Predicting Radiotherapy Cancer Treatment Quality

by

Sander Liebens
Huang-Da Chi

in partial fulfillment of the requirements for the degree of

Bachelor of Science
in Computer Science

at the Delft University of Technology,
in cooperation with
Erasmus MC Daniel Den Hoed - Rotterdam

 Supervisors:  Dr. ir. E. Taskesen  TU Delft
              Ir. W. Schillemans  Erasmus MC
Try not to become a man of success, but rather try to become a man of value.

ALBERT EINSTEIN
Abstract

A person with cancer has several treatment options. One of which is radiotherapy. Radiotherapy is treatment of cancer with radiation. To minimize the damage to healthy tissue, radiation is applied from several directions into the body. When treating cancer with radiotherapy, the organs nearby the tumor are at high risk of getting damaged. In the treatment plan the dose to the organs at risk has to be balanced with the dose given to the target.

These calculations are nowadays done by medical personnel. Although a lot of treatments succeed, without much damage to healthy tissue, a lot of treatments do serious damage to the organs at risk. Can treatment plans be optimized in terms of organ sparing?

To reach optimization, several methods have been executed in order to create groups within a patient set. 115 patients of prostate cancer have been analyzed using Principal Component Analysis and Agglomerative Clustering. The data consist of Overlap Volume Histogram values of the bladder and rectum in a CSV file. Each CSV file contains 201 values. These CSVs are used as an input for both methods.

This led to several figures as results. The principal component analysis showed that 80% of the data is covered by the first principal component and 92% by the first and second. Also, a scatterplot has been made, which shows the transformed data. This scatterplot shows no subgroups can be identified with the bladder and rectum data of the patient. The Agglomerative Clustering method results in six plots. A variation in linkages and connectivity has been used, but all six led to no clear distinction within the data.

These results led to the conclusion that no subgroups are distinguishable based only on OVH data and no prediction can be made that optimizes radiotherapy plans based solely on OVH data of patients.
Preface

This bachelor project was done by two students of the TU Delft Computer Science bachelor in collaboration with the Erasmus MC - Daniel den Hoed cancer clinic, the radiotherapy department. This report contains an introduction about radiotherapy, as well as the research, implementation and the results of our project. Wilco Schillemans, clinical physicist at Erasmus MC and leader of the ICT group at the radiotherapy department, led the project. Erdogan Taskesen, postdoc at TU Delft, bioinformatics department guided us during the project.

We would like to thank a few people. First and most important, Wilco, for the opportunity to work on this interesting project and his guidance when needed. Erdogan, for the help and explanation of subjects which were hard for us to understand.

It was a great experience to work for the Erasmus MC, radiotherapy department. We got a good view of what is done by professionals and what technologies are needed to do the job as good as possible. We hope that our work will be further developed and analyzed, so that it can be used by medical doctors for treating patients.
# Contents

1 Introduction Radiotherapy ........................................ 5

2 Problem ................................................................. 7
   2.1 Problem Definition .................................................. 7
   2.2 Goal ........................................................................ 8
   2.3 Possible Methods and Analysis ...................................... 8
      2.3.1 Shape-based treatment plan optimization .................... 9
      2.3.2 PTV-rectal distance based treatment plan optimization ... 9
      2.3.3 IMRT treatment plan quality control .......................... 10
      2.3.4 Quantitative analysis of the factors which affect OAR dose sparing variation in IMRT plans ...................... 10
   2.4 Software ................................................................... 11
   2.5 Pre-Implementation ................................................... 11

3 Methods ...................................................................... 13
   3.1 Data ......................................................................... 13
   3.2 Principal Component Analysis ..................................... 13
   3.3 Clustering .................................................................. 15
   3.4 Implementation .......................................................... 15
      3.4.1 PCA implementation ............................................. 15
      3.4.2 Clustering implementation ................................... 16
      3.4.3 Testing the implementation .................................... 16
      3.4.4 SIG ................................................................... 18

4 Results .................................................................... 20
   4.1 Data ......................................................................... 20
   4.2 Principal Component Analysis ..................................... 23
   4.3 Clustering .................................................................. 25
   4.4 Feature analysis .......................................................... 27

5 Conclusion & Discussion ............................................... 28

6 Future Perspectives ..................................................... 30

Appendices ....................................................................

A Executive Summary .................................................... 31

B Infosheet ...................................................................... 32
1 Introduction Radiotherapy

Radiotherapy is the treatment of disease with radiation, especially by selective irradiation with x-rays or other ionizing radiation. [2].

