Department of Computing

Computational Methods for Classifying Glaucomatous Visual Field Measurements

Shuanghui Meng

This thesis is presented for the degree of

Doctor of Philosophy

of

Curtin University of Technology

December 2007

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: _____

Date: _____

Abstract

Glaucoma is a common eye disease that affects the optic nerve. It is the second leading cause of visual loss globally and while it can occur in all age groups, it is most common in the elderly. The main symptom of glaucoma is the progressive deterioration of the visual field. Management of glaucoma involves careful monitoring of the progress of disease with regular visual field tests. Accurate identification and early intervention can potentially prevent advanced vision loss. A number of mathematical, statistical, and data mining methods have been proposed to identify glaucomatous progression. However, all criteria used to assess change are hampered by noise that arises from individual visual field measurement. In addition, different clinical trials use different definitions of "progressing". Currently there is no standard method for classifying changes in visual field measurements. The purpose of this thesis is to improve existing methods and to develop new methods for classification of glaucoma.

The thesis first describes a glaucoma modeling software according to a patient's clinical behaviour. The software can handle age-related visual decline, different types and rates of deterioration, and noise. Simulated data is a good resource for testing the efficiency of different methods in detecting progression, and for developing new methods with minimal cost.

The thesis then investigates four classification techniques, including Event Analysis(EA), sequence matching, point-wise linear regression (PLR) and machine learning. For EA methods, the thesis proposed an algorithm "*baseline-follow-up*" for building a 95% (or 99%) confidence interval using a database of repeated Standard Automated Perimetry (SAP) tests of stable visual fields. Experimental results show that the proposed algorithm can improve the sensitivity compared with other EA methods. A major novel contribution is our introduction of sequence matching techniques to the

application of glaucomatous visual field data. Sequence matching techniques typically rely on similarity measure. However, visual field measurements are very noisy, particularly in people with glaucoma, and there is lack of a standard definition of progression. It is therefore difficult to establish a reference dataset including both stable and progressive visual fields. We describe two different matching methods, Weighted Sequence Matching (SM) and Baseline Matching Stable Sequences (BMS). SM uses either the Euclidean or Manhattan distance function to select matches in a stable database R for a given query sequence. BMS uses a *baseline* computed from a query sequence to match means of stable sequences in R. Matches are used to determine whether a query sequence is progressive or not. For PLR methods, the thesis explores the influence of updating a PLR method by adding or deleting an observation, and discusses the application of Kappa statistic for agreement between methods. We finally investigate the application of machine learning methods for the classification of visual field data. Various input features are defined. The feature datasets are extracted from visual field data, in which each patient has been classified by experts. For this study we used the WEKA package, which provides implementations for Decision Tree, Decision Stump, Naive Bayes, and Bayes Network classifiers, as well as Bagging and Boosting methods for applying the classifiers. The accuracy of classification is presented to illustrate the ability of machine learning for classifying visual field change.

Acknowledgements

There are many people I would like to thank for their huge input and ongoing support throughout the completion of this thesis.

My sincere gratitude goes to my supervisors, Dr Mihai Lazarescu, Dr Jim Ivins and Dr Andrew Turpin. Without their invaluable advice, encouragement and constructive criticism, I would never have accomplish this work. I am indebted to Dr Andrew Turpin for his invaluable guidance and superb ideas and suggestions to bettering this research. I am grateful to Dr Lazarescu for accepting the role as my main supervisor when Dr Turpin moved to a new position at RMIT University, and for his wonderful support and encouragements towards finishing this thesis. Many thanks to my co-supervisor Dr Jim Ivins for his patience and advice on each draft, which helped to improve the quality of the thesis tremendously.

I also would like to express my sincere appreciation to Prof. Chris A. Johnson at Devers Eye Institute, Portland, Oregon USA, and Prof. Bal Chauhan at Dalhousie University, Canada for providing real visual field data sets.

Finally, thanks to my husband and children, for their love and emotional support, as well as their understanding and patience.

The work described in this thesis was funded by Curtin University Australian Postgraduate Award (APA).

Publications Based on This Thesis

The work presented in this thesis has been published over the course. The details are as follows:

- Shuanghui Meng, Andrew Turpin and Mihai Lazarescu. Methods for Calculating Glaucoma Change Probability Confidence Intervals. *Proceedings of the 5th Postgraduate Electrical Engineering and Computing Symposium (PEECS04)*, Western University, Perth, Western Australia, Sept. 2004, pp. 159-162.

- Shuanghui Meng, Andrew Turpin, Mihai Lazarescu and Jim Ivins. Classifying Virtual Field loss in Glaucoma through Baseline Matching of Stable Reference Sequences. *Proceedings of 2005 International Conference on Machine Learning and Cybernetics*. Guangzhou, China, August 2005, pp. 3688-3691.

- Shuanghui Meng, Mihai Lazarescu, Jim Ivins and Andrew Turpin. Detecting Change in Visual Fields using Sequence Matching. *Proceedings of the 6th Postgraduate Electrical Engineering and Computing Symposium (PEECS05)*, Edith Cowan University, Perth, Western Australia, Sept. 2005, pp. 12-16.

- Shuanghui Meng, Mihai Lazarescu, Jim Ivins and Andrew Turpin. Monitoring Glaucomatous Progression: Classification of Visual Field Measurements Using Stable Reference Data. *Lecture Notes in Artificial intelligence, Vol 3930: Advances in Machine Learning and Cybernetics, Revised Selected Papers (4th International Conference ICMLC 2005, Guangzhou, China, August 2005)*, Springer Berlin / Heidelberg, 2006, pp. 750-759.

- Shuanghui Meng, Andrew Turpin, Mihai Lazarescu and Jim Ivins. A Comparison of Algorithms for Calculating Glaucoma Change Probability Confidence Intervals. *Journal of Glaucoma*. 2006, 15(5) pp. 405-413.

Contents

List of Figures

List of Tables

List of Abbreviations

ACG	Angle Closure Glaucoma
AGIS	Advanced Glaucoma Intervention Study
SAP	Standard Automated Perimetry
CI	Confidence Interval
CIGTS	Collaborative Initial Glaucoma Treatment Study
CPSD	Corrected Pattern Standard Deviation
dB	decibels
FT	Full Threshold
GCP	Glaucoma Change Probability
GHT	Glaucoma Hemifield Test
LF	Long-term Fluctuation
MD	Mean Deviation
OPG	Open Angle Glaucoma
PLR	Point-wise Linear Regression
PT	Pattern Deviation
PST	Pattern Standard Deviation
SAP	Standard Automated Perimetry
SF	Short-term Fluctuation
TD	Total Deviaiton

Chapter 1

Introduction

Glaucoma is a group of eye diseases that affects the optic nerve. It is the second leading cause of visual loss worldwide (Gupta, 2005; Quigley and Broman, 2006) and while it can occur in all age groups, it is most common in the elderly. Glaucoma is predicted to affect 60.5 million people by the year 2010, and this figure will increase to 79.6 million by the year 2020 (Quigley and Broman, 2006). This is because age is one of the major risk factors of glaucoma, and average life expectancy is on the increase. The *prevalence* [1] of glaucoma in Australia is about 3% of the population according to Mitchell et al. (1996); Rochtchina and Mitchell (2000).

There are several different types of glaucoma including open angle glaucoma (OPG), angle closure glaucoma (ACG) (also called closed-angle glaucoma), congenital glaucoma, and secondary glaucomas. A common feature in all types of glaucoma is that the optic nerve is invariably damaged. In most cases, increased pressure in the eye is the cause of this damage. The optic nerve connects the retina to the brain and is responsible for carrying visual information to the brain. Optic nerve damage therefore causes loss of vision, and may ultimately lead to blindness.

Open angle is the most common type of glaucoma. 74% of patients in the world will have OPG by 2020 (Quigley and Broman, 2006). The development of early glaucoma is a very gradual process, and because the loss of sight is slow and painless, it can be hard to notice (Distelhorst and Hughes, 2003). Damage progresses very slowly and destroys vision gradually, starting with the peripheral vision. One eye covers for

[1] The prevalence of a disease can be defined as the number of persons inflicted with a particular disease or condition in a given population at a designated time (Gupta, 2005)

the other, and the person remains unaware of any problem until a majority of nerve fibres have been lost, and a large part of vision has been destroyed. This damage is irreversible; however, with early detection and timely treatment the damage process can at least be slowed. Therefore, it is crucial to detect the problem as early as possible, to be able to start treatment with as little damage to vision as possible. If a patient continues to lose visual function, the glaucoma is said to be *progressing*, otherwise it is said to be *stable*.

Because early treatment can prevent much advanced vision loss, it is essential to assess a patient's visual function through regular eye examinations. There are three types of simple, painless tests (IGA, 2003).

- **Tonometry**: measuring internal eye pressure, which is one of the factors thought to cause glaucoma.
- **Ophthalmoscopy**: looking at the back of the eye to determine whether damage to the optic nerve can be seen.
- **Perimetry**: if the pressure in the eye is not in the normal range (8 - 21mmHg) or the optic nerve looks unusual, then a special glaucoma test will be done. This **perimetry** or **visual field test** is a mainstay of glaucoma diagnosis and management (Harrington, 1971; Drance and Anderson, 1985; Anderson and Patella, 1999; Gupta, 2005). Perimetry is a measure of the extent of the sensitivity of vision. At present, perimetry is the best test by which to determine the extent of glaucomatous damage to visual function and whether or not visual loss is progressive (Werner et al., 1990). Combining perimetry and computer technology, automated perimetry has been applied in clinical practice for many years. Today, a person's visual field is tested with automated perimeters and standard thresholding algorithms which have improved the ability to quantify visual function. It also has normative data for different groups, and quantitative measurement and analysis methods (Johnson, 1995) as described in Chapter 2.

The purpose of a visual field examination is to detect *defects* [2], determine the specific pattern of visual field loss for diagnosis, and monitor patients for evidence of

[2] defect is an imperfection or absence

visual field progression (Spry et al., 2002). Data obtained by automated perimetry is important not only to aid in disease diagnosis but also to monitor disease progression. Unfortunately, however, data obtained by automated perimetry is typically very noisy. All criteria used to assess change are hampered by noise that arises from learning effects, fatigue, and the inherent variation in the automated tests, especially at locations with progression, all of which make the early identification of progression a difficult task. Thus, one of the major challenges for treating and monitoring glaucomatous patients is to develop a method which can quickly and reliably identify true progression.

1.1 Aims and Approaches

This thesis aims to improve existing methods and develop new methods for classification of glaucoma. It examines point-wise analysis used in traditional methods, as well as the application of artificial intelligence and data mining techniques on perimetry data.

As noted above, early and reliable identification of glaucomatous progression is an important part of the management of glaucoma. A number of mathematical, statistical, and data mining methods have been proposed to determine glaucomatous progression. Some methods focus on analysis of whole visual fields such as *Global Indices* or *Scoring Systems*. These methods are accurate for identifying stable patients; however, for progressive patients, especially those with localized vision loss, these methods are not sensitive. Because a small number of decreasing locations are averaged out across 76 (or 54) locations, and noise is produced by measurement, identifying progressive patients is very difficult. Furthermore, methods based on whole fields do not provide spatial information about where vision loss has occurred.

Some methods focus on point-by-point analysis such as *Glaucoma Change Probability* and *point-wise linear regression*. Although these point-wise analyses are useful for analysing longitudinal glaucomatous data and determining spatial patterns, there is no universally accepted standard against which to validate them (Spry et al., 2002). Different clinical trials use different definitions of "progressing". This lack of a standard definition of progression has led to the development of arbitrary or empirical

criteria for identifying change (Wikins et al., 2006).

The purpose of the work described in this thesis was to distinguish glaucomatous progression from noise as accurately as possible. We first emphasised the use of empirical data to classify a given location in a visual field as progressing or stable. Then we explored statistical methods and machine learning classifiers. In order to evaluate the effectiveness of our methods, we first obtained a large set of real patient data. We then simulated sets of test data according to the behavior of progressive glaucomatous patients and stable subjects. Using a simulation to generate test data has several advantages in this context. In particular, this simulation can provide test data without long delays (real data is obtained by repeated measurements over several years), and also can control visual change and noise. Furthermore, the simulated data are known to be progressive or stable by design. The combination of real and simulated data gave great flexibility in designing experiments.

Of the traditional classification methods, we first consider Glaucoma Change Probability (GCP) analysis which is widely used clinically. The GCP method is commercially available with the Stapac program of the Humphrey Field Analyzer (Carl Zeiss, Meditech, Dublin CA). In order to avoid confusion with the commercially available GCP method which is a part of the Statpac by Zeiss (Carl Zeiss, Meditech, Dublin CA), we rename GCP method now to Event Analysis (EA).

The key question we are asking with this method is how best to establish a 95% or 99% confidence interval by using repeated stable glaucomatous data, because the accuracy of the Event Analysis depends on the range of the confidence intervals. We propose an algorithm to build a confidence interval, called "*baseline-follow-up*". We examine this algorithm not only on simulated data, but also on repeated stable glaucomatous data. Next we investigate linear regression methods to classify field change. Finally, we investigate various machine learning classifiers' ability to detect progressive visual data.

1.2 Significance and Novelty

The significance of the work presented in this thesis rests on the fact that millions people in the world have glaucoma or glaucomatous risk factors. Some affected in-

dividuals are unaware of their field loss, which can delay of treatment. As already noted, glaucoma is the second leading cause of irreversible blindness throughout the world (Gupta, 2005; Quigley and Broman, 2006). Blindness severely restricts a person's normal activities (Spaeth et al., 2006). Therefore, it is important to correctly identify a patient with glaucoma, and to determine whether it is stable or progressive, so that appropriate treatment may be given.

1. We propose a novel algorithm for building a 95% confidence interval for the application of the existing EA method. This algorithm gives higher sensitivity than other EA methods in the literature.

2. We use new matching techniques to classify glaucomatous data. Matching techniques have been successfully used in many areas. Because of noise, however, we cannot use exact matching techniques. We therefore propose approximate matching methods for classification. We implement the matching method and test its performance on real and simulated datasets. To our knowledge, this research presents the first matching technique applied to glaucomatous data.

3. We present two different point-wise linear regression methods for identifying progression. The new methods suggest that high levels of noise make it difficult to identify progressive data.

