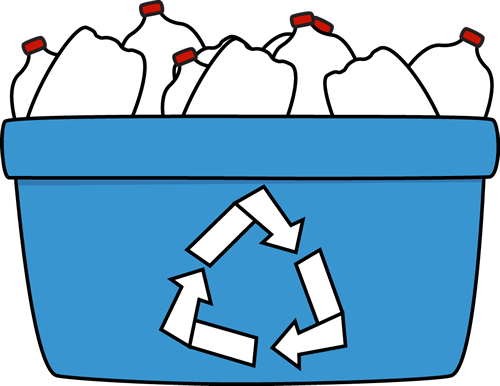
Bioplastic degradation by bacteria

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

410729

BFV3

27-06-2022

Dave R.M. Langers

# Preface

**# Todo**

# 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

Contents

[Preface 2](#_Toc107406741)

[Abstract 3](#_Toc107406742)

[List with abbreviations 4](#_Toc107406743)

[1. Introduction 6](#_Toc107406744)

[2. Theory 7](#_Toc107406745)

[2.1 Lab experiments 7](#_Toc107406746)

[2.2 PHBV 7](#_Toc107406747)

[2.3 Flow Cytometry 8](#_Toc107406748)

[3. Material and methods 9](#_Toc107406749)

[3.1 Materials 9](#_Toc107406750)

[3.2 Method 9](#_Toc107406751)

[3.2.1 Data exploration 10](#_Toc107406752)

[3.2.2 Clustering algorithm 13](#_Toc107406753)

[3.2.2.1 Gaussian mixture model 14](#_Toc107406754)

[3.2.2.2 OPTICS algorithm 15](#_Toc107406755)

[3.2.2.3 DBSCAN 16](#_Toc107406756)

[4. Results 18](#_Toc107406757)

[5. Conclusion and discussion 19](#_Toc107406758)

[Bibliography 20](#_Toc107406759)

[Appendix 21](#_Toc107406760)

# 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 a number of researches 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. 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.

