- *Cluster 0* is composed of cells with high values of area and also of eccentricity and convexity defects, so they have an irregular contour;
- *Cluster 1* is composed of objects with the highest values and convexity defects, but looking at the percentage of objects classified in this cluster, they are very few (namely just 58 objects over 77702). This means that this is a cluster of outliers and for this reason they are discarded from the collective analysis of the evolution of percentage of objects in the clusters for the different treatment;
- *Cluster 2* is characterized by objects with mean size and with an higher value of circularity rather than eccentricity;
- *Cluster 3* is composed by the smallest objects with a round shape, so it can be identified as the cluster of debris;
- *Cluster 4* is composed of small and irregular objects.

The masked objects that belong to the clusters are shown in *Fig.*26, considering as an example the same image of before, referred to the cell sample in the plate B2 without treatment at hour 12h00m.

*Fig.*27 and *Fig.*28 show the results of the evolution of the percentages of elements in each cluster for the different configurations and they are used for a comparison with the results obtained before.*Cluster 1*, which can be considered a cluster of outliers, is discarded.

Looking at the barplots and the linear plots, it is possible to recognize similar trends of the percentages of elements inside the clusters for each treatment as the ones obtained performing AHC with *Ward* linkage, namely *(i)* the percentage of

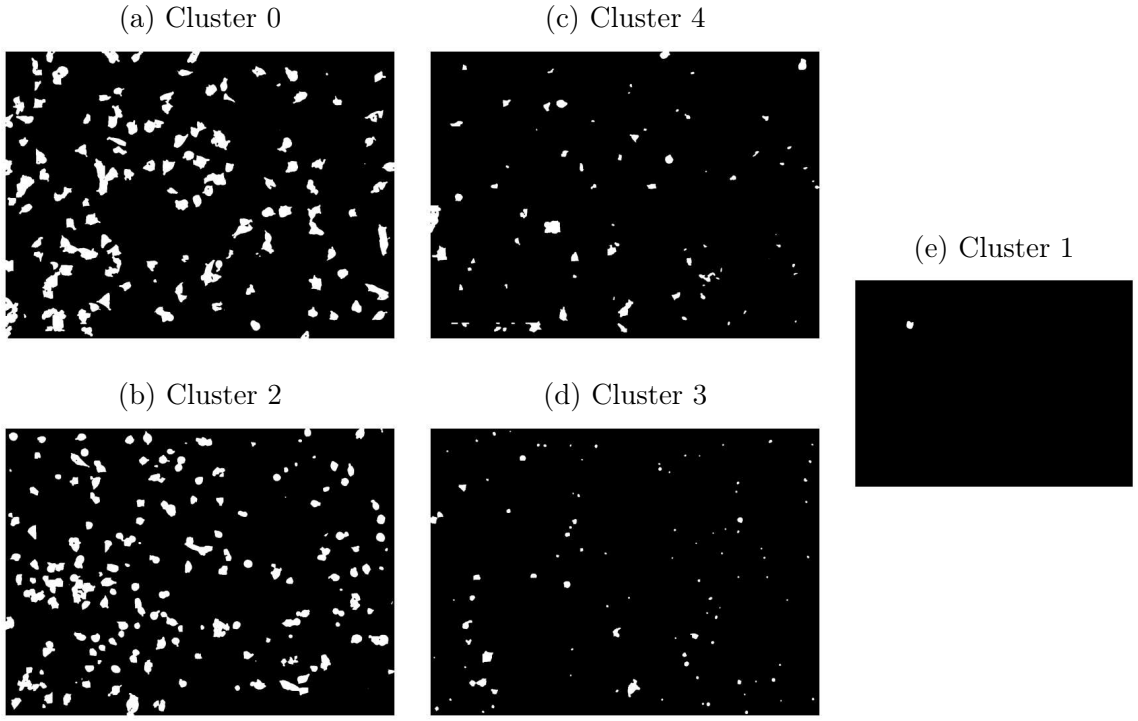


Figure 26: Clusters of objects obtained performing the AHC with *Average* linkage; again it is possible to recognize (a) cluster 0: big and jagged cells, (b) cluster 2: medium-size cells, (c) cluster 3: irregular objects, (d) cluster 4: small and circular objects that can be classified as debris; selecting 5 number of clusters also (c) cluster 1 is obtained, which contains very few objects, characterized by high values of area.

Cluster 0, which refers to the big and jagged cells, increase over time with an higher slope for the configurations with low level of treatment; *(ii)* the percentages in Cluster 2, which is probably the cluster of the medium-size cells, decrease over time; *(iii)* the percentages in Cluster 4, i.e. of small and irregular cells, increases over time also this time; *(iv)* percentages of Cluster 3, which is the one of debris, fluctuate a lot, but at the end are proportional to the amount of treatment used. This means configurations which are subject to low or no treatment show less debris while larger amounts of treatment show more debris after 24 hours. The results of the AHC performed with the two linkage methods described are therefore comparable, thus this suggests that the objects in the found clusters can be classified based on their

