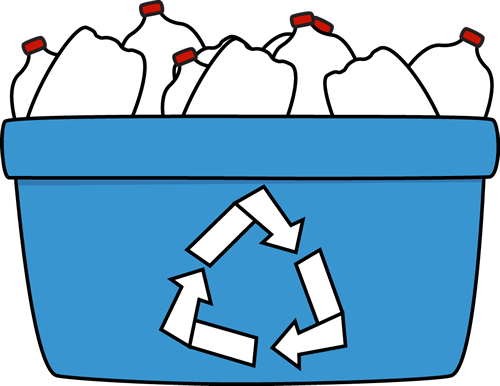
Bioplastic degradation by bacteria

A Machine learning based research

*Calculate the rate of bioplastic degradation with the aid of machine learning*

Afbeelding met tekst, kamer, gokhuis, vectorafbeeldingen

Automatisch gegenereerde beschrijving



Pascal Visser, Rob Meulenkamp, Rienk Heins

410729

BFV3

27-06-2022

Dave R.M. Langers

# Preface

As part of our bioinformatics course, we are interested in multiple biologics and environmental topics. And with global warming and other environmental threats present, this project was a great combination of interest and possible a way to make the world better.

This research can contribute to a way of reducing plastic waste by highlight the characteristics of bioplastic and their superior degradation. With help of teachers and co-students, which are a huge inspiring source, this project is possible.

We like to thank them for the help and guidance.

# Abstract

INTRODUCTION: Plastic debris is a growing environmental threat, with less developed countries getting more developed, there plastic waste is also growing. To counter this problem, most developed countries are shifting towards bioplastics. Bioplastics can easily be broken down by microorganisms. An example of a bioplastic is PHBV (3-hydroxybutyrate-co-3-hydroxyvalerate). Which is a renewable bioplastic, microorganisms also produce it. The degradation rate of this PHBV plastic by bacteria is important. To see which type of bacteria has the highest degradation rate. To measure this degradation, samples with PHBV and bacteria are in timesteps of 2 weeks measure through a flow cytometer. Flow-cytometry is a technique to detect and measure psychical and chemical characteristics of cells or particles. It measures forward and sideward scattering light. With this scatter data, degradation can be calculated.

METHODS: The main programs used in this research are: the statistical programing language R for analyzing the dataset, R version 4.1.2 (2021-11-01). And Python for algorithm testing. Python version 3.9. In short: The flowcytometry data (.fcs data) was loaded into R and transformed in more workable csv format files. Then the data was statistical analyzed by using density plots, PCA and biplots. The results the statistical analysis of showed the strengths and weaknesses of the data. Next, three algorithms were evaluated to determine clusters of plastics and bacteria. With a suitable algorithm, cluster centers can be set and therefore calculated what the shift of the plastic clusters is over time. To make this user friendly, an application is built around the code.

RESULTS: The results of this study are lightly optimistic; it is difficult to rate the degradation but possible. The outcome of the statistical analysis gave many insights is the shape and characteristics of the data. Which lead to density-based algorithms instead of clustering algorithms based on shapes. Also, the biplots showed the significance of certain variables, were in the end only two of the fourteen variables are used. With the application, researchers can easily pass the data to the app and receive statistics and graphics of the results.

DISCUSSION: the main discussion point of the whole research was the data itself. The origin was not form researchers, but from a student. The data was experimental and not consistent in terms of flowcytometry cycles and sample contribution. Which made the data difficult to work with. The struggles with the data costed valuable time, but in the end it was made workable. Future project could be identifying distinct species of bacteria in the samples for better accuracy

# List with abbreviations

**PHBV** = Poly(3-hydroxybutyrate-co-3-hydroxyvalerate)

**.fcs/fcs file** = Flow Cytometry Standard File

**.csv/csv file** = Comma-Separated Values File

**PCA** = Principal component analysis

**GGM** = Gaussian mixture model

Contents

[Preface 2](#_Toc107606196)

[Abstract 3](#_Toc107606197)

[List with abbreviations 4](#_Toc107606198)

[1. Introduction 6](#_Toc107606199)

[2. Theory 7](#_Toc107606200)

[2.1 Lab experiments 7](#_Toc107606201)

[2.2 PHBV 7](#_Toc107606202)

[2.3 Flow Cytometry 8](#_Toc107606203)

[3. Material and methods 9](#_Toc107606204)

[3.1 Materials 9](#_Toc107606205)

[3.2 Method 9](#_Toc107606206)

[3.2.1 Data exploration 10](#_Toc107606207)

[3.2.2 Clustering algorithms 13](#_Toc107606208)

[3.2.2.1 Gaussian mixture model 14](#_Toc107606209)

[3.2.2.2 OPTICS algorithm 15](#_Toc107606210)

[3.2.2.3 DBSCAN 17](#_Toc107606211)

[4. Results 19](#_Toc107606212)

[4.1 Code 19](#_Toc107606213)

[4.2 Pipeline 22](#_Toc107606214)

[4.3 App 24](#_Toc107606215)

[5. Conclusion and discussion 25](#_Toc107606216)

[5.1 Conclusion 25](#_Toc107606217)

[5.2 Discussion 25](#_Toc107606218)

[Bibliography 28](#_Toc107606219)

# Introduction

The world is changing and the search to renewable sources is more important than ever before. In the fight against fossil fuels and plastic, bioplastics are a great alternative to non-biodegradable plastics.

Bioplastic are polymers produced by micro-organisms. An example of a bioplastic is PHBV. PHBV is short for Poly(3-hydroxybutyrate-co-3-hydroxyvalerate). PHBV is a thermoplastic polymer, which is excellent for:

* Controlled release of drugs
* Medical implants and repairs
* Specialty packaging
* Orthopedic devices

PHBV is also bio-degradable which can be used as an alternative to non-biodegradable plastics. And it is renewable. Only drawback is, that is it is now expansive to make. But with new techniques and more funding in this branch, cost may drop exponentially in the (near) future. (Rivera-Briso & Serrano-Aroca, 2018)

This project is about analyzing obtained research data from flow-cytometry and develop, with machine learning, an analysis pipeline that can calculate the degradation of PHBV by bacteria. With this pipeline, researchers in het lab can use it to calculate the degradation by different bacteria spices. The pipeline is supported by an application for user-friendliness. This project is part of the minor High throughput / High-Performance Biocomputing of the Bio-informatics course at the Hanze Hogeschool Groningen. And on behalf of the lab research for bio medical diagnostics.

In the past there are several research conducted in regards of PHBV. Such as the enhancement of the thermal stability of PHBV (Qing-Sheng, Mei-Fang, Wen-Hua, & Zong-Yi, 2009)or study the degradation under anaerobic conditions. (Abou-Zeid, Müller, & Deckwer, 2001). But, noting came close to the analysis that is conducted in the past months.

The outcome of this analysis is of interest for the lab researchers, so that they have a tool to quickly determine what the degradation is of the bioplastic of their interest. The tool will allow them to analyze data and get output statistics to get a clear look of the data. Otherwise, the researchers need to plot each individual piece of data and compare them to each other, which is a time-consuming task.

The project goal is to develop an algorithm to determine the rate of bioplastic degradation by bacteria. And make It easy to uses for lab workers and students.

This report attempts to show how this analysis is conducted and which steps and materials its takes. Supported by background information about the subject and other resources to make this project repeatable and expandable.

# Theory

This section is about the necessary background information regarding conducted lab experiments, bioplastics, and flow cytometry. These concepts are important for understanding analysis processes and the steps that are taken within the project.