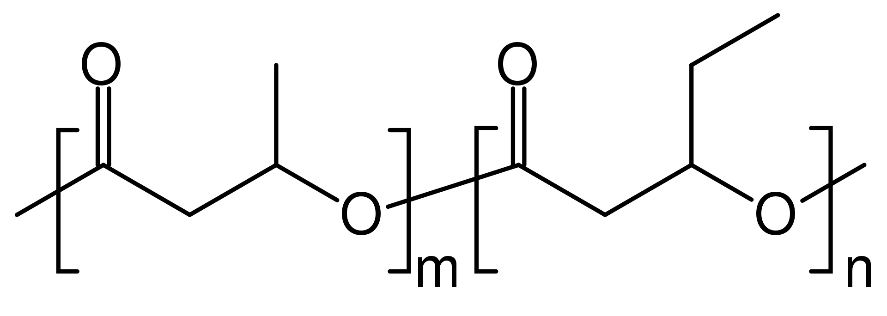


Figure 1: Molecular Structure of PHBV

## 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)

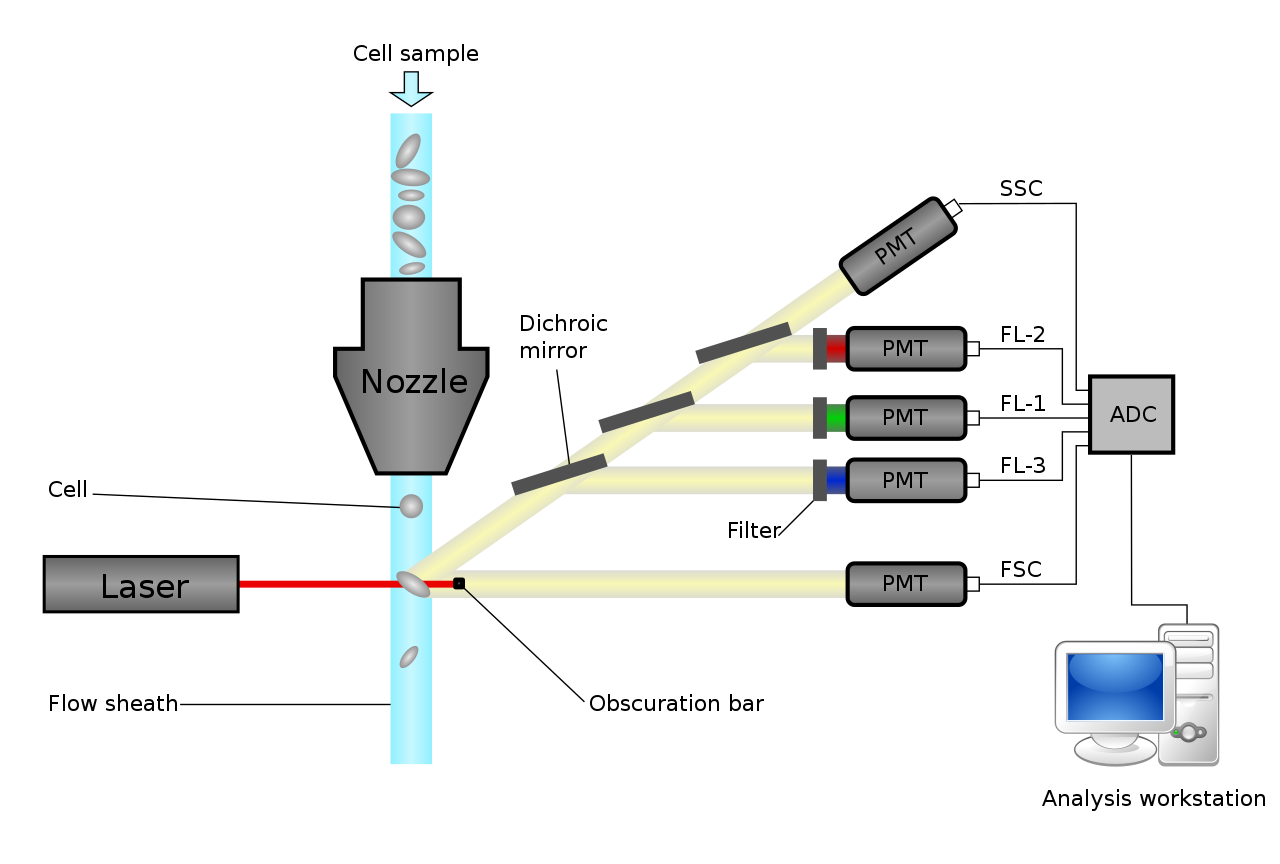


Figure 2: Schematic diagram of a flow cytometer

In above diagram (figure 1) 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 the 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 algorithm

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.

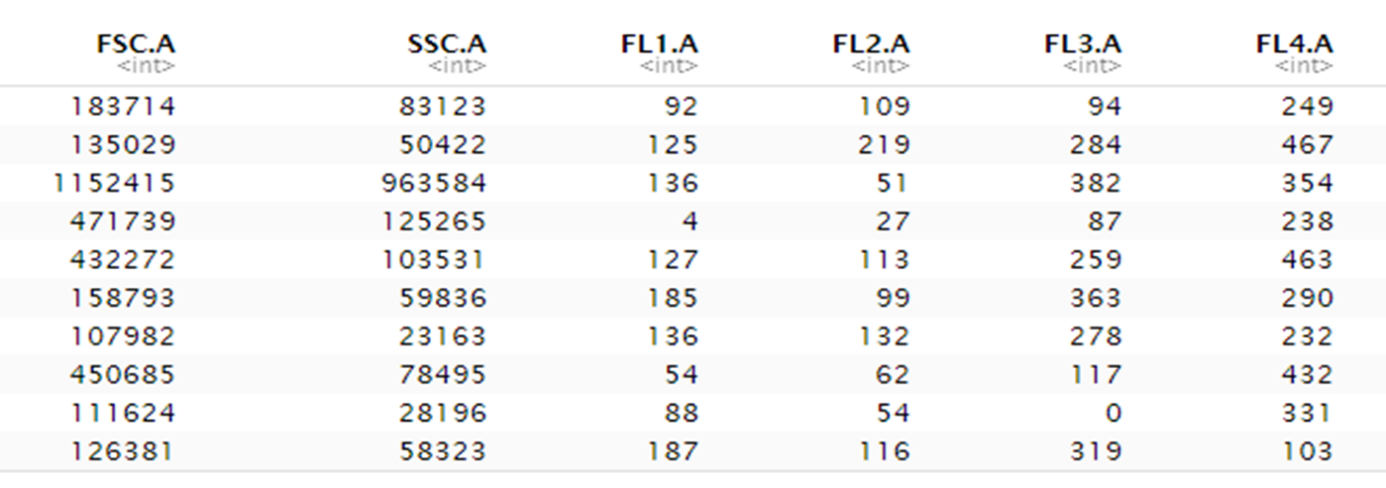
The conversion from .fcs to .csv was more difficult than was expected because R didn’t had build in support for .fcs files. This is where the FlowCore package can in play. This package contains functions to easily open and then save them to workable csv file formats. Below (figure 2) is an example of 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 3).