4. We investigate the application of Artificial Intelligence (AI) techniques for accurate classification of glaucomatous data. In this study, we analyse the strengths and weaknesses of different features and machine learning classifiers, in terms of accuracy for this problem domain.

1.3 Structure of the Thesis

The remainder of the thesis is organised as shown in Figure 1.1.

Chapter 2 reviews work related to classification of glaucomatous data. First, automated perimetry for quantifying visual function is briefly described. Next, methods of classification, including Global Indices, Scoring Systems, Linear regression techniques, bias analysis, machine learning techniques, and the criteria for classifying progression are presented.

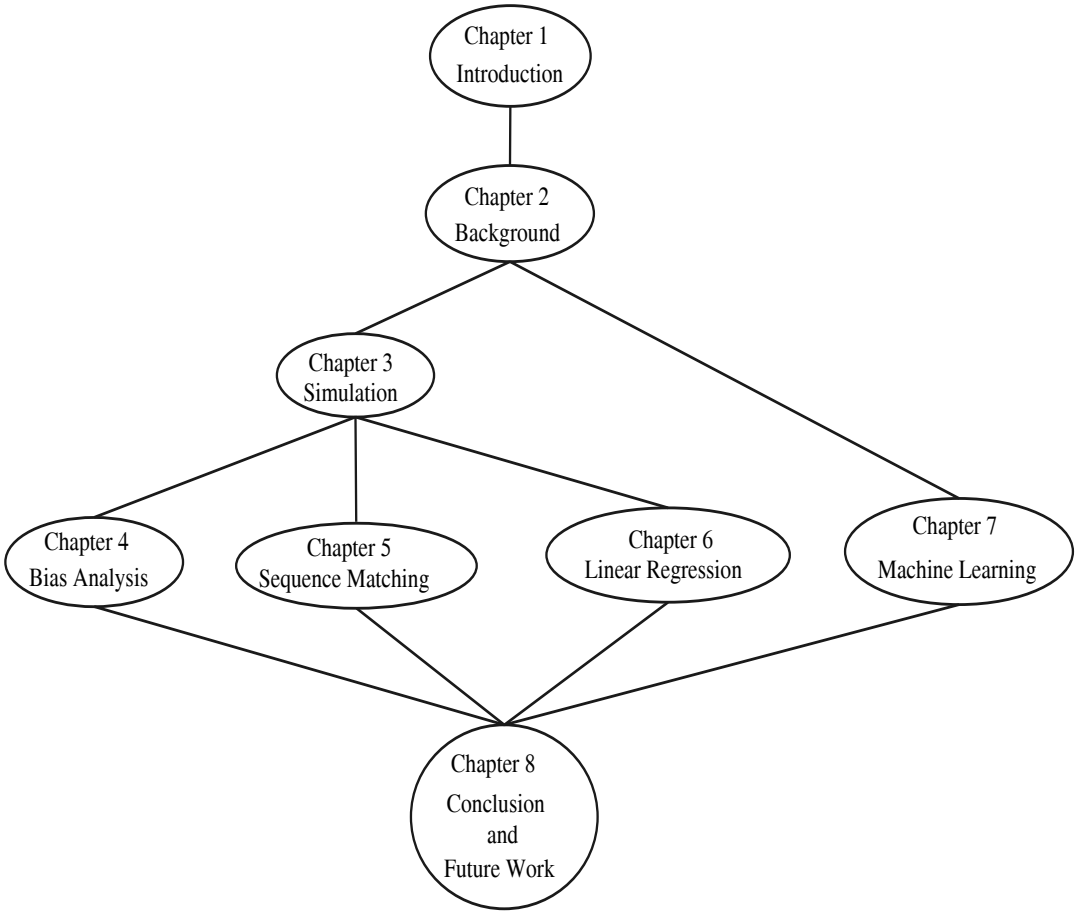


Figure 1.1: Structure of the thesis

Chapter 3 presents a model for simulating glaucomatous data. Computer simulation of visual field progression can offer a fast, cost-effective, controllable data source to examine the effectiveness of methods on a point-by-point basis. This simulation can provide different types of progressive and stable cases. It is also possible to produce any number of tests, and control the noise in the data.

Chapter 4 describes a modification of the method for building a confidence interval that improves the accuracy of the Event Analysis methods. The methods are evaluated on simulated datasets.

Chapter 5 introduces approximate matching techniques for classification of glaucomatous data. A distance function is used to select the closest matches to a query sequence. Matched sequences are then used to determine whether the query sequence is stable or progressive. The methods are evaluated on simulated and real datasets. The results are compared with the EA method.

Chapter 6 focuses on the application of linear regression techniques to classify glaucomatous visual field data. The techniques are designed to add or delete observations in a linear least-squares regression fit. Different methods are compared and the agreement between methods are examined.

Chapter 7 presents the application of machine learning methods for the classification of visual field data. Various input features are defined. The feature data sets are extracted from visual field data, in which each patient has been classified by experts. The accuracy of classification is presented to illustrate the ability of machine learning for classifying change.

Finally, conclusions and future research directions are presented in Chapter 8.

Chapter 2

Background

In this chapter various techniques for determining glaucomatous progression will be reviewed. We begin by explaining the Standard Automated Permetry (SAP) technique which examines a patient's *visual field*. (A visual field is the field of vision for one eye. Note that the visual fields of the two eyes overlap.) We then describe methods that have been developed and used to determine visual field change over time. Sequential visual fields are analysed by using statistical or non-statistical methods.

Procedures for classification of glaucomatous progression can be divided into five broad categories (Turpin et al., 2001; Spry et al., 2002):

- "global indices", which average information across the whole field;
- scoring systems, which assign scores to locations based on their threshold values and those of their immediate neighbours, and then sum the scores into a single value for the whole visual field;
- linear regression methods, which follow test parameters over time to determine the magnitude and significance of patterns within the data;
- Event Analyses, which identify single events of significant change relative to a reference examination; and
- machine learning classifiers, which include artificial neural networks, decision trees, and linear support vectors.

The SAP technique is briefly introduced in Section 2.1. Each category of classification methods is discussed in Section 2.2. Comparison of classification methods is

discussed in Section 2.3. Section 2.4 describes simulation which is used to generate some glaucomatous visual field datasets. A summary is presented in Section 2.5.

2.1 Visual Field Measurement

2.1.1 Thresholds

Automated perimetry is a diagnostic technique used for assessing visual function (Koch et al., 1972; Fankhauser et al., 1972; Portney and Krohn, 1978; Flammer et al., 1985; The Advanced Glaucoma Intervention Study Investigators, 1994; Spry et al., 2002). It is primarily intended for use with patients who have glaucoma or are glaucoma suspects (Werner et al., 1990; Wilson, 2002). Assessing visual function involves quantifying the ability to see a light intensity (visual stimulus) at representative locations in the field. The stimulus intensity can be varied depending on the patient's response (seen or unseen) to previous stimuli. There is a boundary between visibility and invisibility where the patient's responses are typically uncertain or inconsistent. There is a borderline stimulus that sometimes is seen and sometimes is not seen. When a stimulus intensity is not seen at all, the intensity can be adjusted to a level at which the patient can respond 50% of the time. The stimulus intensity at which the patient responds 50% of the time is defined as the *threshold* [1] (Anderson and Patella, 1999), recoded in decibels (dB). Note that the dB scale in automated perimetry represents the *attenuation* of the maximal stimulus intensity, and so 0 dB represents the perimeter's nominal maximum stimulus. Thus a decrease of 1 dB always represents an increase in stimulus intensity.

The measurement of thresholds is time-consuming. However, in clinical perimetry, the threshold is estimated in different ways for different perimetry. For example the weakest stimulus seen is taken as the threshold estimate on the Humphrey Perimeter, by presenting a series of stimulus intensities 2, 3 or 4 dB apart (Anderson and Patella, 1999, p. 18). At present, perimetry is the best test to determine the extent of glaucomatous damage, and whether or not visual field loss is progressive. There are different types of techniques to test visual field such as blue-on-yellow perimetry

[1] We used "threshold" here instead of "threshold sensitivity" to avoid confusion with the concept of the sensitivity of a method

(yellow background and blue lights) (Johnson et al., 1993), Frequency Doubling Technology perimetry (Cello et al., 2000; Maddess et al., 2001; Spry et al., 2001; Turpin et al., 2002), and white-on-white perimetry (Traquair, 1944; Turpin et al., 2001; Tan et al., 2002). In this study, we briefly introduce white-on-white, full threshold Standard Automated Perimetry visual field measurements.

2.1.2 Standard Automated Perimetry

The Humphrey Field Analyzer (HFA Carl Zeiss Meditec, Dublin, CA, USA) is most commonly used in clinical settings for assessing visual function (Hutchings et al., 2000). Standard Automated Perimetry (SAP), also known as white-on-white perimetry, is the standard technique available on the HFA. Standard automated perimetry uses a white stimulus on a white background. It requires subjects to sit still, place their chin on an immobile stand, and fixate on a spot at the center of a half sphere with a white background. Figure 2.1 shows an instrument used for automated visual field measurement.

The visual field is usually tested with a full threshold testing algorithm. That is, an algorithm designed to get threshold values at many locations in the visual field. The central 30 degrees (30°, called 30-2) or 24 degrees (24°, called 24-2) of the visual field is tested, typically at locations 6 degrees apart. Figure 2.2 shows a map of the central 30 degree of the visual field. The central 24 degree of the visual field is shown by the dotted-line. During a test, subjects are asked to press a button whenever they see a flash of light. Lights of varying intensity are shown in each of 54 or 76 locations. Each location in the retina corresponds to a certain direction in the visual field. The 54 or 76 locations make up a grid covering the central 24 or 30 degrees of the visual field. The marginally visible intensity of light is recorded as the *threshold*. Each of the locations within a visual field has a threshold value reported in decibels (dB), printed on a map of the visual field (Anderson and Patella, 1999). For practical purposes, the useful intensity range for white light testing is from 0 to 40 dB. A threshold of 0 dB indicates that the brightest light could not be seen - in other words, the location is blind. Threshold values of 35-40 dB indicate exceptional vision.

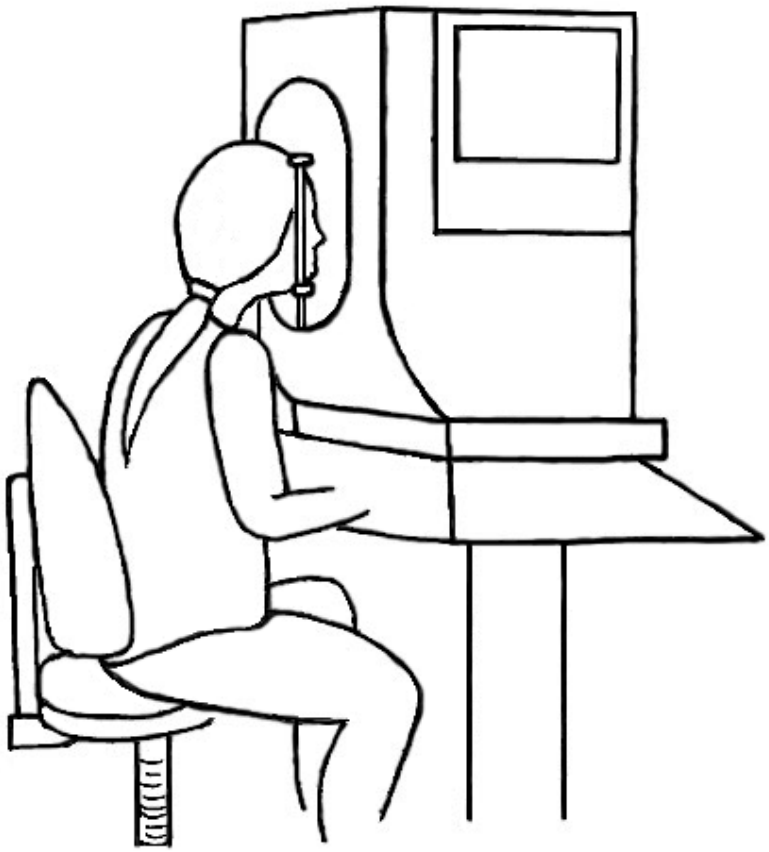


Figure 2.1: An instrument for visual field measurement.

Figure 2.2 shows the results of visual field measurement of the central 30 degrees (30-2) at 76 locations on a right eye with localised vision loss in the top left corner. The filled central circle is the fixation spot. If the central spot is taken as the origin of coordinate axes, there are the two locations near the blind spot (or two physiological blind spots) at (15, -3) and (15, 3), as indicated by small zeros (Anderson and Patella, 1999; Gupta, 2005). Note that for the left eye, the blind spots are at (-15, -3) and (-15, 3). For convenience in data processing, all data from the left eye is converted to right eye format by reflection about the vertical axis.

In clinical use, the measured field is compared and analysed with an age-corrected normal field held in a database in the machine. Using this database, several analyses are performed by the machine, such as "Total Deviation", "Pattern Deviation" and "probability maps". These are defined in the following sections.

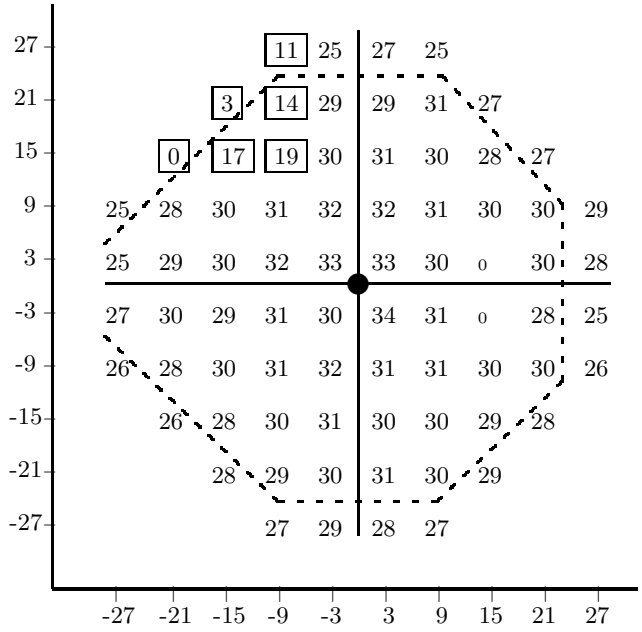


Figure 2.2: A map of the 76 locations in a 30° SAP visual field showing localized loss (in boxes) in the top left quadrant. Small zeros indicate physiological blind spots. The dotted area is the central 24° of the field.

