Technical University of Crete

School of Electrical and Computer Engineering

Diploma Thesis

‘Bio-optical modeling of epithelial neoplasia’

Papadogiannis Sevastianos

Committee Members:

Professor Balas Costas (supervisor)

Professor Garofalakis Minos

Associate Professor Samoladas Vasileios

Chania, 2019

Technical University of Crete

School of Electrical and Computer Engineering

Diploma Thesis

“Bio-optical modeling of epithelial neoplasia”

Papadogiannis Sevastianos

Supervisor: Professor Costas Balas

Abstract

Over the last years, a radical change regarding diagnosis of cervical neoplasia has been prepared. Classic procedure involving biopsy may be replaced by *in vivo* non invasive optical biopsy. In this thesis, we are studying a model that estimates for each combination of 4 biological parameters, the spatiotemporal curves of diffused reflectance versus time, that are produced during the acetowhitening phenomenon. We aim to solve the reverse problem, estimating the bio-parameters given the curve, efficiently in terms of both accuracy and speed. Initially, we used several curve matching algorithms to decide which is optimal for the problem. Then, we used k-means clustering on the model dataset to reduce comparisons and lower procedure duration. Finally, we correlated curve features to bio-parameters and created a decision rule for instant bio-parameter estimation through optimization with decision regions. Judging from the estimation accuracy and the calculations on execution times, we have strong confidence in the future of the method.

Acknowledgements

Firstly, I would like to thank my supervisor, Professor Costas Balas for involving me with this special subject, for his invaluable guidance and his inspiring presence.

I would also like to thank Professors Minos Garofalakis and Vasilis Samoladas for their participation in my examining committee.

Special thanks to Emmanouil Vardoulakis, member of Electronics Laboratory, for our productive cooperation and for passing in me his high work standards.

Moreover, I would like to thank and congratulate all the colleagues that have worked on this subject and especially Giorgos Papoutsoglou whose efforts led to an innovative consideration on cervical neoplasia.

I would like to thank my friends and colleagues for those moments during my undergraduate studies. I particularly wish to Gregory D., Giannis T. and Zisis C. all the best and a bright future!

Finally, I would like to thank my family for everything.

Table of Contents

List of Figures

List of Tables

Chapter 1 – Introduction

1.1 Motivation & Thesis Outline

Motivation

There is a strong motivation to search in this scientific area and this particular problem. Over the years, scientists continue to strive towards a more efficient treatment for Cervical Cancer, making it one of the most preventable. We are blessed to belong in a laboratory with many years of experience and significant achievements in this field.

We hope to discover and implement new ideas that will improve parts of the overall diagnostic and treating procedure and ultimately make a positive impact on the patients' lives.

At the same time, we aim to expand our knowledge in various scientific methods and familiarize ourselves with different fields from those we are to study as Electrical and Computer Engineers.

Thesis Outline

In Chapter 1, we make an introduction to cervical cancer and the present state regarding diagnosis and treatment.

In Chapter 2, we will present pieces of information regarding biology behind neoplasia progress and conclusions and findings from related previous work.

In Chapter 3, we analyze the mathematical background supporting everything we used in this thesis.

In Chapter 4, we describe in detail the methodology behind every part of our work.

Chapter 5 features the results for every technique used and also execution times for the general procedure.

Chapter 6 contains conclusions and related future work.

1.2 Thesis Contribution

In this thesis, we had to deal efficiently with a curve matching problem, in terms of both accuracy and speed,. Higher result quality is equivalent to better estimation regarding the biological parameters of the cervix, while improvement in execution times, up to running in real time, is translated to tremendous value and potential for widespread future usage of the procedure.

Estimating those bio-parameters is considered as crucial in monitoring evolution of the disease. It is currently a task varying from being time consuming to being impossible and this highlights the significance of developing innovative methods like the one we are contributing to.

Contribution

We compared many alternative metrics and concluded on the best algorithm for the curve matching problem. We also used K-means clustering technique and achieved significant improvements on procedure execution times. Finally, we created a test for instant bio-parameter estimation using random variable correlation and decision regions, which offers concurrently an interesting and promising way of reducing the number of curve comparisons and thus accelerating the procedure.

1.3 Introduction to Cervical Cancer

Cancer is a disease featuring abnormal cell behavior. According to [1] and [2], there are common biological capabilities, also known as "Hallmarks of Cancer", which cancer cells have during their reproduction:

1) They can sustain proliferative signals.
2) They ignore anti-growth signals.
3) They resist programmed cell death (apoptosis).
4) They are multiplied without limits (replicative immortality).
5) They request nutrition and waste disposal from the body.
6) They activate Invasion and Metastasis.
7) They reprogram energy metabolism.
8) They evade immune system's identification and elimination.

Cancer is the second leading cause of death in the United States [3] (it causes 25% of deaths), exceeded only by heart disease. In 2016, the latest year for which incidence data are available, 1,658,716 new cases of cancer were reported and 598,031 people died of cancer. This equals to 436 new cases and 156 deaths per 100,000 population.

In 2018, an estimated 1,735,350 new cases of cancer will be diagnosed in the United States and 609,640 people will die from the disease [4].

Cervical Cancer

In 2012 [5], cervical cancer was the fourth most diagnosed cancer worldwide for females with 527,600 estimated new cases and the fourth leading death cause amongst cancers with 265,700 estimated deaths. In many parts of Africa it was the leading cause of death between all cancers.

The American Cancer Society's estimates for cervical cancer in the US for 2019 are [6]:

- About 13,170 new cases of invasive cervical cancer will be diagnosed (8.00 per 100,000 population)
- About 4,250 women will die from cervical cancer (2.58 per 100,000 population)

International Agency for Research on Cancer data for Greece in 2012 are [7]:

- 421 new cases of invasive cervical cancer were diagnosed (7.42 per 100,000 population)
- 208 women died from cervical cancer (3.66 per 100,000 population)

It was reported in 2018[8] by a former Greek official that (Greek) National Neoplasia Archive (EAN) had officially recorded 600 new cases per year of diagnosed cervical cancer (10.75 per 100,000 population)

It is important to keep in mind that cervical pre-cancer (neoplasia) is diagnosed far more often than invasive cervical cancer.

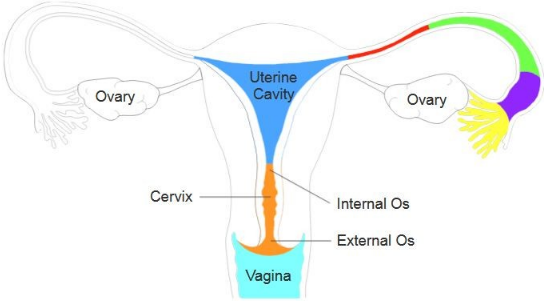
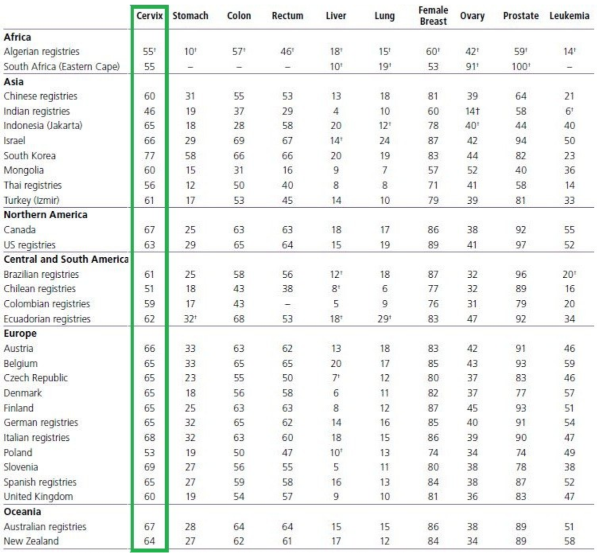


Figure 1.1 Female Reproductive System[9]

Amongst gynecologic cancers, Uterine cancer is the most common, Ovarian cancer causes the most deaths and Cervical cancer is the most preventable; also the only one with a screening test and a vaccine [10].

Relative Survival

The definition or relative survival is the probability of surviving for the overall population divided by the probability of surviving having a disease for a defined time frame in years.



	Cervix	Stomach	Colon	Rectum	Liver	Lung	Female Breast	Ovary	Prostate	Leukemia
Africa										
Algerian registries	55†	10†	57†	46†	18†	15†	60†	42†	59†	14†
South Africa (Eastern Cape)	55	–	–	–	10†	19†	53	91†	100†	–
Asia										
Chinese registries	60	31	55	53	13	18	81	39	64	21
Indian registries	46	19	37	29	4	10	60	14†	58	6†
Indonesia (Jakarta)	65	18	28	58	20	12†	78	40†	44	40
Israel	66	29	69	67	14†	24	87	42	94	50
South Korea	77	58	66	66	20	19	83	44	82	23
Mongolia	60	15	31	16	9	7	57	52	40	36
Thai registries	56	12	50	40	8	8	71	41	58	14
Turkey (Izmir)	61	17	53	45	14	10	79	39	81	33
Northern America										
Canada	67	25	63	63	18	17	86	38	92	55
US registries	63	29	65	64	15	19	89	41	97	52
Central and South America										
Brazilian registries	61	25	58	56	12†	18	87	32	96	20†
Chilean registries	51	18	43	38	8†	6	77	32	89	16
Colombian registries	59	17	43	–	5	9	76	31	79	20
Ecuadorian registries	62	32†	68	53	18†	29†	83	47	92	34
Europe										
Austria	66	33	63	62	13	18	83	42	91	46
Belgium	65	33	65	65	20	17	85	43	93	59
Czech Republic	65	23	55	50	7†	12	80	37	83	46
Denmark	65	18	56	58	6	11	82	37	77	57
Finland	65	25	63	63	8	12	87	45	93	51
German registries	65	32	65	62	14	16	85	40	91	54
Italian registries	68	32	63	60	18	15	86	39	90	47
Poland	53	19	50	47	10†	13	74	34	74	49
Slovenia	69	27	56	55	5	11	80	38	78	38
Spanish registries	65	27	59	58	16	13	84	38	87	52
United Kingdom	60	19	54	57	9	10	81	36	83	47
Oceania										
Australian registries	67	28	64	64	15	15	86	38	89	51
New Zealand	64	27	62	61	17	12	84	34	89	58

Figure 1.2 5-year (2005-2009) Relative Survival Worldwide, for each cancer site and region

5-year (2012-2016) Relative Survival in the US for Cervical Cancer is 67.6% [3] which is 1.1% higher than average Survival Rate of all Cancer Sites for females.

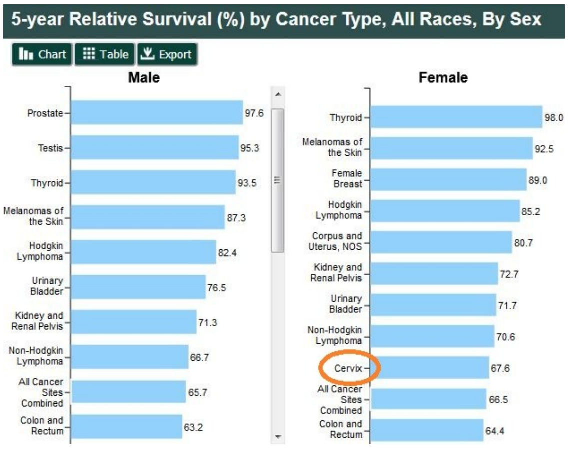


Figure 1.3 5-year Relative Survival in US

Risk Factors [11]

Cervical Cancer is almost only caused by human papillomavirus (HPV) infections [5]. Although everyone is a candidate for HPV infections and most people will have an HPV infection at some point, chances are increased when women become sexually active at an early age and also when they have had numerous sexual partners.

For instance the extremely low cervical cancer rates in the Middle East and in other parts of Asia is attributed to the low HPV infections due to the disapproval of extramarital sexual activity of the respective societies.

In addition, the probability of HPV infection turning into cancer is affected by many factors like smoking, weak immune system, as well as a high number of childbirths (3+) and chronic usage (5+ years) of birth control pills.

Symptoms [12]

If the disease is at an early stage, there may be no signs. When the situation escalates, there may be abnormal, heavier or longer menstrual bleeding and bleeding after sex or between regular periods.

1.4 Cervical Neoplasia Categorization

[13] Usually, a cervical neoplasia passes through a long phase of being pre-invasive. Over the years, disease progression and histological grade were correlated through observation. Furthermore, researchers started considering as continuous, the process of how a normal epithelium gradually evolves into invasive cancer.

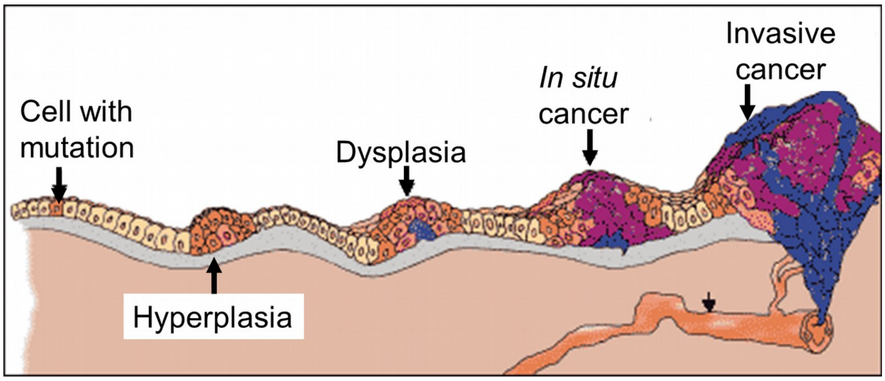


Figure 1.4 General stages of precancer evolving into invasive cancer

The stage in which the disease is in pre-invasive phase is called Cervical Intraepithelial Neoplasia (CIN) followed by a grade showing the extent of cell abnormality and ultimately the severity of the case. As the disease progresses, the thickness of the epithelium layer is gradually affected.