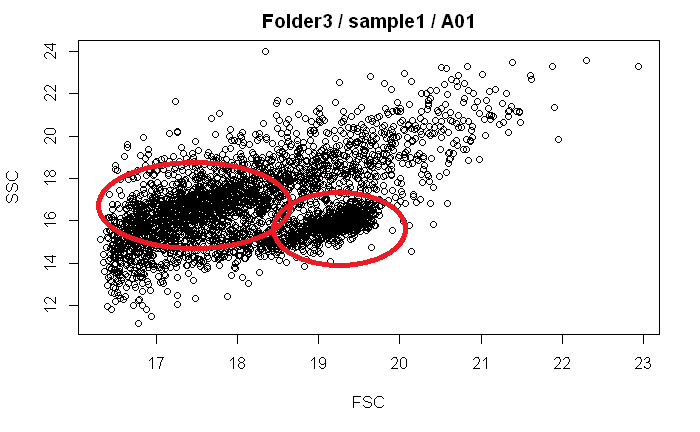


Figure 4: Example of plotting log2 transformed data

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 than 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 4)

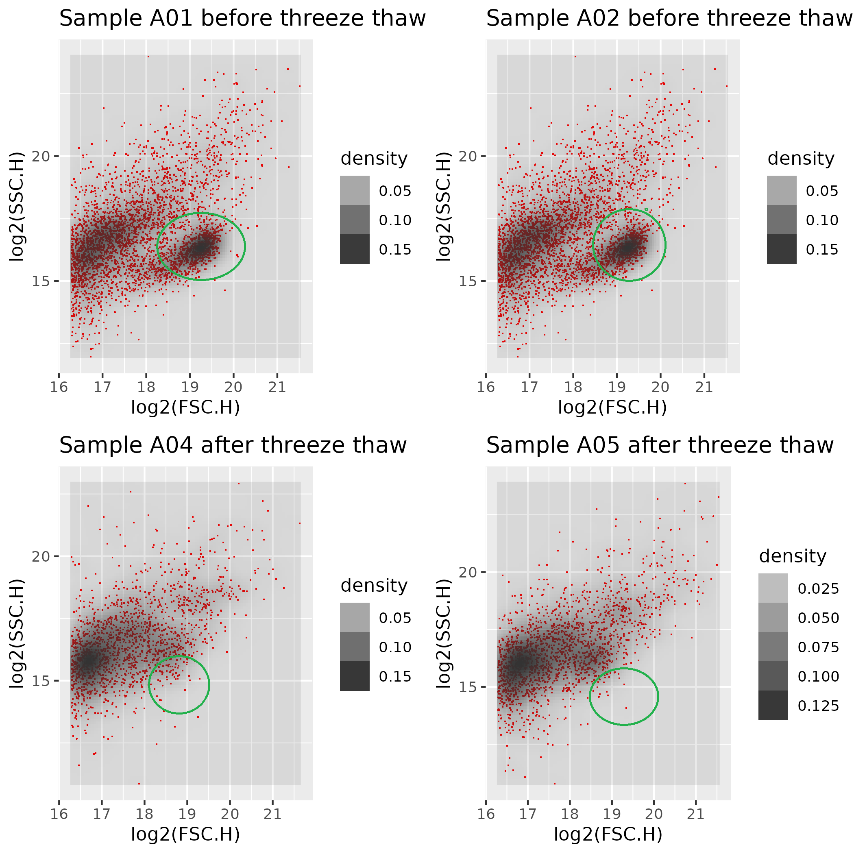


Figure 5: Before and after freeze-thaw experiments

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.