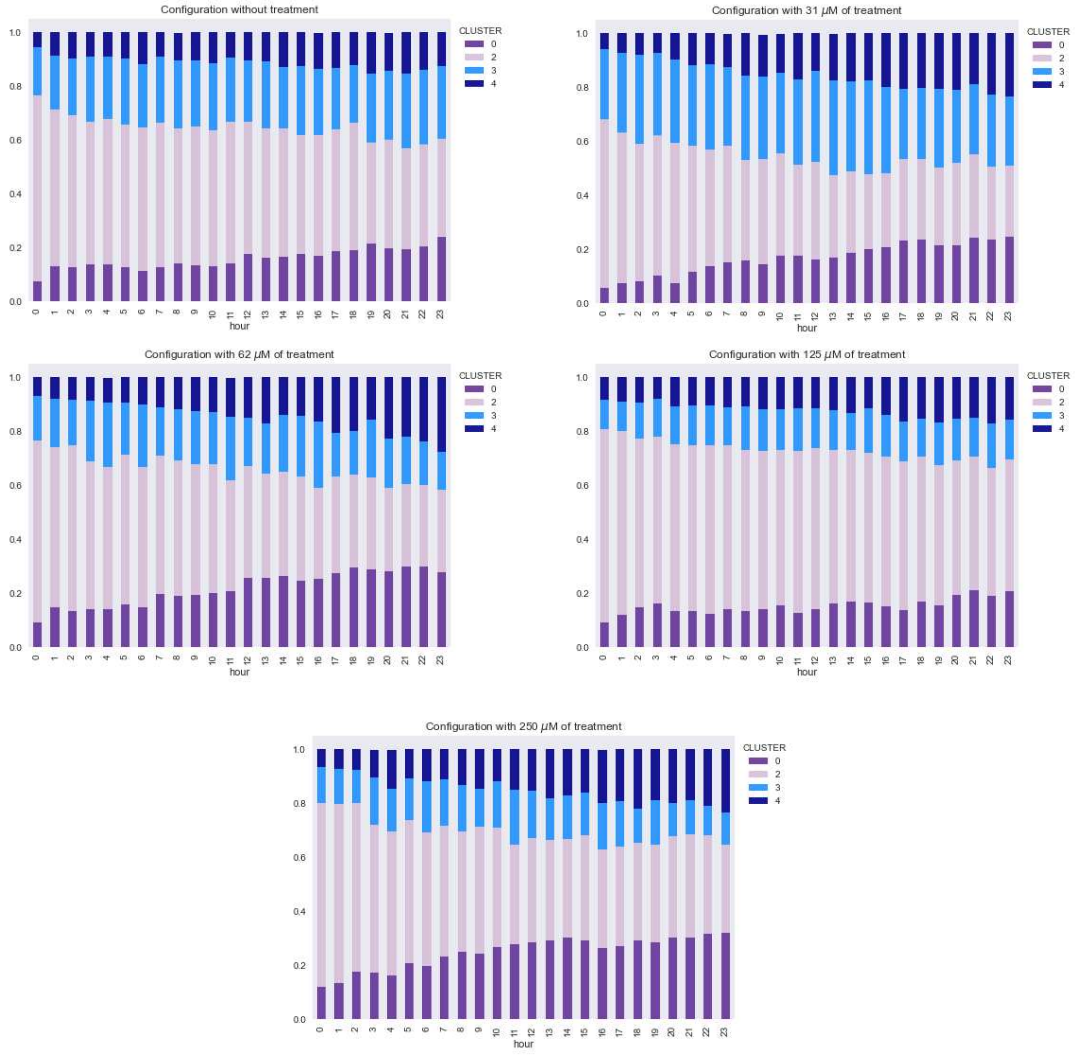


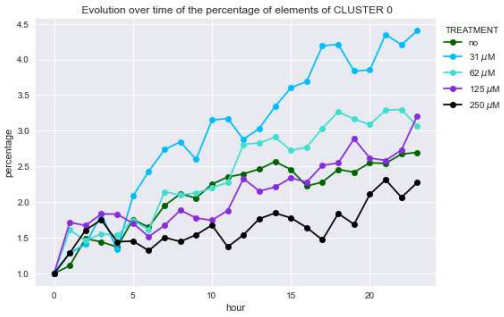
Figure 27: Barplots for the 5 different configurations (samples without and with different level of treatment), each bar represents with different colours the percentages of cells in the clusters.

features.

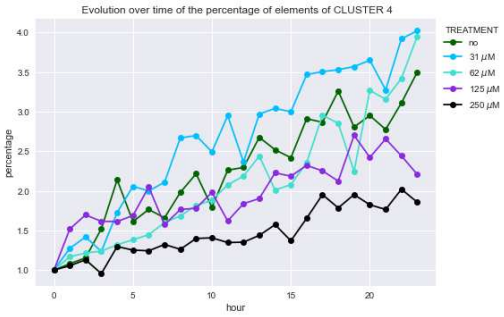
As a trial, the AHC is also performed with *Single* and *Complete* linkage[30]; in the first case the distance corresponds to the distance of the closest pair of elements which belong to the two clusters, while in the second case it corresponds to the distance of the farthest pairwise.

What it is obtained by performing the AHC with *Single* linkage is not consistent, since the score evaluation returns 6 as best number of clusters, but the resulting

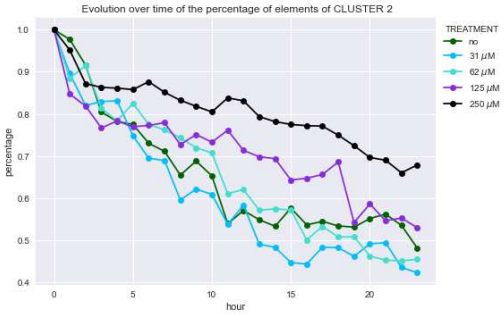
(a) Cluster 0



(b) Cluster 4



(c) Cluster 2



(d) Cluster 3

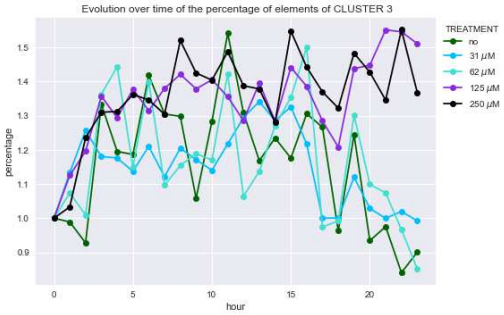


Figure 28: Linear plot of the normalized percentages of cells in the clusters over time. The number of objects in each cluster is normalized with the number of objects at time 00h00m and the values at all the hours are divided for the same number. In this way it is possible to make a comparison of how the percentages of elements of each cluster vary according to the treatment of the sample and to compare how the amount of cells varies respect to different level of treatment.

clusters are composed such that 5 of them are populated by just 1 element and the sixth one collects the remaining objects all together. The AHC performed with *Complete* linkage returns results similar to the AHC with *Ward* linkage, meaning 4 clusters having the same characteristics.

3 Clusters Characterization

Purpose of this section is to find the characteristic features of the clusters found in the previous section performing AHC, in order to define them. To this end, the hypergeometric distribution is employed to evaluate the likelihood that a specific feature is over-expressed within the elements of the cluster[6].

In addition, an alternative approach is implemented, performing the clustering image by image, so not considering anymore the objects of each treatment and each hour all together, but applying the AHC separately hour by hour. Taking into account the results obtained in the previous section, 4 clusters of objects are expected at each hour, which motivates to set the number of clusters equal to 4 every time the AHC is performed for the objects in an image, and then to understand how cluster characteristics change during 24 hours or if a continuity exists, analyzing the characterizing features and comparing the results with the ones obtained previously.