2.1.3 Total Deviation

Total Deviation (TD) values (in dB) are the difference between measured thresholds and age-corrected normal thresholds at each location. That is,

$$(TD)_i = X_i - N_i \qquad i = 1, 2, \ldots, 74$$

where X_i is the measured threshold and N_i is the age-corrected normal threshold at location i. Note that the two locations near the blind spot are excluded. Figure 2.3 shows a measured field, a normal field for a given age group, and the corresponding Total Deviation plot.

The total deviation also can be displayed in a symbol map which indicates the statistical significance of each measured deviation. The symbols increase in darkness as the deviation becomes more significant. The significance limits are derived from a database of normal visual field examinations. For example, if a location is marked with the symbol for $p < 1\%$, that means fewer than 1% of normal age-corrected fields have a threshold value that low (Anderson and Patella, 1999). Figure 2.4 shows an example of a TD probability plot for the field in Figure 2.2. A small dot indicates that a location is in the normal visual field range. Four small dots indicate that a location

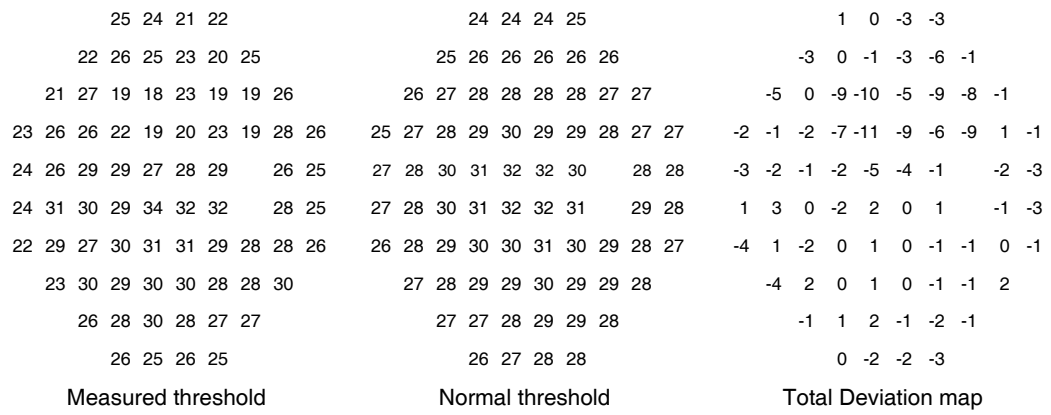


Figure 2.3: Total deviation. The left plot is a measured field (values in dB), the middle plot is an age-matched normal visual field, and the right plot is the total deviation.

is below the 5% significance level. Each symbol indicates a typical significance level and is shown on the right top of the map. "p" in Figure 2.4 stands for significance level or probability.

2.1.4 Pattern Deviation

Pattern Deviation (PD) values are derived from Total Deviation values by adjusting for the General Height (GH). The GH index, in the case of 24-2 (central 24 degree), is defined as the 85th percentile of the distribution of the TD values among the 51 stimulus locations corresponding to the stimulus grid of Program 24-2 (the three stimulus locations in the blind spot are excluded) (Asman and Heijl, 1992a; Asman et al., 2004; Artes et al., 2005). The calculation of GH index is as follows: (1) The TD values are ranked in order from most positive to most negative; (2) the 85th percentile of this ranking, i.e. the seventh-highest value, is taken as the GH index (Figure 2.5).

In the same manner used for Total Deviation probability map, the probability symbols in Pattern Deviation probability map highlight test points where the decibel Pattern Deviation values approach the end of the normal range.

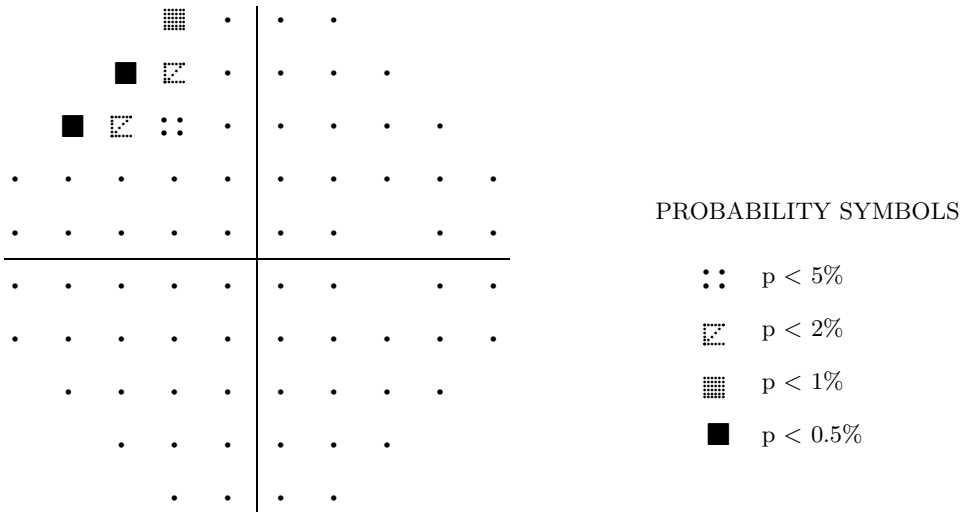


Figure 2.4: A TD probability plot with abnormal locations in the top left quadrant.

2.1.5 Visual field fluctuation

A major problem with automated measurement is that visual field thresholds are not perfectly repeatable, even in normal individuals (Anderson and Patella, 1999; Spry et al., 2002). Many factors can influence the outcome of any given single visual field measurement (Hutchings et al., 2000), such as, for a single visual field test, it is not feasible to spend more than about 10 seconds determining the threshold at each location because of fatigue (Turpin et al., 2001). Fatigue typically causes depression of the threshold, especially in the periphery of the central visual field (Johnson et al., 1988; Hudson et al., 1994). Second, learning effects influence test results. Amjad et al. (2002) and Fujimoto et al. (2002) have evaluated the influence of learning effects. The average threshold in a visual field at the first measure is lower than at the second and third tests. Moreover, learning effects are generally smaller in the central portion of the field than in more peripheral regions (Werner et al., 1990; Heijl and Bengtsson, 1996; Tan et al., 2002). Third, locations with visual field loss increase the measurement variability (Anderson and Patella, 1999). Moreover, a patient may give unreliable response or make inconsistent decisions when responding to the lights (Werner et al., 1990). Finally, the thresholding strategy may cause noise (Spry et al., 2002). For example, the FASTPAC algorithm saves test time: approximately 30% of the time taken by the Full Threshold algorithm. However, FASTPAC offers less accuracy compared with the Full Threshold algorithm (Anderson and Patella, 1999). These factors

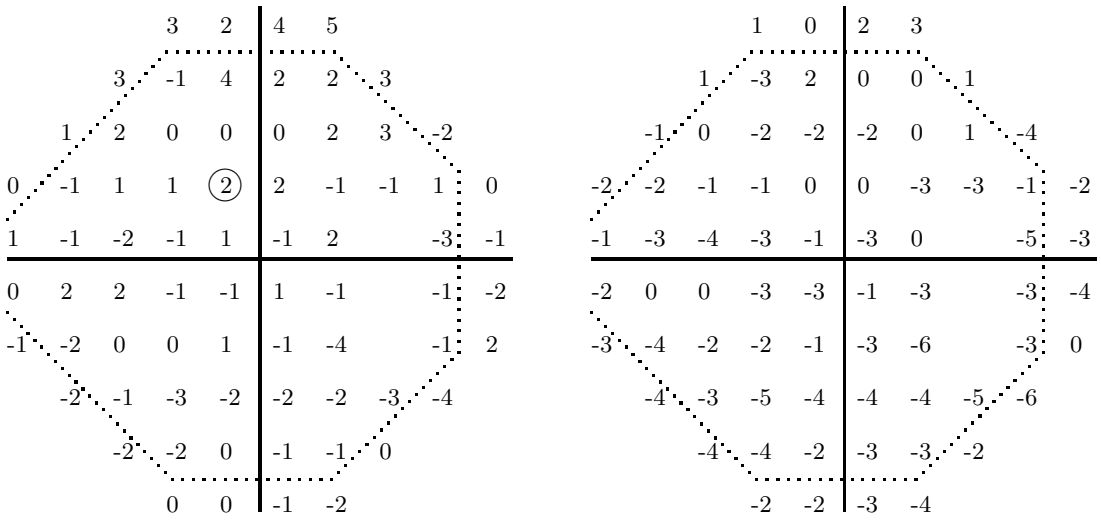


Figure 2.5: Pattern deviation. The left graph is the Total Deviation map. The seventh highest deviation (2 dB) is circled and represents the General Height. The right graph is the Pattern Deviation map formed by subtracting GH.

occur within and between any two successive tests. The variation in the threshold estimate within a test at a given location is termed *intra-test variability*. The amount of intra-test variability is called *Short-term Fluctuation* (SF). Short-term fluctuation represents the variability of threshold measurements repeated at the same location within single visual field test. In contrast, *Long-term Fluctuation* (LF) or *inter-test variability* is the variability in thresholds between two visual field tests. The entire field can also fluctuate from test to test depending on factors such as the patient's mood and alertness, or the physiological state of the patient such eye pressure and heart rate. Long-term fluctuation is divided into a "homogeneous" component in which all locations undergo a change in threshold in unison, and a smaller "heterogeneous" component that accounts for the fact that locations do not all change by the exact same amount (Drance and Anderson, 1985).

In short, visual fields are typically very noisy due to both short- and long-term fluctuation of which there are many sources. Distinguishing genuine change in a visual field from measurement variability is therefore a difficult and complex task (Spry et al., 2002).

2.2 Techniques for Classifying Glaucomatous Progression

Glaucoma patients are usually required to have SAP measurements at half-yearly or yearly intervals. When a series of visual field measurements is obtained, the clinician's task is to diagnose whether visual field loss has occurred, and if so whether it is true progression of the disease or measurement noise. There are several techniques to aid the clinician in this task.

2.2.1 Global Indices

This group of classification techniques includes Mean Deviation (MD), Pattern Standard Deviation (PSD), Short-term Fluctuation (SF), and Corrected Pattern Standard Deviation (CPSD) (Flammer et al., 1984, 1985; Drance and Anderson, 1985; Heijl et al., 1986). Each global index averages thresholds across all locations of the visual field to produce a *single* index value. Each index is calculated based on an age-matched reference normal visual field. The normal threshold at a given location in the visual field is defined as the mean of thresholds measured at that location in disease-free eyes within a given age group (Drance and Anderson, 1985).

Mean Deviation (MD)

Mean deviation (Flammer et al., 1984, 1985; Heijl et al., 1986) is calculated from a weighted mean of Total Deviations (section 2.1.3) of the patient's field. That is:

$$MD = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{X_i - N_i}{S_{1i}^2}}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{S_{1i}^2}} = \frac{\sum_{i=1}^{n}\frac{X_i - N_i}{S_{1i}^2}}{\sum_{i=1}^{n}\frac{1}{S_{1i}^2}} \tag{2.1}$$

where n is the number of tested locations except for the two locations of the blind spot; X_i is the measured threshold at location i; N_i is the corresponding age matched normal reference threshold at location i; and S_{1i}^2 is the variance of the normal measurements at location i.

Assume $(N_i)_1, (N_i)_2, \ldots, (N_i)_m$ to be a series of thresholds at location i in a series of reference normal fields $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$. That is, $(N_i)_1$ is at location i in the field

$\mathcal{M}_1$, $(N_i)_2$ is at location i in the field $\mathcal{M}_2$, ..., $(N_i)_m$ is at location i in the field $\mathcal{M}_m$. Therefore, S_{1i}^2 is calculated as follows:

$$S_{1i}^2 = \frac{1}{m}\sum_{j=1}^{m}((N_i)_j - \bar{N_i})^2 \tag{2.2}$$

where

$$\bar{N_i} = \frac{1}{m}\sum_{j=1}^{m}(N_i)_j$$

Because of the large influence of noise in the periphery (Anderson and Patella, 1999), the locations in the periphery of the reference normal field have a higher variance S_{1i} than the locations in the central field (closer to fixation). Therefore changes at the central points in a field can affect the MD value more than changes at the peripheral points because they have a higher weight in the weighted sum.

As MD is an index of the average of all thresholds, it can be affected by the degree of loss and the number of affected locations. MD is more negative when the defects in the visual field progress. An MD of -4 dB may indicate a 4 dB depression of thresholds everywhere in the field (Anderson and Patella, 1999). If each location in a visual field is depressed, the visual field has *generalized depression*. If only some locations in a visual field are depressed, it has *localized depression*. Both generalized and localized losses affect the MD index.

When a small MD value is obtained, it indicates the measured field is close to normal or the field has only small localized loss. In the right side of Equation 2.1, S_{1i} from the normal reference field does not change, so the MD value is determined by $X_i - N_i$. If each measured X_i is very close to N_i, then $X_i - N_i$ is small. Hence the measured field is normal. If few locations change (that is, only a few values $|X_i - N_i|$ are increasing), the mean of all $X_i - N_i$ may be small (that is, MD may be small), particularly if they are in the periphery where weights $\frac{1}{S_{1i}}$ are small.

The MD index quantifies the degree of overall generalized depression in the absence of localized loss. The main advantage of MD is that when a series of visual field measurements is taken, the series of MD values gives a quick impression of whether the series shows any trend that needs closer inspection. It may be helpful to notice, for example, that the MD is much the same in each measurement, which suggests

that neither progressive generalized depression nor substantial progressive localized loss is occurring (Anderson and Patella, 1999). If a series of MD values is gradually decreasing, then a systematic progressive deterioration of the entire visual field may be evident.

The main disadvantage of MD is that only any subtle localised visual field loss may be incorrectly classified as stable.