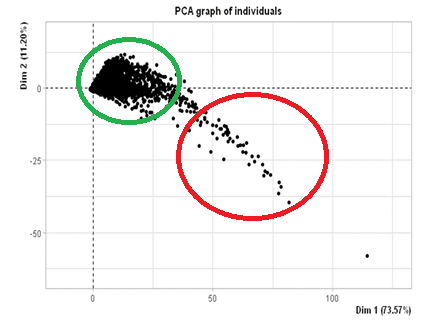
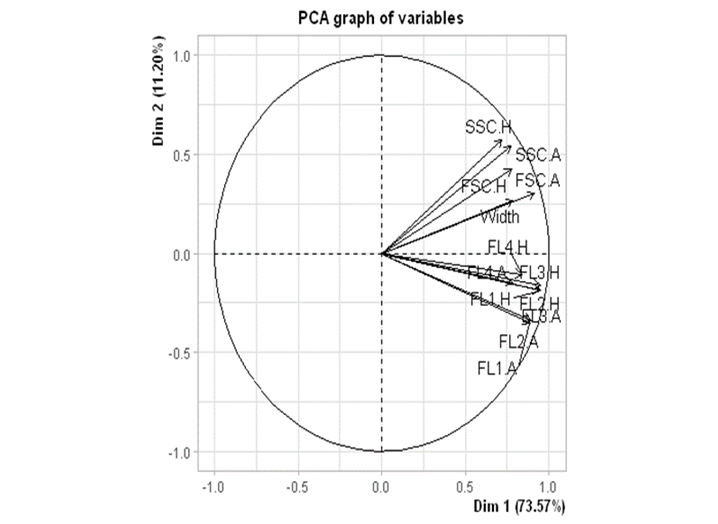
This observation was only concluded of plotting two variables, while there are 12 other variables as well. Plotting all variables against each other is a way of doing this, but it takes much time and effort. To plot all 14 variables on a 2D graph, the dimensions need to be reduced. PCA is an excellent 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 6: PC biplot

Figure 5: PCA plot of PC1 against PC2, sample1 / map1

In the plot (Figure 5), the shape of the points does not look like something with clear clusters anymore. Figure 6 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 5 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 as 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 7) is the density plot of the dataset of folder 1 t/m 4.

