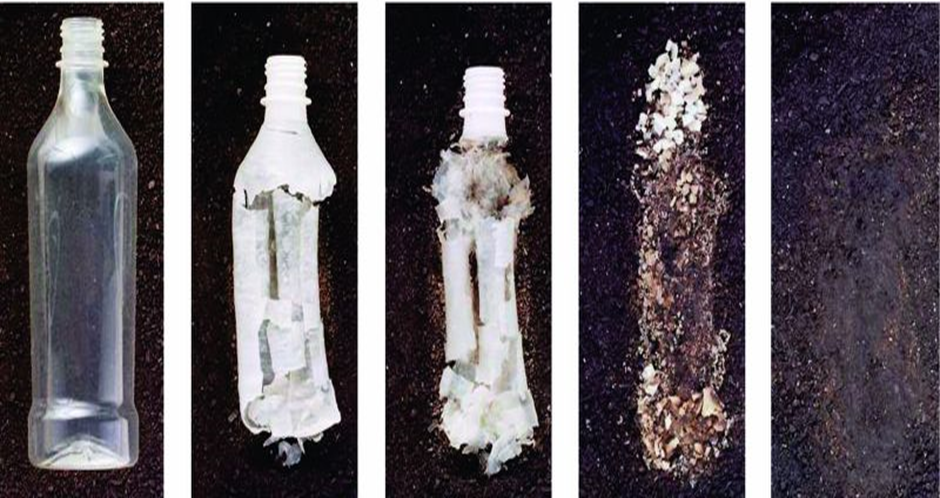
**Bioplastic degradation by bacteria**

A machine learning based research



Meulenkamp LM, Rob

[l.r.meulenkamp@st.hanze.nl](mailto:l.r.meulenkamp@st.hanze.nl)

Heins, RD, Rienk

[r.d.heins@st.hanze.nl](mailto:r.d.heins@st.hanze.nl)

Visser P, Pascal

[p.visser@st.hanze.nl](mailto:p.visser@st.hanze.nl)

Figure 1: bioplastic degradation over time

# 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:

* Brittle
* Low elongation at break
* Low impact resistance

The applications for PHBV are:

* 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.

The goal of the project is to analyze obtained research data from flow-cytometry and develop with machine learning an analysis pipeline that can calculated the degradation of PHBV by bacteria. With this pipeline, researchers can use it to calculate degradation by different bacteria spices.

The obtained data is from 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. The scattering of light means the deflection, by diffraction of light against particles. Bacteria or plastics 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.

# Approach

The main approach can be divided in two parts:

* Data exploration
* Machine learning

In the first part of this project, we were focused on analysing the dataset and to prepare it for machine learning. In short, the first part was entirely spent on data exploration.

To perform the exploration, we had to use some tools and techniques. The main tool that we had, was the statistical programming language R. We have used R version 4.1.2 in the IDE Rstudio (local/online). Furthermore, we used the scrum technique to evaluate and discuss progress and aberrations. We used Trello to keep track of our progress and tasks.

The first challenge was to understand the dataset. The. fsc data retrieved from flow cytometer experiments with bacteria/ PHBV samples, was divided into 4 folders with different files. The file is these folders we sometimes used in experiments such as colouring and freeze thaw. These experiments made the organisation of the files somewhat vague.

This vagueness proved to be the biggest challenge, so multiple conversations took place with the client to better understand the dataset. This resulted in a clearer view of the files and their contents.

After this, several statistical programming steps were made such as: density plots, PCA, finding clusters and plotting different columns against each other.

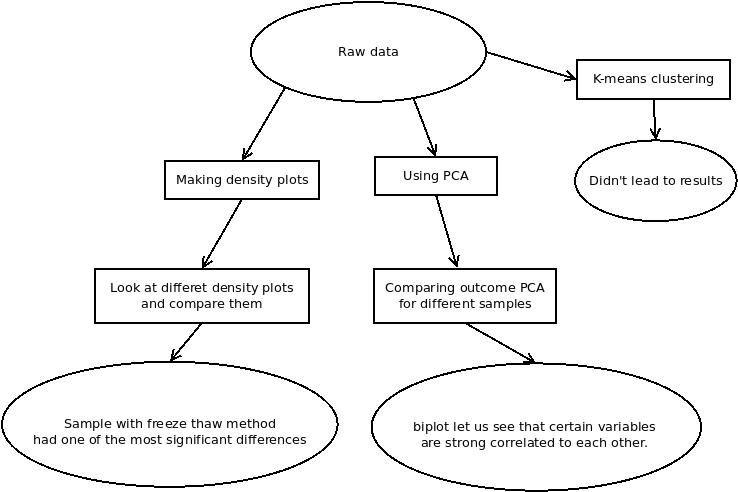


Figure 2: flowchart of approach

The approach for analyse the raw data, was to select a couple of analyse techniques and assign them to a member of the project group. With the assigned technique, plots were made and compared to each other to see if there were any differences or similarities

This comparing was very time consuming but, we had no faster way at the moment. But two of the three approaches resulted into useful plots and graphs.

# Preliminary results

The first results or the big and only result, is al clear prepared dataset for machine learning of with the characteristics are known. To achieve this, several plot and graph were created to visualize this.