## Lab experiments

The collection of the data as taken place in a controlled laboratory environment. The data is harvested using a flow cytometer. Because the experiment was new to the lab technician, several errors and testing as taken place in the experiment.

There were four samples of a bacterial colonies with PHBV particles. Those samples were measured four times with an interval of two weeks. Producing four folders with .fcs data. Throughout the measure days, especially on day two and three, freeze-thaw and coloring testing has taken place.

## PHBV

PHBV is a biodegradable, environmental nontoxic and biocompatible plastic that is naturally produced by bacteria (figure 1). It is a suitable alternative for plastics that has an origin in fossil fuels. The main applications of PHBV are specialty packaging, orthopedic devices and in controlled release of drugs.

PHVB was first commercially manufactured in 1983 but was expensive to produce. So, the market for PHBV was small. Since the growing threat of environmental damage by non-biodegradable plastic, the call of sustainable alternatives is raising. PHBV is an excellent alternative to oil-based plastics because it is produced naturally. PHBV is synthesized by bacteria such as recombinant *Escherichia* *coli* strains, *Paracoccus denitrificans* and *Ralstonia eutropha*. It is synthesized as energy storage compounds in times of growth limiting conditions. It can also be synthesized from genetically engineered plants (Pilla, 2011)

PHBV is the bioplastic what is broken down in this research, all the data consists of degradation of this bioplastic. It is important to know how fast and in which way it degrades the best. For example, in controlled drugs release, gut bacteria and other factors degrades the bioplastic for a controlled release. It is good to know how different bacteria respond at the bioplastic. With lab-controlled tests, this can be sorted out. And in aid of that, the application for automatically calculate the degradation of the bioplastic can really help.

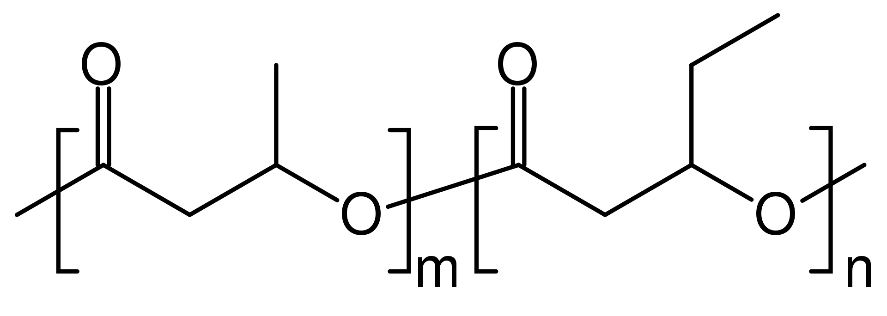


Figure 1: Molecular Structure of PHBV

## Flow Cytometry

The experiments in the lab consisted of analyzing samples of bacteria and PHVB trough a flow cytometer. Flow-cytometry is a technique to detect and measure psychical and chemical characteristics of cells or particles. It measures forward and sideward scattering light with the aid of laser beams. The scattering of light means the deflection, by diffraction of light against particles. The sample with bacteria and bioplastics flow through a microscopically narrow tube and pass through a laser beam. The strength and ratios of the scattered light intensities can be used to assess the nature and characteristics of cell/plastic particles. Additionally, there are four filters that can measure the different wavelengths of the particles. (Nanocellect, 2020)

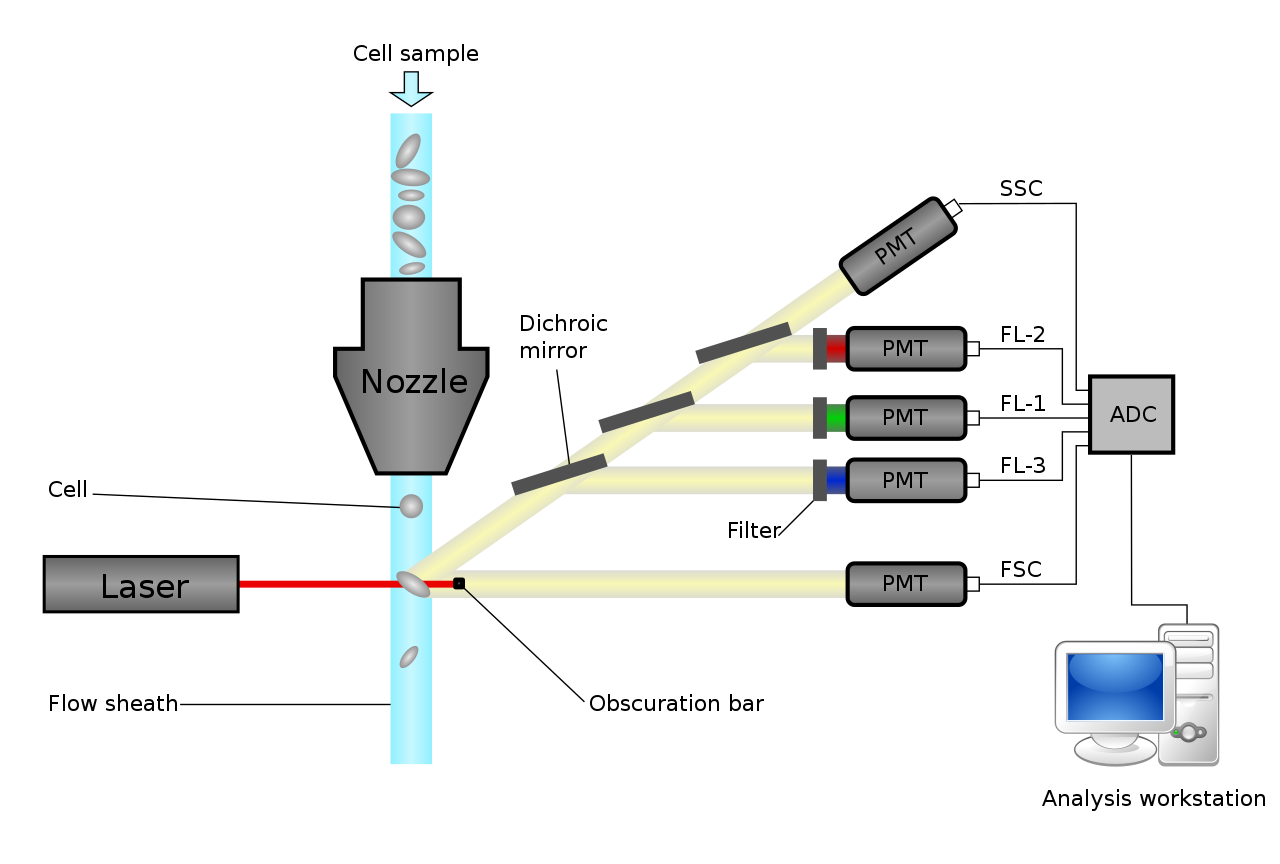


Figure 2: Schematic diagram of a flow cytometer

In above diagram (figure 2) is the working of the flow cytometer seen. It produces six variables per particle in this image: SSC, FL 1 t/m 4 and FSC. All those variables are saved and exported in the .fcs file format. FSC and SSC are the most important variables in this research.

**FSC:**

Forward-scattered light is the scatting, or the lack of scattering on a single particle. It tells something about the size of the particle by the shadow that it produces. The scatting that is left is produced as a FSC value.

**SSC:**

Sideward-scattered light is the scatting of the light in a ninety-degree angel relative to the laser. This value tells something about the surface complexity of the particle. How higher this value, how more complex this particle is.