Pattern Standard Deviation (PSD)

Pattern standard deviation is a weighted standard deviation of the point wise differences between the measured and the normal reference fields. The formula (Anderson and Patella, 1999) is as follows:

$$PSD = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} S_{1i}^2\right)\left(\frac{1}{n-1}\sum_{i=1}^{n}\frac{(X_i - N_i - MD)^2}{S_{1i}^2}\right)} \tag{2.3}$$

The PSD index is the standard deviation around the mean that constitutes the MD index. If all values in the field are equally abnormal according to Equation 2.1, the PSD value is zero calculated by using Equation 2.3. That is, when values in the field are equally abnormal the variance around the mean is not affected. If a field is normal, then the PSD value is small because X_i is close to N_i and the MD value is small, and $(X_i - N_i - MD)^2$ is small. Therefore a small value for PSD indicates that the patient's field is similar to the normal reference field. If only some locations are more affected, then PSD becomes larger. Because S_{1i}^2 from the normal dataset is constant, MD may be small if only a few locations change as discussed for "Mean Deviation (MD)" above, and $X_i - N_i$ is more negative, the value $(X_i - N_i - MD)^2$ becomes large at these locations. Therefore, PSD is an estimate of localised loss.

The advantage of PSD is that after a series of visual fields is measured, PSD increases as some locations first develop deterioration, or the localized depression enlarges (Pearson et al., 1990; Anderson and Patella, 1999), but it may remain at a fixed value during disease being stable. Therefore it is an index for showing localized change in the visual field (Katz et al., 1991). However, once abnormal, the lack of further change of this index should not be taken as a sign that glaucomatous field loss is stable (Anderson and Patella, 1999).

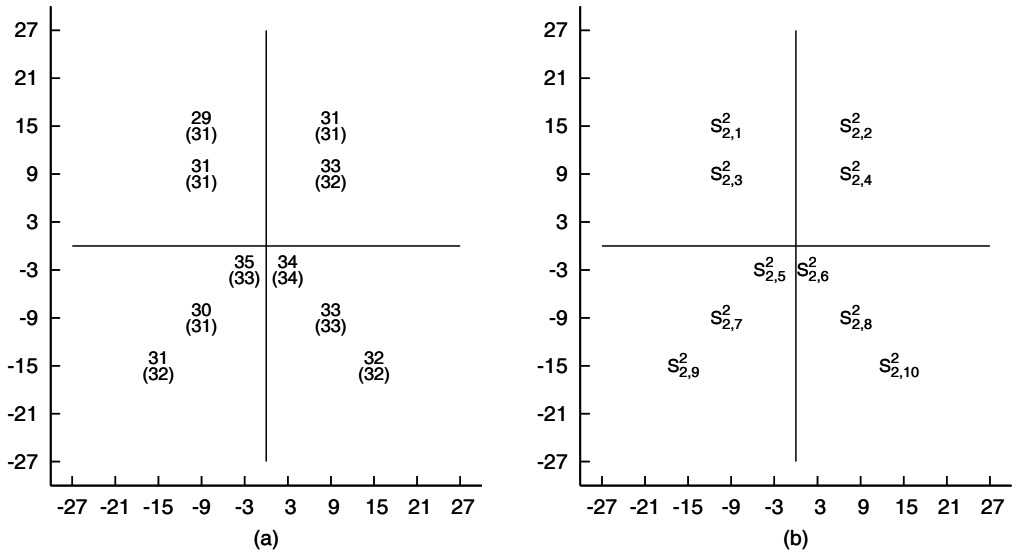


Figure 2.6: A map of SF 10 locations in a visual field. (a) A field in which the threshold is measured twice; (b) the intra-test variances from the normal population at those locations.

The Short-term Fluctuation (SF) Index

The SF is an index of intra-test variability. The short-term fluctuation index is the weighted mean of the standard deviations at a preselected sample of locations, typically ten locations where the threshold is measured twice during a single test (Heijl et al., 1986). Figure 2.6 (a) shows ten preselected locations with twice measured values at each location. Figure 2.6 (b) shows variances of the normal population at those locations. The formula for SF is as follows:

$$SF = \sqrt{\left(\frac{1}{10}\sum_{i=1}^{10} S_{2i}^2\right)\left(\frac{1}{10}\sum_{i=1}^{10}\frac{(X_{i1}-X_{i2})^2}{2\times S_{2i}^2}\right)} \tag{2.4}$$

where X_{i1} is the first and X_{i2} is the second threshold value at location i. The normal intra-test variance at location i is denoted by S_{2i}^2.

A high SF value indicates that the patient gave inconsistent answers to the flashing light, perhaps because of fatigue, inattention or disease. Thus, the SF index is in part

an estimate of measurement error. It is a useful clinical gauge of patient reliability, or the presence of an abnormality, or both (Anderson and Patella, 1999).

As a global index, SF is used to correct the PSD index to produce the CPSD index, as explained below.

Corrected Pattern Standard Deviation (CPSD)

CPSD index is an adjustment of the PSD index which removes the intra-test variability, as represented by the SF index. The formula is:

$$(CPSD)^2 = (PSD)^2 - k(SF)^2$$

where k is 1.28 for the 30-degree field and 1.14 for the 24-degree field (Anderson and Patella, 1999). If $k(SF)^2$ is larger than $(PSD)^2$, the $CPSD$ is assigned a value of 0 dB, since $(CPSD)^2$ is never negative, although $(PSD)^2 - k(SF)^2$ can be negative.

2.2.2 Visual Field Scoring Systems and Cross-Meridional Algorithms

Visual field scoring systems are used in clinical trails of glaucoma diagnosis and treatment as longitudinal measures of visual field progression (Katz et al., 1991; The Advanced Glaucoma Intervention Study Investigators, 1994; Musch et al., 1999). Methods for identifying visual field loss relative to two baseline visual fields have been incorporated into the following large clinical trials: the Advanced Glaucoma Intervention Study (AGIS), the Collaborative Initial Glaucoma Treatment Study (CIGTS), as well as two cross-meridional algorithms - the Glaucoma Hemifield Test (GHT) and the Glaucoma Screening Study (GLASS).

The AGIS and CIGTS assign scores to locations based on their threshold values and those of their immediate neighbours, and then sums the scores into a single value for whole visual field. These scoring systems are comparably scaled. Both AGIS and CIGTS use a 24-2 (central 24 degree) threshold test that includes 54 locations (see Figure 2.2). The outer locations of the 30-2 are not used except for two locations (Katz et al., 1999).

For the AGIS scoring system, the overall visual field score is based on the *Total Deviation (TD)* plot, while the CIGTS scoring system is based on the *Total Deviation (TD)* probability map described in Section 2.1.

It is essential to note that: (1) the two scoring systems are based on very different approaches to assessment of severity of visual field loss; and (2) the criteria for progression differ for each system.

The Advanced Glaucoma Intervention Study (AGIS)

The Advanced Glaucoma Intervention Study (AGIS) score system was designed, as its name suggests, for use with advanced defects (The Advanced Glaucoma Intervention Study Investigators, 1994; The AGIS Investigators, 2000). The AGIS visual field defect score is based on both the extent and depth of clusters of adjacent depressed test locations relative to age-matched normal reference data. The algorithm developed by the AGIS investigators was based on the following concepts: (1) multiple defects can occur in the upper, lower, and nasal hemifields (that is, the field is divided into three areas, as in Figure 2.7); (2) a defect requires 2 or more adjacent abnormal points; (3) the severity of depression must be greater than changes due to variability (that is, depression is identified by using the total deviation plot); and (4) the defect must be caused by glaucoma (Wilson, 2002). The score is increased if either:

1. the numbers of defects increase; or
2. the depth of the defects increases.

The score ranges from 0 (no defect) to 20 (all locations deeply depressed).

As discussed above, the AGIS algorithm is based on the total deviation plot and the region of the field affected (The Advanced Glaucoma Intervention Study Investigators, 1994). The extent of depression considered abnormal varies from -5 to -9 dB depending on the location in the field (see Figure 2.8). The deficit in the periphery is greater than in the center and is larger in the upper than in the lower field.

When depressed locations form clusters of three or more in either hemifield or the nasal region, they are considered as defects and begin to increase the defect score. A

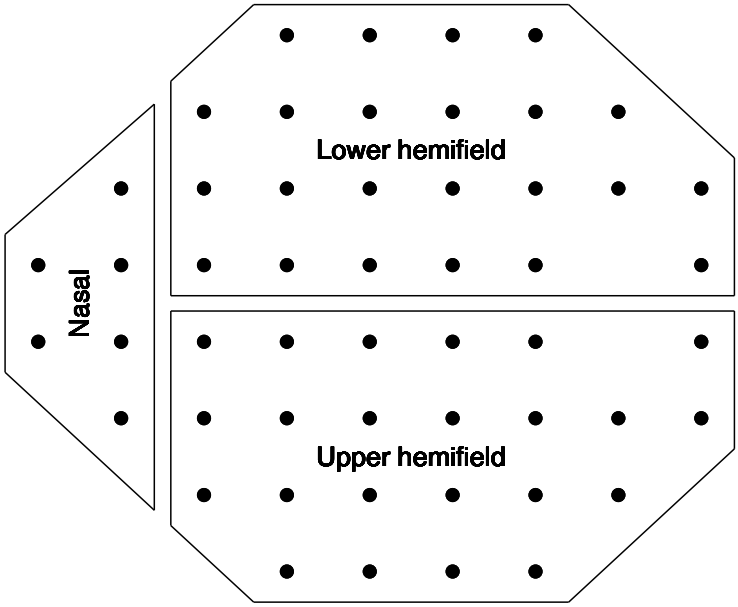


Figure 2.7: Three areas in this AGIS system: nasal region; upper and lower hemifields.

nasal step is defined as one or more contiguous test locations with depressed threshold values in the upper (or lower) nasal area in the absence of any depression in the opposite nasal area (Katz et al., 1999). The scoring procedure is as follows. A score of 2 is assigned to the nasal area if a nasal defect is present, and more than half the nasal test locations have defect depths of 12 dB or greater. If a nasal defect with less depth is present, or a nasal step is present, the score assigned is 1. If no defect is present, then the score is 0 (Wilson, 2002). The scoring rules are summarised in Table 2.1.

Progressive field loss is considered to have occurred if the score has increased by four or more from the baseline reference field (the average of the first two visual fields taken in a short interval), and this change is confirmed by two additional tests (Katz et al., 1999; Vesti et al., 2003).

The Collaborative Initial Glaucoma Treatment Study (CIGTS)

The Collaborative Initial Glaucoma Treatment Study (CIGTS) scoring system is more sensitive to initial glaucoma as its name suggests (Vesti et al., 2003). This system also

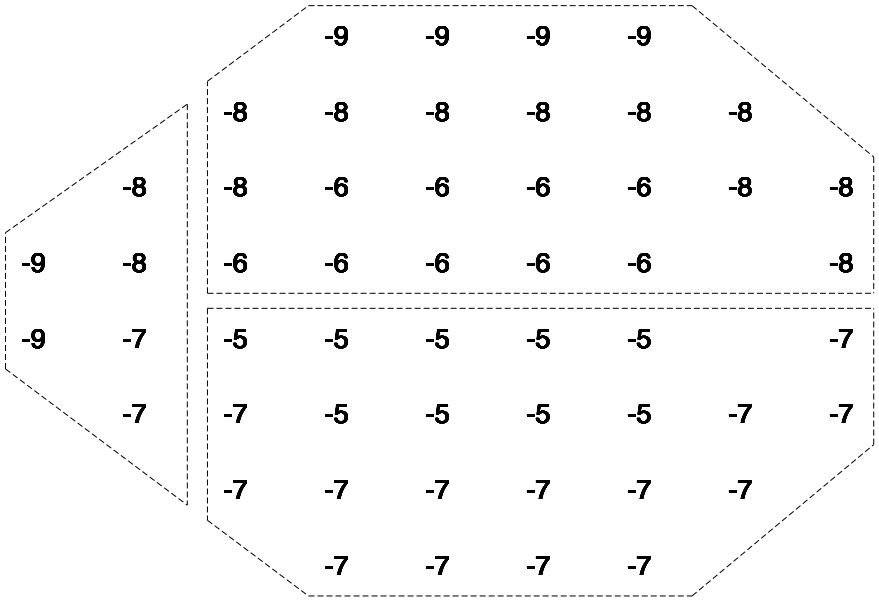


Figure 2.8: The extent of depression (in decibels) in three areas of the 24-2 field in the Advanced Glaucoma Intervention Study for a point to be classified as "abnormal".

uses the total deviation plot and 52 locations (excluding the two blind spot locations), but the overall visual field score is generated from the total deviation probability plot. Neighbour locations are defined as those adjacent to the given location on a side or corner. A depressed location is one with a probability of 5% or less in the same hemifield. If a depressed location has two neighboring depressed locations, then a score between 1 and 4 is assigned to that location. If a depressed location and its two neighboring locations all have $p \leq 5\%$, then this location is assigned a score of 1. If a depressed location and its two neighbouring locations all have $p \leq 2\%$, then this location is assigned a score of 2, (or $p \leq 1\%$, score of 3, or $p \leq 0.5\%$, score of 4). Locations without two neighbouring depressed locations are given a score of 0. For example, if a location is at $p \leq 1\%$ with two neighbouring locations all at $p \leq 5\%$, then this location receives a score of 1. If more than two neighbouring locations are depressed, the score is based on the most depressed neighbours. For example, if a location is at $p \leq 0.5\%$ with two neighbouring locations at $p \leq 1\%$ and one neighbouring location at $p \leq 2\%$, then this location recevies a score of 3. Figure

Table 2.1: The AGIS scoring procedure for visual field defects.

Criterion		Score
In the nasal area		Max = 2
$\geqslant$ 1 depressed location in nasal area and only in 1 hemifield		
or 3 clustered depressed locations		1
4 to 6 clustered locations depressed $\leqslant$ -12 dB		2
In each hemifield		Max = 9
$\geqslant$ 1 cluster of 2 locations with 1 depressed by $\leqslant$ -12 dB		1
$\geqslant$ 1 cluster of 3 to 5 depressed locations		1
$\geqslant$ 1 cluster of 6 to 12 depressed locations		2
$\geqslant$ 1 cluster of 13 to 20 depressed locations		3
$\geqslant$ 1 cluster of $>$ 20 depressed locations		4
If 50% of depressed hemifield locations are depressed by		
	$\leqslant$ -12 dB	+1
	$\leqslant$ -16 dB	+2
	$\leqslant$ -20 dB	+3
	$\leqslant$ -24 dB	+4
	$\leqslant$ -28 dB	+5
Sum the score for the nasal area and each hemifield		Max = 20

2.9 shows an example of a CIGTS visual field in which some locations are assigned scores.