3.1 P-value Test for Over-Expressed Features

The purpose now is to check if some features in the clusters found before are over-expressed, i.e. if there are some features that are characterizing in each cluster.

To implement the cluster characterization, the values of each feature are split in three intervals, with the help of their standardized distributions computed in the second chapter. For example, it is possible to look at the boxplots in *Fig.*18 and assign a feature the label *"1"* if its value is included among the lower whiskers and the lower quartile, the label *"2"* if it is included among the lower and the upper quartile (the box), and the label *"3"* if it is included among the upper quartile and the upper whisker. Therefore, features with the label 1 refer to low values of that feature, with label 2 medium and with label 3 high (e.g. if one object is characterized by the feature "area 1" it means it has a small size and so on). Since each feature is split each in three intervals, they are now 12 features in total.

To check if a feature is over-expressed among the elements of a cluster and therefore characterizes that cluster the procedure described in the article: *"Community characterization of heterogeneous complex systems*, Journal of Statistical Mechanics Theory and Experiment (2011)"[6] is followed. It is briefly presented here.

The overall number of objects is denoted as N and the number of objects within a cluster C is denoted as N_C.The total number of features is denoted as N_A - in this case, it is equal to 12. For each feature $Q \in N_A$, a test is conducted to determine if Q is over-expressed in the cluster C. This involves checking if the number $N_{C,Q}$ of objects in the cluster C with attribute Q is significantly larger than what it would be obtained by randomly selection of N_C objects from all the N objects of the system.

The probability that X objects, randomly selected in the cluster C, have the feature Q under the null hypothesis, is given by the *hypergeometric distribution*[31]:

$$H(X \mid N, N_C, N_Q) = \frac{\binom{N_C}{X}\binom{N-N_C}{N_Q-X}}{\binom{N}{N_Q}}$$

where, as said, N is the total number of objects, N_C is the number of objects in the cluster C and N_Q is the number of objects in the system with the attribute Q (see *Fig.*29).

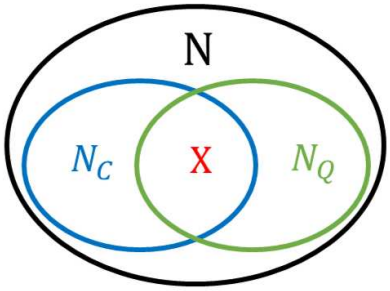


Figure 29: Scheme of the parameters that appear in the hypergeometric distribution used to test the over expression of the features in a cluster. N is the total number of objects in the system, N_C is the number of objects in the cluster C, N_Q is the number of objects in the system that own the feature Q and X is the number of objects in the cluster C which have the feature Q.

The *p-value* is then associated to the number $N_{C,Q}$ computed as:

$$p(N_{C,Q}) = 1 - \sum_{X=0}^{N_{C,Q}-1} H(X \mid N, N_C, N_Q)$$

if the value of $p(N_{C,Q})$ is smaller than the threshold equal to $p_t = 0.01$, the feature Q is recognized as an over-expressed feature which characterizes the cluster C.

The test is performed for all the features, a data frame for the results is created where the number of the cluster (labeled from 0 to 3) in which they result characterizing is collected together with the respective value of $p(N_{C,Q})$. As an example, *Table*IV shows how the data frame looks like for the procedure applied to the clusters found selecting the *Ward* linkage.

Some features are characterizing in more than one cluster and one cluster can also be characterized by more than one feature.

features	clusters
area_1	3, 1
area_2	2, 1
area_3	0
circularity_1	0
circularity_2	2, 1
circularity_3	2, 3
convexity_defects_1	1, 3
convexity_defects_2	1, 0, 2
convexity_defects_3	0, 2
eccentricity_1	2, 3
eccentricity_2	1, 0*
eccentricity_3	0, 1

Table IV: Table with the clusters for which the features are characterizing. The p-values are equal to zero for all the characterizing features except that for eccentricity_2* in cluster 0 where a p-value of 0.002 is obtained.

It is also possible now to characterize the clusters found in the previous chapter, performing the AHC with *Ward* and *Average* linkage.
The obtained results are listed in the first case:

- *Cluster 0*, whose percentage of elements increases during time and which was identified as the cluster of cells that are going to split, is characterized by the features:

  area 3,circularity 1,convexity defects 2,convexity defects 3,eccentricity 2,eccentricity 3

  this means that the objects in this cluster have a big size, jagged contours and elliptic shape;

- *Cluster 1*, which, analyzing the mean values, was identified as a cluster of small and irregular objects, is characterized, as expected, by the features:
  area 1,area 2,circularity 2,convexity defects 1,convexity defects 2,eccentricity 2,eccentricity 3;

- *Cluster 2*, which was classified as the cluster of the cells with a round shape,

the ones observed at the beginning when they are seeded in the well plate and whose percentage of elements decreases over time, is characterized by the features:

area 2,circularity 2,circularity 3,convexity defects 2,convexity defects 3,eccentricity 1

therefore they show a big values of circularity (and low values of eccentricity) and medium size;

- *Cluster 3*, which finally was identified as the cluster of debris, is characterized by the features:

  area 1,circularity 3,convexity defects 1,eccentricity 1

  namely a cluster of small, circular objects with a regular contour.

The results of the analysis repeated using (*Average* linkage) are also listed:

- *Cluster 0*, is characterized by the features:

  area 3,circularity 1,eccentricity 3, convexity defects 2,convexity defects3

  therefore, except for the value "eccentricity 2", it corresponds to cluster 0 obtained selecting *Ward* linkage and so this is the cluster of the splitting cells;

- *Cluster 1*, is the cluster of outliers that was discarded in the previous analysis and it is characterized by the features:

  area 3,circularity 3,eccentricity 1,convexity defects 3;

- *Cluster 2*, is characterized by the features:

  area 2,circularity 2,eccentricity 2,convexity defects 2,convexity defects 3

  this combination of features is most similar to the one of cluster 2 obtained before, but they differ in the values: "circularity 3", which is not present in this case and "eccentricity 1" that in this case is "eccentricity 2". This means that the method with *Ward* linkage classifies objects better which are elements in the cluster with rounded and medium sized cells;

- *Cluster 3*, is characterized by the features:

area 1,circularity 3,eccentricity 1,convexity defects 1

so it perfectly corresponds to cluster 3 obtained with the first method and therefore it is the cluster of debris;

- *Cluster 4*, is characterized by the features:

  area 1,area 2,circularity 1,eccentricity 3,convexity defects 1

  it shows the most similar combination with the cluster 1 obtained before, but they differ in the values "circularity 2", that here is "circularity 1", "eccentricity 2" and "convexity defects 2" that here are not present. This means that this method rather than the method with *Ward* linkage, classifies objects better which are elements in the cluster of small or medium size objects with more regular or elliptic shapes.