There are also more variables that derive from the .fcs file format, but only the FSC and SSC are used in this project. Later is explained why. In this research, FSC and SSC tells most of the degradation. There are bioplastic particles and bacteria. The lab experiment has taken place over time. So, it is expected that bacteria will grow, and plastic particles will degrade. Bacteria tent to get a higher FSC, the bioplastics gets not evenly broken down around the shape of the bioplastic, so the SSC will increase.

# Material and methods

This chapter is about the used software and sources, also the method of completing the project is broken down and analyzed.

## Materials

The main programming language used was the statistical programming language R. R was used in the IDE RStudio (local/online) together with several libraries. Also, Python was used for testing algorithms and develop the application. Python and Snakemake, a workflow management system, were used for developing the pipeline.

Versions of used software as on 28-06-2022:

* R = 4.2.1
  + FlowCore = 2.6.0
  + Bioconductor = 1.30.18
  + Cytolib = 2.6.2
  + Dbscan = 1.1-10-1
  + Dplyr = 2.2.0
* Python = 3.9.13
* Snakemake = 6.15.5

Furthermore, the scrum technique was easy to use to evaluate and discuss progress and aberrations. Trello was used to keep track of the progress and task management.

## Method

The method for the project can be broken down in two parts:

* Data exploration
* Clustering algorithms

Where the exploration focuses more on understanding the data, and the machine learning is the use of the data to a practical tool with calculation of degradation scores.

The whole project was divided in so called sprints, which is part of the scrum method. A sprint is a time period where certain goals are set and strive to be achieved. At the end of a sprint, a presentation is held to show the progress and present the future goals.

### Data exploration

The project started with four data folders, a report from the lab experiments and an assignment to develop. First the data needed to be read into R for data analysis. The data came, as said, in four folders with each multiple .fcs files. These files in FCS format are not workable in R. Therefore, the files needed first to be converted to CSV file format, which is more common and much better to work with.

Afbeelding met tafel

Automatisch gegenereerde beschrijvingThe conversion from .fcs to .csv was more difficult than was expected because R didn’t have built in support for .fcs files. This is where the FlowCore package can in play. This package contains functions to easily open and then save them to workable csv file formats. Below (figure 3) is an example of a part of the flow cytometer data

Figure 3: Flow Cytometer data

The rows represent a single particle, and the columns are the variables corresponding to the different values generated by the flow cytometer. In total there are 14 variables.

There are four directories with .fcs files. Each directory stands for a time period. By merging the samples for the different time periods together, four datasets were formed. Each dataset corresponds to a folder. Dataset 1 is the zero measurement, the fresh measurement. Dataset two is two weeks later, Dataset 3 is four weeks and dataset four is six weeks.

Now there are different dataset to work with. The first objective was to get a feeling of the data, knowing what we are working with. But after some test plotting the data turned out to be difficult. As mentioned earlier, some experiments were done on the samples between the measurement days. This resulted in some vague data parts in the datasets. Which turned out to be difficult to plot.

After a clear up talk with the creator of the data, we decided to normalize the clear parts of the data, to get hopefully get better results. The data was log2 transformed to normalize the variables. By plotting FSC vs SSC a shape can be seen (figure 4).

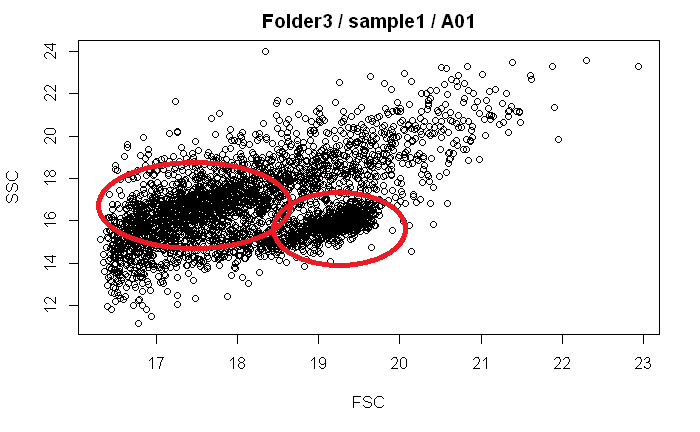


Figure 4: Example of plotting log2 transformed data

In above figure, two vaguely clusters can be seen circled in red. Because of these visually clear clusters, the idea was born to follow the clusters over time to calculate the movement of these clusters. By clustering these datasets, centers could be assigned and followed. The movement could then be rated as a number. This number can than correspond to the degradation rate of a certain bacteria species.

To achieve this, it was necessary to understand which cluster is what. What are the bacteria and what are the plastics? Luckily one of the experiments, freeze-thaw, gave the opportunity to see what happens when bacteria die. With the freeze method, bacteria will die and will not be present trough the next flow cytometer measurement cycle. In below image can be seen what happens before and after freeze-thaw. (Figure 5)

Afbeelding met tekst, schermafbeelding, ontvangstbewijs

Automatisch gegenereerde beschrijving

Figure 5: Before and after freeze-thaw experiments

In the figure clearly can be seen that a certain spot falls away after the freeze-thaw. This must be the bacteria cluster. Which makes sense because the assumption is that bacteria have roughly the same size and shape. So expected is that bacteria have a small densely cluster in the plot.

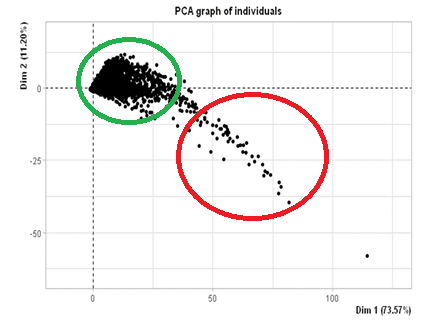
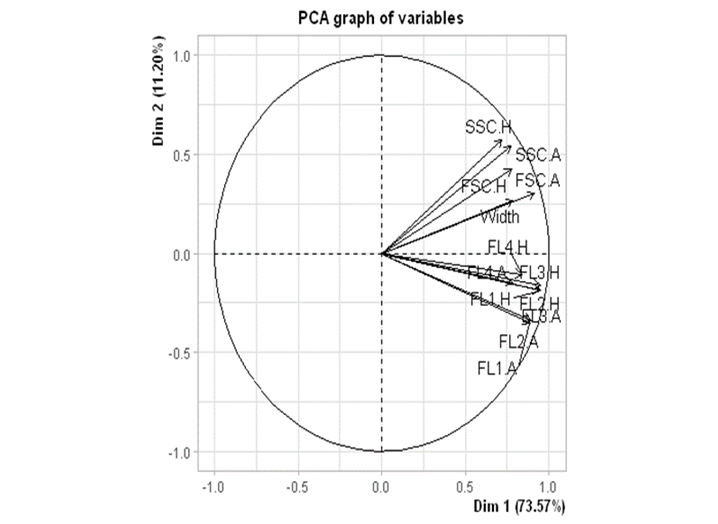
This observation was only concluded of plotting two variables, while there are 12 other variables as well. Plotting all variables against each other is a way of doing this, but it takes much time and effort. To plot all 14 variables on a 2D graph, the dimensions need to be reduced. PCA is an excellent choice for this. PCA is a useful statical analysis technique to find hidden correlations between variables. With PCA all variables are reduced to Principal components (PC) which represent a part of variance in the data. By plotting PC1 and PC2 against each other, maybe new insights are shown.