Each of the 52 locations in the visual field is given a score ranging from 0 to 4. The scores for all 52 locations are summed. The minimum of summed scores is 0, and the maximum is 208 ($= 52 \times 4$). This sum is divided by 10.4 so that the overall visual field score is scaled to a range from 0 (no depressed locations) to 20 (all locations depressed at $p \leq 0.5\%$) (Musch et al., 1999; Gillespie et al., 2003). The scoring procedure is summarised in Table 2.2.

Progressive field loss is considered to have occurred if the score has increased by three or more from the baseline reference field, and this change is confirmed by two additional tests (Katz et al., 1999; Vesti et al., 2003).

The advantages of the AGIS and CIGTS scoring systems are that "*test results*

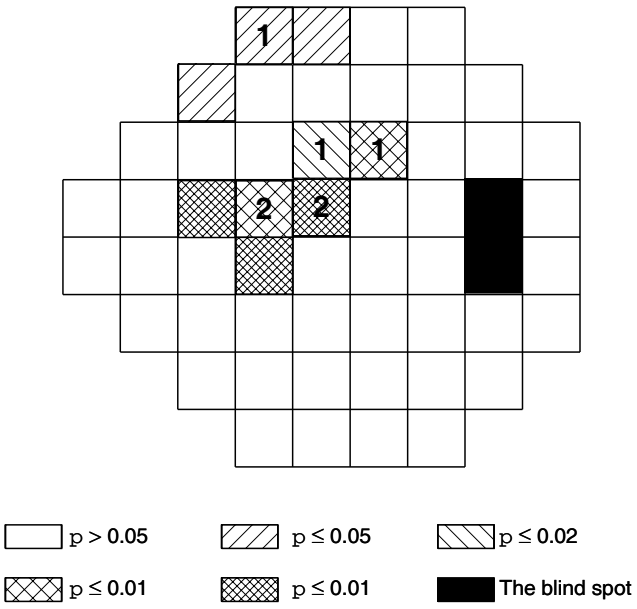


Figure 2.9: Example of a CIGTS visual field score calculation.

are immediately stratified into broadly similar defect magnitudes, interpretation is relatively simple, and progression can be easily defined as worsening of the score over time" (Spry et al., 2002, p. 164).

However, a disadvantage of both scoring systems is that they do not provide spatial information about visual field defects. For example, if a score of 10 is obtained for the visual field, neither scoring system indicates which regions in the field have progression. Another disadvantage is that the score is not scaled linearly. For example, a change from 0 to 3 may not be equal to a change from 10 to 13 (Spry et al., 2002).

Comparisons of AGIS and CIGTS scoring systems using common longitudinal data has shown that the AGIS criteria for detection of visual field progression are more conservative than the CIGTS criteria. The CIGTS system identifies progression at twice the rate as the AGIS system, and detects confirmed progression earlier when both are applied to the same longitudinal data set (Nouri-Mahdavi et al., 1997; Katz, 1999; Spry et al., 2002; Vesti et al., 2003).

Table 2.2: The CIGTS scoring procedure for visual field defects.

1. Score each of the 52 test locations		
Probability value	Neighboring locations	Score
$p > 5\%$		0
$p \leq 5\%$	0 or 1 at $p \leq 5\%$	0
$p \leq 5\%$	2 to 8 at $p \leq 5\%$	1
$p \leq 2\%$	2 to 8 at $p \leq 2\%$	2
$p \leq 1\%$	2 to 8 at $p \leq 1\%$	3
$p \leq 0.5\%$	2 to 8 at $p \leq 0.5\%$	4
2. Summing all scores from 52 locations		Max = 208
3. Divided by 10.4		Max = 20

Cross-meridional Algorithms

For well-known anatomic reasons, glaucomatous field loss most commonly is not symmetric across the horizontal meridian (Sommer et al., 1987; Boden et al., 2002). Two cross-meridional algorithms are the Glaucoma Hemifield Test (GHT) and the Glaucoma Screening Study (GLASS).

Glaucoma Hemifield Test

The Glaucoma Hemifield Test analysis groups points in a visual field along the paths of nerve fiber bundles (see Figure 2.10) (Asman and Heijl, 1992b; Asman et al., 1992) and thus it is more finely tuned to detecting a path of early glaucomatous visual field loss when compared to the global indices (Anderson and Patella, 1999). The GHT is provided for 30-2 (76 locations) and 24-2 (52 locations) tests. In the GHT, five clusters are in the upper field and five clusters are in the lower field. The clusters are constructed in the approximate patterns of retinal nerve fiber (see Figure 2.11).

Each of locations in clusters is assigned a score according to the pattern deviation probability map. If a location is not significant, then this location has a score of 0. If a location is at $p < 0.05$, then this location receives a score of 2 (or $p < 0.02$, score of 5, or $p < 0.01$, score of 10). The sum of probability scores is calculated in each cluster in both hemifields. For each of the five corresponding mirror image pairs of clusters, the up-down difference between these sums are determined (Asman et al.,

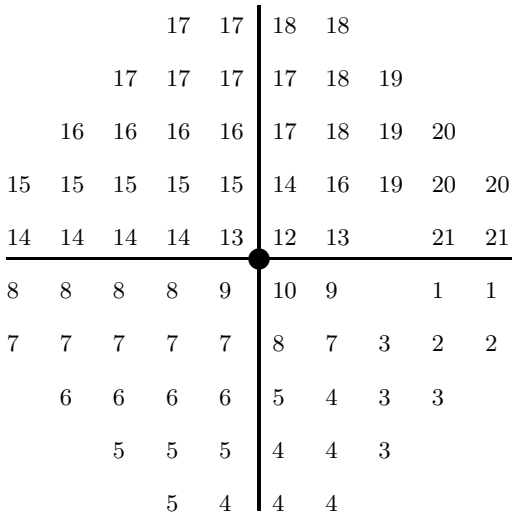


Figure 2.10: A map identifying the nerve fiber bundles for each visual field location.

1992; Asman and Heijl, 1992a).

The scoring algorithms (GHT) compared the two hemifields based on the five pairs of clusters. A visual field is considered as abnormal if the any mirror image difference fall outside the 0.5% limits found in normal visual fields. There are other variants of this scoring method such as Borderline, General reduction of sensitivity and abnormally high sensitivity, but we do not detail them in this thesis.

The Glaucoma Screening Study Cross meridional and cluster algorithm GLASS (Duggan et al., 1985; Sommer et al., 1987) defines visual field loss from 30-2 test. Four pairs of crossing the horizontal meridian clusters {(1)(5)}, {(2)(6)}, {(3)(7)}, and {(4)(8)} in a 30-2 visual field are shown in Figure 2.12.

The algorithm is based on the threshold measurement values. The threshold values in four pairs of clusters are first summed. Then the absolute difference of sums in each pair of clusters is calculated. The visual field is considered to be abnormal if the absolute difference at any pair of clusters exceeds a specified cutoff or sum of threshold values in all clusters in the superior (or inferior) hemifield is lower than a specified cutoff which are listed in Table 2.3.

Two cluster and cross meridional algorithms are equally effective at separating patients with glaucoma and normal subjects, compared with Global indices (Katz

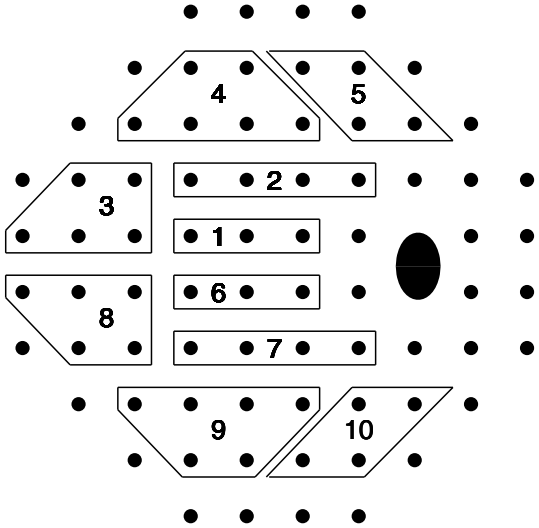


Figure 2.11: Glaucoma Hemifield Test clusters (or zones) based on the nerve fiber bundle map in Figure 2.10.

et al., 1991). However, the choice of different zones may have been a more likely explanation for the lack of concordance between the two methods GHT and GLASS (Katz et al., 1991).

2.2.3 Linear regression analyses

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. Applications of regression are numerous and occur in almost every discipline, including detection of progression of glaucomatous visual fields. This section first briefly describes linear regression and parameter estimation, and testing of goodness of fit. The application of linear regression in glaucomatous data analysis is then discussed.

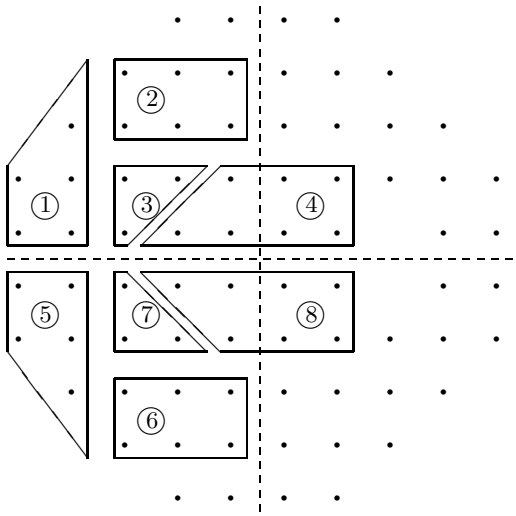


Figure 2.12: Four pairs of crossing the horizontal meridian clusters {(1)(5)}, {(2)(6)}, {(3)(7)}, and {(4)(8)} in a 30-2 visual field.

Table 2.3: The GLASS criteria for classifying visual field abnormal.

Absolute differences of four pairs of clusters	Cutoff
$\|(1)-(5)\|$	30
$\|(2)-(6)\|$	30
$\|(3)-(7)\|$	15
$\|(4)-(8)\|$	35
Sum of thresholds in clusters(1), (2), (3) and (4)	630
Sum of thresholds in clusters(5), (6), (7) and (8)	630

Linear regression and parameter estimation

Regression analysis on glaucomatous data mostly employ a simple linear regression model. That is, a model with a single variable (or regressor) x:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2.5}$$

where β_0 and β_1 are the intercept and the slope respectively, both of which are unknown constants, and where ϵ is a random error component.

In linear regression analysis, the random errors are assumed to:

1. have a mean of zero and unknown equal variance σ^2 of normal distribution;

2. be uncorrelated, such that the value of one error does not depend on the value of any other error (Montgomery and Peck, 1992).

Figure 2.13 presents a graph of a linear regression model showing observed value for each point x.

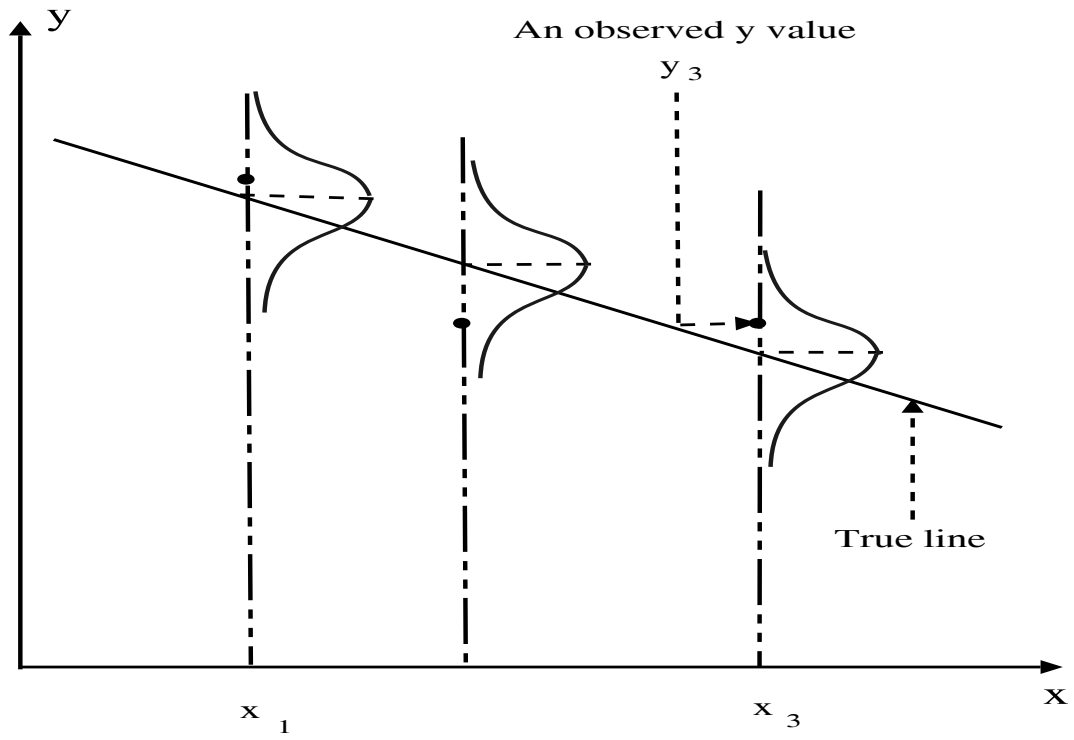


Figure 2.13: A graph of a linear regression model and observed value for each point x. The response (or random) variable y for variable x is assumed to be normally distributed with a mean value of $\beta_0 + \beta_1 x$, and the same variance σ^2.

The mean value of the random error is denoted by $E(\epsilon)$ and the variance is denoted by $V(\epsilon)$; hence $E(\epsilon) = 0$, and $V(\epsilon) = \sigma^2$. The variable x is controlled by the data analyst and measured with negligible error. The response y is a random variable due to random variable ϵ from Equation 2.5. That is, there is a probability distribution for y at each possible value x (see each curve at each point x in Figure 2.13).