3.2 Trends of Cluster Attributes over Time

In *Chapter 2* the AHC was performed collecting the features of all the objects of all the treatments of 24 hours, and at the end 4 clusters (classes) were obtained that can be then analyzed image by image, checking the elements that belong to the clusters at a certain hour. The data frame in the previous case was of ~ 80000 rows, which collected the information of the objects of 120 images all together.

Now an alternative approach is built, which performs the AHC separately image by image and characterizes the obtained clusters. Since image by image the population in the classes can change, and so the features that characterize them, the purpose is to check if there are some conserved combinations of over-expressed features during the 24 hours, that in this way would allow to recognize if there are recurrent characterized clusters in the images and analyze how they evolve during time.

In order to perform AHC, the same features of the objects selected during the

preprocessing of the data in *Chapter 2*, are collected in a data frame image by image, namely: "area", "eccentricity", "circularity" and "convexity defects". This way, each object is again defined by a vector of 4 values of features.
To implement the algorithm for each hour, the "Euclidean" distance and the "Ward" linkage are set and also the number of clusters is fixed equal to 4. This time, repeating the procedure image by image for 24 hours for each treatment, the data frames are now 120 with different length.

Once the test was performed for all the 24 hours, it is possible to check what are the characterizing features of the 4 clusters detected with the AHC image by image. It is now built a data frame where the characterizing features are collected (*Table* V).

	Img 0	Img 1	Img 2 ...
CLUSTER 0	area_2,circularity_2,circularity_3...	area_2,circularity_2,circularity_3...	area_2,circularity_2,circularity_3...
CLUSTER 1	area_2,circularity_1,eccentricity_3...	area_3,circularity_1,eccentricity_3...	area_3,circularity_1,eccentricity_3...
CLUSTER 2	area_1,circularity_3,eccentricity_3...	area_1,area_2,circularity_2...	area_1,circularity_3,eccentricity_1...
CLUSTER 3	area_3,circularity_1,eccentricity_3...	area_1,circularity_3,eccentricity_1...	area_1,circularity_1,eccentricity_3...

Table V: First three instances in the time series of AHC.

Since the AHC algorithm assigns the labels completely random to the clusters when it is performed in different images, this means that the combination of features that characterizes a cluster at time t can characterize a cluster labeled with a different number at time *t+1*. Moreover, a combination of features that characterizes a cluster in an image at time t can change at time *t+1*, since the population of objects can change image by image.
For these reasons, what happens is that the selected number of clusters is equal to 4 every time the AHC is performed in one image, but the total validated combinations of features during the all selected period of 24 hours are much more than

4, because they change over time. Then it is necessary to check if, among all the possible validated combinations, there are some recurrent ones or, at least, if there are similar groups of combinations.

Therefore, the purpose is to analyze if there are recurrent combinations of features over time, and they will be recognized as *characterized clusters*. Further, what can happen over time is that some combinations can differ in a few features, so that the objects included in those clusters are very similar, although the clusters are not validated as the same. For this reason what is done is to consider the list of all the possible validated combinations of features obtained so far, and perform again the AHC of the elements of the list (which are therefore strings of features) in order to find the most similar combinations of clusters. This procedure is equivalent to perform a clustering of clusters and allows to order the list of the validated combinations according to their similarity and will help to recognize the clusters over time, as it will be described later more in details.

To this end, the *Hamming matrix*[32] is computed, whose columns are indexed by the 12 features, while its rows are indexed by the combinations of features. The entries of the matrix are equal to 1 if the feature is present in the combination and equal to 0 if it is not. This matrix is used as metric distance to perform the AHC. A clustering of the validated combinations is performed, which groups "similar" combinations, based on how many features they have in common. When the AHC is performed, the selected number of clusters is also 4, because is what is expected, based on the previous analysis.

As an example, *Fig.*30 shows the result of a list of validated combinations of features which is obtained for a treatment on a set of 24 images, ordered according to the dendrogram leaves.

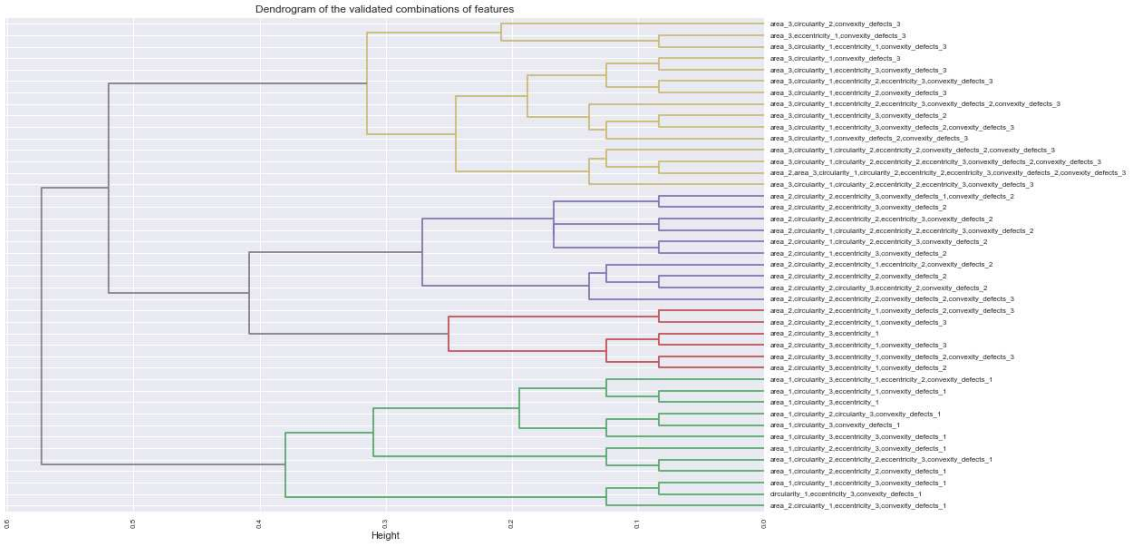


Figure 30: Dendrogram for the list of validated combinations of features; it is obtained performing the AHC with the hamming metric distance and a number of clusters equal to 4. In this way similar validated combinations, i.e. combinations with a certain number of features in common, are grouped in the same cluster and the list is ordered according to the leaves of the dendrogram.