Figure 7: PC biplot

Figure 6: PCA plot of PC1 against PC2, sample1 / map1

In the plot (Figure 6), the shape of the points does not look like something with clear clusters anymore. Figure 7 shows the biplot. The correlations between variables in the dataset. All the FSC and SSC value tent to have a high correlation to each other, while the fluorescence variables also cluster together. In figure 6 circled in red are the FSC and SSC values, and in green the FL values. Based on this , there is decided to not include FL variables, because they are not very related to the scatter variables. Between the FSC and SSC, there are .H and .A extensions. These stand for Height and Area. Comparing area and height to each other barely made a difference, so it was decided to just go with the first FSC and SSC variables, .A / Area.

With sights of clusters and sets variables to use, it was time to define the clusters. Knowing that the bacteria formed a densely spot on the plot. It was useful to make density plot of the data. A density plot is a graph were regions with a high concentration of points are differently colored. Form this plot, the crowded parts of the data can be recognized. In below graph (figure 8) is the density plot of the dataset of folder 1 t/m 4.

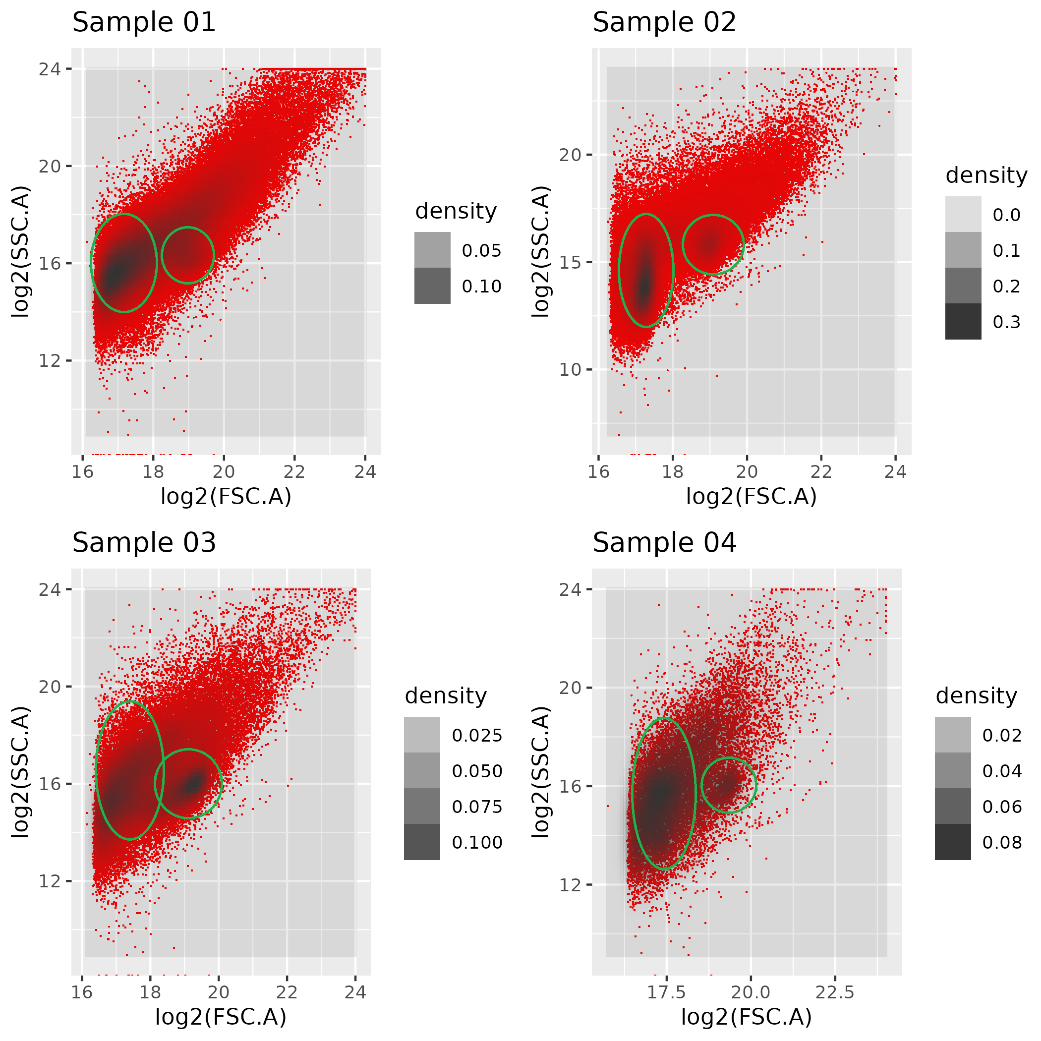


Figure 8: Density plot of the datasets

In above plot, the dense centers are circled in green. In all the samples there are two clear cluster. One bacteria cluster and one bioplastic cluster. Although the plastics have a clear dense center, they have a ‘tail’ with less dense plastic regions.

Overall density shows possibilities because in this difficult data some dense clusters are present. There is decided to go with density clustering for an algorithm, because this is the most promising. The idea is to use these cluster central to calculate the movement over time.

### Clustering algorithms

With a clear idea of how to handle the data, now it was time to make something that can distinguish bacteria from plastics. To achieve this, a good clustering algorithm was necessary. After some research there were three candidates: Gaussian mixture model, OPTICS algorithm and DBScan. Each algorithm has their own advantages and disadvantages.

#### Gaussian mixture model

Gaussian mixture model (GMM) is a joint model that implies the data is spread as a Gaussian distribution. The algorithm is used for example clustering or density estimation. GMM is used to discover clusters in data where it is unclear how many clusters there are. Another feature of GMM is to give an estimation of the probability that a new point relates to a cluster. Gaussians are somewhat robust to outliers. This means that Gaussian produces accurate results if some points do not fit to any cluster. This is the reason why Gaussian a powerful and flexible algorithm is for clustering data. Gaussian is a probabilistic model where Gaussian distributions are presumed for each class, and they consist of a means and covariances which determine the parameters. GGM contains two pieces - mean vectors (μ) & covariance matrices (Σ). Gaussian distribution describes a continuous probability distribution that accepts a bell-shaped curve. Gaussian distribution is also known as a normal distribution. (Kumar, 2022)

**Expectation-maximization (EM) method**

Gaussian algorithm uses a strong tool named expectation-maximization method. This is used for computing the parameters of Gaussian mixture model. Expectation (E) discovers Gaussian parameters which are used to portray each component of the Gaussian model. Maximization (M) is engaged in deciding if new data points can be an addition or not. EM is a two-step iterative algorithm that switches between doing an expectation step which calculate expectations for each data point using existing parameter estimates and then optimize to produce a new Gaussian. This is followed with a maximization step where the Gaussian mean are updated based on the maximum likelihood estimates. EM operate with initializing the parameters of the Gaussian mixture model, then repeatedly improving the estimates. With each iteration, the expectation step computes the expectation of the log-likelihood function with the parameters. The expectation utilizes the likelihood in the maximization step. This operation is repeated until it converges. (Kumar, 2022)

The Gaussian mixture model wasn’t suited to collect the result. Before executing the Gaussian algorithm, the user must determine the number of components. This provide always result, but it could be the wrong result. Sometimes the user doesn’t know the number of components then the Gaussian mixture model can’t be used.

#### OPTICS algorithm

One of the algorithms that was tried is the OPTICS algorithm. This algorithm was used because it is density based with the big advantage that it is specialized to be better at classifying data with varying density, which seems to be the case in this project.