CIN 1 is associated with mild dysplasia when approximately one third of the epithelium layer of the cervix is affected. CIN 2 indicates moderate dysplasia meaning that around two thirds of the epithelium layers affected. CIN 3 means severe dysplasia up to Carcinoma In Situ.

Cytology screening (Pap test) reports its findings of abnormal squamous cells using a system called The Bethesda System.

Result in The Bethesda System	Description
Atypical squamous cells (ASC)	Cells do not look normal
ASC – undetermined significance (ASCUS)	Some cells do not look completely normal. It is unclear what the cell changes mean.
Low-grade squamous intraepithelial lesion (LSIL)	Cells do not look normal, but usually they are not precancerous. Considered mild abnormality.
ASC – cannot rule out high-grade squamous intraepithelial lesion (ASC-H)	Cells do not look normal. It's unclear what the cell changes mean. HSIL can't be ruled out. The abnormal changes may be pre-cancerous.
High-grade squamous intraepithelial lesion (HSIL)	There are abnormal or precancerous cells present. The cells may develop into cancer if they are not treated.
Squamous cell carcinoma (SCC)	There are cancerous cells present.

Table 1.1 Pap test result categorization

In case of HSIL, proceeding to immediate treatment is imperative. The 3 histological grades are connected with the cytology grades; CIN 1 with LSIL and CIN 2 and CIN 3 with HSIL.

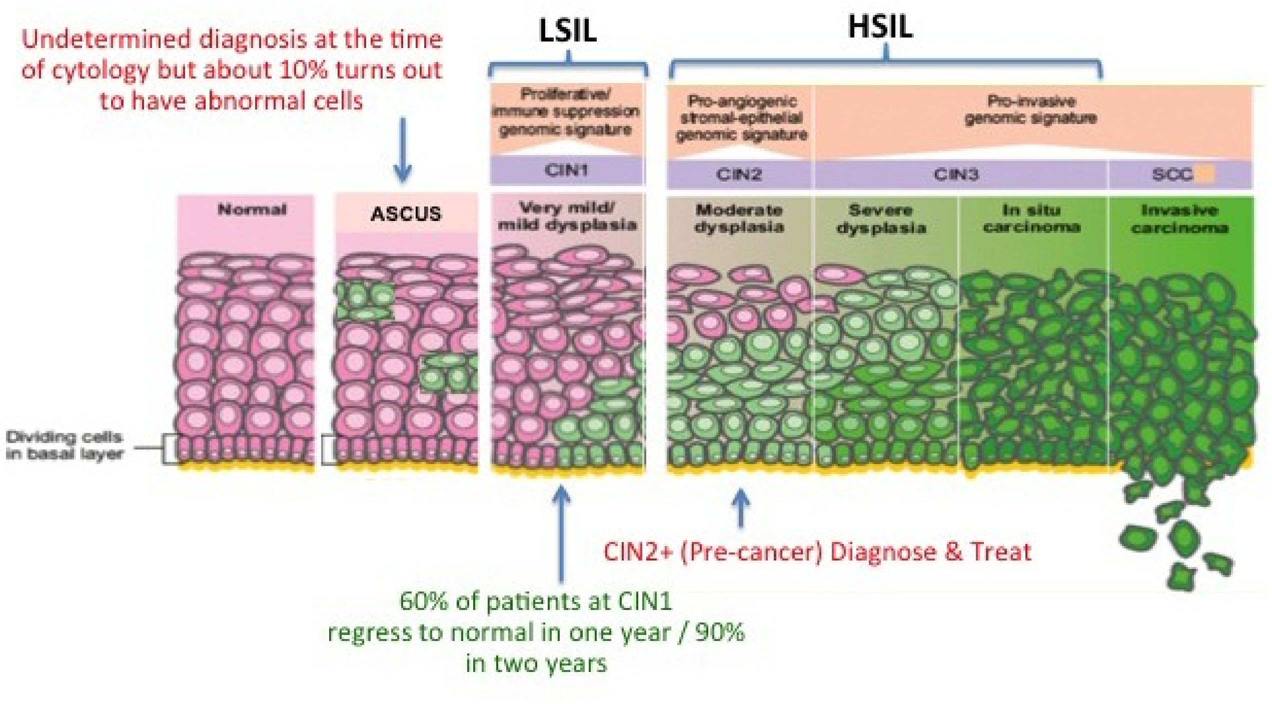

Undetermined diagnosis at the time of cytology but about 10% turns out to have abnormal cells
LSIL
HSIL
Proliferative/ immune suppression genomic signature
Pro-angiogenic stromal-epithelial genomic signature
Pro-invasive genomic signature
CIN1
CIN2
CIN3
SCC
Normal
ASCUS
Very mild/ mild dysplasia
Moderate dysplasia
Severe dysplasia
In situ carcinoma
Invasive carcinoma
Dividing cells in basal layer
CIN2+ (Pre-cancer) Diagnose & Treat
60% of patients at CIN1 regress to normal in one year / 90% in two years

Figure 1.5 Stages of Cervical Intraepithelial Neoplasia (CIN) [14]

It is important to note that some cases show regression over time, some show persistence and some progress into a worse case. This happens with different probabilities depending on the CIN category as the following table containing statistical results from studies between 1950 and 1993 shows:

CIN category	Regression	Persistence	Progression to CIN 3	Progression to invasive cancer
CIN 1	57%	32%	11%	1%
CIN 2	43%	35%	22%	1.5%
CIN 3	32%	56%	-	12%

Table 1.2 Probabilities of CIN Evolution

Respectively, there are statistics showing the probabilities of regression and progression over a period of 24 months, of each category of cytology result. They come from a meta-analysis of 27,000 cases back in 1999[15]:

Cytological Abnormality	Regression to normal	Progression to HSIL	Progression to invasive cancer
ASCUS	68,2%	7,1%	0,3%
LSIL	47,4%	20,8%	0,2%
HSIL	35,0%	23,4% (persistence)	1,4%

Table 1.3 Natural history of SIL

The overall conclusion of studies in the 1990s is that most low-grade lesions (CIN 1) are transient and regress to normal given a short time period, or do not progress to more sever forms. On the contrary, CIN 2 and CIN 3 are much more likely to progress into an invasive cancer stage despite the fact that some cases regress or persist.

1.5 Diagnosis and Treatment

The American Cancer Society recommends[16] that women follow a screening procedure with its main goal being to detect cervical cancer early. This can ease the treatment and avoid severe measures for the patient.

Ages	Screening Method	Time Frame
21-29	Pap Test	Every 3 years
30-65	Pap Test + HPV Test (recommended)	Every 5 years
30-65	Pap Test (sufficient)	Every 3 years
Abnormal Pap Test	Follow-up Pap Test in 6-12 months (doctor may recommend combining with HPV Test)	
65+	Stop testing unless they have had serious precancer	

Table 1.4 Testing Recommendations for Early Detection

Vaccinating against HPV does not make screening unnecessary.

Incidence rates in the US for the disease dropped by more than 50% between 1975 and 2015 due in part due to an increase in screening, which can detect cervical changes before they turn cancerous[17]. The Pap test, when combined with a regular program of screening and appropriate follow-up, can reduce cervical cancer deaths by up to 80%.

[18] If there are certain symptoms that are worrisome for cancer, or if the Pap test detects abnormal cells, it is followed by referral to colposcopy, which determines the location of the most severe dysplastic region for biopsy sampling.

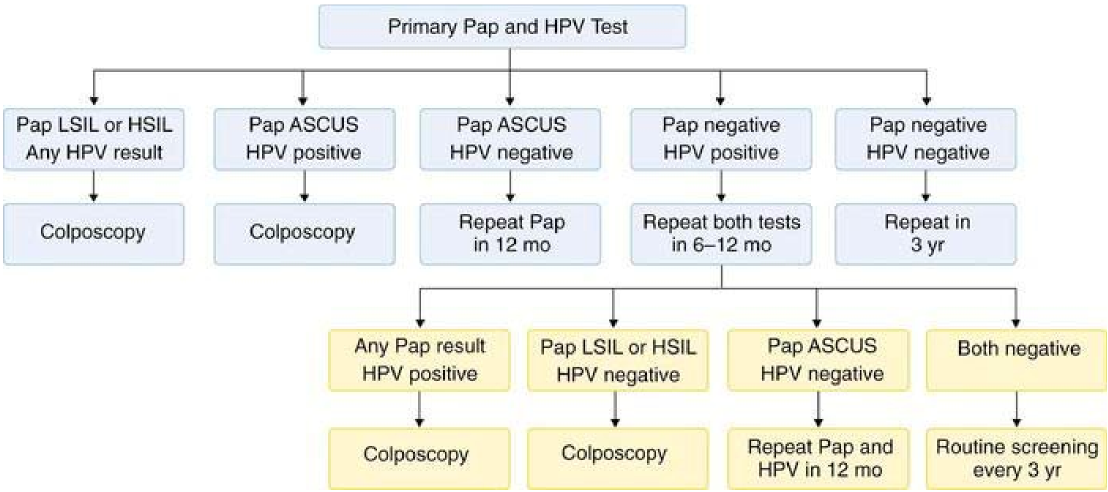

Primary Pap and HPV Test
Pap LSIL or HSIL Any HPV result
Pap ASCUS HPV positive
Pap ASCUS HPV negative
Pap negative HPV positive
Pap negative HPV negative
Colposcopy
Colposcopy
Repeat Pap in 12 mo
Repeat both tests in 6–12 mo
Repeat in 3 yr
Any Pap result HPV positive
Pap LSIL or HSIL HPV negative
Pap ASCUS HPV negative
Both negative
Colposcopy
Colposcopy
Repeat Pap and HPV in 12 mo
Routine screening every 3 yr

Figure 1.6 Next actions according to Pap and HPV Test results

During colposcopy, the doctor will place a speculum in the vagina to see the cervix and apply a weak solution of acetic acid to the cervix to make abnormal areas more visible. If an abnormal area is seen, a biopsy (removal of a small piece of tissue) will be performed.

A biopsy is said to be the best way to tell for certain whether an abnormal area is a pre-cancer, a true cancer or neither. Cervical biopsy may cause discomfort, cramping, bleeding or even pain in some women.

Main problems of colposcopy are:

- Low sensitivity (55-65%) and specificity (70-90%)
- 52% of screening failures, including missed lesions
- Unnecessarily repeated tests and diagnostic delays [19]

These are added to the low sensitivity problem of Pap test (59%).

These weaknesses of colposcopy are critical to diagnostic reliability because the biopsy will inform the doctor on the condition of a small sample tissue representing the area of the colposcopy suggestion. It is up to the colposcopy to identify accurately the most dysplastic region.

The performance of colposcopy can be improved by the assistance of Dynamic Spectral Imaging (DSI) as we will later present.

Treatment

[20] Treatment of Cervical Cancer can be performed in several ways depending on its type and how spread it is. Treatments include surgery, chemotherapy, cryotherapy and radiation therapy.

- Surgery is an operation in which doctors will remove the cancer tissue.

- Chemotherapy is a procedure involving special medicines to kill the cancer or shrink its size. The patient receives the treatment through oral or intravenous adninistration, or sometimes both.

- Cryotherapy uses liquid nitrogen of around -50^{o} C temperature, to destroy precancerous cells on the cervix [46].

- Radiation can be used to kill the cancer by using high-energy rays (similar to X-rays).

Realizing the value of early detection so as not to end up using extreme treating means is imperative.

1.6 Dynamic Spectral Imaging

As we mentioned previously, the quality of colposcopy affects the reliability of the biopsy and is crucial for the whole diagnostic procedure. Dynamic Spectral Imaging System (DySIS^{TM}) is a system offering enhanced colposcopy.

Colposcopy uses the acetowhitening phenomenon that occurs when a weak (3-5%) solution of acetic acid is applied to the cervix. Its key element is that it temporarily changes optical properties of the area from the interaction of the acid with abnormal cells. Conventional colposcopy uses observation to estimate the most dysplastic region. On the other hand, DySIS measures the changes of reflective properties and it can produce substantially more accurate findings with this systematic approach.

Firstly, before applying the acetic solution, DySIS measures the black body radiation for normalization reasons. Then it applies the acid and takes snapshots of the cervix every 7 seconds (29 images in total).

The important part of the phenomenon lasts around 3-4 minutes, during which the reflective properties of the lesions are affected from the acid and that is translated to changes in the optical signal (pixel luminosity) on the received images.

Then the system processes the images and measures the intensity of Diffused Reflectance (DR) versus time for every pixel which represents a small area of the cervix. To correctly do that for every small area, it ‘smartly’ aligns pictures to remove spatial ‘noise’ caused by patient movement during the procedure. Now the output accurately describes the dynamic changes due to the acetowhitening phenomenon. Eventually, the measurement is normalized and a 1x29 vector of DR versus time is created for each pixel.

The DySIS colposcope produces a high resolution output (1024*768 = 786432 pixels - curves).

Curve analysis has led to a connection with disease categories that have been tested with biopsy results. Below we can see centroid curves for different categories [21].