Once the combinations of features were ordered in this way, it is possible to show the heatmap of the clusters over time. The heatmap presents vertically hour by hour 4 clusters as selected. The color of each cluster is referred to the percentage of elements in that cluster according to the legend in the bar in the left. In the right there is the corresponding combination of characterizing features of the cluster at that hour. Since the combinations of features are ordered as in the dendrogram, it is possible to follow which of them are preserved over time.The obtained results will be compared for the samples from the configuration without treatment to the one with the highest level (250μM).

For the configuration without treatment (*Fig.*31), which is the most heterogeneous over time, what is obtained is one combination of features that is persistent over time, with the characterizing features: *"area 1", "circularity 3", "eccentricity 1", "convexity defects 1"*. This is exactly the configuration of features obtained performing the AHC with the objects all together and for both *Ward* and *Average*

linkage methods, namely the cluster of debris.

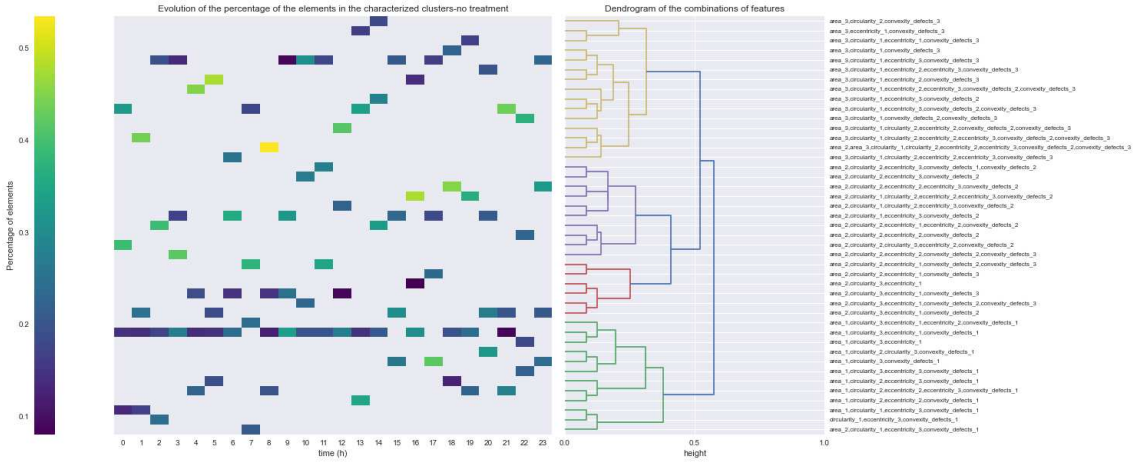


Figure 31: Configuration without treatment: this is expected to be the most heterogenous configuration and indeed over the 24 hours, the obtained clusters are characterized by different combinations of features; the only one which can be followed over time and that keeps constant the characterizing attributes is the leave in the green group of the dendrogram, with the features "area 1", "circularity 3", "eccentricity 1", "convexity defects 1", i.e. small objects with a regular rounded shape, which can be classified as debris.

For the configuration with the lowest level of treatment, *Fig.*32, two conserved over time combinations of features were obtained (green and red leaves of the dendrogram), the first one is again the one referred to the cluster of debris, the second one instead *"area 3", "circularity 1", "eccentricity 3", "convexity defects 3"*, which is the most similar to the combination of features of cluster 0, obtained with the previous method with *Ward* and *Average* linkage; it differs only in the features "eccentricity 2", "convexity defects 2", but it can be identified as the same cluster of cells with big sizes and irregular, jagged contours.

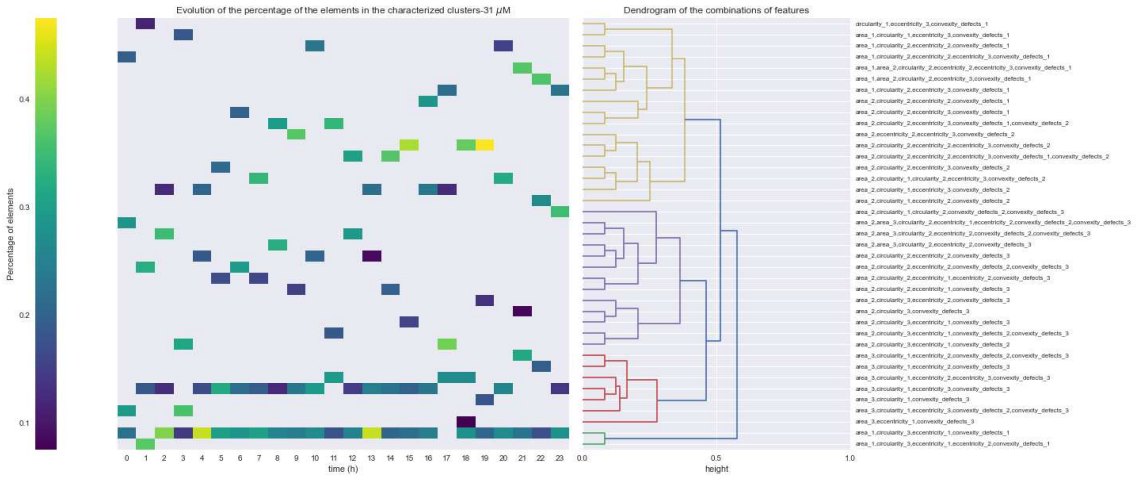


Figure 32: Configuration with 31 μM of treatment: in this case two conserved over time combinations of features can be recognized, namely the one equal to the previous case (configuration without treatment) which identify the debris cluster, the other composed by "area 3", "circularity 1", "eccentricity 3", "convexity defects 3", i.e. cells with a big sizes and irregular, jagged contours. These are the cells which can be identified as the ones that attach at the bottom of the well plate and enlarge their size until they split.