The Erasmus Medical Centrum (Erasmus MC) Daniel den Hoed, located in Rotterdam - the Netherlands, is a cancer treatment location in the Netherlands. The Daniel den Hoed location specializes in only treating cancer, but in several ways. Some examples are hyperthermia (heating of tissue to 45 degrees Celsius), chemotherapy (chemical substances), surgery, and of course radiotherapy.

At the Erasmus MC, a multidisciplinary approach is being used. Firstly, surgery is done to remove the tumor if possible. Afterwards, chemotherapy and radiotherapy (with additional hyperthermia in selected cases) is used to remove the metastasis of the tumor and make sure the parts that surgery did not remove are being treated as well. About 50% of the patients treated at the Erasmus MC are treated with radiotherapy.

![Figure 1: Figure of the treatment of a patient by radiotherapy](image)

Figure 1: Figure of the treatment of a patient by radiotherapy

As can be seen in figure 1 during radiotherapy treatment, a patient lies on a
couch and should remain in this state during fractions of about twenty minutes. Fractions of the treatment are spread over multiple days. During the treatment, the linear accelerator moves in a circle around the patient, to make sure that the healthy tissue is being spared as much as possible, while the tumor still gets the full treatment. This is done because radiation loses power when passing through obstacles. Skin tissue fortunately will not get any radiation, because of the skin-spearing effect. Tissue about two cm below the skin of a patient will always get 100% of the radiation provided, while the tumor, which is deeper inside the patient, will only receive about 60% of the radiation. This is of course different for every patient, because each will be located differently.

It is also very important that the patient stays as still as possible during a fraction, because even the slightest movement could cause a partial miss of the tumor by a radiation beam. Two strategies are being used. The first strategy is adding margin to the target area, this is done to most of the patients. The second strategy is making use of technology that tracks body movement. This is done by following the movement of the tumor, while the patient breaths. Whenever the machine thinks it knows the patients movement rhythm, it will start radiating automatically, making the same moves and additions as the tumor does during breathing. This strategy is only being used for a small group of patients.
2 Problem

This section contains the research question. In addition, the software used by the Erasmus MC will be explained and some thoughts about the implementation will be shared.

2.1 Problem Definition

In the department of Radiotherapy, treatment planners have to be very precise to minimize damage to healthy cells when producing radiotherapy plans while not diminishing the constraints on tumor radiation. When a patient is diagnosed with cancer, an X-RAY Computed Tomography scan (X-RAY CT) and a Magnetic Resonance Imaging (MRI) scan will be taken to position the tumor. These scans are then judged by the medical doctors (MD).

Can we assist the MDs and planners in deciding whether the plan is optimal or not? Can treatment plans be optimized in terms of organ sparing?

The tumor contours with added margin is called the Planning Target Volume (PTV). The margin is added to the contours of the tumor, because patients move unintentionally during radiotherapy sessions, which could take up to half an hour. If margin is not added, certain parts of the tumor could be missed because of unintentional patient movement. This will not result in a cured patient, because parts of the tumor that do not receive radiation are able to grow.

When widening the margins of the marked tumor, organs near the prostate (the bladder and rectum) or breast (heart and lung) will receive more radiation, which could result in negative side effects such as diarrhea or rectal bleedings for prostate cancer and heart diseases for breast cancer. Even the chance of getting cancer in the bladder, rectum or lung increases when the radiation exceeds certain limits. The organs near a tumor, which need to be taken into account when drawing this line are called organs at risk (OAR).

When the planners are done, the planning will go to the radiotherapists. A radiotherapy treatment will take place in a number sessions that are spread over a certain time period. This can vary from only one session to 40 sessions.

A challenge is whether a treatment plan can be optimized in terms of OAR sparing. Suboptimal plans may be inadvertently approved or dosimetrist may waste time trying to optimize a plan that is already maximally optimized. If the treatment planners were aware of the achievable degree of OAR sparing initially, plans could be generated in more efficient fashion, potentially resulting in less toxicity and possibly saving time during the planning process [5].
2.2 Goal

A method is desired, which could predict the optimal OAR sparing in a radiotherapy plan, given the PTV and OAR geometry. The method will compare a new patient with a data set of old patients to suggest an optimal achievable radiation dosage of a radiotherapy plan for any particular cancer patient. This method has to comply with the coding standards set by the radiotherapy department of the cancer institute of the Erasmus MC: Daniel den Hoed.

An optimal achievable radiation dosage plan is one that spares the OAR as much as possible without altering the dosage constraints to the radiotherapy plan, possibly even increasing the PTV without increasing the dosage to the OAR.