At first the OPTICS algorithm seemed to have some problems classifying the points as seen in figure (optics plot 1). While it does find some points at the right places it only classifies a very small portion of the points. So, the same data was classified with different settings.

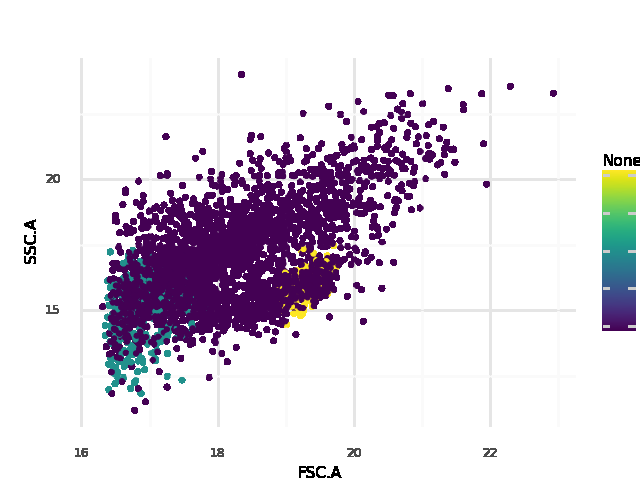


Figure 9: OPTICS first plot, with default settings

After some testing an OPTICS run with a higher number of points needed to form a cluster and lower xi value the results seen in figure (OPTICS 2) were created. As seen in the figure it classifies the bacteria very well and finds the density cluster of the plastics. The only real problems being the still big amount of noise classified points and no points in the tail end of the plastics found. Yet these results would be good enough to do statistics on.

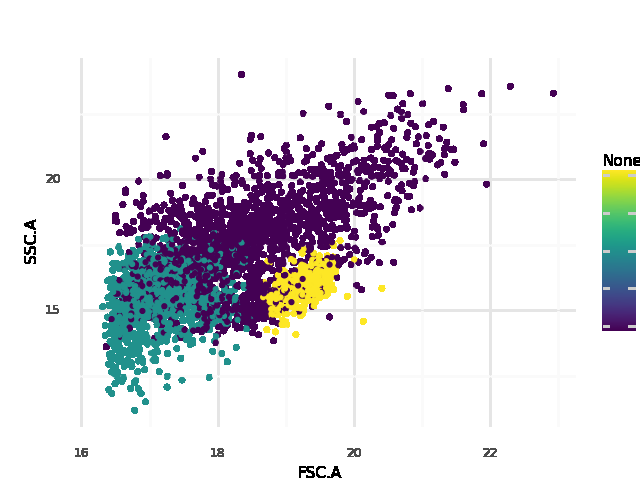


Figure 10: Second OPTICS plot with lowered xi value and higher points needed for cluster

While the results of figure (OPTICS 2) were promising, there was still more testing that needed to be done as this was a dataset with two clear clusters while the full dataset has quite some overlap and files that don't have distinct clusters. Because of this the algorithm was tested on another dataset with less clear distinction seen in figure (OPTICS 3). As seen the algorithm fails to detect the bacteria here and doesn't do a great job on the plastics. Even after more tweaking with the data this problem couldn't be removed and altered the results on the classification of earlier files.

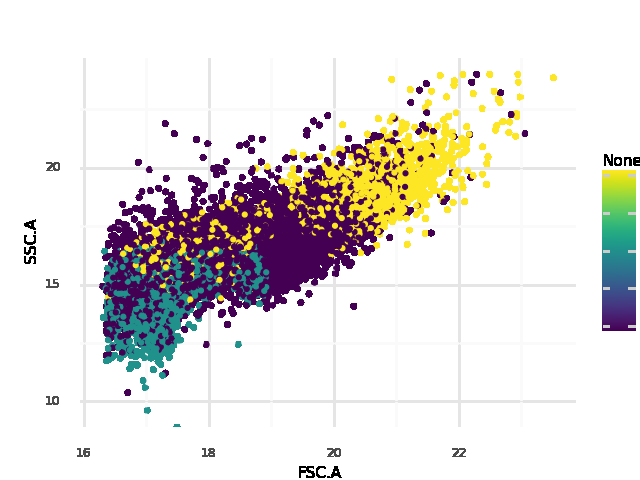
Finally, the algorithm was still tested on a full map of data from time point one. This was to see if it still might have some potential and the results can be seen in figure (OPTICS 4). As seen also here the algorithm doesn't find the bacteria, it even doesn't find a second cluster at all is seems. Even the plastics it does find seem to not have their center at the right place. Because of this it was decided that OPTICS would not be used as the final algorithm.

Figure 11: Third OPTICS plot on dataset with less distinct clusters

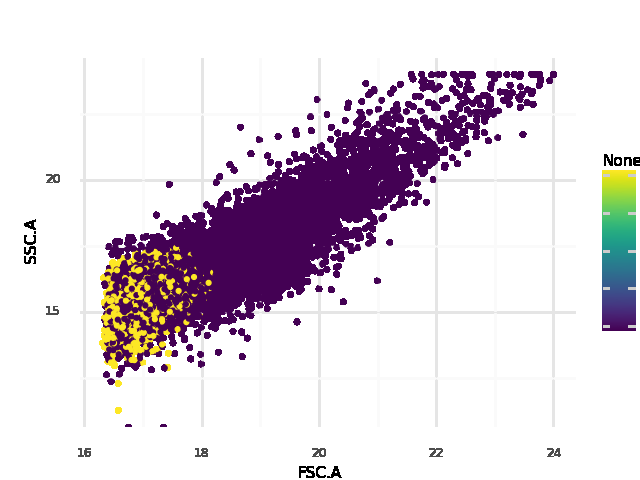


Figure 12: OPTICS classification on full dataset

#### DBSCAN

DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. It Is a clustering algorithm that can handle noise, and as well non-linearly separable clusters. It works at the base of a range and a given number of nearly neighbors.

In comparison with k-means: a plot with a dot and a ring around it, k-means will separate the shapes by a straight line. Where DBSCAN will cluster the neighbors and get the shapes right (figure 13)

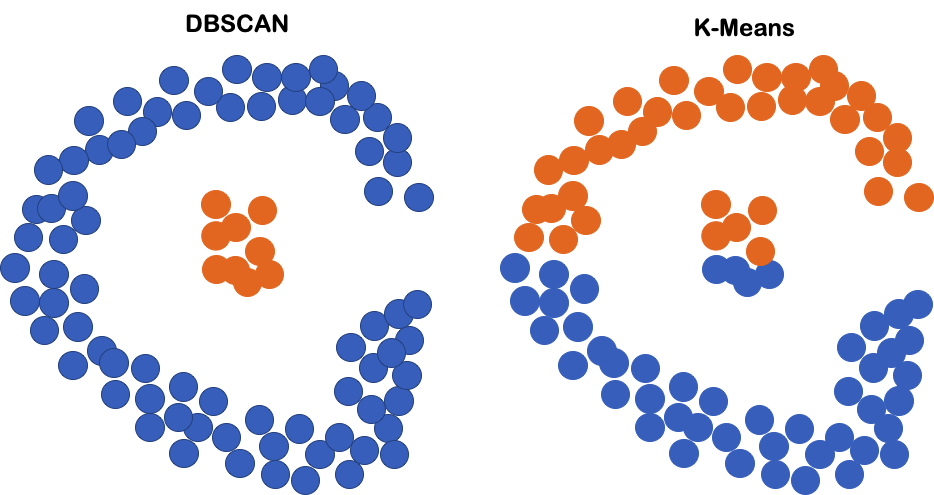


Figure 13: DBSCAN vs K-means

Because this right clustering of non-linearly separable clusters and density, DBSCAN looks perfect for the datasets. It assigns the cluster with the following method (Figure 14):