*Fig.*33 and *Fig.*34 show the results: again is obtained a continuity of the combination of features that can be identified as the cluster of debris and the cluster of big cells with elliptic shape and irregular, jagged contour that can be classified as the cells glued to the well plate that are going to split.

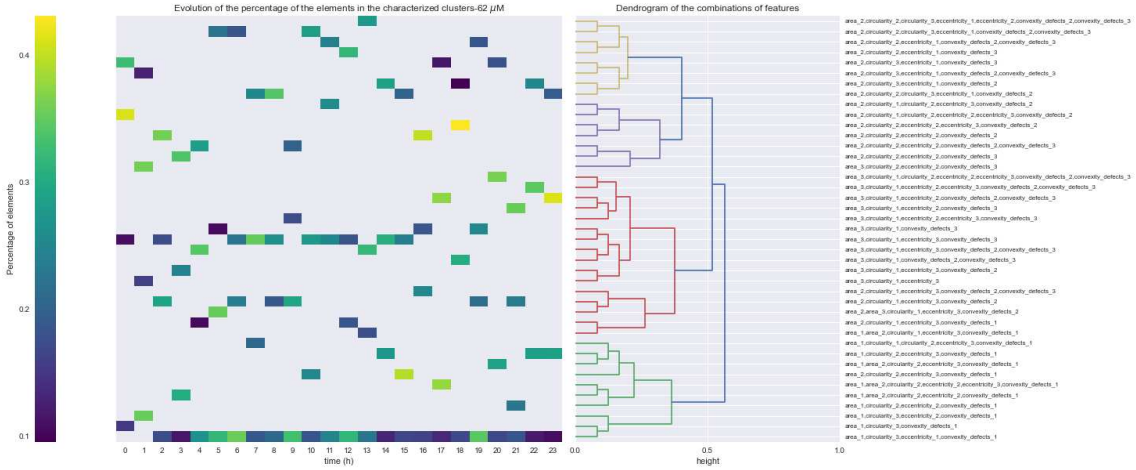


Figure 33: Configuration with 62 μM of treatment: also in this case the cluster of the debris can be recognized and a certain continuity of the combinations of features which identify the big and jagged cells.

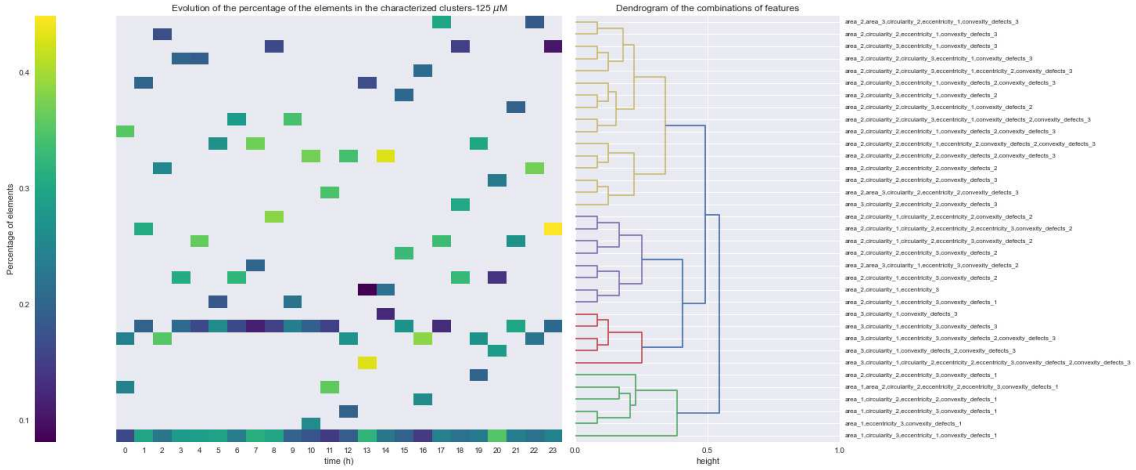


Figure 34: Configuration with 125 μM of treatment: again the cluster of the debris can be recognized and a certain continuity of the combinations of features which identify the big and jagged cells.

Finally the results for the configuration with the maximum level of treatment are provided; in this case again the two clusters revealed before can be observed, but in addition a certain continuity with high percentage of elements exists also for the combination of features that correspond to the green leaves of the dendrogram in *Fig.*35, i.e. *"area 2", "circularity 2", "circularity 3", "eccentricity 1", "eccentricity 2", "convexity defects 2"*. This combination of features is the most similar to the combination of features of the cluster 2 obtained before with *Ward* and *Average* linkage; they differ respectively in the first case for the features " convexity defects 3" and "eccentricity 2", while in the second case for "circularity 3", "eccentricity 1", "convexity defects 3". This group can be therefore identified as the cluster of rounded shape cells that is observed at the beginning (time 00h00m) and decreases over time . This is compatible with the results obtained in the previous section where it was observed that for the treatment with the highest level of H_2O_2 the percentage of this type of cells decreases over time with the lowest slope and so it is probable to observe them during the 24 hours..