With clustering, our goal was to distinguish bacteria from plastic particles. But this proved no easy task, there were no annotations in the dataset of what was a bacterium or a plastic particle. In the freeze thaw experiments, bacteria were killed due the cold environment. So, by clustering the before and after. Differences had to been seen. In the graphs below the results are shown.

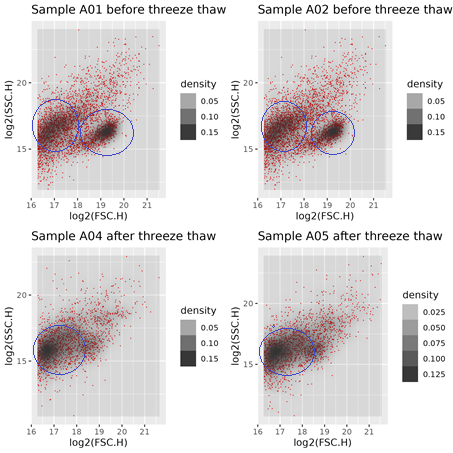


Figure 3: freeze thaw plots

There is clearly one cluster missing after the freezing, which indicates that, these cluster must be the bacteria.

Furthermore, we have used principal component analysis (PCA) to sees ‘hidden’ correlation between all the variables. PCA has let us see, that the expected variables are stronger correlated. And that some variables are not correlated at all.

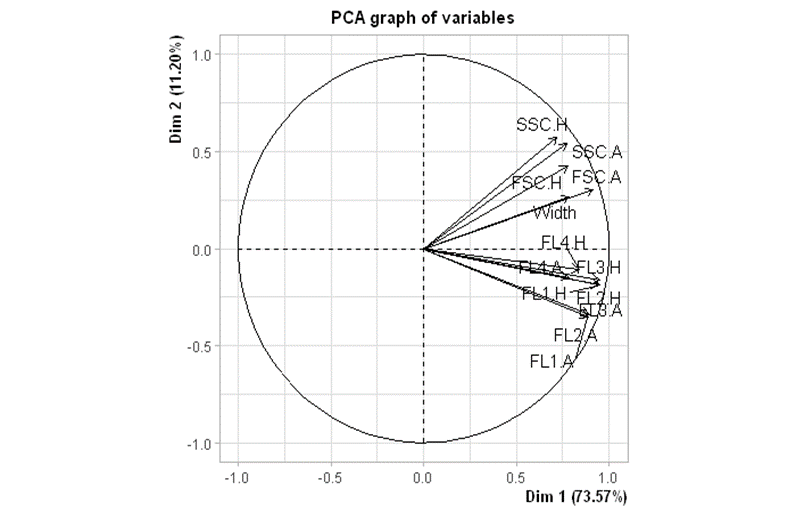
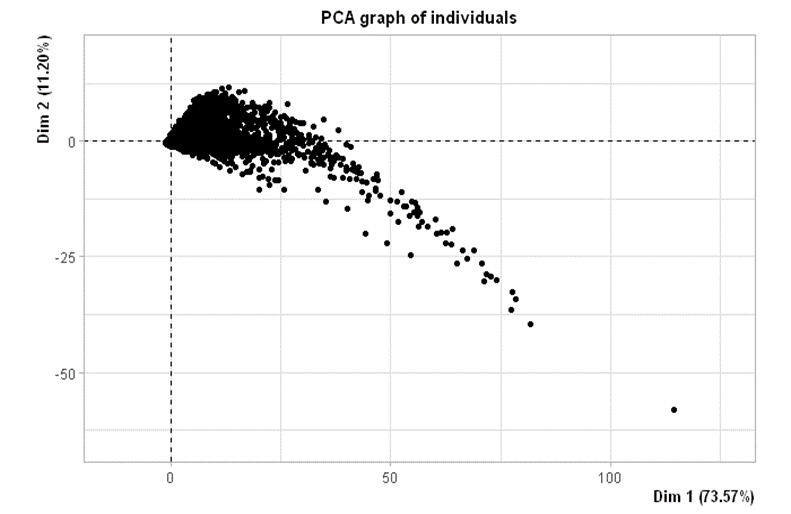


Figure 5: PC1 vs PC2

Figure 5: PCA biplot

The biplot shows that certain variables are heavy correlated, where other are loosely correlated or not at all. The plotting of PC1 against PC2 showed that there are not clear clusters in this sample. But, known is that the ‘tail’ of the cluster consists of FL variables, which are not correlated to the scatter variables, as is shown in the biplot. So, these group together.

Also in early stages is k-means clustering preformed, but this showed nothing special.

# Discussion and outlook

The biggest flaw in this first half, was the understanding of the dataset. If we earlier had known the structure and organization of the data. Some unknows could be filtered out sooner. This was a case of a certain lack of clear communication between the client and us. This point is and big one to improve in the future.

At our opinion, we spent to long on the data exploration. Our initial goal was to do the exploration within one sprint in stead of two. But these things happen, and it was a good and educational moment for our as project group.

The strength of this first half was the determination of going trough the difficulties along the way. And trying to get a result form, what was for us, a messy dataset. And we are very please with the outcome so far.

Following up to the next sprint, we first going to talk with the client to see if we are on the same track. After that, machine learning is the one to go with. Our goal at the end of sprint three is to create a concept machine learning model which we can optimize in sprint four.