In order to do this, dose-volume histograms (DVH) will be predicted for new cancer patients with the optimal achievable radiation dosage. A DVH is a histogram, plotting the amount of Gray a part of the body receives at a certain volume. For example, as can be seen in figure 2: 38.2% of the bladder receives 40 Gray of radiation, while 73% of the rectum receives 40 Gray of radiation.

![Figure 2: A Dose Volume Histogram, showing the bladder (yellow), rectum (brown) and prostate (pink).](image)

2.3 Possible Methods and Analysis

Three state-of-the-art methods are described in literature that can predict DVHs as accurately as possible. These methods are summarized in the next three sections.
2.3.1 Shape-based treatment plan optimization

This method proposes a model to assess the quality of an Intensity-Modulated Radiation Therapy (IMRT) treatment plan using data of prior cancer patients that underwent radiotherapy. An important parameter in this method is the proximity to, or the overlap of the OARs with the PTV. The overlap volume histogram (OVH), is an essential tool to describe this relationship. The $OVH(r)$ is a one dimensional function that describes the fraction of the OAR volume that is encompassed by a uniform expansion or contraction of the PTV by a distance $r$.[5]

Achievable dose-volume objectives (i.e., points on the DVH curve) are predicted by searching in the database of all prior patients for patients where the OAR were more difficult to spare (i.e., which required a smaller PTV expansion distance to cover the same volume). The minimum achieved dose among the selected patients is determined and used as approximation of the lowest achievable dose to the OAR of the new patient.

According to the research report of Petit et al., the achieved doses using this method yielded results that differed by less than 0.82 gray (Gy, $Gy = J/kg = m^2/s^2$) from the predicted lowest achievable dose to the OAR.[5]

Also, the predicted dose for the liver was achievable within 1 Gy for 90% of the cases and for 98% of the cases within 2 Gy. For the kidneys the percentages were 82% within 1 Gy and 94% within 2 Gy. The predictions were thus achievable in most of the cases.[5]

It was demonstrated that the predicted doses were achievable within 2 Gy for >94% of the patients and that they were a good approximation of the best achievable dose.[5]

According to the results of the research done into this method by Petit, the mean liver dose decreased on average by 1.4 Gy (range 0–4.6 Gy) and the mean kidney dose decreased 1.7 Gy (range 0–6.3 Gy) after re-planning. This demonstrates the relevance of this model for organ sparing.[5]

This method’s main advantage is the ease of implementation and the positive results.

2.3.2 PTV-rectal distance based treatment plan optimization

This method is a quality assurance method (QA-method), which is based on the distance between tumor and OARs.

Assume that the dose in a voxel depends on the distance of the voxel to the PTV. A large distance to the PTV implies a low dose. This method might not always be realistic, because it implies that the percentage of an organ that
receives most radiation is also the percentage of the organ that is closest to the PTV. If for two patients, the closest percentage of the voxels lies closer to the PTV of one patient than the other, it would mean that the minimal dose should be lower for the first patient. [7].

2.3.3 IMRT treatment plan quality control

Quality control is done by patient geometry-driven information and relies on searching in a database of related cancer patients, comparing the geometry of the area of the tumor with the data of prior radiotherapy patients. This is done during the process of making radiotherapy or after so it can be determined how much dosage decrease could have been possible [8].

The DVH and OVH of the related patients are then compared to determine how much the patients are related and if there could be improvements to the current radiotherapy plan. Basically, this method is a manual version of the Shape-based treatment plan optimization method and is more widely used as a means of quality control.

2.3.4 Quantitative analysis of the factors which affect OAR dose sparing variation in IMRT plans

This method, is an evidence-based approach to quantify the effects of patient anatomical features of the PTVs and OARs and their spatial relationships by learning from high-quality prior plans [9].

In this paper by Yuan, the dependence of OAR DVHs on patient anatomical factors is analyzed and learned from prior plans by a stepwise multiple regression method. Two major groups of anatomical features are considered to be of importance in this study. The volumetric information and the spatial information. The geometry of OARs relative to PTV is represented by OVHs. Important anatomical and dosimetric features are extracted from OVHs and DVHs by Principal Component Analysis. The results show that the most significant patient anatomical factors contributing to OAR dose sparing are:

1. Median distance between OAR and PTV (spatial)
2. Portion of OAR volume within an OAR specific distance range (spatial)
3. Fraction of OAR volume which overlaps with PTV (volumetric)
4. Portion of OAR volume outside the primary treatment field (volumetric)

The $R^2$ coefficients for predicting the first PC score are more than 0.68 and more than 0.53 for the second PC score. Thus the above set of anatomical features capture a significant portions of the DVH variations [9]. These factors can be incorporated into evidence-based learning models as effective features to provide patient-specific OAR dose sparing goals.
This method has not yet been used at the Erasmus oncology institute and show that major improvements could be made to the existing platform. Therefore, this method was decided to be implemented.