The parameters β_0 and β_1 are estimated using n pairs of sample data: (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) as

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{2.6}$$

and

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \tag{2.7}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $S_{xy} = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Therefore, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimators of the intercept β_0 and slope β_1. The fitted simple linear regression model is then

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2.8}$$

When the data represents a random sample, the least squares line by Equation 2.8 is the line of "best" fit because the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are the unbiased estimates of the parameters β_0 and β_1 (Draper, 1981; Montgomery and Peck, 1992).

Hypothesis testing on the slope β_1

Given n pairs of data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, and an underlying linear model $y = \beta_0 + \beta_1 x + \epsilon$, then a linear regression function

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

can be calculated by using equations 2.6 and 2.7. The problem is whether the variable x and the random variable y are correlated. If so, the slope β_1 in Equation 2.5 must not be zero (otherwise, y is said to be independent of x). Therefore, given a significance level α (say $\alpha = 0.01$ or 0.05), it is important to test a hypothesis whether the slope $\beta_1 = 0$.

The appropriate hypotheses are

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

If the null hypothesis H_0: $\beta_1 = 0$ is true, the statistic

$$t_0 = \frac{\hat{\beta}_1}{Q}\sqrt{(n-2)S_{xx}} \tag{2.9}$$

where $Q = \sqrt{S_{yy} - \hat{b}S_{xy}}$ and $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$, can be proved to be a t-distribution with $n-2$ degrees of freedom. The statistic t_0 is used to test H_0: $\beta_1 = 0$ by comparing the observed value of t_0 from Equation 2.9 and a critical value ($t_{\alpha/2,n-2}$) obtained from a t-distribution with $n-2$ degrees of freedom. If $|t_0| > t_{\alpha/2,n-2}$, then the null hypothesis is rejected (which means that the slope β_1 is not zero), so the linear relationship between variable x and variable y is statistically significant. On the other hand, if accepting the null hypothesis, the linear relationship of x and y is not statistically significant. This may imply either that the change of y does not depend on x, that is $\hat{y} = \bar{y}$, or that the true relationship between x and y is not linear (see Figure 2.14).

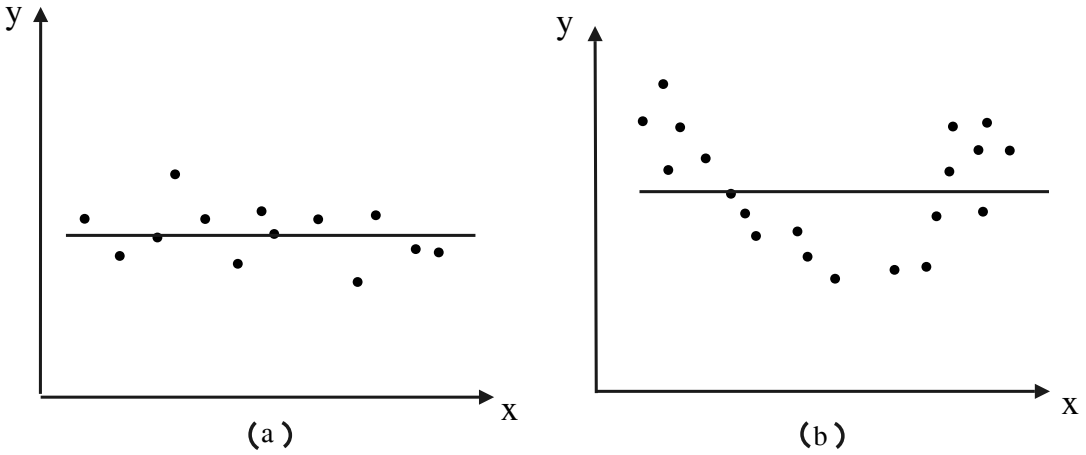


Figure 2.14: Examples in which the hypothesis $H_0 : \beta_1 = 0$ is not rejected.

Application of linear regression analyses to glaucomatous data

Linear regression analysis on Mean Deviation and individual locations is currently in clinical use (Heijl et al., 1986; Fitzke et al., 1996). Many researchers have investigated classification of visual fields using linear regression which we will discuss shortly. In addition, some researchers also used linear regression to analyse spatial and temporal processing of threshold data for reducing fluctuation in classification (Spry et al., 2002).

(1) **Linear regression based on Global Indices**

Some researchers applied linear regression on whole visual fields to analyse the change. Werner et al. (1988) used the mean value of each visual field against time to fit the best straight line. If the slope was negative at the $\alpha < 0.05$ level, the visual field was considered to have progressed.

Some researchers used the values of Global Indices. *Change Analysis* is a statistical function for the analysis of visual field changes, that is a part of Statpac package described in detail by Heijl et al. (1986) (or Statpac 2), and performs a linear regression of the Mean Deviation values. For sequential tests, the mean deviation of each test is calculated. The values of MDs against follow-up time are used to calculate the slope. If five or more tests are available, the slope of a model fitted to Mean Deviation values can be automatically calculated by Equation 2.7. If the slope is negative with a given significance level α, then visual field loss is identified (Heijl et al., 1986; Smith et al., 1996; Nouri-Mahdavi et al., 1997; Heijl et al., 2003).

Katz et al. (1997) and Smith et al. (1996) not only used linear regression analysis of Mean Deviation, but also used linear regression analysis of Corrected Pattern Standard Deviation (CPSD). The slope estimating the linear change in CPSD over time was considered non-zero if the significance level $\alpha \leq 0.025$.

Both MD and CPSD are calculated from a weighted sums involving age-matched normal reference fields, linear regression analysis of these global indices is calculated with reference to age-matched normative data. Thus, any negative slope may be considered as indicating progressive visual field loss not physiological age-related vision loss. However, the disadvantage of these approaches is that MD and CPSD quantify overall visual field loss; therefore, any subtle localised visual field loss cannot be identified. Many researchers have reported that linear regression on global indices is not sensitive to glaucomatous visual field loss (Chauhan et al., 1990; Smith et al., 1996; Katz et al., 1997).

(2) **Linear regression based on Glaucoma Hemifield Test zones**

In the Glaucoma Hemifield Test (GHT), five clusters (or zones) are in the upper field and five clusters are in the lower field (see Figure 2.11). Katz et al. (1997) and Smith et al. (1996) used linear regression analysis based on GHT zones for detecting

visual field loss. For each zone of each visual field, the simple mean of thresholds is calculated. When a series of visual fields for one patient is obtained, a series of means for each zone can be calculated. Thus, the series of means for one zone can be plotted against follow-up time. Using Equations 2.6 and 2.7, the fitted straight line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is obtained. In the evaluation of the slope for significance, the criteria used by Katz et al. (1997) and Smith et al. (1996) is that, for any one zone, it is considered significant if $\alpha < 0.005$. Werner et al. (1988) used $\alpha < 0.05$ to analyse the change.

As Boden et al. (2002) pointed out, early glaucomatous visual field loss rarely crosses the horizontal meridian. Therefore, a linear regression analysis based on each zone independently is more sensitive than that based on MD or CPSD (Katz et al., 1997; Nouri-Mahdavi et al., 1997; Smith et al., 1996).

(3) **Linear regression based on individual test locations**

A visual field change depends on threshold change at each location. Linear regression analysis of point-wise threshold data is commercially available in the *Progressor* software package (Moorfields Eye Hospital, London, UK / Medisoft Ltd., Leeds, UK). Progressor calculates the relationship between threshold values against follow-up times for each test location when visual field sequences include both 30-2 and 24-2 results. The program applies linear regression analysis to individual test locations to determine whether there is statistically significant progression.

A number of researchers have investigated the application of linear regression based on threshold values for each test location. Noureddin et al. (1991) required the test location to have significant ($\alpha < 0.05$) slopes < -2.4 dB per year to be considered progressive. A visual field was considered to be progressive if at least the two consecutive slopes of one or more tested locations were less than -2.4 dB ($\alpha < 0.05$).

Some researchers employed the criterion of slope < -1 dB per year with a significance level $\alpha < 0.05$ (Birch et al., 1995; Viswanathan et al., 1997; Nouri-Mahdavi et al., 1997; Spry et al., 2000; Artes et al., 2005) while many other researchers used the criterion of slope < -1 dB per year with a significant level $\alpha < 0.01$ (Wild et al., 1997; Gardiner and Crabb, 2002b,a; Vesti et al., 2002; Nouri-Mahdavi et al., 2004,

2005). Katz et al. (1997) defined progressive loss as any significant negative slope with $\alpha < 0.001$. Wikins et al. (2006) have used a fixed critical slope of -1 dB/year and with a range of significance level α from 0.001 to 0.05.

Manassakorn et al. (2006) have investigated the application of linear regression based on total deviation and pattern deviation values for each test location. They used the criterion of slope < -1 dB per year with a significant level $\alpha < 0.01$ for detection of visual field progression.

Definition of progression for a location also depends on number of consecutive tests. For example, Hitchings et al. (1994) defined a sequence of measurements $x_1, x_2, \ldots, x_{n-1}, x_n$ for a location as progressive if both $x_1, x_2, \ldots, x_{n-1}$ and $x_1, x_2, \ldots, x_{n-1}, x_n$ satisfied the criterion of slope < -1 dB per year with a significant level $\alpha < 0.01$. Membrey et al. (2000) defined progression for a location as the presence of a significant regression slope ($\alpha < 0.01$) showing slope < -1 dB per year at the same location with addition of two out of three successive tests. Gardiner and Crabb (2002a) defined a sequence of measurements as progressive if it satisfied the criterion of slope < -1 dB per year with a significant level $\alpha < 0.01$, continues to satisfy this criterion when the last value in the sequence is deleted and a new measurement is added into the sequence.

In short, it appears that there is no consensus as to the point-wise analysis outcome that is best for determining progressive visual field loss. Table 2.4 summarises point-wise linear regression analyses used for detection of visual field progression. Values in the first column are analysed by using point-wise linear regression. The criteria of classifying visual field progression are listed in the second and third columns.

The advantage of point-wise linear regression analysis is that important spatial relationships between locations are maintained. The disadvantage is that visual field loss is detected later than by other methods such as the AGIS and CIGTS scoring systems and the Glaucoma Change Probability analysis described in the next section (Vesti et al., 2003). In practice, it is suggested that a minimum of 7 or 8 visual fields (which is equivalent of $3.5 \sim 4$ years minimum) are required to achieve reasonable levels of *sensitivity* and *specificity* (Katz et al., 1997; Spry et al., 2000). Sensitivity is the proportion (or percentage) of glaucoma patients correctly classified as having glaucoma.

Table 2.4: Summary of point-wise linear regression analyses for detection of visual field progression. TD is Total Deviation, and PD is Pattern Deviation.

values (y_i)	Slope	α	Authors
Threshold values	-2.4	0.05	Noureddin et al. (1991)
	-1	0.05	Birch et al. (1995); Viswanathan et al. (1997); Nouri-Mahdavi et al. (1997); Spry et al. (2000); Artes et al. (2005)
	-1	0.01	Wild et al. (1997); Gardiner and Crabb (2002b); Vesti et al. (2002); Gardiner and Crabb (2002a); Nouri-Mahdavi et al. (2004, 2005)
	-1	[0.001, 0.05]	Wikins et al. (2006)
	0	0.001	Katz et al. (1997)
	negative	0.01	Fitzke et al. (1996)
TD and PD values	-1	0.01	Manassakorn et al. (2006)

Specificity is the proportion (or percentage) of stable individuals correctly classified as normal (Anderson and Patella, 1999; Goldbaum et al., 2002).

2.2.4 Event Analysis

Event analysis is valuable because it attempts to identify single events of significant change relative to a reference examination. Event analysis looks for statistically significant differences between one examination and another (Spry et al., 2002). The Glaucoma Change Probability (GCP) method is one type of event analysis. However, GCP is also a name for the proprietary piece of software that is a part of the Zeiss Statpac (Carl Zeiss, Meditech, Dublin CA). To avoid confusion with the GCP software, we refer to the GCP method and algorithm as Event Analysis (EA).

Event Analysis determines change on a point by point basis (Heijl et al., 1989; Katz, 2000; Boden et al., 2004). EA calculates the difference between a threshold and a *baseline* which is defined as $(X_1 + X_2)$ / 2 where X_1 and X_2 are measured in a short time period for each location. EA chooses a pre-determined *confidence interval* (CI) which is established from a database of *test-retest* stable glaucomatous visual fields (Heijl et al., 1989; Morgan et al., 1991; Anderson and Patella, 1999). Test-retest means that several tests for one patient are taken in a short period, usually once per

week. EA then determines whether the difference falls inside or outside a 95% or 99% confidence interval (Anderson and Patella, 1999; Katz, 2000; Spry et al., 2002). The algorithm for classifying visual field progression using EA is as follows.

Algorithm 1. *The EA algorithm.*

For each location i in a series of visual fields $X_1, \ldots, X_n$

Calculate the baseline value $b_i = (x_{1i} + x_{2i})/2$

Calculate the difference between the baseline and current value $\delta = x_i - b_i$

Calculate the average of the first two total deviation (TD) or pattern deviation (PD) value to select a CI

if $\delta <$ the lower limit of CI,

the location i is progressive;

else it is stable

Some investigators selected confidence intervals according to the baseline defect which equals $(x_{1i} + x_{2i})$ / 2 - n_i where x_{1i} and x_{2i} are measured in short time period and n_i is an age-matched normal value for the location (Heijl et al., 1989; Morgan et al., 1991; Anderson and Patella, 1999). The baseline defect = $(x_{1i} + x_{2i})$ / 2 - n_i = $((x_{1i} - n_i) + (x_{2i} - n_i))/$ 2, hence the baseline defect is the average of the first two Total Deviation values. Total Deviation values are therefore used as indices to select confidence intervals. Some investigators used *Pattern Deviation* values as indices to select confidence intervals (Katz, 2000; Boden et al., 2004).

The procedure of building a confidence interval

Heijl et al. (1989) used an algorithm for building a confidence interval, where given a sequence of five test-retest values in a stable database at a single location $y_1, y_2, \ldots, y_5$, the difference is calculated between each value y_i and the corresponding age-matched normal threshold value n. The quantity $y_i - n$ is the *deviation*. The average of the first two deviations in the sequence is taken as the *defect* of the sequence. All sequences are classified according to defect and eccentricity. Eccentricity can be one of: "inner", the 30 central points of the field excluding the two blind spots

$(15°, \pm 3°)$; "middle", the 20 points directly bordering the inner field; or "outer", the 24 points on the edge of the 30-2 pattern as shown in Figure 2.15.