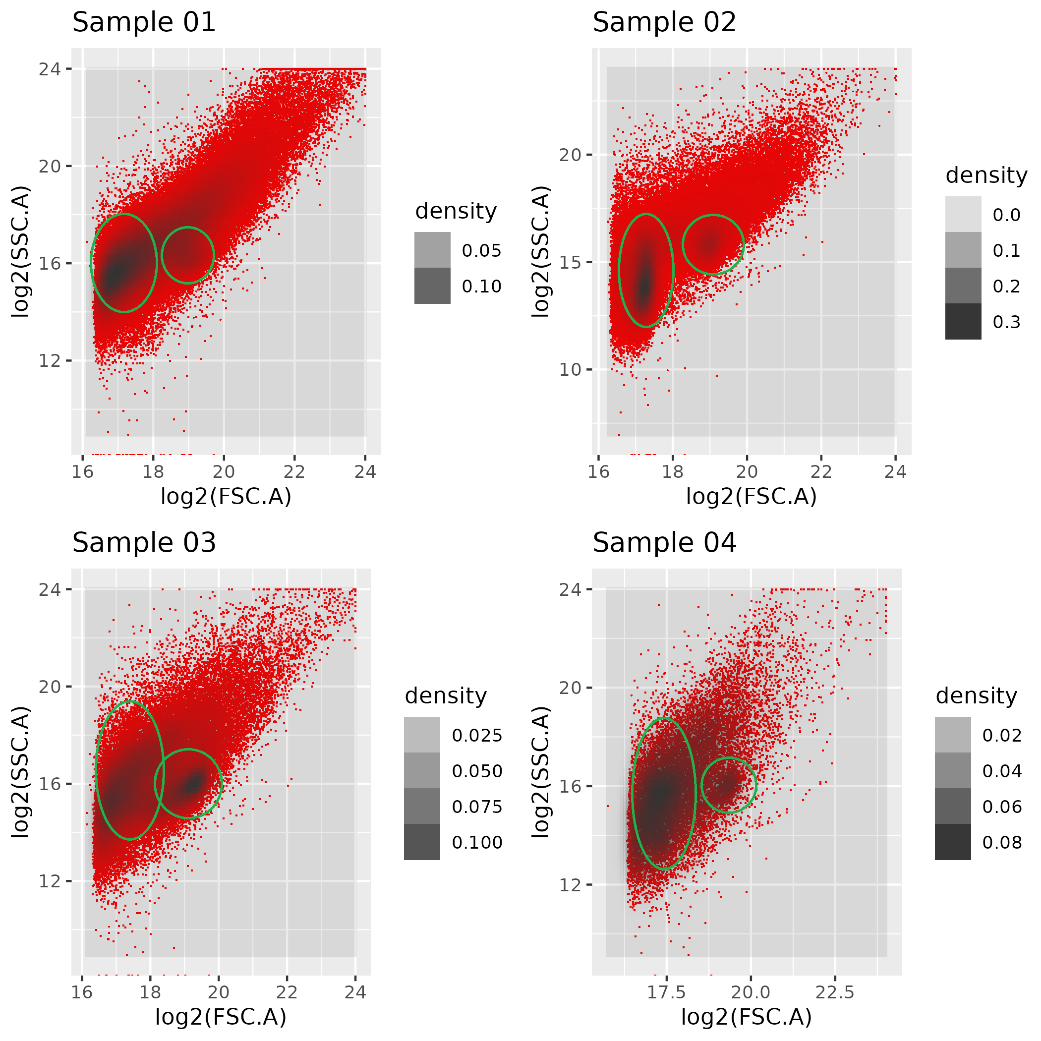


Figure 7: Density plot of the datasets

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 algorithm

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 there own advantages and disadvantages.

#### Gaussian mixture model

#### OPTICS algorithm

#### 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 ?)

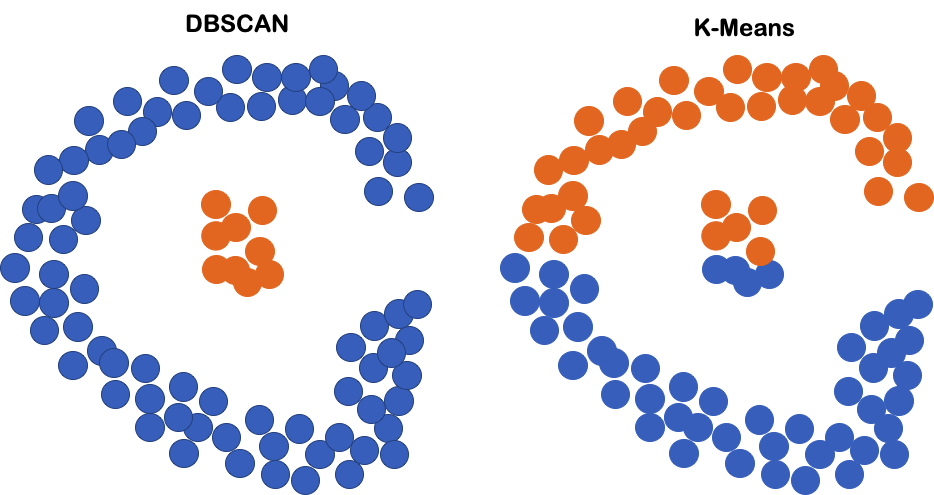


Figure ?: DBSCAN vs K-means

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 ?):

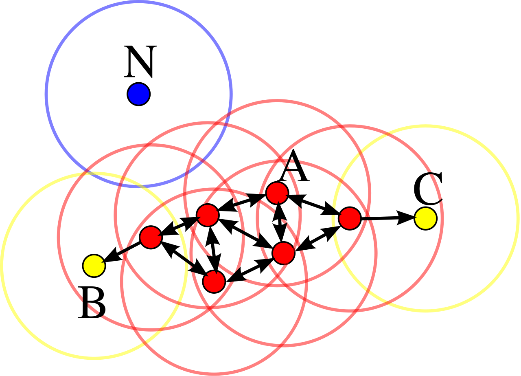
The DBSCAN takes two arguments: Epsilon (*ε* ) and minPts

Figure ?: 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.