2.4 Software

The software that is used at the Erasmus MC: Daniel den Hoed is Matterhorn. This is an application written by developers at the Erasmus MC. The software was written in C++ and runs Python 2.7.5 Plug-ins for Matterhorn. The Python scripts can be put into the Matterhorn install folder and will then be loaded by the system as a menu option. Whenever this menu-option is being selected, the python runnable will load and the plug-in is started. The Erasmus MC already has a small library of plug-in methods which are able to calculate the OVH or DVH, or other important measures.

Matterhorn is used to analyse patient data. It is able to load patient images from X-Ray CT scans or MRI scans. Beside that, the software is also able to draw contour lines of the tumor, PTV and OARs and show these contour lines in 3D by combining different slices. Another feature is that it can show dose effects within tissue of a patient, to show analysts how much radiation each part of the body has gotten.

Beside these features, Matterhorn has the possibility to load plug-in methods written in Python or C++. The Erasmus MC already has a small library of plug-in methods which are able to calculate the OVH or DVH, or other important measures.

2.5 Pre-Implementation

The method, written in a plug-in, will consist of an algorithm which is able to predict DVH values for a new patient, based on OVHs and body geometry of a set of already treated patients. The algorithm will be split up into several scripts, each with their own runnable function. Therefore, one can not complete the prediction process unless all steps of the algorithm complete successfully. Each step, the results will be written to a CSV file, so that each file can be checked for quality during the process and physicists are able to verify that everything is working as it should be.

For this method, the Quantitative analysis of the factors which affect OAR dose sparing variation in IMRT plans section 2.3.4 will be implemented. This method is chosen, because the it is a method that has not yet been tried out by the Erasmus oncology institute and shows promising results by Yuan’s paper. The method will be implemented in Python 2.7.5 For each runnable and any script related to the runnable, tests will be written. The Erasmus MC has a build server which compiles and runs all software and tests. Beside unit tests, integration tests are run as well, but for now, only unit tests are important. When the software is proven to be valuable for the Erasmus MC and is working
As can be seen in the following diagram, the input will consist of data from old patients and a new patient. This will be analyzed using PCA, after which a comparison is made with the new patient. The method will then decide which old patient(s) will suite the new patient best according to distance in a graph and suggests to look at or use the same treatment plans as have been used with these patients. Possible improvements to this approach might be the use of a regression model.

![Figure 3: A boxdiagram of the plug-in method to be implemented.](image)
3 Methods

This section will explain the methods that have been used during the research and the data that has been used. It will shortly explain what the method does and afterwards discuss why we have chosen to use that method.

3.1 Data

The Erasmus MC provided us sets of anonymized data from 115 patients that we used in our research. Each patient had two data sets, one for rectum data, the other for bladder data. Each data set contained 230 CSV files containing OVH data. 115 files containing OVH towards PTV plus a little margin and 115 files containing OVH towards PTV plus the seminal vesicle and a little margin. In this case, each CSV file was treated as one sample, so there were 230 bladder samples and 230 rectum samples.

![Figure 4: Figure of the Matterhorn software, which shows the geometry of the patient, showing its PTV and the amount of Gray in a 3D picture.](image)

3.2 Principal Component Analysis

Principal Component Analysis (PCA) is a mathematical algorithm that reduces the dimensionality of the data, while retaining most of the variation in the data set. It accomplishes this reduction by identifying directions, called principal components, along which the variation in the data is maximal. By using a few components, each sample can be represented by relatively few numbers instead of by values for thousands of variables. Samples can then be plotted, making it possible to visually assess similarities and differences between samples and determine whether samples can be grouped.\[6\]
These principal components are linear combinations of the original variables and are used to display the pattern of similarity of the observations and of the variables as points in maps [3, 1]. Principal component analysis will help us in determining the most important distinguishing feature of OAR sparing goals, which will in turn help predict the new optimal DVH for the new patient.

PCA uses 4 steps, after the data has been normalized:

1. Calculate the covariance matrix. The covariance matrix is needed to see the relationship between two parameters and to calculate the eigenvector and eigenvalue.

2. Calculate eigenvectors and eigenvalues of the covariance. Each dimension has its own eigenvector and eigenvalues, all perpendicular to each other. These eigenvectors will tell useful information such as what datapoint has most variance.