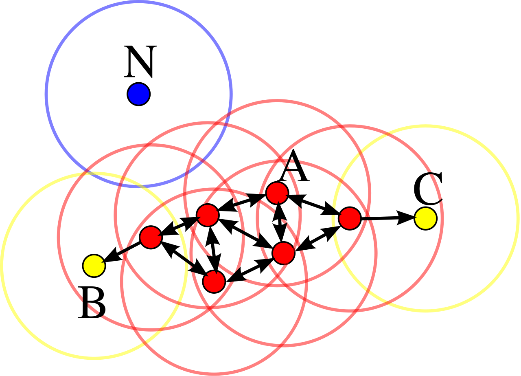
The DBSCAN takes two arguments: Epsilon (*ε* ) and minPts

Figure 14: Density-based clustering

Where *ε* is the radius of neighborhood, and minPts the number of minimum points. In this example minPts = 4

Starting at point A: In the range of *ε* around point A, DBSCAN looks for four datapoints. If its finds four datapoints of more it assigns it as a core point. From this core point it looks by all neighbors of A if they qualify as core point. If they are, then DBSCAN look beyond these points as well.

When the DBSCAN reach point B and C, they fall in the range of the nearby core point, but themselves are not considered core points, because they fail to meet the requirement of four minimum neighbors, B and C are called non-core points. All core points and non-core points form a cluster. Where a cluster has at least one core point. Non-core points are considered the edge of a cluster.

N stand for Noise, this are points that don’t fall in a cluster because they are not in reach of a core point and don’t have four neighbors themself.

The DBSCAN as many advantages over the other algorithms:

* It does not require specific number of clusters (unsupervised)
* DBSCAN can find arbitrarily shaped clusters, even cluster around a cluster (single-link effect)
* DBSCAN can deal with noise and notice it
* In Big complex datasets, DBSCAN only requires two parameters

The only disadvantage with the current datasets is dealing with evenly spread data and very small data. And it is difficult to estimate which parameters to use.

For this problem there is the HDBSCAN. It extends DBSCAN by converting it into a hierarchical clustering algorithm, and then using a technique to extract a flat clustering based in the stability of clusters This variant uses a flexible ***ε****,* to optimize clustering. So, it only needs a minimum number of points to work with.

Because of this effectiveness and simplicity for large datasets, there is decided to use HDBSCAN above the other three clustering algorithms.

With some interpretation with the data and R. The coding followed, and the adjusted algorithm was born. In the next chapter, the results of the algorithm will be talked over and explained step by step.

# Results

This chapter is about the final version of the code, pipeline, and app. All the subjects will be supported with a well-directed link to the directory in the repo.

## Code

The final HBDSCAN is a coding script with two functions. This script is the core function of the project because it converts the Flow cytometer data to density clusters by means of the HBDSCAN algorithm. The script contains two functions:

* Db\_clustering
* Points\_calc

**DB\_clustering:**

The partial goal of the project was to cluster flow cytometer data by an algorithm. This was achieved with help of the HDBSCAN algorithm. This function contains data preparation and visualization.

1. First the data is checked on number of data points. With too many datapoints, the HDBSCAN generates later a to big vector that exceeds the memory of most systems. To tackle this problem, data with points above the 10.000 get random sampled in a set number. By doing that, the memory will never exceed the 4GB limit, which should run smoothly on most systems.
2. Next, the FSC and SSC columns are filtered out of the data.
3. The data contains occasionally some NAs or values with zero. This can give possible errors, so the zeros and NAs are converted to the mean of that column.
4. For normalization, all the data is log2 transformed
5. The clean log data gets scanned by the HDBSCAN with a set number of minPts
6. The function return statistics and a plot of the log data with colored cluster found by the HDBSCAN.

**Points\_calc:**

This function is to correctly calculate the number of points to use in the minPts parameter in the HDBSCAN. The function work as follow:

1. It takes the number of rows as count of datapoints and takes 1.5% of this number.
2. It looks at the sum of datapoints, if it is lower than 1000, it returns number op points to use as 15.

The number of points to see as optimal is tested thoroughly with multiple runs with different percentage of points. 1.5% gave overall the best results. 15 was the golden number for small datasets. With a to small dataset at 1.5%, no clusters are found.

Link to file = …………….?

For example, if we take the same sample as in figure 5, sample A01, and use this in the HDBSCAN function. Which looks as following: db\_clustering(sampleA01, points\_calc(samlpleA01). It returns the plot and statistics:

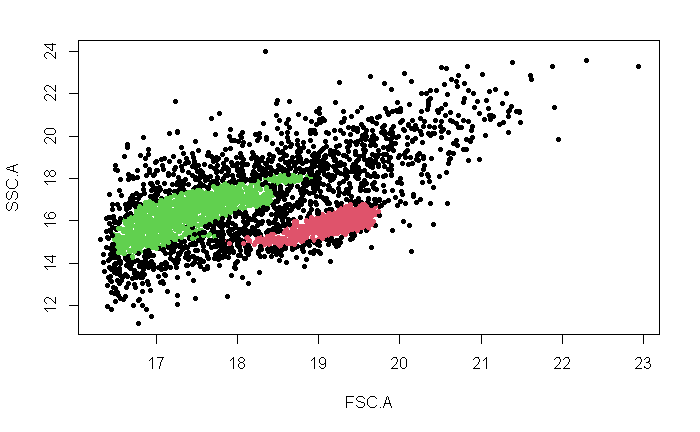


Figure 15: Clustering by HDBSCAN

Figure 15 shows the result of the HDBSCAN. A clear image with two plots. One plastic dense region and a dense region with bacteria. From the knowledge of freeze-thaw, we know that the red cluster are the bacteria. Which means that the green cluster are the plastics. It is important for the calculations.

After the clustering, the calculations are made. By taking the sum of all FSC values of the plastic cluster divided by the total length of FSC column, you get the center of a cluster. This middle point is marked and located. By calculating all the center for the four directories, you get the movement