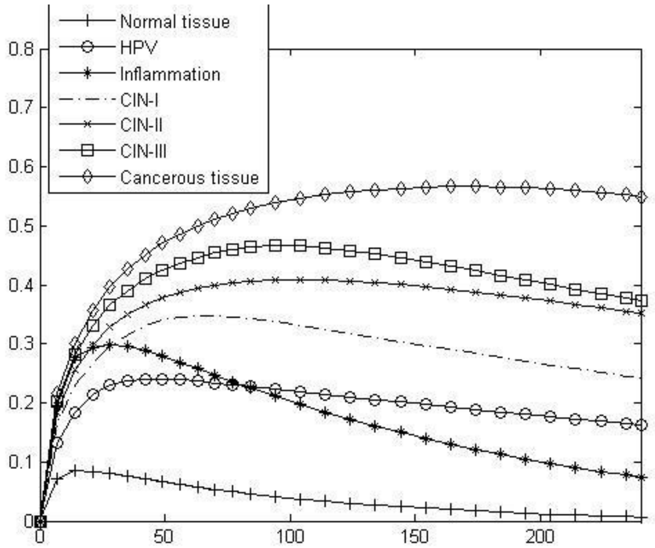


Figure 1.7: Representation of seven clinical reference centroids reflecting the clinical trends.

The system processes the curves in real time and creates a DySIS map which highlights in detail, using pseudocoloring, the estimation for the disease category of each cervix region and how spread the damage is. This explicit output describing the areas of interest, helps doctors tremendously and the system's high sensitivity makes misjudging a rare occasion.

If a high grade neoplasia is detected and there is need for a biopsy follow-up, DySIS map guidance will contribute decisively to the procedure by improving the spatial accuracy.

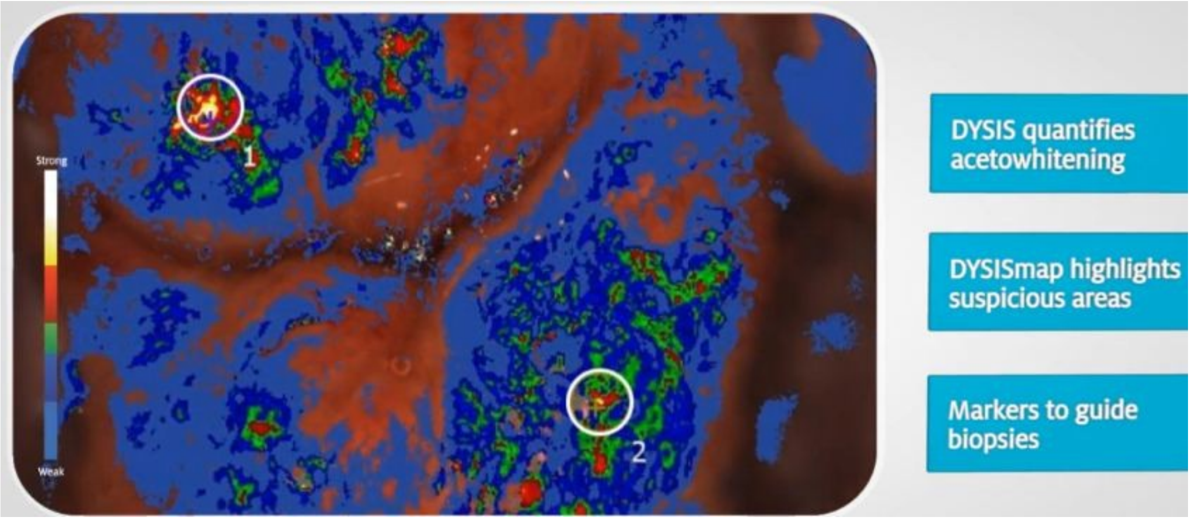

Strong
Weak
1
2
DYSIS quantifies acetowhitening
DYSISmap highlights suspicious areas
Markers to guide biopsies

Figure 1.8: DySIS map, result of colposcopy[22]

DySIS had its diagnostic value confirmed in large international clinical trials [23] [24], where it performed 63% better in sensitivity over Pap test and colposcopy.

	DySIS	Colposcopy	Cytology
Sensitivity	79%	49%	53%
Specificity	76%	89%	86%
Diagnostic Odds Ratio	11.81	7.91	6.88

Table 1.5 The Sensitivity and Specificity values of DySIS™ vs values of Pap and Colposcopy.

To sum up, DySIS colposcopy clarifies the current health situation, offers a quality guidance for biopsy, leads to reliable diagnosis for patients and provides the health system with an efficient way to detect cervical neoplasias with regard to both sensitivity and cost.

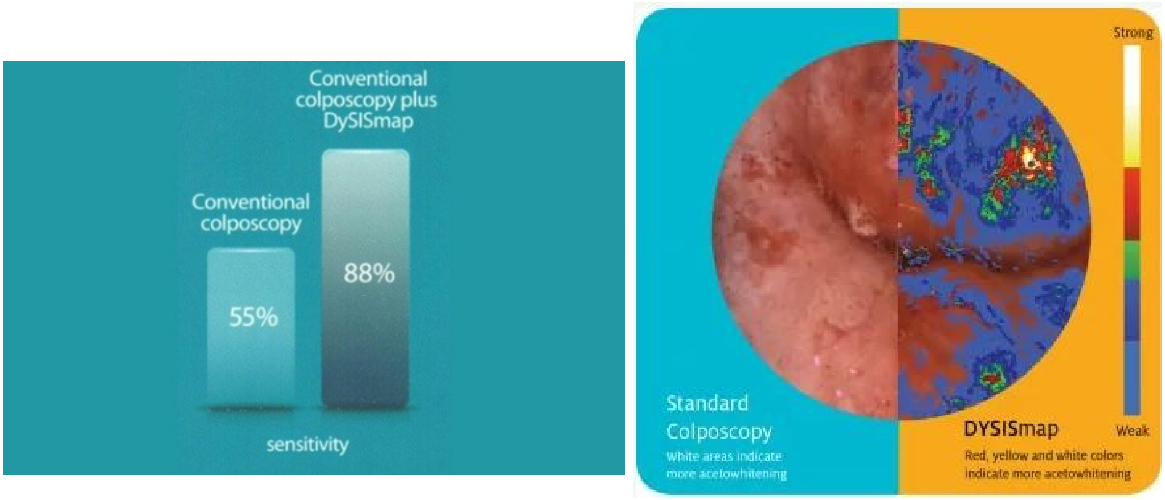


Figure 1.9: Comparing colposcopy with and without DySIS map [25]

1.7 Prediction Model

Relevant research to the topic showed that the DR versus time curves are shaped according to the values of a combination of biological parameters of the cervix and that by having knowledge on those values, we can accurately predict the respective curve. We will present information on that research and its conclusions in Chapter 2.

This brings us to the next considerable step in this project and what this thesis is all about. That is exploring the reverse route of the curve prediction model and determining whether the accurate estimation of cervical bio-parameters given the curve, is feasible.

As the values of experimental bio-parameters form a continuous space, the produced curves will differ from those of our model so we are looking for the closest match from the of model curves to them. It has to be noted, that the model has suitably small step in each bio-parameter value to cover the continuous space efficiently.

Optimally, figuring out the closest match and thus the bio-parameter combination, has to be done in real time to help not only patients but also the health system.

Chapter 2 – Theory and Related Work / Prior Art

2.1 Biological Features of Neoplasia

Research on biology regarding the cervical area and about parameters and conditions promoting neoplasia growth, has increased our knowledge and enables us to focus our attention to biological mechanisms and chemical concentrations and develop new techniques to diagnose the patients' state. Similar techniques can be also used to monitor the progress on neoplasia growth and provide doctors with a full image regarding each case.

Now we will present pieces of information acquired by this research over the years:

[26] , [27] Epithelial cells consider maintaining the intracellular pH close to neutral (7-7.5) as vital for their existence and use a set of short and long term mechanisms to restore this value after an acute acid load.

Short term mechanisms last for a second or less and they perform three passive processes: 1) Physicochemical buffering, 2) Metabolic processing and 3) Organellar buffering.

Long term mechanisms are called ion pumps and are activated to aid the passive processes in pH regulation. They should achieve their goal within a few minutes.

Under high acidic load conditions, short term mechanisms are proven to have limited capacity to successfully regulate the pH and the extrusion of the acid is done actively by buffers yielding back H+ from previous consumption, leading to pH level stabilization.

Tumor cells have almost the same intracellular pH as normal cells, however their extracellular pH is lower. This is associated with the production is lactic acid under anaerobic conditions and to the hydrolysis of adenosine triphosphate in an energy-deficient environment [28] , [29]. Extracellular pH at normal epithelium is around 7.3 and in tumors 6.8.

The intracellular pH regulating mechanisms are surprisingly not affected by the progress of neoplasia while the extracellular ones are severely influenced.

Furthermore, it has been shown experimentally that H+ flow to peritumoral normal tissue provokes extracellular matrix degradation and normal cell necrosis. Concurrently, tumor cells develop resistance to acid-induced toxicity during carcinogenesis and this allows spreading and invasion over the damaged normal tissue.

Lastly, promotion of cancer progression and metastasis has been associated with alteration in the tight juction complexes [34].

2.2 Model Analysis

We know neoplasia growth creates problems in functional (extracellular acidity) and structural (extracellular space and number of abnormal layers) parameters. The changes have been correlated undoubtedly with increased intensity of diffused reflectance (DR) versus time during the acetowhitening (AW) phenomenon[30].

There have been significant efforts to identify all the biological parameters regarding the AW effect on the optical properties of cervix that are translated to changes in DR versus time curves. Gradually, a model was invented, the equations of which are in accordance with Fick's Law and Goldman-Hodgkin-Katz constant field equation.

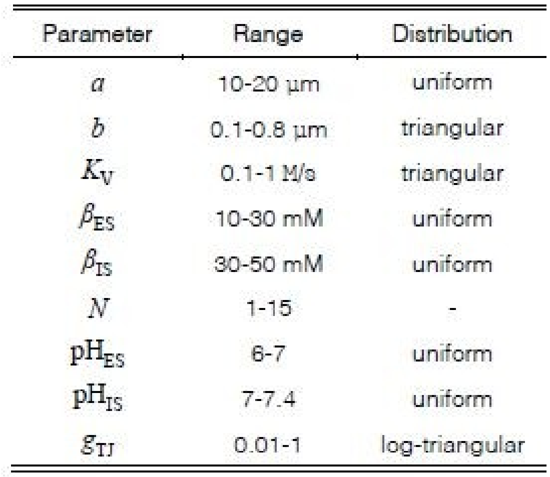
The equations are too brilliant to be analyzed here. However, we will present the scientific processes that were used to identify the neoplasia-specific biological features that decisively determine the changes on the characteristics of the optical signal. That requires combination of theory and algorithms quantifying the impact each bio-parameter has. Ultimately, low-impact parameters will be removed if the estimation does not lose its accuracy.

Theory suggests that parameter N, number of dysplastic epithelial cell layers, is known to be correlated with neoplasia growth; it is even the biopsy measurement for grading the lesions. In addition, parameters N and b (Extracellular Space) are known to be increasing with neoplasia progress [31].

In addition, parameter pH_{ES} (ES = Extracellular Space) has been recently linked, according to the "acid-mediated tumor invasion model" with the ability of tumor cells to form invasive cancers [32].

After including the facts analyzed in chapter 2.1, extra parameters were added to the model such as b_{ES} and b_{IS} which are ion buffering parameters, K_V that refers to solute diffusion rate at the tissue basement membrane interface and g_{TJ} which stands for the porosity fluctuation due to the alteration of tight junction geometrical properties [33].

A set of 9 in total parameters affecting the optical signal was initially introduced.



Parameter	Range	Distribution
a	10-20 µm	uniform
b	0.1-0.8 µm	triangular
K_V	0.1-1 M/s	triangular
β_{ES}	10-30 mM	uniform
β_{IS}	30-50 mM	uniform
N	1-15	-
pH_{ES}	6-7	uniform
pH_{IS}	7-7.4	uniform
g_{TJ}	0.01-1	log-triangular

Table 2.1 The initial set of model parameters

It is important to note that both pH_{ES} and β_{IS} which are determining factors are functional parameters and cannot be measured *in vitro* because the biopsy samples are dead tissues. This fact highlights even more the need for optical biopsy and the estimation of their values through solving the inverse problem, going from DR versus time curves to bio-parameters [35].

Global Sensitivity Analysis (GSA)

The purpose of GSA is to quantify the effect of each parameter on the DR curves and ultimately construct a subset of the initial set containing as few parameters as possible without losses in the accuracy. That will occur only if all the high impact parameters are included on the subset.

Generally, GSA does not require a priori knowledge of a relationship between input and output and it is suitable for providing sensitivity inferences for complex and possibly non-linear systems [36]. In cases whrere the model is a highly nonlinear algorithm and cannot be expressed with a single analytical function, GSA methods are much more superior than classical local sensitivity that has linearity and analytiticy limitations [37].

In this complex model, parameter N is more than an independent variable of the differential equation system, it also defines the number of differential equations.

To continue, variance-based GSA methods are particularly suitable for the problem because of their model-independent approach. Those methods split the model's uncertainty and associate it to the input parameter variations causing it.

Furthermore, variance-based GSA performs 2 distinct processes, factor prioritization and factor fixing. The goal of the first is to rank the parameters according to their importance by measuring each input factor to the output variance. Highly ranked parameters should be dominant in the formation of the model's output line shape, which, theoretically, maximizes the possibility for their identification. On the other hand, factor fixing process is defining the low impact parameters that are more or less unrelated with the output variance and can be fixed to their nominal values.