3. The eigenvector with the highest eigenvalue is the principal component of the dataset. The second highest is component two, etc. This is in order of highest variance; the more components you leave out, the less varied your analysis will be. Now these components are put together in a matrix called featurevector.

4. Deriving the new dataset. This is simply done by multiplying the transpose of the featurevector with the transpose of the dataset. The new dataset gives us the original data solely in terms of eigenvectors we chose.

By analyzing the variances and the covariance matrix and describing the principal components, it was possible to show the data in 2D or 3D plots. This method had been chosen, because it contains some main factors that are needed to show relations between patients.

The first factor is the explained variance. The explained variance is build up from the eigenvalues, showing the cumulative percentage the principal components reflect the data. For example, the first two components could be able to reflect 87% of the data. Therefore, we can deduce that the data has little variation.

The second factor are the principal components. These components tell which dimension of the input data is of most value. For selling houses, this could for example be in which neighbourhood the house is placed. In the OVH data used in this research, the dimensions are the features of the patients. In order to be able to predict future DVHs, it is important to know which patient in the old data set is of most significance by its OVH.
3.3 Clustering

Another method to analyze the data is by clustering the old data set. To do this, Agglomerative Hierarchical clustering has been used. In agglomerative clustering, a "bottom-up" approach is used, which means that each point in the plot has its own cluster, after which pairs of clusters merge as one moves up the hierarchy.

Each cluster has several options that could be set. Firstly, \( n \), the amount of clusters, which records how many clusters one desires. Secondly, the connectivity. Connectivity states that points close to each other are more related to one another than points far away from each other.

Lastly, three linkages are evaluated.

1. Complete linkage: again, each point starts its own cluster and clusters are combined based on the shortest distance between clusters. In complete linkage, the distance between clusters is equal to the distance between the two elements in each cluster that are the farthest away from each other. The shortest distance of these clusters is then being combined.

2. Average linkage: each point starts its own cluster, after which clusters are combined using the average of the distance of all points in one cluster to another cluster. The smallest average clusters are then being combined.

3. Ward linkage: Instead of distance, this linkage method is based on variance. Each point starts out as if being its own cluster, after which the Error Sum of Squares and the \( r^2 \) are calculated. Error Sum of Squares is a function where one sums up all variables and all the points within each cluster. Each point is then compared to the cluster means. The smaller the Error Sum of Squares, the higher the suggestion that the new point belongs to that cluster. \( r^2 \) is being interpreted as the variation explained by a particular cluster. Clusters are then combined using the minimal error sum of squares or the maximum value of \( r^2 \). [4]

3.4 Implementation

This section explains the implementation of the PCA and Clustering methods that have been made. It will show a test scenario including sinus data. Beside that, the feedback of the Software Improvement Group will be discussed.

3.4.1 PCA implementation

A condition to use the PCA method, is that the data needs to follow a normal distribution. In Python, a method has been written that was able to plot the combined matrix, showing a plot of the OVH data. In MatLab, a histogram was made of the OVH data.
The PCA implementation was made, using the NumPy library. This implementation required the NumPy library and was determined to be capable of reducing the dimensionality of our data set, which are the OVH and DVH data points of the sample patients and calculate the variability in the data. The explained variance can be displayed per principal component cumulatively.

For further use of the output of the PCA, a complete method has been written, that seperately calculates the scores and the coefficients of the input matrix.

### 3.4.2 Clustering implementation

The implementation of the clustering method is based upon the sklearn package of Agglomerative clustering. Firstly, the labels are created, after which the array of patients is created. After that, the Agglomerative clustering method is executed, which returns a model in which the array is fitted. When the patients array is fit in the model, a plot can be made of each of the three linkages.

### 3.4.3 Testing the implementation

Before continuing, knowing the PCA implementation worked was a must. Therefore, some testdata was generated, using a CSV generator. The generator was programmed to create 100 files including fifteen rows and two columns, containing the values of different sines. These values were used to plot the PCA and check the scores and loadings.

![Plot of sinus data set](image)

Figure 5: Plot of the sample sinus data set. Left figure plots the sinus values, right figure plots the PCA data.

Figure 5 shows the outcome of a PCA analysis on sinus data. The sinus points are created to differ only a tiny bit. Therefore the outcome is as expected. The
right plot shows a perfect circle because the sinus has moved 1 spot at a time and therefore reaches its beginning at some point. The left plot shows the loadings, the biggest variance in the sinus data. In further analysis this could be used to plot the new graph, in which the loadings would apply as x-axis and y-axis.

As can be seen in figure 6, the explained variance shows that two principal components are able to subscribe the entire data set. This is exactly as expected, because the sinus data set is a two dimensional set. In this case, one principal component could subscribe half the data set. That could mean that only one axis is an infinite line describing only the x- or y-axis.