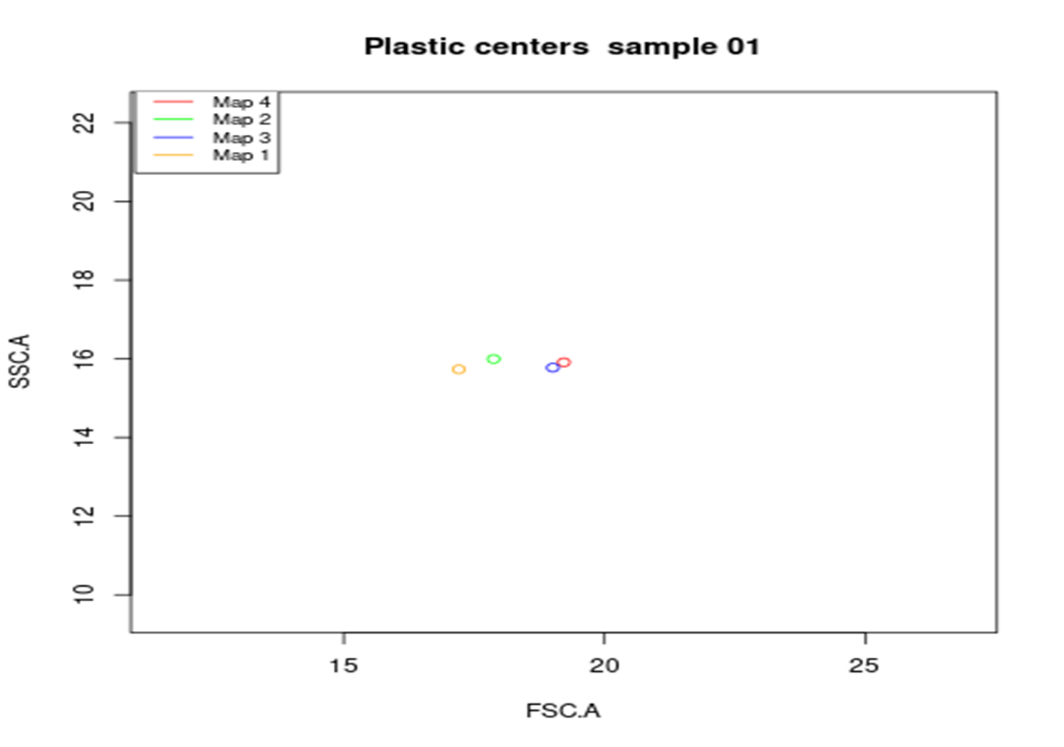


Figure1 6: plastic cluster centers

In above figure 16 the results of the cluster centers of time phase 1 till 4 is shown. Here is where the vague data come in play, it will be discussed over later. Starting by the red dot, followed by the yellow, green, and blue. How farther a dot moves to the left, means that the FSC is getting smaller, less shadow, smaller particle center, more degradation.

It looks like there is first a huge jump in degradation, but then it starts to build up again? This is where the difficult lies in the data, because of the experiments. The code work and produces results, but is this result the same with different data? Lies the fault in the data or the algorithm?

## Pipeline

A pipeline is a piece of code to execute the smaller source codes in an ordered way, which is a fast and efficient. Our pipeline consists of executing command, which activates the R script. This are the scripts in order:

1. Library
2. Convert
3. Merge
4. HDBSCAN
5. Statistics
6. First necessary libraries are downloaded the run the code, if they already exist, then this will be skipped
7. The .fcs file will be converted to .csv files
8. The .csv file per time phase are merged into one data frame
9. The data frames will be passed through the HDBSCAN and clustered
10. The cluster centers will be plotted, even as the clusters per map. Also, the statistics are outputted.

The figure below will show the pipeline (figure 17)

Diagram

Description automatically generated

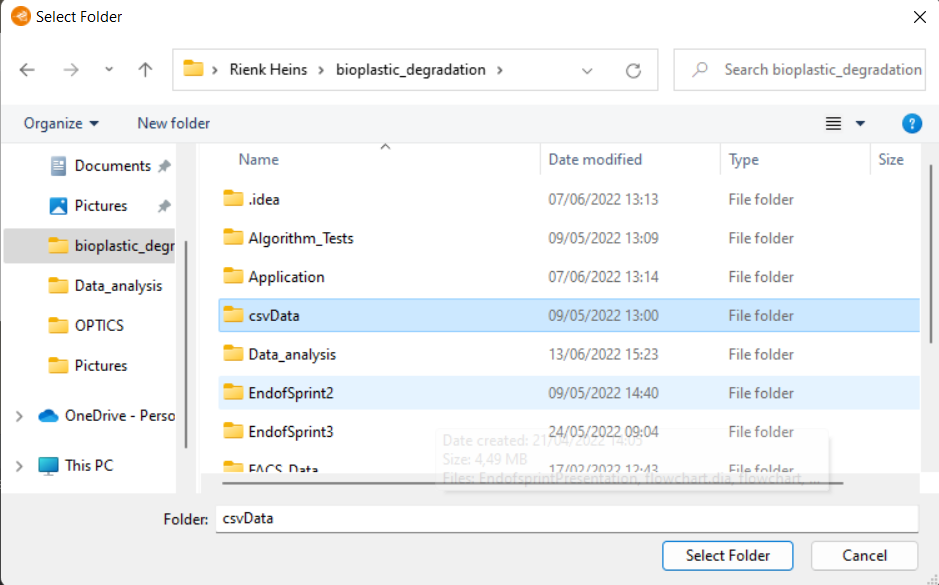
Figure 17: Pipeline

The pipeline is a great way to easily execute multiple codes in series. The outputs are collected in different directories. These are overwritten by a future run.

Repo link = …….?

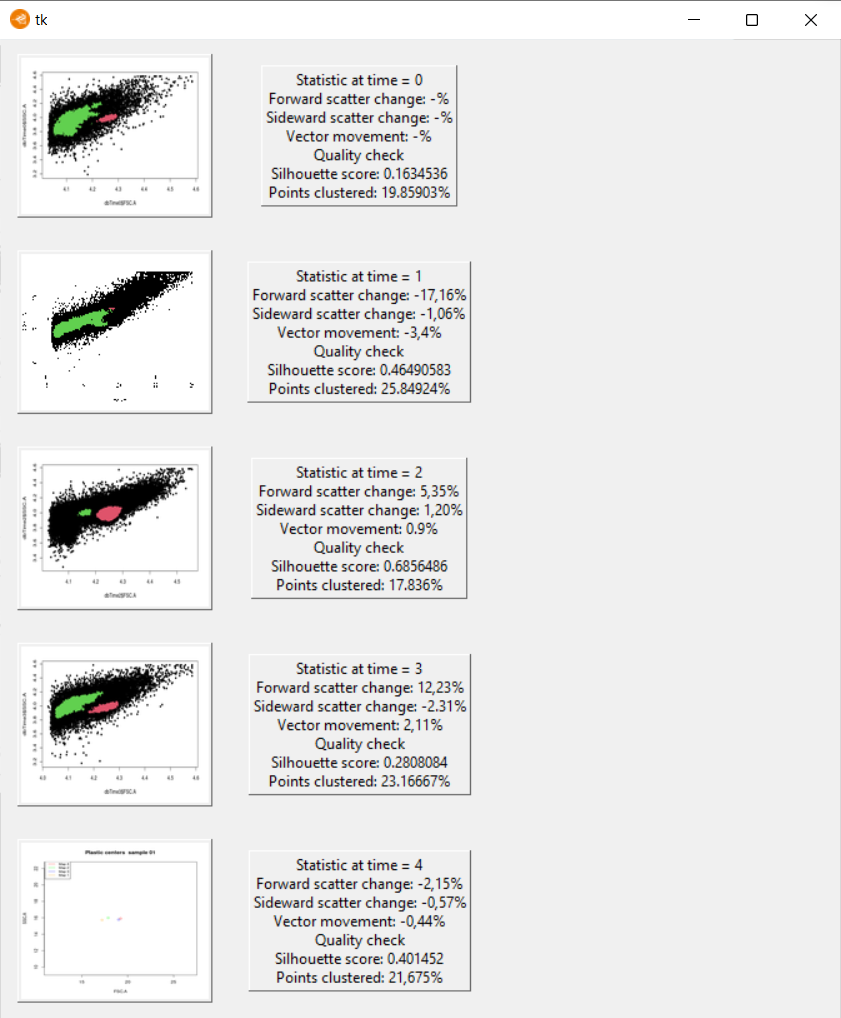
## Application

At the end of the project an algorithm was made in python using tkinter. It functions by asking the user for a working directory with .fcs data, and then using this to activate the pipeline with this data. Then the results of the pipeline will be used to output all the results as seen in figure 18.



The .fcs data can easily be selected by a lab worker, and thrown into the pipeline by the app. The app will then generate the output.