The GSA calculations performed is the model creation were based on the improved Sobol's method [38] which considers the variances of a specific model equation as multidimensional integrals and estimates them using a quasi-Monte Carlo algorithm. [40] Computation of variances using Monte Carlo method creates an error in the estimations which is inversely proportional to the number of simulations. For this reason, Sobol introduced the “probable error” as an estimate of variability.

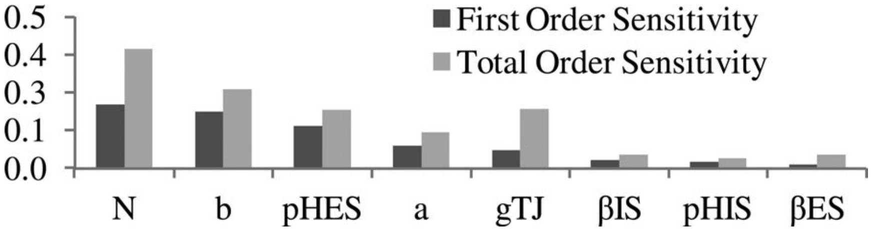


Figure 2.1 First and Total Order sensitivity of the set parameters [37]

This result shows that the number of dysplastic layers (N), the size of Extracellular space (b), pH_{ES} , the size of intracellular space (a) and the tissue's porosity (gTJ) are highly ranked. It also shows potential for dimensionality reduction as the β_{IS}, β_{ES} and pH_{IS} do not pass the combined high sensitivity and identifiability criteria and can be kept constant at around their nominal values without accuracy dropping. K_V was excluded because its Total Order Sensitivity was found 0 at all time points.

The results match well with the theoretical knowledge because there is no evidence suggesting the latter 3 bio-parameters are changing with neoplasia growth [39]. On the other hand, the correlation of the other parameters with neoplasia progress has been ascertained. As a result, the dimensionality reduction is not expected to affect negatively the potential information gain.

GSA outcomes can be used as an input to figuring out the level of interactions and correlations between the identifiable parameter sets. High level of interactions and correlations will result in poor estimability.

As a result of GSA methods, the initial nine input parameter of the model are now 4. Estimation of them using Differential Evolution algorithm achieves 99% accuracy [42].

The 4 bio-parameters affecting the DR versus time curves most are:

- N, the number of cervix layers where there is cancerous activity.
- ECS, the size of extracellular space (also known as parameter b). It is known to increase with the neoplasia progress. Cancerous cells occupy less space, thus higher ECS equals to a worse situation.
- pH_{ES} , the acidity level of the extracellular space. It tends to be more acidic around lesions due to anaerobic glycolysis.
- TP, Tissues Porosity, (known as parameter g_{TJ}). In general, Cervical Intraepithelial Neoplasia (CIN) carcinogenesis disrupts the state of the tissue adhesion structures which has been associated with increased tissue permeability [16]. A healthier tissue will dispose the acetic acid more efficiently and will have a low TP value.

Bio-parameters	Min Value	Max Value	step	# of values
N	1	10	1	10
ECS	$4*10^{-7}$	$8*10^{-7}$	$0.1*10^{-7}$	41
pH_{ES}	6	7	0.1	11
TP	$7*10^{-10}$	$18*10^{-10}$	$0.5*10^{-10}$	23

Table 2.2 List of the important set of bio-parameters and their value space in the model

The model has predicted curves for every possible combination of bio-parameters and has created a dataset of 10 * 41 * 11 * 23 = 103730 curves, each one represented by a 1x29 vector.

Behavior of curves with high pH_{ES}:

- Rise until peak ~ (t = 50 sec)
- Higher decrease when higher PH

Behavior of curves with low pH_{ES}:

- Rise until peak ~ (t = 50 sec)
- Slight decrease when higher PH (6.2-6.4) no decrease or even slight steady increase for very low (6.0-6.1) PH; peak is at the final value

In all curves, the peak decided primarily by N and secondarily on ECS as we will see in chapter 5.3.

N for CIN 1 case is 1-4, for CIN 2 case is 5-8 and for CIN 3 case is 9+

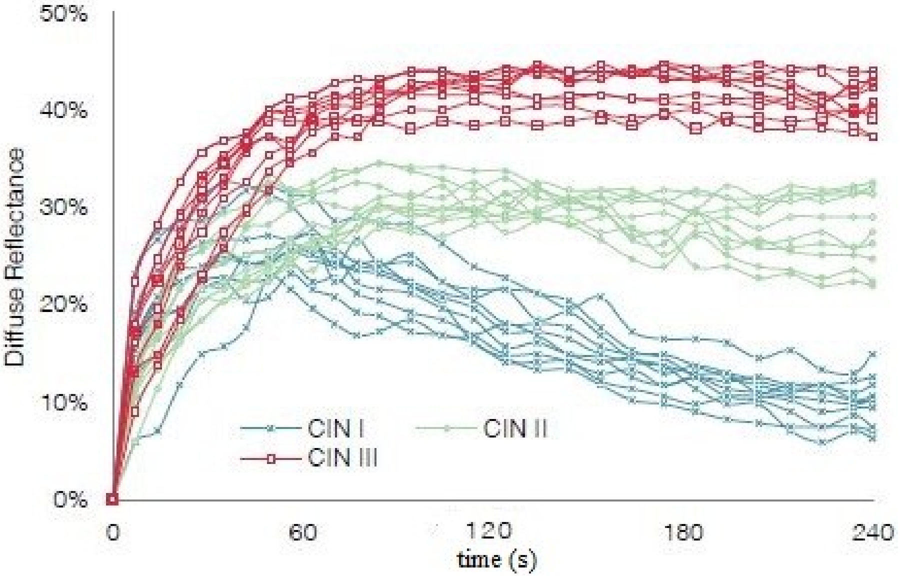


Figure 2.2 Diffure Reflectance versus time for CIN cases

Chapter 3 – Mathematical Background

3.1 Curve Matching Algorithms

In our thesis, one of the main tasks is to find the optimal way of comparing random curves to those of our model and accurately connect them to their best matches. To do that, we used several curve matching algorithms.

Each algorithm calculates a metric and considers as best match the maximum or minimum metric value, depending on the logic behind its construction. Each has advantages and disadvantages, scientific regions and applications that strives on and is used most.

The two elements that characterize the algorithms for our purpose are the quality of results and the comparison speed. The first has to do with how relevant in reality the most similar curves to the original are, which for this thesis ultimately is how close the bio-parameter combination of the random curve to its best matches is. The second mainly depends on the algorithm complexity, however, some level of complexity is essential to have quality in the results. There is a clear tradeoff between those two criteria for the performance evaluation of each metric.

We evaluated the following metrics:

1) Euclidean Distance

2) Pearson Correlation Coefficient

3) Cosine Distance

4) Adaptive Wiener Normalization (AWN)

5) Spectral Angle Mapper (SAM)

6) Spectral Correlation Mapper (SCM)

We will now briefly present some information about each algorithm.

Euclidean Distance

Formula:

$$\sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Matlab Command:

pED = norm(x-y);

Euclidean Distance is one of the most widely used metrics. Also known as L2 norm, it calculates distances between vectors. It has a simple formula and is one of the fastest algorithms of its kind. It can be used as reference to evaluate the speed of other metrics.

Pearson Correlation Coefficient

Formula:

$$\boxed{\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}}$$

Matlab Commands:

```
C = cov(x,y);                          % returns a 2x2 matrix with 1,2 element being C_xy
pPCC = C(2) / (std(x)*std(y));   % C(2) = C(1,2) , std is standard deviation
```

Pearson Correlation Coefficient considers the 2 vectors as random variables and the metric represents the linear correlation between them. It takes values in space [-1,1]. Values close to 1 or -1, mean there is a strong positive or negative correlation and close to 0 mean no correlation. Similar curves should be highly correlated.

Cosine Distance

Formula:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Matlab Command:

```
pCOS = pdist([x;y],'cosine');
```

Cosine Distance or Cosine Similarity is connected to the dot product of two vectors which is:

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos\theta$$

The Matlab command calculating the metric, returns the value 1 – cos(x,y). If the vectors x and y are similar, their angle is close to 0° and the cos(x,y) goes to 1 so the best match has its metric value close to 0 (if x = y then metric equals to 0).

Adaptive Wiener Normalization (AWN)

Matlab Command:

```
a = 0.5; % parameter for AWN
xNorm = x./sum(x);
yNorm = y./sum(y);
pAWN = a*mean(abs(xNorm-yNorm))+(1-a)*max(abs(xNorm-yNorm));
```

AWN metric is related to spectral imaging field. To use AWN, one has to normalize the compared vectors. This has the advantage of focusing on each vector shape rather than amplitude. The result of AWN is the middle point between average and max absolute difference of those normalized vectors x and y.

Spectral Angle Mapper

Formula:

$$SAM(\mathrm{x}, \mathrm{y}) = \arccos\left(\frac{<\mathrm{x}, \mathrm{y}>}{||\mathrm{x}||_2||\mathrm{y}||_2}\right)$$

Matlab Command:

```
pSAM = acos ( dot (x,y) ./ (norm(x).*norm(y) ) );
% acos(x) = cos^-1(x), the reverse function of cosine;
% 'dot' produces the dot product shown above in Cosine Distance details.
```

SAM algorithm is widely used when comparing spectra aiming to find the best similarities. It returns the angle between the compared vectors x and y in radians with the lower metric value showing the higher similarity.

Note: The result of SAM is effectively the same with that of Cosine Distance with a different translation of the same finding. However, SAM is much faster because of the difference in Matlab commands used (the simpler commands are executed faster).

Spectral Correlation Mapper (SCM)

Formula:

$$SCM(\mathrm{x}, \mathrm{y}) = \frac{\sum_{i=1}^{n}(x_i - \overline{\mathrm{x}})(y_i - \overline{\mathrm{y}})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{\mathrm{x}})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{\mathrm{y}})^2}}$$

Matlab Commands:

```
a = x - mean(x);
b = y - mean(y);
pSCM = sum(a.*b)./(sqrt(sum(a.^2)).*sqrt(sum(b.^2)));
```

SCM calculates the Pearson's Correlation Coefficient, a metric analyzed above. Again, there is significant difference in speed because of the different Matlab commands.

3.2 K-means Clustering

When facing problems like curve matching, clustering can be particularly helpful with reducing execution time. The basic idea of clustering is to organize the model's dataset into teams/clusters and to compare the random curves with a few of them, which are a small fraction of the total, rather than pointlessly wasting time comparing curves that differ significantly.

K-means is a clustering algorithm that has been extensively used in many scientific problems. It has an iterative logic and converges to the final cluster structure after a number of repetitions. Initially, K-means executes the following steps, given a clustering algorithm and the number of clusters, K:

1) Randomly (or selectively if user has prior knowledge) pick K curves (or points in a different problem) and consider them cluster centroids, meaning they represent their cluster in comparisons.

2) Compare all curves with every centroid using the selected clustering algorithm and find the closest match according to that metric.

3) Place each curve into the cluster having the centroid with the closest match. After this step all curves belong somewhere and the initial structure is complete.

4) Optimize the teams by choosing the optimal centroid from the set of curves belonging there. The aim is to reduce the inner distance of the curves to their centroid, thus making each team more balanced and the current structure stronger.

After these steps the first iteration is done. Now, K-means repeats steps 2, 3 and 4 with the new centroids. After a number of iterations depending on the complexity and size of the problem, the repetition does not lead to a better structure because K-means has converged to that final result.

It is important to remember that K-means may produce different structures given the same clustering algorithm and number of clusters due to the random initialization. This is the reason there is need to repeat the whole procedure 3 to 5 times for each K to safely determine the structure's strength and consequently the optimal K.

There are multiple ways to evaluate the structure produced by K-means and determine optimal K. In this work, we used Silhouette index and Elbow method.

Silhouette index [43]

The idea of silhouette index is to compare the similarity of each curve with curves of its assigned cluster and the similarities with curves of other clusters. The Silhouette Index is defined as:

$S = (b-a) / \max (a,b)$

'a' is the average distance of the curve with the elements of the same cluster and 'b' is the minimum average distance of all different clusters for that curve.

Each curve receives a value inside [-1,1] depending on how well placed it is. Higher value not only means the curve is well placed but also that changing cluster for the curve would cause a poor placement. The mean value of all curves is assigned to the structure and describes how well built it is.

In case of a low Silhouette index, improvement to structure's strength might happen from increasing or decreasing K.

Silhouette Index	Structure Strength
0.71 – 1.0	A strong structure has been found
0.51 – 0.7	A reasonable structure has been found
0.26 – 0.50	The structure is weak and could be artificial. Try additional methods of data analysis
< 0.25	No substantial structure has been found

Table 3.1 Structure Strength given Silhouette Index

Elbow method [44]

The idea of the Elbow method is to estimate the optimal number of clusters by calculating the Sum of Squared Error (SSE) of each curve from its centroid, for all clusters. SSE is defined as:

$$SSE = \sum_{i=1}^{K} \sum_{x \in c_i} dist(x, c_i)^2$$

C_i is the centroid of cluster i. Low SSE means the cluster elements are close to their respective centroids and thus, it is an indication for a strong structure.

As K increases, SSE decreases due to the fact that a higher number of clusters enables “smaller” clusters having centroids continuously closer to the rest of cluster curves. The improvement of SSE also declines until a point where increasing K improves the overall result very slightly.

The SSE vs K plotted diagram has the shape of the elbow and the optimal point is right on the “edge” of the elbow. That is the K value whose increase offers less from that point on.