Figure 7 and figure 8 show the outcome of the described clustering method above. Figure 7 shows ten clusters without connectivity. It can be seen that the points that are very close to each other form a group, which means the clustering succeeded. The same goes for figure 8, this figure shows three clusters with connectivity. Again, three clear groups are shown which have most in common.
3.4.4 SIG

In the fifth week of the project, our code had to be submitted to the Software Improvement Group (SIG) for an all-round quality check of the code that has been written. The results were as expected: because of the research that has been done, our code base was not big enough to be put to a good test. SIG evaluated our code personally, and came to the conclusion that most of our code was written in a main function and it should be split up into several functions. This would make it easier to maintain and a lot easier to test. Which led to their second remark, there were no unit tests yet. SIG pointed out facts that were known, but with only three weeks to go, it became clear that some time had to be spend on making our code maintainable. As our code has grown exponentially since then, a lot of time has been spend on changing the code base to maintainable code, including tests, before the second submission. After
the second submission, SIG recommended that methods should be even more divided so that they are easier to maintain. They also mentioned that the first recommendations were used in writing the rest of our code. In future code writing, these recommendations are taken into account. Methods will be minimalistic, so that they are easy to maintain and a more test-driven approach will be done, so failures are easily noted.
4 Results

This section will show and explain the results of the PCA and Agglomerative Clustering method. All data has been plotted and used as input in these methods. The results are figures that are explained in this section explicitly.

4.1 Data

The OVH data of the 115 patients of both the bladder and rectum are all visualised in figure 9(a), figure 9(b), figure 9(c) and figure 9(d). All of them put together form figure 10. With the exception of 2 samples, there is a general trend in the data. As told in the methods above, we treated the first PTV and second PTV of the patients as two samples.

Figure 9
Figure 10: All OVH data plot

Plotting the data in a histogram yields figure 11. The data is densely concentrated around 0-25, and closely resembles a normal distribution.
After normalizing the data for principal component analysis, the data distribution look as follows in figure 12.

Figure 11: All OVH data distribution
4.2 Principal Component Analysis

The results of the principal components analysis are found in figure 13 and figure 14. The data is primarily located in one group. No conclusions can be drawn from this. The first principal component has an explained variance of 83 percent and the first two have a combined explained variance of 92 percent, which means that there is not a lot of variation in the data that is not covered by the first two principal components, even after normalizing the data.
Plotting the PCA output in a 3D plane as in figure 15 and adding labels to what data points are of the rectum and what data points are from the bladder.
OVHs, we can see that the data is still concentrated in one group. The x, y and z axis of the graph show the first, second and third principal component. The bladder and rectum outliers are different but the amount of points is so little that we cannot distinguish it from each other.

Figure 15: Hierarchical rectum and bladder PCA in 3D

4.3 Clustering

The results of agglomerative clustering are shown in figure 16 and figure 17. First off, we see that clustering with or without a connectivity matrix has no significant impact on the clusters that are found. We can see that there is one big group of data, which is not positive for categorizing subgroups of data.
Furthermore, the amount of outliers that are not part of the group of points in the middle is so low that it can’t be used as a different category within the OVHs.

This was not what we were expecting to see because the OVHs of both the bladder and rectum were used to cluster. Expected were at least two significant groups found, one that matches the bladder profile and one that matches the rectum profile. However, after plotting the variance in data of both the bladder and rectum OVHs (as seen in figure 4), the results made sense. There is little to differentiate between the bladder and rectum OVHs and even less so if the rectum or bladder data was used separately, because the samples are so close to each other and form one general trend.
4.4 Feature analysis

Clustering the features instead of the samples together yields figure 18, which explains why it was hard to classify the data into subgroups. Most of the feature data is grouped together in one big group. In fact, if we were to classify the samples in subgroups, it would be very difficult as most samples features are a lot alike. So the amount of features that are uniquely distinguishable is so small that it is hard to classify some samples that share these features as a subgroup and not as an outlier.

All in all, the results show that no claims can be made that there are subgroups in the patient data based on the OVH data provided, which is geometric data of the PTV, and the OAR around it.
5 Conclusion & Discussion

As the project started, a great application was to be produced. But after the first two research weeks, it has been noticed that the product might not lead to the result desired by the Erasmus MC. Before trying to build a prediction tool, it was desired to know whether prediction on OVH results was even possible. Therefore, it has been decided to move slightly from the original destination, and do research in whether OVH results could be analysed in such way, that future prediction could be possible.