# Results

*Presenteer je resultaten in een logische volgorde.*

* *Benoem hier de eigen (gemaakte) materialen, methoden en resultaten. De materialen, methoden en resultaten uit de literatuur benoem je niet hier, maar in het vorige hoofdstuk.*
  + Beschrijf wat je zelf in het kader van dit onderzoek hebt ontwikkeld aan methodologieën. Dat kan een filterstrategie zijn, of parallellisatie op een rekencluster, maar ook workflow beschrijvingen horen hier thuis.
  + Geef de naam van scripts/programma’s en waar deze (in je repo) te vinden zijn.
  + [Zorg ervoor de je repo logisch is ingericht met goede Readme document(en)]
* *Beschrijf wat er te zien is en werk toe naar het beantwoorden van je doelstelling.*
* *Beschrijf indien mogelijke de logica van opeenvolgende fases.*
* Structureer je uitleg in paragrafen. Behandel in elke paragraaf één kerngedachte of onderwerp. In elke paragraaf worden de feitelijke conclusies gegeven (bv ‘de vergelijking laat zien dat het gemiddelde van groep A significant afwijkt van het gemiddelde van groep B, met een p-waarde van … Dit komt niet overeen met de in de literatuur (REF) beschreven waardes.’)
  + Als je het woord significant gebruikt, of aanzienlijk, of duidelijk, hoort er een statistische test bij!
* *Gebruik zo veel mogelijk figuren om je informatie uit je resultaten over te brengen, gebruik tabellen wanneer figuren minder geschikt zijn. In de tekst worden figuren en tabellen geïntroduceerd, besproken en de belangrijkste aspecten toegelicht.*
* *Voorzie je figuren van een nummer en een beschrijvende titel. Zorg voor correcte as-labels (eenheid en grootheid), legenda en bijschrift. Controleer of alles nog leesbaar is wanneer ze zijn opgenomen in het verslag!*
* ***Boven*** *de tabel geef je een titel/literatuurverwijzing inclusief tabelnummer.*
* ***Onder*** *de grafiek geef je een titel/literatuurverwijzing inclusief grafieknummer.*
* Verwijs naar en bespreek in de tekst altijd je figuren en tabellen – die staan nooit op zichzelf.
* *Neem hier een schets op van wat je hebt gemaakt, een activiteitendiagram en eventueel een serie screenshots om zicht te geven op wat je hebt gerealiseerd.*

3.1 Data acquisitie en preprocessing

* Beschrijf de data preprocessing en de resultaten ervan.

3.2 Validatie en performance

* Beschrijf hoe je de betrouwbaarheid van je eigen methodologieën en code hebt geborgd.

3.3 Onderzoeksresultaten

* Beschrijf de eigenlijke onderzoeksresultaten.

# Conclusion and discussion

* *Formuleer je conclusie door eerst in te zoomen op je eigen data en daarna uit te zoomen.*
* *Zoom in door je resultaten samen te vatten. Zoom uit om de waarde van je werk te beoordelen, door je bijvoorbeeld de volgende vragen te stellen:* 
  + *Kunnen mijn resultaten gebruikt worden in het werkveld? Wat betekenen ze voor het werkveld?*
  + *Zijn mijn data betrouwbaar?*
* *Bespreek de resultaten zodanig dat je ze ter discussie stelt, wees kritisch.* 
  + *Vergelijk je resultaten met de literatuur of eerder ontwikkelde data.*
* *Geef aanbevelingen voor een vervolg en staaf je aanbevelingen door de impact op wetenschappelijk of maatschappelijk vlak te beschrijven.*

*4.1 Samenvatting van de resultaten*

*• Herhaal hier letterlijk je doelstelling en koppel je resultaten eraan vast.*

*• Vat de resultaten samen en geef jouw interpretatie ervan.*

*• Bespreek de consequenties van je resultaten.*

*• Bespreek de betrouwbaarheid van je resultaten.*

*4.2 Discussie*

*• Vergelijk jouw resultaten met gegevens uit de literatuur en bespreek verschillen*

*hiermee.*

*• Bespreek opvallende en verrassende resultaten.*

*• Bespreek zwaktes van je onderzoek en hoe deze aangepakt zouden kunnen*

*worden.*

*4.3 Algemene conclusie en perspectief*

*• Bespreek wat je doel was, wat het eindresultaat is en hoe vanuit hier verder*

*gewerkt zou kunnen worden.*

*• Bespreek de sociale en/of wetenschappelijke implicaties van je werk.*

# 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

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/

# Appendix

**# Code?**