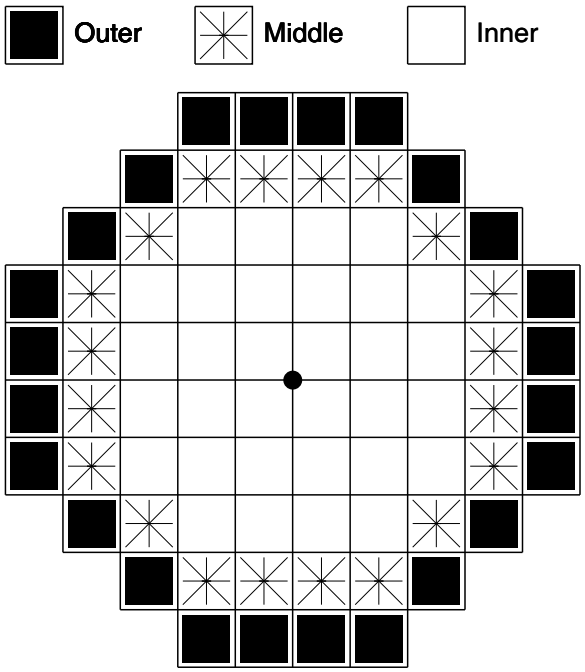


Figure 2.15: Inner, middle and outer rings of the visual field.

For each class of sequences, the test-retest differences in each sequence, $y_i - y_{i+1}$, $1 \leq i \leq 4$ are calculated and sorted from the smallest to the largest. Once classification and differencing have been performed, the 2.5th and 97.5th percentiles are computed for each group to form a 95% confidence interval about the defect for that group.

The advantage of EA is that it may identify test locations that appear progressive early (Spry et al., 2002). Vesti et al. (2003) showed that EA identifies visual field loss earlier than linear regression (at least 1 year early for underlying all variabilities).

However, there is a disadvantage with this approach as the baseline should be chosen carefully (Anderson and Patella, 1999). The baseline visual field measurement(s) establishes the condition at the beginning of the follow-up period. Sometimes, the first measurement may not be the most accurate due to learning effect (Anderson and Patella, 1999; Amjad et al., 2002; Fujimoto et al., 2002), as described in Section 2.1.5.

Similar to the application of linear regression on glaucomatous data, there is no consensus on the significant change at the level α <0.05 or 0.01 (for 95% or 99% confidence interval) and how many locations which are confirmed for determining progressive visual field loss by using the EA analysis. For example, Chauhan et al. (1999) used four or more locations at 5% level occurring in two of three consecutive fields. Vesti et al. (2003) used three consecutive fields.

The Early Manifest Glaucoma Trial (EMGT)

The Early Manifest Glaucoma Trial (EMGT) is a clinical trial to assess the effectiveness of reducing intraocular pressure in early, previously untreated open-angle glaucoma (Heijl et al., 2003). The Glaucoma Change Probability map (EA) based on the pattern deviation is used for the EMGT to define the progression. Locations that have significantly changed at the level $\alpha < 0.05$ are flagged. Definition of visual field progression is that three or more locations are flagged and confirmed in two consecutive tests, the locations need not to be contiguous (Heijl et al., 2003).

2.2.5 Machine Learning Techniques

Machine learning is generally taken to encompass automatic computing procedures that learn a task from a series of examples (Fisher, 1987). Recent methods for detection of glaucomatous visual field defects have concentrated on the application of machine learning classifiers such as neural networks (Henson et al., 1996, 1997; Goldbaum et al., 2002; Lin et al., 2003; Sample et al., 2004), support vectors (Turpin et al., 2001) and decision trees (M. Lazarescu, 2002; Lazarescu and Turpin, 2003). Henson et al. (1997) addressed how the machine learning technique can be used to accommodate the noise within the data. However, most researchers used machine learning classifiers to classify raw visual field data (Goldbaum et al., 1994; Brigatti et al., 1996; Sample et al., 2005) and compared their performance with Global Indices (Chan et al., 2002; Goldbaum et al., 2002). These experiments have successfully discriminated normal and abnormal visual fields. In this section, we briefly describe Bayesian classification and decision tree techniques, and then review work in applying these techniques to classification of glaucomatous visual fields.

Bayesian Classification

Bayesian classification is based on Bayes theorem. Let X be a data sample whose class is unknown. Let H be some hypothesis, such as that data sample X belongs to a specified class C. Bayes Theorem is as follows:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

The probability $P(X)$, $P(X|H)$ and $P(H)$ may be estimated from the given data. The posterior probability $P(H|X)$ can then be calculated using Bayes Theorem.

The *naive Bayesian classification* is the application of Bayesian theorem. According to Han and Kamber (2000), the naive bayesian classifier works as follows:

- Each data sample is represented by an $n-$dimensional feature vector X. That is $X = (x_1, x_2, \ldots, x_n)$. Each x_i is called an attribute.
- Suppose that there are m classes, $C_1, C_2, \ldots, C_m$. Given an unknown data sample, X (i.e., having no class label), the classifier will predicate that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naive Bayesian classifier assigns an unknown sample X to the class C_i if and only if

  $$P(C_i|X) > P(C_j|X) \qquad for \quad 1 \leqslant j \leqslant m, \quad j \neq i$$

  By Bayesian theorem, each $P(C_k|X)$ is calculated as follows:

  $$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)}$$

Naive Bayesian classifiers assume that each attribute x_i is independent of other attributes. This assumption simplifies the calculation of the posterior probability. *Bayesian networks* are graphical models which, unlike naive Bayesian classifiers, allow the representation of dependencies among subsets of attributes. Bayesian networks can be used for classification. Tucker et al. (2004) used a special Bayesian network that assumes dependencies between variables based on some spatial neighbourhood.

For classification of glaucomatous visual fields, Jansonius (2005) used Bayes' the-

orem to calculate the *positive predictive value* ($= \frac{true\ positive}{true\ positive\ +\ false\ positive}$) of progression, in which, three different prior probability values 0.05, 0.10, and 0.20 are used.

The advantages and disadvantages of Bayesian learning methods for classification are comparable with other learning algorithms such as decision tree and neural network (Mitchell, 1997; Han and Kamber, 2000). Additionally, Bayesian learning methods require initial knowledge of the prior probabilities of classes ($P(H)$), and dependencies in a Bayes Network. Another problem is that given data sets with many attributes, it is computationally expensive to compute $P(X|C_k)$.

Decision trees

Decision tree learning is one of the most widely used machine learning methods. A decision tree can be used to classify instances by sorting them down the tree from the root to a leaf node. The leaf node represents the classification of the instance. Each node in the tree specifies a condition based on some *attribute* of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute.

In decision tree learning, the key issue is how to construct the tree. Mitchell (1997) and Quinlan (1993) used a statistical method to evaluate each attribute and determine how well it alone classifies the training examples. The best attribute is selected and used as the test at the root node of the tree. A descendant of the root node is then created for each possible value of this attribute, and training examples are allocated to the appropriate descendant node. This procedure is repeated using the training examples at each node to create the descendant nodes.

According to Mitchell (1997), decision trees are usually suitable for problems that exhibits the following traits.

- *Instances are represented with attribute-value pairs.* The instances are represented with a fixed set of attributes (e.g., *temperature*) and their values (e.g., *hot, cold*).

- *The target function has discrete output values.*
- *The training data may contain errors, and/or missing attribute values.*

Many real world problems fit these characteristics. For example, decision tree learning has been applied to classify medical patients by disease. Lazarescu and Turpin (2003) used decision tree to classify glaucomatous patients to determine if their visual fields are progressing or stable.

The advantage of decision tree learning for classification is that the trees are easy to understand without significant knowledge of machine learning techniques. This makes them suitable for clinical application. However, one drawback is that accuracy of classification depends on attribute selection.

2.3 Comparison of Classification Methods

Detection of glaucomatous visual field loss and determination of subsequent visual field progression are the most important aspects of glaucoma management (Spry et al., 2002). However, the ability of a specific method to detect visual field progression may be affected by the degree of the initial glaucomatous visual field loss and the variability of intra- and inter-test measurements. In clinical settings, different methods use different criteria for determining visual field progression. Some methods may be sensitive for detecting progression of early losses but less sensitive for detecting progression when the visual field loss is more advanced (Vesti et al., 2003) or vice versa. Many researchers compared the performance of different methods for determining visual field progression using the same longitudinal data (Nouri-Mahdavi et al., 1997; Katz et al., 1999; Lee et al., 2002; Nouri-Mahdavi et al., 2005). Vesti et al. (2003) compared methods currently available in clinics. They used the three criteria of EA, two criteria of point-wise linear regression methods, and the AGIS and the CIGTS scoring schemes on the same simulated data sets (no-, moderate- and high-variability). The criteria of EA and point-wise linear regression methods for determining visual field progression are based on different numbers (4 or 8) of progressive locations and consecutive fields (two of three, or three of three). The results show that the EA methods detect progression earlier than other methods. The EA methods detected

about 60% of progressive visual fields underlying no- and moderate-variability conditions. Although the EA methods can reach about 80% accuracy for detecting visual field progression underlying high-variability, they sacrifice about 70% specificity. The AGIS and CIGTS scoring systems detected less than 40% progressive visual fields. The point-wise linear regression methods offer more accuracy for detecting visual field progression. However, the drawback of these methods is that they required the longest time to detect progression (5.5-6.5 years) (Vesti et al., 2003).

2.4 Simulation

Computer simulation techniques have been used extensively to evaluate the performance of thresholding algorithms (Johnson et al., 1992; Chauhan and House, 1994; Glass et al., 1995; Turpin et al., 2002) in addition to methods for the evaluation of progression (Hitchings et al., 1994; Spenceley and Henson, 1996; Spry et al., 2000; Gardiner and Crabb, 2002a; Vesti et al., 2002, 2003). In order to evaluate progression detection methods, a series of simulated whole visual fields is generated using some underlying model of change over time.

Spry et al. (2000), Spry et al. (2002) and Vesti et al. (2002) used computer programs to simulate sets of serial visual fields between initial and final tests (that is, simulation was based on whole fields). Simulated visual field data could be presented in 30-2 format (76 locations) or 24-2 format. The simulation provided for linear, episodic, and bilinear decay of data values, and also for no progression (i.e. initial and final visual fields the same). Long- and short-term fluctuation as noise were added separately into the simulated data.

Gardiner and Crabb (2002a) used simulation model to simulate 1000 virtual stable eyes and 1000 virtual deteriorating eyes, all with two tests per year over a 6-year follow-up (13 fields in total). For the virtual stable eye, only a normal age-related decline of 0.1 dB was subtracted from a Humphrey 30-2 visual field. For the virtual deteriorating eye, six locations in the virtual stable eye are replaced with starting values 32, 28, 24, 20, 16 and 12 dB. These six locations were given a rate of loss of 2 dB per year. A functional value provided by Henson et al. (2000) was added as noise in simulation, instead of long- and short-term fluctuation.

Gardiner and Crabb (2002b) also simulated different frequency of testing for detecting visual field progression. They evaluated the outcome of frequency of testing, and pointed out that a location is identified as progressive quicker when more tests are carried out each year.

The advantage of simulation based on whole visual fields is to simulate real patient cases. However, for *point-wise analysis* (i.e. the EA methods), methods for determining visual field loss depend on the number of progressive locations and consecutive visual field test. There is no consistent number of locations and consecutive visual field test for determining change. Therefore, simulation based on individual test locations rather than whole visual fields is a valid method for assessing point-wise methods. Therefore, simulation based on individual test locations is adopted in this thesis, and explain in detail in the next chapter.

2.5 Summary

In this chapter, we reviewed Standard Automated Perimetry, including threshold values, total deviation, pattern deviation plots and probability symbol maps. Existing methods of detecting progressive glaucomatous visual field loss including global indices, scoring systems, linear regression analyses, Glaucoma Change Probability, and machine learning classifiers were reviewed. Previous comparisons of methods for speed and accuracy of classification shows that if methods work well with stable fields, they may work poorly with progressive cases or vice versa. Hence it is very difficult to find a technique to have high sensitivity and specificity. These benefits and drawbacks were considered as we developed the new techniques to improve upon these methods presented in later chapters. As a first step towards this goal, simulation of visual fields is described in the next Chapter, including noise, age decline, and types of progression.

Chapter 3

Simulation

3.1 Introduction

This chapter describes the approaches used to simulate sets of threshold sequences for individual test locations. A major problem with glaucoma is that acquiring clinical data is difficult and expensive. This is because (1) the data collection is time consuming, and must be repeated over several years; and (2) the collected data needs to be classified by experts (which also is time consuming and costly). A further problem with clinical data is that there is no agreed "gold standard" (or "ground truth") for diagnosis of glaucoma or glaucomatous progression. An alternative method to obtain large, longitudinal data sets is computer simulation, in which simulated visual fields are known to be progressing or stable by design. The purposes of using computers to simulate visual field data are as follows:

1. *Simulated data can be used to test the efficiency of different methods to detect progression, or to develop new methods with minimal cost.* Analysing the effectiveness of different methods to detect progression requires large, longitudinal sets of visual fields. As noted above, collecting these data is usually difficult, time consuming and expensive;

2. *Change in visual fields can be controlled by computer simulation.* There is no universally accepted definition of glaucomatous progression for evaluating the performance of different analysis methods. A variety of studies have used different methods to define and classify glaucomatous visual field defects, and agreement between methods is low - between 22% and 35% (Vesti et al., 2003). Therefore,

detecting change in visual fields by different methods can give different results. However, simulated visual fields are known to be progressing or stable by design. As long as the simulated data reflects real world data, results on simulated data can be applied to clinical data with confidence. Vesti et al. (2002) validate simulated data against real data. 0, 1, 2, 4, and 6 dB short-term variability and 0, 1, 2, and 4 dB long-term variability were used in the simulation. The results shown that short-term rather long-term fluctuation is the most important factor determining the variability of threshold.