As time passed by, and several methods were tried and executed, it became clear that this project was going to be much harder than firstly predicted. Yuan[9], who did research in DVH prediction, clearly stated that it is possible to predict DVH by using PCA. Therefore, in consultation with ir. Schillemans, the same method was tried. Research was done to the usage and implementation of PCA, and several data have passed the analysis. Only to conclude that there was no direct link in the OVH results, and the principal components did not lead to a definite dimension on which to focus.

After some time analyzing the results of the PCA, in consultation with Dr. Taskesen another way of analysis was tried. Agglomerative clustering. This clustering technique is used to split data into groups, starting bottom-up. That would mean that clusters could be formed, based on three linkages.

Research of ways to implement the Agglomerative Clustering led to the sklearn package. After analysing the output plots of the clustering, it was noted that there were no actual clusters. All points were included in one big scatter plot, without any clear clusters. As the labels were printed it became even more clear: bladder and rectum results were too much a like, to make a clear difference. That resulted in the same conclusion as was noted after the PCA results.

After a consultation with Dr. Taskesen it had been stated that our research did indeed work out as predicted. A prediction algorithm could have been build, but without knowing whether the data was suitable for prediction, it could lead to false results. Ir. Schillemans noted several times that he did not expect any outcome at first. If the prediction algorithm was to become a success, that would have been great. But if it were not, and the Erasmus MC had a start for further research, to make prediction possible in the future, they are satisfied.

So after ten weeks of research and trying over and over again, a lot has been learned. Firstly, learning to work with such fully-scheduled people was hard at first. For the first time, progress was in our own hands. Waiting for feedback before continuing was not an option, because that would increase the delay to deliver a final result. So while working on a new part of the project, feedback of the previous results were processed.
Besides that, a more important learning point has been achieved: it is OK if a research project does not lead to the expected results. It was very hard to consider the idea that there might not be a product to deliver, just a small paper stating the research had other results than expected. This research resulted in the fact that prediction is not possible with OVH data of the bladder and rectum, based on the results of PCA and Agglomerative Clustering. So the research indeed succeeded, but without the results that has been hoped for. The Erasmus MC now has a result to start further research in future treatment prediction.

To conclude this paper, a valuable lesson has been learned. It is not always about the winning or succeeding in everything that has been done. A baby will never learn to stand up again, if it has never fallen on the ground. Therefore, we would like to end with a quote by Albert Einstein:

*Try not to become a man of success, but rather try to become a man of value.*
6 Future Perspectives

During this project, we conducted research using the Principal Component Analysis method and Agglomerative Clustering. These methods were implemented using python and could be used by the Erasmus for further research in the same or a different field.

In this case, research was conducted using the OVH data of prostate cancer patients. There were no subgroups that could be classified with this data in the timespan we were provided. However this does not mean that there are no potential subgroups that could be classified when using it on other data. For example, if breast cancer patients were used, this could have lead to interesting results. Breast cancer can manifest in the left, right or both breasts. If our methods are to be used, maybe it could potentially identify these groups and therefore lead to quicker identification of the breast cancer type.

Furthermore, according to the results of the research we did, there are no subgroups among prostate cancer patients. We mainly used the OVH data of the cancer patients, which means that we cannot classify subgroups based on geometric information. The sample pool was large enough (115 prostate cancer patients) to conclude this.

This could have at least 2 possible explanations. The first one is that prostate cancer patients are as specific as possible as of now and no specific radiotherapy treatment can be given to different prostate cancer patients. This means that a more general radiotherapy improvement should be researched for prostate cancer patients instead of searching for potential subgroups.

Another possible explanation is that potential subgroups cannot be classified using only OVH geometric data. Research can be done using geometric data of OVH’s in conjunction with other types of geometric data.

In retrospect, we could have significantly sped up our research period if we were accustomed to doing data analysis and would have created at least a prototype of the desired prediction tool. This would have been a far more pleasant conclusion of the project, but given the circumstances, it was not unexpected to not have done so. Even so, the skills we gained during this project will be a fine enhancement of our arsenal, which will help us in future works.
Appendices

A Executive Summary

A person with cancer has several treatment options. One of which is radiotherapy. Radiotherapy is treatment of cancer with radiation. To minimize the damage to healthy tissue, radiation is applied from several directions into the body. When treating cancer with radiotherapy, the organs nearby the tumor are at high risk of getting damaged. In the treatment plan the dose to the organs at risk has to be balanced with the dose given to the target. These calculations are nowadays done by medical personnel. Although a lot of treatments succeed, without much damage to healthy tissue, a lot of treatments do serious damage to the organs at risk. Can the achievable organ sparing be predicted using geometrical parameters?