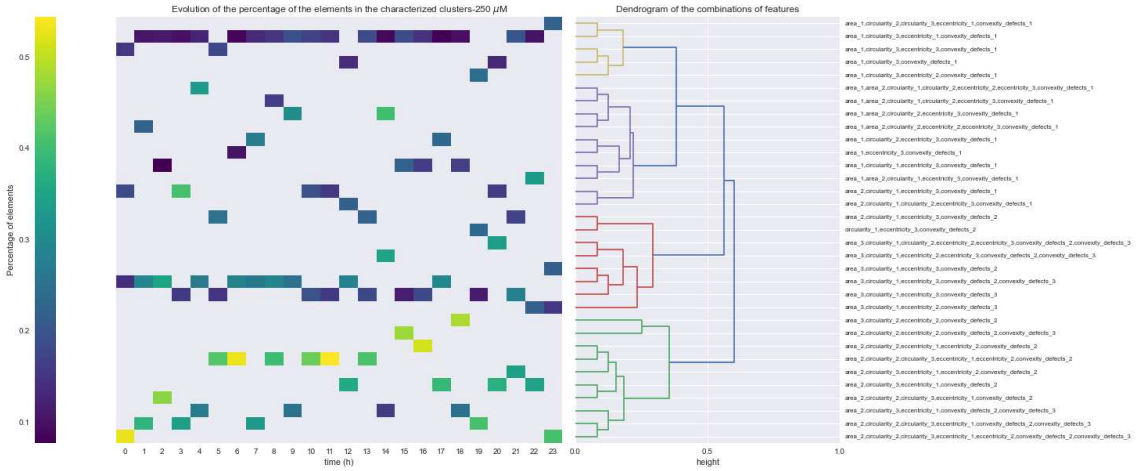
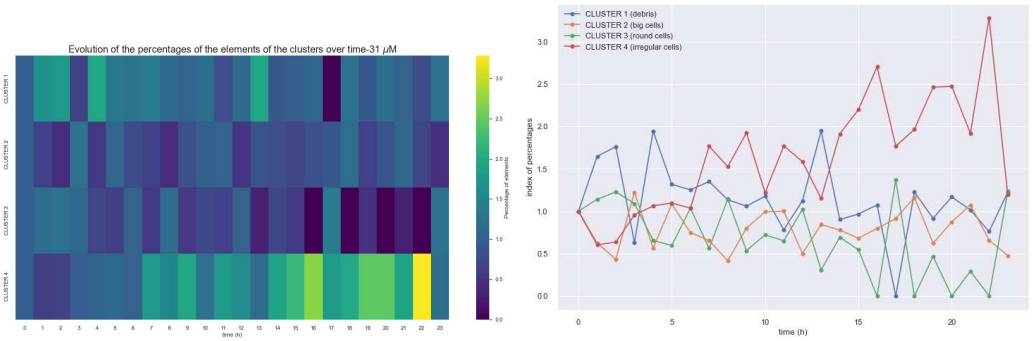


Figure 35: Configuration with 250 μM of treatment: in this case, over the two combinations of features which are conserved for all the other configurations, it can be observed a high percentage of cells in the green group of the dendrogram below with these features: "area 2", "circularity 2", "circularity 3", "eccentricity 1", "eccentricity 2", "convexity defects 2"; this implies that they are cells of medium size and a more circular than eccentric shape (cells with original morphology).

Looking now at the evolution of the percentages of elements for each cluster, the trends are not clear as in the case of the AHC performed on the data frame of objects grouped all together because they fluctuate a lot and it is no more possible to recognize patterns as before. A probable explanation of this result could be that in this case the clustering is performed separately for each image and the objects contained in a single image are not sufficient to return clusters as precise as before.

As an example, *Fig.*36 shows the trends obtained for the configuration with low level of treatment, after that all the percentages corresponding to the 4 clusters of the dendrogram are summed and normalized dividing by the first value of that cluster.



(a) (b)

Figure 36: Example of the results of the trend for the configuration with $31\mu M$ of treatment: (a) sum of the percentages of the elements of the 4 groups of leaves of the dendrogram, (b) linear plot of the normalized percentages. It is possible to see that there is no net behaviour of the percentages over time, as it was revealed with the method where the clustering was performed with the $\sim$ 80000 objects all together. For this reason probably, the trends fluctuate a lot.

Conclusions

The results obtained in each chapter are now summarized. In *Chapter 1* the data set is pre-processed, segmenting the objects in the images and analyzing the trends of the number of cells and the median of areas over time. What is obtained is consistent with what is expected, i.e. the slope of growth (*Fig.*12) of the number of cells is higher for the configurations with none or low level of treatment and it keeps decreasing for higher levels, while the median of the areas (*Fig.*9-11) for all the configurations increases in the first range of hours, when the cells attach at the bottom and are going to duplicate and then decreases when the number of cells reaches the plateau because the cells have less space to enlarge and duplicate further. This is therefore an index that the segmentation and detection was made properly.

In *Chapter 2* useful features to perform *Agglomerative Hierarchical Clustering* are collected, in particular the morphological features that can be extracted directly from the objects in the masked images. In order to perform the clustering a data frame (*Table*I) is built with all the extracted features of all the objects in the images of all the configurations from time 00h00m to 23h00m. Thus there is an unique data frame of 77702 rows (objects) and the AHC algorithm was applied to it. Performing the clustering with different linkage methods it results that the objects can be classified in 4 clusters and the evolution of the percentages of elements in each cluster can be analyzed during the selected period of 24 hours; again, the results are consistent with what is expected, since the percentage of elements in the cluster of rounded and medium size objects (*Fig.*24c and *Fig.*28c), which is the shape that the cells have at the beginning when they are seeded in the well plate, decreases over time, because they change their morphology, and the (negative) slopes are higher for the configurations with low level of treatment, where the cells are more vital and so over time the system is expected to become more heterogeneous, and lower for the configurations with higher level, where instead the cells over time change their

morphology more slowly due to the fact that they are poisoned with the treatment. In correspondence it was obtained that the percentage of elements of the cluster of cells with large values of area, jagged contours and elliptic shape (*Fig.*24a and *Fig.*28a), which are the type of cells observed after few hours when they attach at the bottom of the well plate and are going to split, increases over time with higher values of slope for the configurations with low level of treatment and lower values for the configurations with high level. The same result was also obtained for the cluster of small objects with irregular shape, which could be classified as the divided cells and/or debris (*Fig.*24b and *Fig.*28b). Finally a constant percentage of elements is observed during the 24 hours for all the configurations for the cluster of debris which is what is observed in the images where they are always present (*Fig.*24d and *Fig.*28d).

In *Chapter 3* the clusters are characterized, analyzing with a p-value test, according to if and which are the over-expressed features; the results corresponds to the mean values of the features that were computed in the previous chapter and used to identify the clusters. An alternative approach to detect the clusters is also implemented, performing the AHC separately image by image and then following which combinations of characterizing features are constant during time. In this way, for each configuration, two clusters recognizable over time are detected, which correspond to the cluster of debris and the cluster of splitting cells identified with the previous method. However, looking at the evolution of the percentages of elements in those clusters (*Fig.*36), it is no more possible to observe net trends as before, probably because the performance of the clustering algorithm is not as efficient as in the case were it was applied on the data frame of about 80000 objects, since this time it is applied separately image by image, therefore in groups of the order of some hundreds of elements.

These results can be considered as a first stage of the analysis of the data set.

Many improvements in the analysis can be done. In the Introduction was already mentioned that artificial intelligence methods might be used in order to detect and classify cells. Furthermore, one could also improve the efficiency of the clustering algorithms by considering not only the morphological features, but also by working directly on the original images rather than on the masked ones. The advantage would be that of extracting directly from the images optical properties of the cells, or other information such as the presence and the size of nuclei inside them.