3.3 Correlation and Decision Regions

Correlation

Theory says that when 2 random variables are strongly correlated, they tend to increase or decrease together if the correlation is highly positive, or one increases as the other decreases if there is highly negative correlation between them.

Mediocre correlation means that there are exceptions to that rule and the uncertainty regarding curve behavior grows as correlation drops.

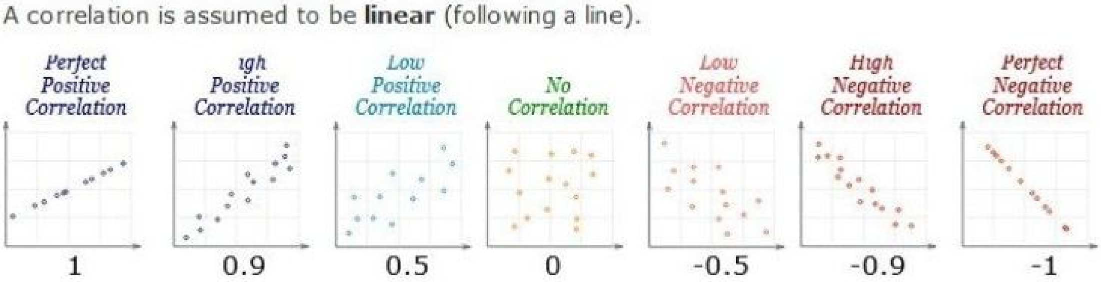


Figure 3.1 Shape of dataset points for various correlation values

To calculate the correlation coefficient between random variables X and Y, we should know the following quantities:

μ_X and μ_Y, which are the mean values of X and Y respectively

σ_X and σ_Y, which are the variances of X and Y respectively

Correlation Coefficient is given by this formula:

$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

E[ ] is the mean value of what is inside the brackets.

Decision Regions

I came across the way of estimating through Decision Regions in the Telecommunications field. The basic problem they are involved in, is the following:

There are 2 random variables, possibly with a different value space. One has a known distribution (or we make an assumption about it) and we also know that the 2 random variables are linked; usually one represents the input and the other one the output.

Decision Regions method is about estimating the values of the unknown random variable for every value of the known variable. To do that efficiently, we want to maximize the probability of being correct. In the following equation, we decide the value of random variable y, given x; it is the y value that has the highest chance of being correct amongst all y values, given x:

$X_y = \{ x \mid p(y|x) \geq p(y'|x) \text{ for all } y' \neq y\}$.

And here is how to combine what we presented above:

We can list some statistics regarding curve behavior (amplitude, rhythm of increase/decrease, integral) and calculate the statistic values for each curve. As a result, numerous random variables are formed (one for each statistic) and by measuring the correlation coefficient of them to the bio-parameters, we can see if there is a connection.

If there is a high correlation between a random variable representing a curve statistic and a bio-parameter, we can gradually create a test that will eventually be splitting the value space of the selected statistic into regions and associate each region with an estimated bio-parameter value.

Finally, for every curve we can calculate the value of that statistic and immediately decide on the bio-parameter with the prediction accuracy increasing with the strength of the correlation.

Chapter 4 – Methodology

4.1 Evaluating Algorithms

First task of the project is to determine the best curve matching algorithm for future curve comparing. As mentioned earlier, each curve is a 1x29 vector that has unique association with a set of 4 bio-parameters.

Our target behind curve matching is to estimate the bio-parameters of the sample curve accurately. Therefore, we define the best algorithm as the one showing as most similar curves, those that have similar bio-parameters. Furthermore, the task must be executed at a rapid pace, so speed is another important factor along estimation accuracy.

Each algorithm is tested using the same procedure. Firstly we choose 1,000 random curves, that is, almost 1% of dataset; they are to be tested by all algorithms. Then we compare each curve with the whole dataset, storing values of similarity in a vector and measuring execution time. Then we sort the vector concurrently with a vector of initial positions to keep track of the reference of similarity values to the original curves.

To proceed, we check if the algorithm rates closer matches with lower or higher metric value and we pick curves from the beginning or the end of the sorted vector accordingly. Normally, the closest match is the same curve as the original as the sample curves exist in the compared set of curves. We can pick any number of similar curves we want; we choose the top 20 most similar curves.

Now it’s time to check how accurate the results are. We compare each of the 4 bio-parameters of the sample curve to those of the top 20 and calculate accuracy percentages. Highest score would be 100%, 20,000 / 20,000, because there are 1000 sample curves and 20 most similar to each one. Lowest score would be 5%, 1000 / 20,000, because each top 20 contains at least the original sample curve.

In the end, we calculate accuracy for each bio-parameter “by majority” when at least 11 of the 20 most similar curves have the correct parameter value and by “strong majority” for 14 correct out of 20 and execution time for each algorithm. The most accurate and fast algorithm is declared the most suitable for the problem.

4.2 K-means Clustering

Now the task is to use K-means to cluster the dataset and see how that affects the bio-parameter estimation accuracy. It is well known that clustering improves execution speed decisively because each sample curve will be compared to a fraction of the dataset.

To begin with, we choose Euclidean Distance and Cosine Distance as algorithms with whom to cluster the dataset. Then we perform clustering on the dataset for various K values and create different structures each time.

It is important to remember from Chapter 3.2 that the K-means clustering command might produce different structures and thus we need to repeat the clustering 3-5 times for each K to be certain about the value of a structure with K clusters.

The structures have different "strength"; that is how well placed in clusters the curves are. The more strength, the more successful the clustering is. To rate the structure's strength, we use the silhouette index metric and compare the values for each structure to determine the optimal K.

To continue, we handle a second way to estimate the best number of clusters, the elbow method. In each structure we sum the distance of each curve from the cluster centroid for every cluster. As K increases, the sum decreases, but with a slower rhythm. We use theory behind the elbow method; we plot the diagram of distances for each K and a second estimation of optimal number of clusters is ready.

Finally, we test the impact of clustering with the most practical way. We compare curve matching results when the dataset is clustered vs. when non-clustered. When there is no clustering, we use the methodology described in Chapter 4.1. If clustering has been performed, we initially compare each of the 1,000 sample curves to every centroid curve. Then we compare the sample curve with the collection of curves that belong to the team having the most similar centroid curve.

Again we choose the top 20 most similar curves, however this time they all belong to the same cluster. Finally, we compare the quality of the results of each case; if there is no difference, then clustering has only positive impact on overall performance.

The procedure was repeated a second time with 5,000 other samples to ascertain the findings.

4.3 Creating Decision Regions

The process of creating Decision Regions starts like this:

Initially we study theory behind the curve prediction model and focus on curve features and how they are affected by the bio-parameters. Then we make a list of statistics related with the DR vs. time curves.

We continue by editing and extending the Bio-parameter vector which is initially 103,730x4; we add extra dimensions, one for every statistic we have chosen. In those dimensions we assign values that are calculated for each curve. For instance, let us assume 5th dimension refers to max curve value. We calculate the peak value of curve x and store it in position x,5. The calculations continue for all curves and all statistics that are associated with the new dimensions. After that, we have created the new Parameter vector whose dimensions are 103,730 x (4 + number of statistics).

Then we proceed by calculating the matrix containing all correlations between random variables. This is done by using the following Matlab command [45]:

R = corrcoef (A) % returns the matrix of correlation coefficients for A, where the columns of A represent random variables and the rows represent observations.

The output matrix is rectangular with dimensions (4 + number of statistics) x (4 + number of statistics) and its elements show how strong the correlation between bio-parameters and statistics is. If there is strong correlation between a bio-parameter and a statistic, we can continue by splitting the dataset into teams with curves having that bio-parameter as common value.

Next step is measuring minimum, maximum and average statistics for each team. This is followed by the optimization part. Here, we begin by splitting the value space of the statistic, making initial heuristic Decision Regions. Then we receive each dataset curve and calculate its statistic value. We decide on what the bio-parameter value should be based on the current Decision Region and see how accurate we are for every curve.

Then we automatically move frontiers between regions and test every time if we are improving or not. The plan is to improve the accuracy by every single change and in the end we will have the optimal Decision Regions that offer maximum accuracy. This iterative optimization changes in every repetition almost all frontier values and eventually will end when there is no room for further improvement.

We followed the above methodology selecting the statistic **max value / 28th value** which was found to be strongly correlated to bio-parameter pH_{ES}. After defining the initial Decision Regions heuristically, we kept improving them for many iterations and slowly the frontier values converged into the final ones.

Later, we calculated 2 matrices; one is the distribution of probability of estimation given the true value and the other and more useful, the probability of distribution of true values given the estimation of the test which is based on the Decision Regions. The later matrix determines which teams of curves are safely excluded each time our test makes a decision. This helps substantially with reducing compared dataset as we will see later in Chapter 5.3.

Chapter 5 – Results

5.1 Algorithmic Performance Evaluation

We obtained the following results with regard to (relative) speed (CPU is an i5 @ 3.4GHz):

Algorithm	Comparisons per second	Full dataset curve comparison (103730 curves)
Euclidean	**791,832**	**0.131 sec**
Pearson	8,793	11.80 sec
AWN	60,029	1.727 sec
Cosine	10,856	9.56 sec
SAM	**76,610**	**1.354 sec**
SCM	30,172	3.438 sec

Table 5.1 Curve comparison speed per algorithm

As expected, the simple Euclidean Distance is significantly faster than the other algorithms. Spectral Angle Mapper (SAM) is next, followed by Adaptive Wiener Normalization (AWN). The rest of them lack in speed for this speed-demanding procedure / application.

We obtained the following results with regard to the bio-parameter values of top 20 most similar curves:

Algorithm	-3 or less	-2	-1	**Correct**	+1	+2	+3 or more
Euclidean	1,29%	4,55%	16,54%	**56,86%**	15,42%	4,29%	1,07%
Pearson	0,03%	0,33%	4,36%	**91,82%**	3,40%	0,06%	0%
AWN	0,02%	0,36%	5,05%	**90,54%**	3,89%	0,16%	0%
Cosine	0,03%	0,32%	4,06%	**92,36%**	3,19%	0,07%	0%
SAM	0,03%	0,32%	4,06%	**92,36%**	3,19%	0,07%	0%
SCM	0,03%	0,33%	4,36%	**91,82%**	3,40%	0,06%	0%

Table 5.2 Bio-parameter N of top20 curves, 1000 samples

Difference	-16,7% or less	-15%.. -11,67%	-10% … -6,67%	-5% ... -1,67%	**Correct**	+1,67%... +5%	+6,67%... +10%	+11,67%...+15%	+16,7% or more
Algorithm	-1.0 or less	-0.9...-0.7	-0.6...-0.4	-0.3...-0.1	**Correct**	+0.1...+0.3	+0.4...+0.6	+0.7...+0.9	+1.0 or more
Euclidean	15,10%	4,64%	3,69%	17,74%	**15,57%**	18,27%	3,93%	5,04%	16,02%
Pearson	9,67%	10,52%	14,37%	14,90%	**5,61%**	14,63%	13,75%	9,41%	7,15%
AWN	8,87%	10,26%	14,52%	15,24%	**5,69%**	15,18%	13,98%	9,26%	7,00%
Cosine	9,75%	10,71%	14,09%	14,63%	**5,55%**	14,49%	13,62%	9,70%	7,46%
SAM	9,75%	10,71%	14,09%	14,63%	**5,55%**	14,49%	13,62%	9,70%	7,46%
SCM	9,67%	10,52%	14,37%	14,90%	**5,61%**	14,63%	13,75%	9,41%	7,15%

Table 5.3 Bio-parameter ECS of top20 curves, 1000 samples

As we explained in Chapter 4, the procedure was to find the closest 20 curves from the dataset to the sample curve. Those curves must have something different to the original as every bio-parameter combination is unique. This means that when estimating the four parameter combination, one of them must be inaccurate so that the other 3 parameter estimations are close to perfection.

Undoubtedly, all the algorithms consider ECS to be the least impactful to the curve's shape and thus they choose curves with same bio-parameters N, PH and TP. 5% correct ECS is the lowest possible percentage for the values top 20 most similar curves as every algorithm finds the original curve as the perfect match.

Difference	-4,65% or less	-3,1%	-1,55%	**Correct**	+1,55%	+3,1%	+4,65% or more
Algorithm	-0.3 or less	-0.2	-0.1	**Correct**	+0.1	+0.2	+0.3 or more
Euclidean	0,47%	1,44%	7,91%	**80,85%**	7,75%	0,98%	0,60%
Pearson	0%	0%	0,08%	**99,79%**	0,12%	0,01%	0%
AWN	0%	0%	0,31%	**99,46%**	0,22%	0,01%	0%
Cosine	0%	0%	0,09%	**99,75%**	0,15%	0,01%	0%
SAM	0%	0%	0,09%	**99,75%**	0,15%	0,01%	0%
SCM	0%	0%	0,08%	**99,79%**	0,12%	0,01%	0%

Table 5.4 Bio-parameter pH of top20 curves, 1000 samples

We know from theory and is confirmed later in table 5.17, that pH_{ES} strongly affects the DR vs. time curve's derivative after the peak.

All the algorithms find most similar curves, the ones having the pH and N the same as the original sample curve, sensing correctly that those 2 parameters affects the most the shape of the curves.