Figure 18: Selecting input



The output of the application is plots with cluster and scores. Also, the plot with the plastic centers is shown.

Figure 19: Output of application

The statistics that are returned are the time, FSC change, SSC change, vector movement, silhouette score and clustered points percentage.

The FSC and SSC changes are relative to the previous time phase. So, 1 is compared to 2, and 3 is compared to 2. The vector movement is also relative to the pervious time.

The silhouette score shows contrast of clusters with each other. And the percentage of points clustered is how much of the data is fitted and gives a statistic for how accurate the result is.

# Conclusion and discussion

## Conclusion

To conclude, an app is made which runs a pipeline that calculates centers cluster of .fcs files to determine the movement over time. This is a tool in a sort of development state that can be used by the lab, under so restriction regarding functionality.

To review the project goal: develop an algorithm to determine the rate of bioplastic degradation by bacteria. And make It easy to uses for lab workers and students. It happened, but it is not optimal. Why Is discussed below

## Discussion

At the end of the project, it was concluded that there was a small degradation of plastics over the 6-week period, with a big degradation at the start and some build up again in the next two time periods. Yet this does isn't the conclusion of the original study. There are multiple problems that might explain this result. Starting with the application itself.

**Application**

One of the problems of the application is that it only classifies around 28 percent of the total points. While it does find the density well this still gives quite some points that might be relevant for the center of the plastics. Maybe the speed of degradation is mostly seen in the breaking down of bigger particles that now aren't clustered. And even with the real center probably being around the density because of the number of points here this still gives a lot of info that is lost. The reason of this problem may lie in the algorithm itself as it was created later than was planned at the start of the project because of the delay in statistical analysis. More time would mean more tests and might have helped perfecting the algorithm to classify points. Yet at the same time the Gaussian Mixture Model and the OPTICS algorithm also had a hard time classifying the data. This might suggest that the data is quite hard to classify, or that the density-based approach just isn't the best one. This also was a problem standing from the later start of the machine learning process as because of this fewer machine learning algorithms were tested. More testing with different algorithms might have given a better end algorithm.

**The Data**

Another source of problems was the data provided. To start flow cytometry data seems to be quite difficult to use to determine the degradation of plastics. As its main attributes to do so, forward scatter and sideward scatter, give a score from the cytometer which doesn't correlate to a certain weight or size. So, it of course can be used to see if plastics are degraded, but it doesn't give a good view of how much plastics this is, how many grams for instance.

Furthermore, there were also problems with the data itself. For starters as pointed out previously, the data was said to be obtained from the same base sample. Yet as seen in the results section there seems to be quite some difference between the four samples in the data. Also, the weird shape of the sample two cluster with the area of the sideward and forward scatter suggests that something went wrong with this data, as it is completely different from all the other samples.

Another problem is that there are a lot of samples that have some sort of extra variable, for example the red staining in the colored data samples or the samples where the freeze taw method was used. But also, one extra variable that was used which hasn't been named yet is the dilution of data. This has been done in a pattern that can be seen as random when looking at the data as it has been given. The problem that this gives in some samples is that there are a lot fewer points. This problem goes away when looking at the full map but made it harder to test the algorithm on single files which was done in the testing phase and for consistency’s sake in the data it would have helped if the dilution would stay equal over the data. All these factors make sense for the original research as the goal there was to find any degradation at all, and the freeze taw method has been very useful in this project to find out which particles are bacteria. But as stated earlier the freeze taw samples had to be removed as the bacteria remnants would create a lot of extra noise points which might be classified as plastics. Because there is no way of knowing which points are remnants and which are plastics it would have been too big of a gamble to keep them in, but with this a lot of data files were removed.

**End Result**

With all of this in mind there is a decent change that the result is incorrect, yet there is no real way of confirming this. One of the biggest problems with the project is that there is nothing to compare the results to. There are some more experiments like the original one being done, yet these results weren't ready by the time of writing this report. Having these other results would have helped a lot in grading the result, as there is degradation but no way to really grade this. Maybe it is a lot of degradation and maybe it isn't. Having more results could give a better way to grade the degradation as you could create a scoring matrix based on multiple results giving a better view of what is a lot of degradation and what is a little. Furthermore, about the buildup of plastics after time 1, maybe this did occur. As the plastics are an energy source for the bacteria they might build it up again over time when a certain threshold is reached. If this is the case it would be seen over an experiment spanning a longer time, this would also give more centers to give a clearer course of degradation over time. Yet this build up might also just be an error created by flaws in the data or algorithm, but this can't be determined without more data to compare it with. If all other experiments show a continued degradation over time the results of this project are wrong but for the time being there is no way to tell. Yet when looking at the results of the original experiment, the problems found in the data and the creation of the algorithm it would be assumed that the result has some flaws in it.

**Recommendations for further research**

The recommendation would be to recreate this project using the data obtained from the new experiments. This new project could focus a lot more on creating a machine learning algorithm that classifies most of the data and to determine what a rapid and slow degradation of plastics is. The statistical analysis of this project has shown a way to quickly change the .fcs data to a .csv dataset which can directly be used for machine learning purposes as the .fcs aren't built to be used in the algorithms. Also, this project has already determined where the plastics and bacteria are in the plots. Because of this the new project can skip most of this statistical analysis and completely focus on the grading of degradation and creation of an algorithm. Further the multiple different datasets could give a much-needed quality control to the project, meaning that problems with the data could much easier be seen. Further it would be recommended to create a dataset with no side variables like freeze taw or coloring and a dilution that stays constant over the data. This way it can be made sure that no side variables disturb the classification.

A further project would be to create a better application. The one created for this project was created in two weeks’ time and is very basic because the priority lay on getting a good result from the clustering. The app only asks for a folder and gives a small output window, yet this could be made a lot more user friendly and uncluttered. This is important as the application is to be used by lab technicians that might not have a lot of computer experience and could be somewhat frightened with the current use of starting the app on the command line or in PyCharm.

# Bibliography

Abou-Zeid, D.-M., Müller, R.-J., & Deckwer, W.-D. (2001, March 30). *Degradation of natural and synthetic polyesters under anaerobic conditions.* Retrieved from ScienceDirect: https://www.sciencedirect.com/science/article/pii/S0168165600004065

Kumar, A. (2022, April 14). *Gaussian Mixture Models: What are they & when to use?* Retrieved from Vitalflux: https://vitalflux.com/gaussian-mixture-models-what-are-they-when-to-use/

Nanocellect. (2020, Februay 22). *How Does Flow Cytometry Work?* Retrieved from Nanocellct: https://nanocellect.com/blog/how-does-flow-cytometry-work/#:~:text=Flow%20Cytometry%20is%20a%20technique,into%20the%20flow%20cytometer%20instrument.

Pilla, S. (2011). *Handbook of Bioplastics and Biocomposites Engineering Applications.* John Wiley & Sons.

Qing-Sheng, L., Mei-Fang, Z., Wen-Hua, W., & Zong-Yi, Q. (2009, january 1). *Reducing the formation of six-membered ring ester during thermal degradation of biodegradable PHBV to enhance its thermal stability.* Retrieved from ScienceDirect: https://www.sciencedirect.com/science/article/pii/S0141391008003327

Rivera-Briso, A. L., & Serrano-Aroca, Á. (2018, Juli 3). *Poly(3-Hydroxybutyrate-co-3-Hydroxyvalerate): Enhancement Strategies for Advanced Applications*. Retrieved from NCBI: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6403723/