To research if any prediction can be made, several data analysis methods have been implemented in order to identify subgroups within a patient set that either are more difficult or easier to treat. The dataset consists of 115 patients. Each patient had two data sets, one for rectum data and the other for bladder data. Each set contains of 230 CSV files containing Overlap Volume Histogram values. These CSV files are used as input for the research.

Two methods have been used: Principal Component Analysis (PCA) and Agglomerative Clustering. PCA is a mathematical algorithm that reduces the dimensionality of the data, while retaining most of the variation in the data set. It accomplishes this reduction by identifying features where the variation in the data is as high as possible, called principal components. By using a few components, each sample can be represented by a few features instead of thousands. Samples can then be plotted, making it possible to visually assess similarities and differences between samples and determine whether samples can be grouped.

Besides PCA, Agglomerative Hierarchical clustering has also been used. In agglomerative clustering, a "bottom-up" approach is used, which means that each point in the plot has its own cluster, after which pairs of clusters merge as one moves up the hierarchy. Each cluster has several options that could be set. Firstly, \( n \), the amount of clusters, which records how many clusters one desires. Secondly, the connectivity. Connectivity states that points close to each other are more related to one another than points far away from each other. Lastly, three linkages are evaluated: complete, average and ward.

This led to several figures as results. A scatterplot has been made, which shows the transformed data. This shows that the data is primarily located in one group and shows no subgroups can be identified with the data. The first principal component has an explained variance of 83 percent and the first two have a combined explained variance of 92 percent, which means that there is not a lot of variation in the data that is not covered by the first two principal components, even after normalizing the data.

The Agglomerative Clustering method results in six plots. A variation in link-
ages and connectivity has been used, but all six led to no clear distinction within
the data. These results led to the conclusion that no subgroups are distinguishable based
only on OVH data and no prediction can be made that optimizes radiotherapy
plans based solely on OVH data of patients.

B Infosheet

Predict Radiotherapy Plan Quality

Erasmun MC - Department of Radiotherapy

26 June 2015

A person with cancer has several treatment options. One of which is radiother-
apy, treatment of cancer with radiation. To minimize the damage to healthy
tissue, radiation is applied from several directions into the body. When treating
cancer with radiotherapy, the organs and tissue nearby the tumor are at high
risk. The organ-sparing calculations are done by medical personnel. Although
a lot of treatments succeed, without much damage to healthy tissue, a lot of
treatments do serious damage to the organs at risk. Can the achievable organ
sparing be predicted using geometrical parameters? To research if any predic-
tion can be made, several data analysis methods have been implemented in
order to identify subgroups within a patient set that either are more difficult or
easier to treat. Several new techniques were learned, of which Principal Compo-
nent Analysis and Agglomerative Clustering where used in the implementation.
Besides that, in the research phase it became clear that 10 weeks was too short
for the whole plugin to be created. Therefore, it has been decided to only do
research in whether certain sub-groups can be found and with these sub-groups
organ sparing can be predicted. Every week, a meeting with our TU supervisor,
dr. ir. E Taskesen was planned. He reviewed our work and discussed how to
continue. Every two weeks, ir. W. Schillemans from the Erasmus MC was at
the meeting as well. We kept on discussing how to continue after results be-
came clear. This was a great approach, because it was easy to adapt to ones
needs. The product consists of analysis implementation of PCA and Agglom-
erative Clustering, which has been tested on patient data from the Erasmus
MC. This product will be used to do further research on predicting plan quality
for future patients. The project team consists of: Sander Liebens, a student
in computer science who likes learning new methods and approaches to tackle
problems. Does not excel in a specific programming language, but likes to know
his way around with several and Huang-Da Chi, a student in computer science
who especially likes web-programming. Likes high-level programming languages
as they can do much with little implementation. All coding has been done with
pair-programming principle, therefore it is hard to highlight who has really done
what. Same goes for the final report and research phase. All writing has done
while both members were present and working. Continuously reviewing and adapting each others work made this project team come as far as it did. It was important to do everything together, as we learned a lot from each other and gave new insights in techniques and working methods. Client: Ir. Wilco Schillings, clinical physicist at Erasmus MC Daniel den Hoed - Rotterdam. TU Coach: Dr. ir. Erdogan Taskesen, postdoc at Bioinformatics department - TU Delft. Students: Sander Liebens - Computer Science - sander.liebens@gmail.com; Huang-Da Chi - Computer Science - huang-da.12@hotmail.com. The final report for this project can be found at [http://repository.tudelft.nl](http://repository.tudelft.nl)

References