Difference	-28% or less	-20% or -24%	-12% or -16%	-4% or -8%	**Correct**	+4% or +8%	+12% or +16%	+20% or +24%	+28% or more
Algorithm	-3.5 or less	-2.5 or -3.0	-1.5 or -2.0	-0.5 or -1.0	**Correct**	+0.5 or +1.0	+1.5 or +2.0	+2.5 or +3.0	+3.5 or more
Euclidean	11,47%	6,82%	10,20%	16,30%	**12,74%**	15,72%	9,63%	6,22%	10,90%
Pearson	1,49%	1,03%	1,45%	5,91%	**82,30%**	4,74%	1,17%	0,86%	1,05%
AWN	2,70%	1,32%	1,17%	6,64%	**79,35%**	4,79%	1,02%	1,07%	1,94%
Cosine	1,58%	1,00%	1,27%	5,74%	**82,82%**	4,46%	1,09%	0,90%	1,14%
SAM	1,58%	1,00%	1,27%	5,74%	**82,82%**	4,46%	1,09%	0,90%	1,14%
SCM	1,49%	1,03%	1,45%	5,91%	**82,30%**	4,74%	1,17%	0,86%	1,05%

Table 5.5 Bio-parameter TP of top20 curves, 1000 samples

The tables 5.2 – 5.5 show that the algorithms will estimate N PH and TP with a high chance of being right and their estimation on ECS is purely irrelevant.

We chose Spectral Angle Mapper (SAM) as the best algorithm in terms of accuracy and speed combination and proceeded with determining the estimation accuracies for each bio-parameter:

Majority correct (11/20 or more)	Big Majority Correct (14/20 or more)
97,30%	90,60%

Table 5.6 Estimation of bio-parameter N (#dysplastic layers), top 20 curves, 1000 samples

Majority correct (11/20 or more)	Big Majority Correct (14/20 or more)
100%	99,80%

Table 5.7 Estimation of pH_{ES}, top 20 curves, 1000 samples

Majority correct (11/20 or more)	Big Majority Correct (14/20 or more)
89,10%	76,10%

Table 5.8 Estimation of TP (Tissue's Porosity), top 20 curves, 1000 samples

Estimation of ECS (Extracellular Space) is meaningless, as we mentioned earlier.

We know from theory and is confirmed in table 5.17, that N strongly affects the DR vs time curve's maximum amplitude.

Every algorithm apart from Euclidean Distance performs well in estimating N problem. That happens because the other algorithms are more suitable for the curve features that include varying derivatives. Their better accuracy is a tradeoff for their lower speed.

The algorithms, except Euclidean Distance, estimated Tissue's Porosity with a high accuracy, around 90%.

Algorithm	Speed	N estimation	ECS estimation	pH estimation	TP estimation	Usefulness
Euclidean	Excellent	Low	Bad	Good	Bad	Some
SAM	Good	Excellent	Bad	Excellent	Good	High (Best)
AWN	Good	Excellent	Bad	Excellent	Good	High
SCM	Low	Excellent	Bad	Excellent	Good	Mediocre
Cosine	Bad	Excellent	Bad	Excellent	Good	Low
Pearson	Bad	Excellent	Bad	Excellent	Good	Low

Table 5.9 Assessment of overall algorithmic performance

Conclusions:

The optimal combination of accuracy and speed is given by Spectral Angle Mapper (SAM) due to the high performance and the best speed amongst the other high accuracy algorithms.

Adaptive Wiener Normalization was a bit worse than SAM in every estimation category, and was 21.6% slower in curve comparison.

Spectral Correlation Mapper, Cosine Distance and Pearson Correlation do not improve the bio-parameter estimation so they are left aside due to their low speed caused by their higher algorithmic complexity.

Euclidean Distance's comparison speed (10 times more than SAM) might prove useful for an earlier stage, when deciding if there is need for further examining an experimental curve. That is left for future work.

5.2 K-means Structure Evaluation

Silhouette Index

We used Silhouette index to rate the clustered structure and to determine the best K, the number of clusters. We obtained the following results:

Euclidean Distance:

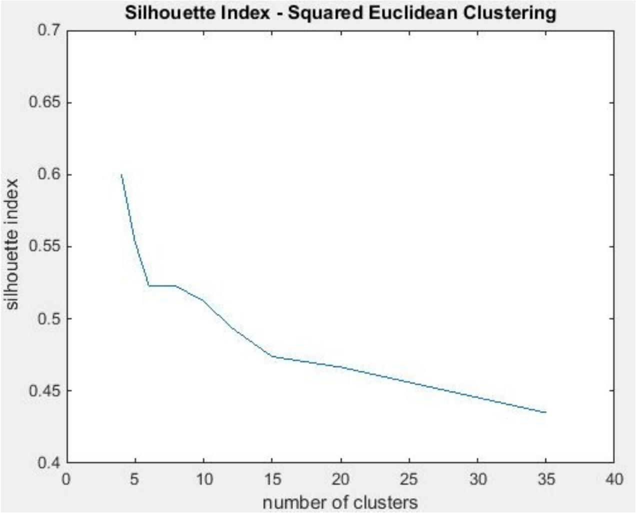


Figure 5.1 Silhouette Index for Euclidean Distance Clustering

Cosine Distance:

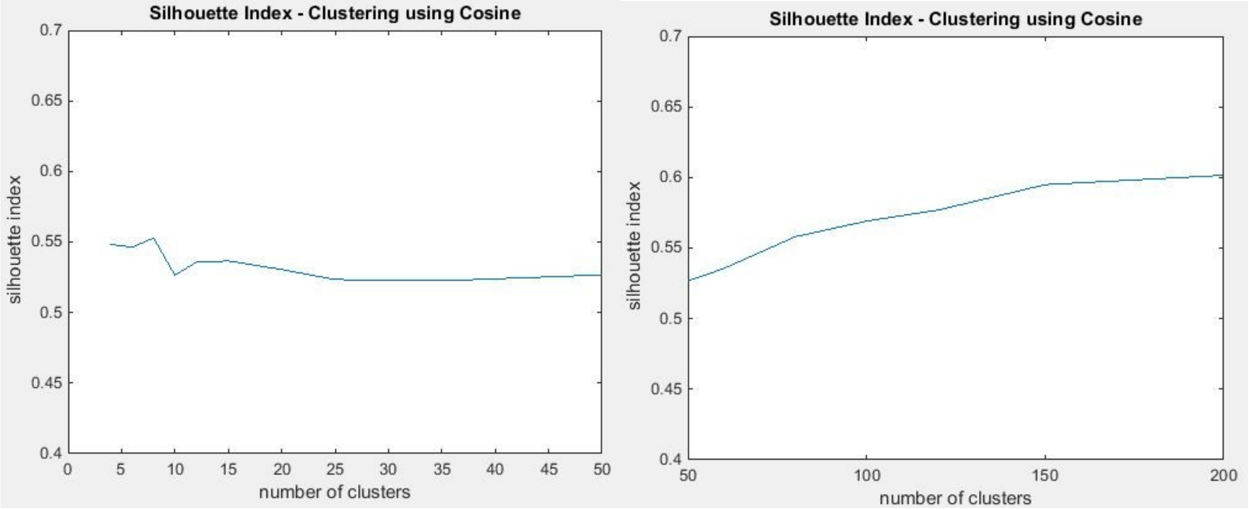


Figure 5.2 Silhouette Index for Cosine Distance Clustering

Silhouette Index	Structure Strength
0.71 – 1.0	A strong structure has been found
0.51 – 0.7	A reasonable structure has been found
0.26 – 0.50	The structure is weak and could be artificial. Try additional methods of data analysis
< 0.25	No substantial structure has been found

Table 5.10 Structure Strength given Silhouette Index

According to the silhouette index metric that rates the structures produced by K-means clustering for various K, no structure was rated “strong” when Euclidean and Cosine distance were used.

Elbow method

We used Elbow method to determine the best number of clusters and we obtained the following results:

Euclidean Distance:

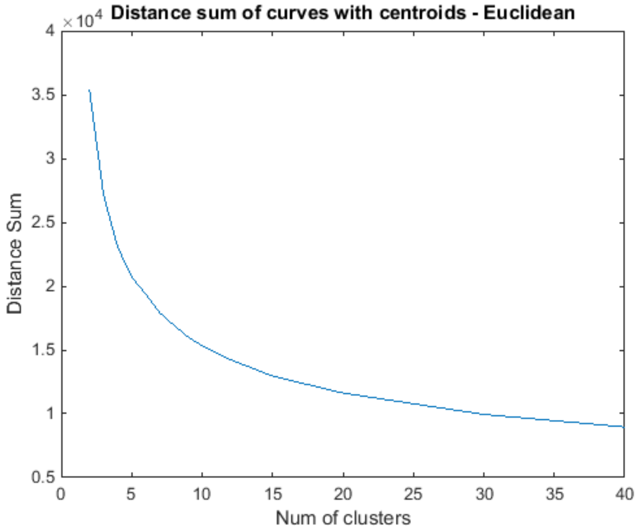


Figure 5.3 SSE for K-means Clustering using Euclidean Distance

According to Elbow metric, Euclidean Distance optimal K equals to 10-12.

Clustering with Euclidean Distance is often efficient at low number of clusters (8-12).

Cosine Distance:

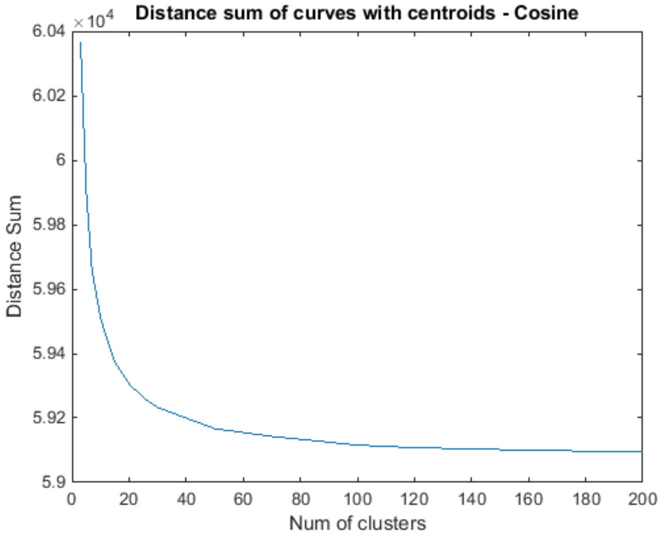


Figure 5.4 SSE for K-means Clustering using Cosine Distance

According to the Elbow metric, Cosine Distance optimal K equals to 25-30.

Clustering using Cosine Distance or an algorithm focused on correlation of curve features, will normally be efficient at higher number of clusters because of the more complex nature of the clustering algorithm.

Another factor that makes higher K more suitable is the 6 digit number of curves.

We ommited clustering using Euclidean Distance because of the low estimation accuracy of the algorithm and we calculated the accuracy of bio-parameter estimation for Cosine Distance clustering with several number of clusters K. Curve matching is done by using **Spectral Angle Mapper** for higher speed, which is equivalent to Cosine Distance (see chapter 3.1):

Number of Clusters K	Majority correct (11/20 or more)	Strong Majority Correct (14/20 or more)
1 (unclustered)	97,30%	90,60%
30	97,03%	89,80%
40	97,03%	89,80%
50	97,07%	90,23%
70	96,73%	89,67%
100	96,90%	89,63%
150	96,90%	89,80%

Table 5.11 Estimation of bio-parameter N (#dysplastic layers), top 20 curves, 1000 samples

Number of Clusters K	Majority correct (11/20 or more)	Strong Majority Correct (14/20 or more)
1 (unclustered)	100%	99,80%
30	100%	99,80%
40	100%	99,67%
50	99,87%	99,63%
70	99,90%	99,70%
100	99,90%	99,70%
150	99,60%	99,20%

Table 5.12 Estimation of pH_{ES}, top 20 curves, 1000 samples

Number of Clusters K	Majority correct (11/20 or more)	Strong Majority Correct (14/20 or more)
1 (unclustered)	89,10%	76,10%
30	88,80%	75,10%
40	88,50%	74,80%
50	88,40%	75,20%
70	88,13%	74,73%
100	87,83%	74,40%
150	88,40%	74,60%

Table 5.13 Estimation of TP (Tissue's Porosity), top 20 curves, 1000 samples

After confirming that the clustering using cosine distance is beneficial, we repeated the procedure for 5000 other samples:

Number of Clusters K	Majority correct (11/20 or more)	Strong Majority Correct (14/20 or more)
30	97,20%	89,26%
40	96,98%	88,90%
50	97,20%	89,34%
70	96,92%	88,98%
100	96,86%	89,28%
150	96,98%	88,74%

Table 5.14 Estimation of bio-parameter N (#dysplastic layers), top 20 curves, 5000 samples

Number of Clusters K	Majority correct (11/20 or more)	Strong Majority Correct (14/20 or more)
30	99,90%	99,80%
40	99,94%	99,78%
50	99,90%	99,72%
70	99,92%	99,72%
100	99,94%	99,78%
150	99,86%	99,58%

Table 5.15 Estimation of pH_{ES}, top 20 curves, 5000 samples

Number of Clusters K	Majority correct (11/20 or more)	Strong Majority Correct (14/20 or more)
30	90,16%	75,44%
40	88,82%	75,10%
50	89,94%	75,28%
70	89,68%	74,88%
100	89,64%	75,20%
150	89,66%	74,50%

Table 5.16 Estimation of TP (Tissue's Porosity), top 20 curves, 5000 samples

We see that we can cluster the dataset with Cosine Distance using high number of K; for instance 50-100, without trading much of the upper bound of estimation accuracy.