In this preliminary investigation , the analysis was focused on the first 24 hours of evolution of the cells, since this is a "stationary" phase of the system. In fact, this is a phase where division or merging among the objects are very limited. However, the whole data set is composed of images where the samples are tracked for 72 hours. Therefore the next step could be to follow the trajectory of each cell, and to do this the objects have to be label image by image, even in the regime where they start to split. This data could be used in a twofold way. On one hand, the information collected about the cells trajectory and kinetics could be used in order to optimize the clustering algorithm and eventually to validate the results obtained by machine learning approaches[33]. On the other hand, such information could be used for detecting possible collective behaviors of cells during their time evolution through a correlation-based network approach and a subsequent cluster analysis. Also the stability of clusters over time could be investigated. Moreover, it would be possible also to detect contagion effects amongst cells[34]. An approach could be to characterize whether and how senescent cells affect the neighboring cells and by means of Granger causality tests[35] the possible existence of causal relationships among senescent cells and normal ones could be revealed.

References

[1] F. Rodier and J. Campisi. "Four faces of cellular senescence". *J Cell Biol.* 192 (4) (2011), pp. 547–556.

[2] http://www.fondazionerimed.eu/Content/home.aspx.

[3] T Soni Madhulatha. "An overview on clustering methods". *arXiv preprint arXiv:1205.1117* (2012).

[4] M. K. McKinnon. "Flow Cytometry: An Overview". *Curr Protoc Immunol.* 120 (2018), pp. 5.1.1–5.1.11.

[5] William HE Day and Herbert Edelsbrunner. "Efficient algorithms for agglomerative hierarchical clustering methods". *Journal of classification* 1 (1984), pp. 7–24.

[6] Michele Tumminello, Salvatore Micciche, Fabrizio Lillo, Jan Varho, Jyrki Piilo, and Rosario N Mantegna. "Community characterization of heterogeneous complex systems". *Journal of Statistical Mechanics: Theory and Experiment* 2011.01 (2011), P01019.

[7] https://www.sartorius.com/download/930502/incucyte-s3-technical-specification-sheet-8000-0527-c00-en-s-1–data.pdf.

[8] https://www.synthego.com/hek293hek293Origin.

[9] J.C. Caicedo et al. "Data-analysis strategies for image-based cell profiling". *Nature Methods* 14 (2017), pp. 849–863.

[10] Tony J Collins. "ImageJ for microscopy". *Biotechniques* 43.S1 (2007), S25–S30.

[11] Pauli Virtanen et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". *Nature methods* 17 (2020), pp. 261–272.

[12] Xiangyang Xu, Shengzhou Xu, Lianghai Jin, and Enmin Song. "Characteristic analysis of Otsu threshold and its applications". *Pattern recognition letters* 32 (2011), pp. 956–961.

[13] Ivan Culjak, David Abram, Tomislav Pribanic, Hrvoje Dzapo, and Mario Cifrek. "A brief introduction to OpenCV". In: *2012 proceedings of the 35th international convention MIPRO*. IEEE. 2012, pp. 1725–1730.

[14] https://www.geeksforgeeks.org/python-thresholding-techniques-using-opencv-set-3-otsu-thresholding/.

[15] Jesús Angulo, Dominique Jeulin, et al. "Stochastic watershed segmentation." In: *ISMM (1)*. 2007, pp. 265–276.

[16] Stefan Van der Walt et al. "scikit-image: image processing in Python". *PeerJ* 2 (2014), e453.

[17] Frank Nielsen. "Hierarchical clustering". *Introduction to HPC with MPI for Data Science* (2016), pp. 195–211.

[18] C. Coronnello, M. Tumminello, F. Lillo, S. Micciche, and R. N. Mantegna. "Sector identification in a set of stock return time series traded at the London Stock Exchange". *arXiv preprint cond-mat/0508122* 36 (9) (2005), pp. 2653–2679.

[19] Oliver Kramer. *Machine learning for evolution strategies*. Vol. 20. Springer, 2016, pp. 45–53.

[20] https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/.

[21] Sahar Sohangir and Dingding Wang. "Improved sqrt-cosine similarity measurement". *Journal of Big Data* 4 (2017), p. 25.

[22] M. R. Berthold, C Borgelt, F. Höppner, and F. Klawonn. *Guide to Intelligent Data Analysis*. Springer, 2010.

[23] A. Airola. "Data Understanding I". Lecture of Data Analysis and Knowledge Discovery, University of Turku, Department of Computing, 2022.

[24] Rosario N Mantegna and H Eugene Stanley. *Introduction to econophysics: correlations and complexity in finance*. Cambridge university press, 1999.

[25] Israel Cohen et al. "Pearson correlation coefficient". *Noise reduction in speech processing* (2009), pp. 1–4.

[26] P. J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65.

[27] T. Caliński and J. Harabasz. "A dendrite method for cluster analysis". *Communications in Statistics* 3 (1972), pp. 1–27.

[28] D. L. Davies and D. W. Bouldin. "A Cluster Separation Measure". *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1 (2) (1979), pp. 224–227.

[29] Byoungwook Kim, JaMee Kim, and Gangman Yi. "Analysis of clustering evaluation considering features of item response data using data mining technique for setting cut-off scores". *Symmetry* 9 (2017), p. 62.

[30] D. Müllner. "Modern hierarchical, agglomerative clustering algorithms". *arXiv preprint arXiv:1109.2378* (2011).

[31] W. Feller. *An Introduction to Probability Theory and Its Applications, , Volume 1, 3rd Edition*. Wiley, 1968.

[32] Mohammad Norouzi, David J Fleet, and Russ R Salakhutdinov. "Hamming distance metric learning". *Advances in neural information processing systems* 25 (2012).

[33] E. Meijering, O. Dzyubachyk, and I. Smal. "Methods for Cell and Particle Tracking". *Imaging and Spectroscopic Analysis of Living Cells* 504 (9) (2012), pp. 183–200.

[34] C. Castellano, S. Fortunato, and V. Loreto. "Statistical physics of social dynamics". *Reviews of Modern Physics* 81 (2009), pp. 591–646.

[35] Steven L Bressler and Anil K Seth. "Wiener–Granger causality: a well established methodology". *Neuroimage* 58.2 (2011), pp. 323–329.