This is normal because there are 11 different pH teams and 10 different N teams and SAM thinks they are well separated; each cluster may have unique pH and N values, 11*10 = 110 clusters. For higher K values, bio-parameter estimation becomes less reliable.

Conclusion:

For this application, K-means clustering has tremendous beneficial impact by lowering procedure time of curve comparing because of leaving a 1/K fraction to be compared. This substantial improvement is done without affecting much the estimation quality.

5.3 Bio-parameter Estimation Accuracy through Decision Regions

We obtained the following results regarding correlation between curve statistics and Bio-parameters:

	Correlation with 5)	Correlation with 6)	Correlation with 7)
1) N	0,0454	**0,8716**	**0,8484**
2) ECS	0,0016	0,3810	0,3807
3) PH_{ES}	**0,8280**	0,0288	0,1044
4) TP	0,0850	0,1604	0,1875
5) max value / 28th value	1		
6) max value		1	
7) integral			1

Table 5.17 Correlation of Curve Statistics and Bio-parameters

The measured correlations are strong enough to proceed as the following table suggests:

Random Variable Correlation Coefficient	Description
< 0,3	little if any (linear) correlation
0,3 – 0,5	low correlation
0,5 – 0,7	moderate correlation
0,7 – 0,9	high correlation
0,9 – 1,0	very high correlation

Table 5.18 Description of Correlation Strength given Coefficient

We continued the analysis selecting the statistic **max value / 28th value** which is strongly correlated to pH and shows how much the curve's amplitude decreases after its peak.

We continued by splitting the dataset into teams according to methodology, obtaining the following results:

Team's pH Value	Team's Minimum Statistic	Team's Maximum Statistic	Team's Average Statistic
6,0	1	1	1
6,1	1	1,0100	1,0010
6,2	1	1,0415	1,0146
6,3	1	1,0942	1,0483
6,4	1	1,1728	1,1025
6,5	1	1,2789	1,1781
6,6	1	1,4202	1,2765
6,7	1,0015	1,6015	1,4008
6,8	1,0155	1,8217	1,5569
6,9	1,0440	2,1050	1,7510
7,0	1,0868	2,4692	1,9900

Table 5.19 Statistic Values of each Curve Team

By using optimization that was described in Chapter 4.3, we obtained the following splitting of value space into Decision Regions:

Statistic (max / 28th) Value Region	pH Value Decision
1	6,0
1,0000001 – 1,007717	6,1
1,007718 – 1,040007	6,2
1,040008 – 1,093706	6,3
1,093707 – 1,171033	6,4
1,171034 – 1,274265	6,5
1,274266 – 1,411475	6,6
1,411476 – 1,588933	6,7
1,588934 – 1,805585	6,8
1,805586 – 2,086017	6,9
2,086018 – 2,4693	7,0

Table 5.20 pH Decision Regions

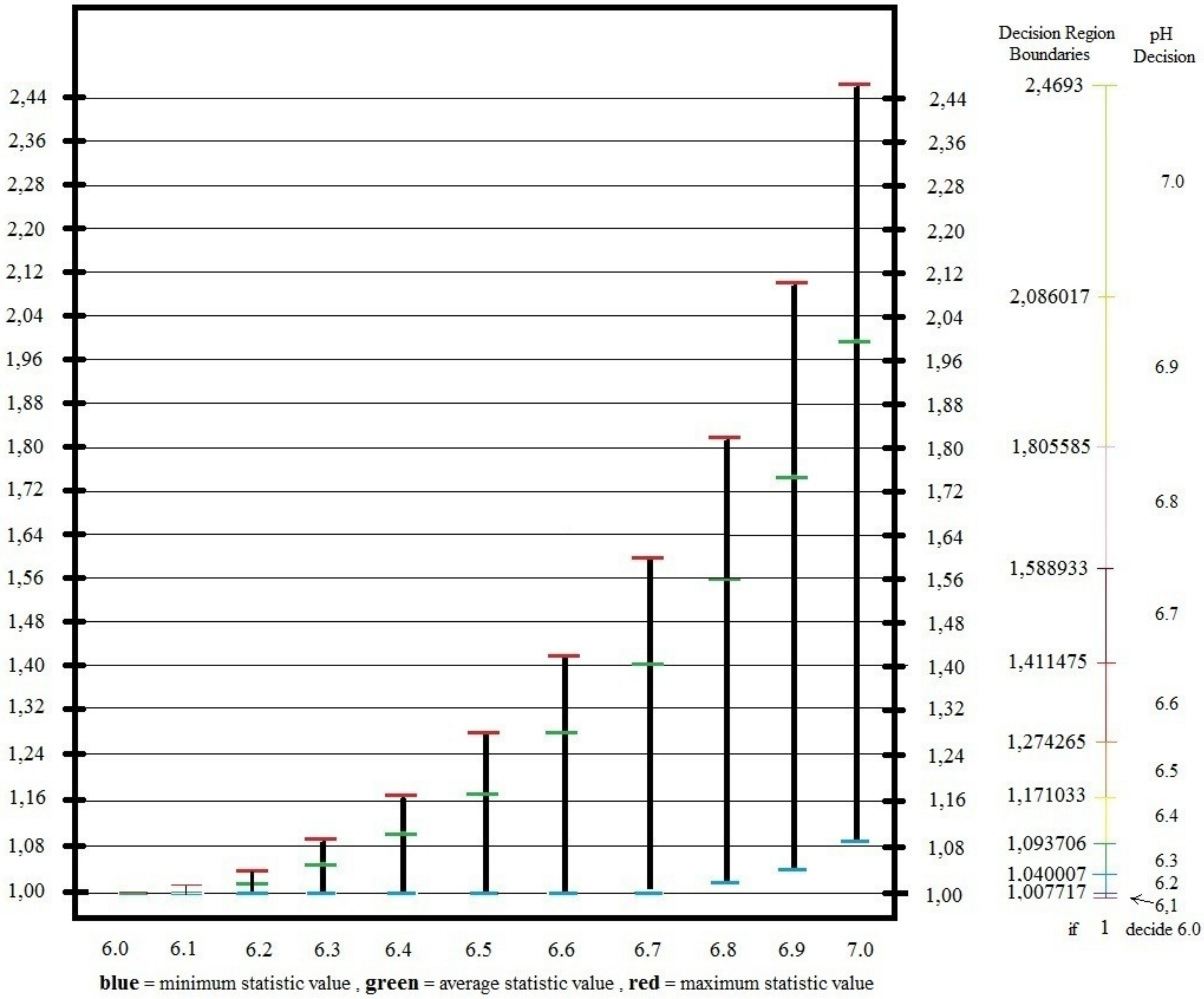


Figure 5.5 Statistic Value Distribution of each Team and pH Decision Regions

We obtained the following results in accuracy for instant pH estimation using the Decision Regions:

Real Value minus Estimated Value	-0.1	**(correct)** 0	+0.1	+0.2	+0.3	+0.4	+0.5
Percentages	0,57%	**59,04%**	20,76%	8,73%	4,82%	2,91%	2,33%

Table 5.21 Accuracy of pH estimation with Decision Regions

Method Advantages:

- Highest estimation speed.
- Will not undersell the situation. Estimated value is almost the lower bound.
- Can be extended as a method by combining statistics.

Method Disadvantages:

- In rare cases, the estimation might be far off the real value.

In the following chart we see the distribution of real values for every estimated value from the test:

Estimated Values	**Real Values**										
	6.0	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
6.0	**42,13%**	27,01%	13,31%	7,88%	4,43%	3,53%	1,72%				
6.1		**56,95%**	16,06%	10,04%	8,87%	1,94%	5,14%	1,00%			
6.2		1,51%	**64,39%**	14,10%	5,78%	4,21%	2,47%	6,20%	1,35%		
6.3			0,31%	**64,37%**	15,65%	7,76%	2,81%	3,05%	4,20%	1,41%	0,45%
6.4				0,05%	**63,92%**	15,74%	8,39%	3,95%	3,74%	2,76%	1,45%
6.5					0,20%	**63,58%**	17,13%	7,14%	4,94%	5,29%	1,72%
6.6						0,77%	**58,48%**	21,07%	8,04%	5,82%	5,83%
6.7							0,83%	**56,88%**	24,18%	10,63%	7,49%
6.8								0,83%	**59,57%**	26,98%	12,62%
6.9									1,43%	**66,33%**	32,24%
7.0										1,96%	**98,04%**

Table 5.22 Distribution of Real Values for every Estimated Value

In the following chart we see the distribution of estimated values from the test for every real value:

Real Values	**Estimated Values**										
	6.0	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
6.0	**100%**										
6.1	64,11%	**34,52%**	1,37%								
6.2	31,60%	9,73%	**58,37%**	0,30%							
6.3	18,70%	6,09%	12,78%	**62,39%**	0,05%						
6.4	10,51%	5,38%	5,24%	15,16%	**63,51%**	0,20%					
6.5	8,39%	1,18%	3,82%	7,52%	15,64%	**62,64%**	0,82%				
6.6	4,07%	3,12%	2,24%	2,73%	8,34%	16,87%	**61,81%**	0,83%			
6.7		0,60%	5,62%	2,96%	3,92%	7,03%	22,27%	**56,86%**	0,73%		
6.8			1,22%	4,07%	3,71%	4,87%	8,49%	24,17%	**52,41%**	1,06%	
6.9				1,37%	2,75%	5,21%	6,15%	10,63%	23,73%	**49,22%**	0,95%
7.0				0,43%	1,44%	1,70%	6,16%	7,49%	11,10%	23,92%	**47,75%**

Table 5.23 Distribution of Estimated Values for every Real Value

5.4 Improvement on Comparing Procedure

We now turn our attention to what knowledge Table 5.14 offers. Basically, it says that for each pH estimated value, some teams are automatically excluded from the similarity search. For instance, if the prediction is 6.7, teams with 6.0 to 6.5 are excluded and the search will follow the path where the probabilities of finding the correct team are higher.

It is important to note that, as we saw in chapter 5.1, the good algorithms show all most similar curves having the same pH value as the original. This means, after adopting the knowledge from the Correlation into Decision Regions idea, **we only need to determine the correct team** and the most similar curves will all belong there!

Now we quantify the improvement from this knowledge using the number from table 5.14. Each team has 9430 curves. We will calculate how many teams we need to check every time:

Probability of Correct Estimation	Number of Curve Comparisons	Average Comparisons (Number * Probability)
42,13%	9430	3973
27,01%	18860	5094
13,31%	28290	3765
7,88%	37730	2972
4,43%	47150	2089
3,53%	56580	1997
1,72%	66010	1135
		Total: 21026,07

Table 5.24 Average Number of Curve Comparisons for Estimated pH = 6.0

Probability of Correct Estimation	Number of Curve Comparisons	Average Comparisons (Number * Probability)
56,95%	9430	5370
16,06%	18860	3029
10,04%	28290	2840
8,87%	37730	3346
5,14%	47150	2424
1,94%	56580	1098
1,00%	66010	660
		Total: 18766,64

Table 5.25 Average Number of Curve Comparisons for Estimated pH = 6.1

Probability of Correct Estimation	Number of Curve Comparisons	Average Comparisons (Number * Probability)
64,39%	9430	6072
14,10%	18860	2659
6,20%	28290	1754
5,78%	37730	2180
4,21%	47150	1985
2,47%	56580	1397
1,51%	66010	997
1,35%	75440	1018
		Total: 18063,17

Table 5.26 Average Number of Curve Comparisons for Estimated pH = 6.2

Probability of Correct Estimation	Number of Curve Comparisons	Average Comparisons (Number * Probability)
64,37%	9430	3973
15,65%	18860	5094
7,76%	28290	3765
4,20%	37730	2972
3,05%	47150	2089
2,81%	56580	1997
1,41%	66010	1135
0,45%	75440	339
0,31%	84870	263
		Total: 17362,52

Table 5.27 Average Number of Curve Comparisons for Estimated pH = 6.3

Probability of Correct Estimation	Number of Curve Comparisons	Average Comparisons (Number * Probability)
63,92%	9430	6028
15,74%	18860	2969
8,39%	28290	2374
3,95%	37730	1490
3,74%	47150	1763
2,76%	56580	1562
1,45%	66010	957
0,05%	75440	38
		Total: 17179,57

Table 5.28 Average Number of Curve Comparisons for Estimated pH = 6.4

Probability of Correct Estimation	Number of Curve Comparisons	Average Comparisons (Number * Probability)
63,58%	9430	5996
17,13%	18860	3231
7,14%	28290	2020
5,29%	37730	1995
4,94%	47150	2329
1,72%	56580	973
0,20%	66010	132
		Total: 16676,01

Table 5.29 Average Number of Curve Comparisons for Estimated pH = 6.5

Probability of Correct Estimation	Number of Curve Comparisons	Average Comparisons (Number * Probability)
58,48%	9430	5515
21,07%	18860	3974
8,04%	28290	2275
5,83%	37730	2199
5,82%	47150	2744
0,77%	56580	436
		Total: 17141,85

Table 5.30 Average Number of Curve Comparisons for Estimated pH = 6.6

Probability of Correct Estimation	Number of Curve Comparisons	Average Comparisons (Number * Probability)
56,88%	9430	5364
24,18%	18860	4560
10,63%	28290	3007
7,49%	37730	2825
0,83%	47150	391
		Total: 16147,93

Table 5.31 Average Number of Curve Comparisons for Estimated pH = 6.7

Probability of Correct Estimation	Number of Curve Comparisons	Average Comparisons (Number * Probability)
59,57%	9430	5617
26,98%	18860	5088
12,62%	28290	3570
0,83%	37730	313
		Total: 14589,15

Table 5.32 Average Number of Curve Comparisons for Estimated pH = 6.8

Probability of Correct Estimation	Number of Curve Comparisons	Average Comparisons (Number * Probability)
66,33%	9430	6255
32,24%	18860	6080
1,43%	28290	405
		Total: 12739,93

Table 5.33 Average Number of Curve Comparisons for Estimated pH = 6.9

Probability of Correct Estimation	Number of Curve Comparisons	Average Comparisons (Number * Probability)
98,04%	9430	9245
1,96%	18860	370
		Total: 9614,83

Table 5.34 Average Number of Curve Comparisons for Estimated pH = 7.0

Then we calculate the probability the test estimates each value to find average number of comparisons:

Estimated Value	Estimation Probability
6.0	21,58%
6.1	5,51%
6.2	8,24%
6.3	8,81%
6.4	9,03%
6.5	8,96%
6.6	9,61%
6.7	9,09%
6.8	8,00%
6.9	6,75%
7.0	4,43%

Table 5.35 Distribution of Test Estimated Values for every Real Value

We see that the test estimates pH value equal to 6.0 far more often than the rest values because many curves have their statistic equal to 1, meaning they never decline and their peak is right at the final value.

Estimated Value	Estimation Probability	Total Average Number of Comparisons per Estimated Value	Weighted Average Number of Comparisons
6.0	21,58%	21026,07	4537,43
6.1	5,51%	18766,64	1034,04
6.2	8,24%	18063,16	1488,40
6.3	8,81%	17362,52	1529,64
6.4	9,03%	17179,57	1551,32
6.5	8,96%	16,676,01	1494,17
6.6	9,61%	17141,85	1647,33
6.7	9,09%	16147,93	1467,85
6.8	8,00%	14589,15	1167,13
6.9	6,75%	12739,93	859,95
7.0	4,43%	9614,83	425,94
			Total: 17203,19

Table 5.36 General Average Number of Curve Comparisons

Conclusion:

If we search for the best curve similarities on the whole dataset, we need 103730 comparisons. If we take this test into account and use the knowledge it offers, we need on average 17203 comparisons which represent a mere 16,58% of the dataset (83,42% less comparisons).

That improvement happens because we search with specific order (higher probability of appearing decides the order) and we also exclude teams with PH_{ES} with 0 probability of appearing.

If we smartly pick the correct team with few comparisons, we might achieve the lower bound of number of comparisons of this method which is 9430 plus the few extra comparisons to determine the correct team. To do that we need tested similarity thresholds to ensure the quality of the result.

Then we can also optimize the comparisons inside the team by first choosing smartly the correct space value of bio-parameter N and possibly lower the comparisons into around 2000 in total (1886 is the number of curves with 2 different N values inside a team, so as not to miss something) in total which would be the best achievable target for this procedure.

5.5 Use Cases – Indicative Timings

We will now present statistics about execution times for the predefined procedures / use cases and estimations for more generic future procedures.

We studied the case in which we need to compare a sample curve and efficiently, both in accuracy and speed, determine it's bio-parameters by listing the most similar curves to our sample. We suggested two ways to improve this procedure and we will now measure and compare them.

Method 1:

Use clustering on the dataset with Cosine Distance (we know it is effectively the same with Spectral Angle Mapper which is the one we consider best for curve comparison regarding this problem). The sample curve is compared with the team having the most similar centroid to that curve.

Depending of number or clusters K, the procedure becomes faster. For high K, the accuracy declines as we showed in Chapter 5.2. We also showed there that, for cosine distance, picking a high K value without trading the speed gain to accuracy loss is feasible, for instance K = 50.

Every curve is compared with approximately K + 103730 / K curves = 50 + 103730 / 50 = 2075 comparisons

Method 2:

Alternatively, we showed we can use what correlation and decision regions combination can offer and lower the number of comparisons. Unoptimized result is 17200 comparisons on average while estimated feasible lower bound is 2000 comparisons per curve.

Relative (to processor and programming enviroment) comparison speed of SAM is 76610 comparisons per second. So the time required to perform the procedure for a single sample curve for each method is:

Method 1 : 2075 / 76610 = 27,1 ms
Method 2 (unoptimized, current upper bound) = 17200 / 76610 = 224,5 ms
Method 2 (optimized, estimated lower bound) = 2000 / 76610 = 26,1 ms

Generic Procedure

What we analyzed before is just a part of the big picture of the general problem. When we deal with the full problem, estimating bio-parameters for all the cervix area, we just need to calculate efficient how many little one curve problems we have to solve. We will try to analyze future optimizations and estimate the execution time of the full problem. We consider as input a future full HD DySIS Map image (1920x1080 = 2,073,600 pixels = curves).

Initially we may use euclidean distance and quickly exclude around 95% of the curves as areas of no interest. This requires a T_1 execution time. To continue, in the interesting areas we can do a sampling; only in case future work proves no critical information is lost. That may be true if we assume that little areas very close to each other have the same bio-parameters That sampling may reduce the 5% of the curves left to 1% of the total number of curves.

Then we have 2,073,600 * 1% = 20,736 curves to compare. By method 1 we need 20,736 * 27.1 ms = 561.5 sec, by method 2 unoptimized, we need 20,736 * 224.5 ms = 4655 sec and by method 2 optimized, we need 20,736 * 26.1 ms = 541.2 sec.

That can be further improved with the use of GPUs. Clearly, this application has the potential to run in real time.

Chapter 6 – Epilogue

6.1 Summary

This thesis was aimed at exploring possibilities of estimating cervical bio-parameter from curves produced during acetowhitening phenomenon.

Although the analysis is concerning only theoretical model curves, we are pretty confident by the results we demonstrated in the algorithmic evaluation, that parameter estimation can be done accurately and, with some future improvements, in real time. Spectral Angle Mapper (SAM) offers the best combination of quality and speed for this problem.

In addition, the results from clustering are promising for further performance optimization.

Finally, the problem handling with decision regions offered worthy results showing the potential many interesting scientific methods have in this innovative topic.

6.2 Future work

This rich area/field of research has a lot of interesting dimensions and multiple problems to focus on in the future. Some of them are:

- Testing procedure with experimental data.

- Finding a cutoff regarding DR curves for healthy/unimportant areas, possibly by using the speed of Euclidean Distance.

- Estimating % of spatial reduction for the cervix area. Close pixels should have same bio-parameters and produce exactly the same curves and there is no need to test them all especially in the absence of abnormalities.

- Calculating speed acceleration when using GPUs (helps substantially with vectors).

References

Papers

[1] D.Hanahan, R.A.Weinberg, ”Hallmarks of Cancer”, Cell, vol 100, page 57-70, 2000

[2] D.Hanahan, R.A.Weinberg, ”Hallmarks of Cancer: The Next Generation”, Cell, vol 144, page 646-74, 2011

[15] Melinkow *et al.*, 1998

[19] Raab SS, Grzybicki DM, Zarbo RJ, et al. Frequency and outcome of cervical cancer prevention failures in the United States. AmJClin Pathol 2007;128:817^24.

[21] Ioanna Vourlaki, Costas Balas, George Livanos, Manos Vardoulakis, George Giakos, Michalis Zervakis
Bootstrap Clustering Approaches for Organization of Data: Application in improving grade separability in cervical neoplasia, 2017

[23] J. Louwers, A. Zaal, M. Kocken, W.A. ter Harmsel, G. Graziosi, J. Spruijt, J. Berkhof, C. Balas, E. Papagiannakis, P. Snijders, C. Meijer, F. van Kemenade and R. Verheijen, “Dynamic spectral imaging colposcopy: higher sensitivity for detection of premalignant cervical lesions,” BJOG, vol. 118, no. 3, pp. 309-18, Feb, 2011.

[24] W. P. Soutter, E. Diakomanolis, D. Lyons, S. Ghaem-Maghami, T. Ajala, D. Haidopoulos, D. Doumplis, C. Kalpaktsoglou, G. Sakellaropoulos, S. Soliman, K. Perryman, V. Hird, C.H. Buckley, K. Pavlakis, S. Markaki, R. Dina, V. Healy, and C. Balas, “Dynamic spectral imaging: improving colposcopy,” Clin Cancer Res, vol. 15, no. 5, pp. 1814-20, Mar 1, 2009.

[26] C. Balas *et al. “ In Vivo* Molecular Imaging of Cervical Neoplasia Using Acetic Acid as Biomarker”, 2008

[27] C. Balas *et al. “*Modelling of epithelial transport phenomena related with the acetowhitening optical characteristics: potential for the in-vivo diagnosis of cervical neoplasia”, 2008

[28] I. F. Tannock and D. Rotin, “Acid pH in tumors and its potential for therapeutic exploitation,” *Cancer Res.*, vol. 49, pp. 4373–4384, 1989.

[29] S. D. Webb, J. A. Sherratt, and R. G. Fish, “Mathematical modelling of tumour acidity: Regulation of intracellular pH,” *J. Theor. Biol.*, vol. 196, pp. 237–250, 1999.

[30] C. Balas *et al. “*Modelling of epithelial transport phenomena related with the acetowhitening optical characteristics: potential for the in-vivo diagnosis of cervical neoplasia”, 2008

[31] D. C. Walker *et al.*, “A study of the morphological parameters of cervical squamous epithelium,” *Physiological Measurements,* vol. 24 pp. 1–15, 2003.

[32] R. A. Gatenby, and E. T. Gawlinski, “The glycolytic phenotype in carcinogenesis and tumor invasion: insights through mathematical models,” *Cancer Res,* vol. 63, no. 14, pp. 3847-54, Jul 15, 2003.

[33] G. Papoutsoglou, “In vivo Molecular Imaging of Epithelial Pre-Cancer based on Dynamic Optical Scattering Modeling”, 2014

[34] I. J. Latorre, K. K. Frese, and R. T. Javier, “Tight junction proteins and cancer,” in *Tight Junctions*, L. Gonzalez-Mariscal, Ed. New York: Springer US and Landes Bioscience, 2006, pp. 116–134.

[35] C. Balas *et al. “In vivo* Dynamic Imaging, *in silico* Modeling and Global Sensitivity Analysis for the Study and the Diagnosis of Epithelial Neoplasia”, 2011

[36] A. Saltelli *et al.*, *Global Sensitivity Analysis: The Primer.* Chichester, U.K.: Wiley, 2008.

[37] C. Balas *et al. “*Estimation of Neoplasia-Related Biological Parameters Through Modeling and Sensitivity Analysis of Optical Molecular Imaging Data”, 2012

[38] A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola, “Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index,” *Comput. Phys. Commun.*, vol. 181, no. 2, pp. 259–270, Feb. 2010.

[39] P. Swietach, R. D. Vaughan-Jones, and A. L. Harris, “Regulation of tumor pH and the role of carbonic anhydrase 9,” *Cancer Metastasis Rev.*, vol. 26, no. 2, pp. 299–310, 2007.

[40] C. Balas *et al. “*Estimation and Mapping of Cervical Neoplasia-Related Parameters through the Comparison of the *in vivo* Measured with the *in silico* Modeled Dynamic Bio-Optical Characteristics”, 2014

[41] G. Sobel, et al., “Increased expression of claudins in cervical squamous intraepithelial neoplasia and invasive carcinoma.,” Hum. Pathol., vol. 36, no. 2, pp. 162–9, Feb. 2005.

[42] T. Giakoumakis “Multiband, Dynamic and Molecular Imaging”, 2016

Other Reference

[3] https://gis.cdc.gov/Cancer/USCS/DataViz.html

[4] https://www.cancer.gov/about-cancer/understanding/statistics

[5]https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/global-cancer-facts-and-figures/global-cancer-facts-and-figures-3rd-edition.pdf

[6] https://www.cancer.org/cancer/cervical-cancer/about/key-statistics.html

[7]https://www.iatropedia.gr/eidiseis/o-karkinos-stin-ellada-sokaroun-oi-arithmoi-pliri-statistika-kai-provlepsi-dekaetias/50941/

[8] https://www.youtube.com/watch?v=7CpG63zdCPo&t=64

[9] https://www.smartdraw.com/reproductive-system-diagram/examples/female-reproductive-system-diagram/

[10] https://www.cdc.gov/cancer/gynecologic/quiz/index.htm

[11] https://www.cdc.gov/cancer/cervical/basic_info/risk_factors.htm

[12] https://www.cdc.gov/cancer/cervical/basic_info/symptoms.htm

[13] Cervix, Cervical Cancer and Colposcopy introduction, (chapter 2) An introduction to Cervical Intraepithelial Neoplasia.

[14] https://oncogenesisdx.com/cancer-etiology/

[16]https://www.cancer.org/cancer/cervical-cancer/prevention-and-early-detection/cervical-cancer-screening-guidelines.html

[17] https://www.cancer.net/cancer-types/cervical-cancer/statistics

[18] https://www.cancer.org/cancer/cervical-cancer/prevention-and-early-detection/abn-pap-work-up.html

[20] https://www.cdc.gov/cancer/cervical/basic_info/diagnosis_treatment.htm

[22] https://dysismedical.com/clinicians/dysis-in-action/

[25] https://downriverobgyn.com/obgyn-services/colposcopy/dysis-advanced-cervical-screening/dysis-colposcopy/

[43] https://www.mathworks.com/help/stats/silhouette.html
[44] https://www.quora.com/How-can-we-choose-a-good-K-for-K-means-clustering
[45] https://www.mathworks.com/help/matlab/ref/corrcoef.html

[46] https://www.healthline.com/health/cervix-treatment-cryosurgery