3. *Short- and long-term fluctuation can be controlled by computer simulation.* In some patients, initial measurements can result in unreliable or unusually low threshold values because of learning effects, fatigue and physiological factors. In the clinic, for some unreliable initial measurements, some patients are required to do more tests to establish a baseline value. In contrast, threshold values at the beginning of a simulated sequence are valid by definition, which assists in evaluation of methods which depend on baseline values for classification.

Unlike simulation of visual fields by Spry et al. (2000) or Gardiner and Crabb (2002b) discussed in Chapter 2, we use simulation to generate sequences at individual locations. This is because we wish to evaluate point-wise analysis. The simulation program used in this thesis generates two types of data sets; one is sets of progressive sequences, and the other is a set of stable sequences. The program can simulate noise like Spry et al. (2000) and noise like Gardiner and Crabb (2002b). Spry et al. (2000) used short- and long-term fluctuation as noise to simulate measured values, and Gardiner and Crabb (2002b) used a formula to calculate noise based on a given true value.

The structure of this chapter is as follows: Section 3.2 details the simulation; experimental data is shown in Section 3.3; and the discussion and summary are in Section 3.4 and 3.5 respectively.

3.2 Sequence Simulation

The simulation program permits generation of simulated sequences with chosen levels of fluctuation, type and rate of progression, and frequency of measurement. A schematic overview of the simulation is shown in Figure 3.1.

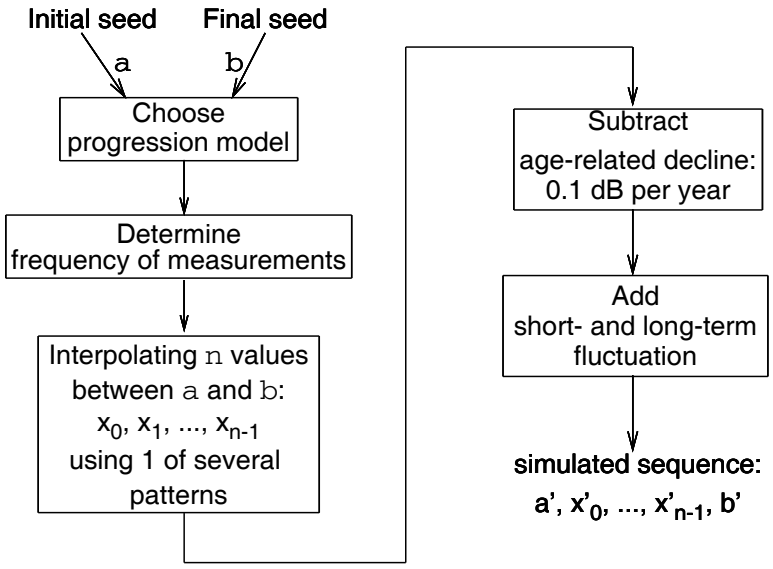


Figure 3.1: Schematic representation of the simulation for generating threshold sequences based on an initial value a and a final value b

3.2.1 Input Data and Interpolated Values

To generate a sequence of threshold values for a single location in the visual field, an initial seed value a between 0 and 39 (since $0 \leqslant \mathit{threshold} \leqslant 39$) can either be chosen randomly, or a location in a real normal or a stable visual field can be used. The final value b can be obtained by subtracting a constant from a for a progressive sequence, or by duplicating a (that is $b = a$) for a stable sequence. Thus, $b \leqslant a$ for both stable and progressive sequences.

Given the initial and final values for a single location in the visual field, the program generates a sequence of values between these seed values. More formally, given an

initial value a and final value b, a sequence of $n+2$ values (including the initial and final seed values) a, x_0, x_1, ..., x_{n-1}, b is obtained by using the program to interpolate the middle n values. The age-related decline is subtracted from a, x_0, x_1, ..., x_{n-1}, b, and then short- and long-term fluctuations are added. Finally, simulated sequence a', x'_0, x'_1, ..., x'_{n-1}, b' is obtained.

For the first interpolated value x_0, if a classification method requires a baseline measurement, x_0 is the same as the initial value a (until noise is added). This is because the average of the first two values in a sequence is taken as the baseline measurement, and in practice baseline measurements are obtained *in a short period* (weeks at most), so no change occurs. If a classification method does not require a baseline measurement, the first interpolated value x_0 is not the same as the initial value a.

3.2.2 Simulated Sequences

Patterns of loss over time as described by Spry et al. (2002) are: linear (49%), curvilinear (20%), episodic (7%), and non-progressing (24%). Therefore, the simulation included linear, bi-linear, curvilinear, and non-progressing patterns. Examples of these simulated patterns are shown in Figure 3.2. Figure 3.2 shows the first interpolated value x_0 is the same as the initial value a before noise is added. The first value in each graph (at visit 0) is the baseline (the average of the first two measurements, that is $\frac{a+x_0}{2}$), x_i is the ith visit measurement $i = 1, 2, ..., 9$, and b is the 10th and final visit.

In the *linear* case, the simulation function for interpolated values x_i is

$$x_i = a + (b-a)/n \times i \qquad i = 0, 1, 2, \ldots, n-1 \tag{3.1}$$

where $x_0 = a$ if a method requires the baseline measurement (e,g, EA method), or

$$x_i = a + (b-a)/(n+1) \times (i+1) \qquad i = 0, 1, 2, \ldots, n-1 \tag{3.2}$$

where $x_0 \neq a$ if a method does not require the baseline measurement (e,g,linear regression).

For example, if $a = 30$, $b = 20$, and the baseline measurement is required, the following sequence can be produced by linearly interpolating 10 values between the

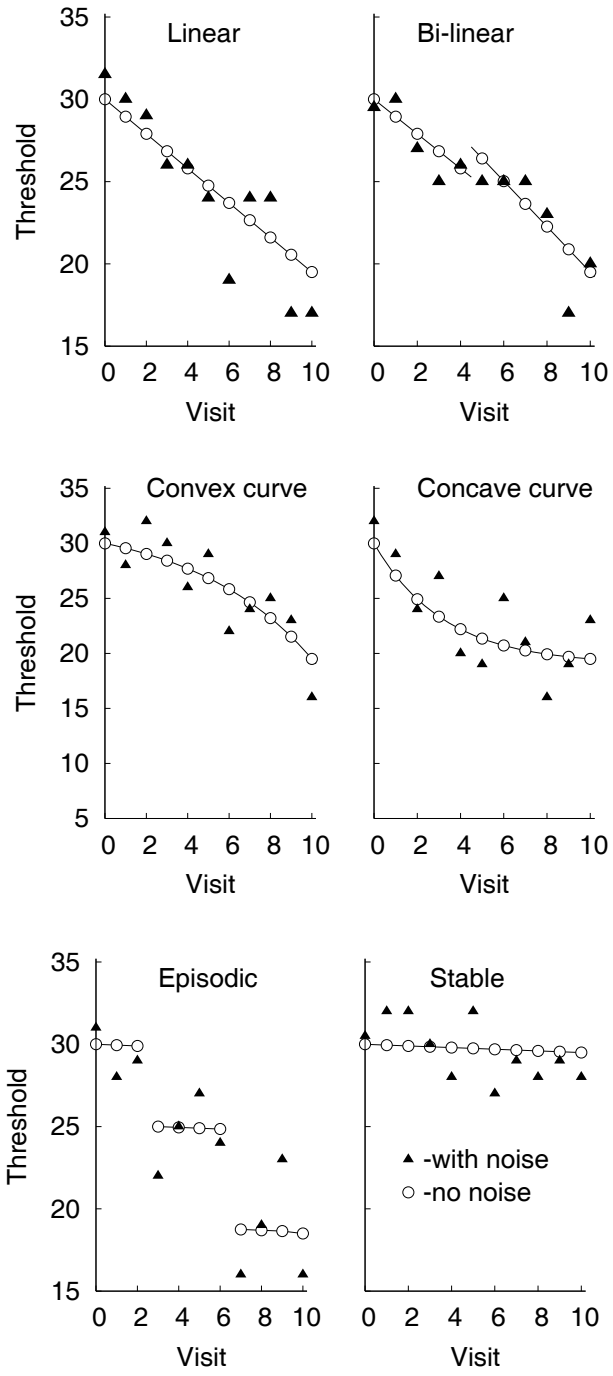


Figure 3.2: Simulated threshold sequences. The first value (visit 0) in each graph is the baseline (the average of the first two measurements). Lines with small circles indicate that only age-related decline is included. Isolated triangles indicate that both age-related decline and noise are included.

seed values: 30, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20. If no baseline measurement is required, the sequence produced is: 30, 29.1, 28.2, 27.3, 26.4, 25.5, 24.5, 23.6, 22.7, 21.8, 20.9, 20.

In the *curvilinear* case, we provide two types of curves: *convex* and *concave*. For the convex curve, if the baseline measurement is required, the interpolated value x_i is given by

$$x_i = A - (A - a)e^{c \times i} \tag{3.3}$$

where $\forall A > a$, and $c = \frac{1}{n} log_e \left(\frac{A - b}{A - a} \right)$. If the baseline measurement is not required,

$$x_i = A - (A - a)e^{c \times (i+1)} \tag{3.4}$$

where $\forall A > a$ ($\forall A > a$ means that we can choose any one value of being greater than the initial value a). $c = \frac{1}{n+1} log_e \left(\frac{A - b}{A - a} \right)$.

For example, if $a = 30$ and $b = 20$, to interpolate five new values ($n = 5$) values and a baseline measurement is required, and choosing $A = 32.01 > a = 30$, then the convex function is:

$$x_i = 32.01 - (32.01 - 30)e^{c \times i}, \text{ where } c = \frac{1}{5} log_e \left(\frac{32.01 - 20}{32.01 - 30} \right)$$

Thus, the sequence produced by the convex function is 30, 30, 29.1, 27.9, 26.1, 23.6, 20.

For the concave curve, if the baseline measurement is required, the interpolated values x_i are given by:

$$x_i = B + (a - B)e^{-c \times i} \tag{3.5}$$

where $\forall B < b$ and $c = \frac{1}{n} log_e \left(\frac{a - B}{b - B} \right)$. If the baseline measurement is not required,

$$x_i = B + (a - B)e^{-c \times (i+1)} \tag{3.6}$$

where $\forall B$ is less than the final b, $c = \frac{1}{n+1} log_e \left(\frac{a - B}{b - B} \right)$.

In the *bi-linear* case, the first half of the values are fitted with one line, and the rest are fitted with another. For example, given a and b, to interpolate 10 values between these seed values, the bi-linear method (with baseline required) is:

$$x_i = \begin{cases} a + (b - a)/10 \times i & \text{if } i < k \\ 2 + (b + a)/2 + ((b - a)/2 - 2)/k \times (i - k) & \text{otherwise.} \end{cases} \tag{3.7}$$

where bend point is at the k^{th} test.

In the *episodic* case, the program can simulate many different segments in a range, and the number and position of episodes are randomly chosen. At most two bend points are used, because in clinical trials, patients' eyes are not measured many times. (If there are many bend points, the episodic function may be replaced by a convex or concave function.) For example, given a and b, to interpolate 10 values between these seed values by using the episodic function, x_i is given by:

$$x_i = \begin{cases} a & \text{if } i < 3 \\ c & \text{if } 3 \leqslant i < 7 \\ b & \text{otherwise} \end{cases} \tag{3.8}$$

where c is a constant such that $a > c > b$.

In the *stable* case, the simulation function for interpolated values is simply

$$x_i = a \tag{3.9}$$

3.2.3 Frequency of Measurement

In clinical work, eyes with glaucoma are sometimes measured once per year (Spry et al., 2000), but more usually twice per year (though Gardiner and Crabb (2002b) examined more than twice per year). In this study, we simulated measurement(s) once or twice per year.

3.2.4 Age-related decline

An age-related decline of 0.1 dB per year is added to each sequence of measurements (Gardiner and Crabb, 2002a; Spry et al., 2002) . This is to simulate natural degeneration over time.

For example, consider a sequence of 10 values 30, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20. If the simulated measurement frequency is twice per year, then 0.05 dB per visit (half year) is subtracted from the third and subsequent values: 30, 30,

28.95, 27.9, 26.85, 25.8, 24.75, 23.7, 22.65, 21.6, 20.55, 19.5. This sequence represents a 2 dB decrease with natural age-related decline, without noise, over five years of measurements. If the simulated frequency of measurements is once per year, then 0.1 dB per visit (one year) is subtracted from the third subsequent values: 30, 30, 28.9, 27.8, 26.7, 25.6, 24.5, 23.4, 22.3, 21.2, 20.1, 19. This sequence represents a 1 dB decrease with age-related decline for 10 years of measurements.

3.2.5 Noise (short- and long-term fluctuation)

As the variability (noise) of standard automated perimetry is complex, the key to any accurate visual field simulation is how well this variability is estimated (Gardiner and Crabb, 2002a).

Noise is added to simulate short- and long-term fluctuation in two different ways. In the first method, short-term and long-term fluctuation are separately added as described by Spry et al. (2000). The other method is similar to that described by Gardiner and Crabb (2002a) which uses a function described by Henson et al. (2000).

Representation of short-term and long-term fluctuation. A measured value x is usually influenced by short- and long-term fluctuation. That is:

$$x = true\ value + \varepsilon_l + \varepsilon_s \tag{3.10}$$

where ε_l is a random variable repsrenting long-term and ε_s repsrenting short-term fluctuation. Both fluctuations have Gaussian probability density functions (Spry et al., 2000): $\varepsilon_s = \mathrm{G}(\mu_1, (SD_s)^2)$ $\varepsilon_l = \mathrm{G}(\mu_2, (SD_l)^2)$, where $\mu_1 = \mu_2 = 0$, and SD_s and SD_l are the means and standard deviations respectively.

The standard deviation, SD_l, of long-term fluctuation is set using a constant value equal to 1 according to experimental results on the difference between real and computer-simulated visual fields (Vesti et al., 2002). This quantity represents the amount of homogeneous fluctuation that occurs between one test and the next (Spry et al., 2000).

The standard deviation SD_s of short-term fluctuation varies according to two factors at a particular location.

The first factor is the threshold value at the location. Deviation SD_s increases by 0.08 dB per 1 dB deviation from the age-matched threshold: $0.08 \times |t - n|$, where t is the threshold at a location, and n is the age-matched normal value for that location (Spry et al., 2000).

The second factor is the position in the visual field. Each visual field can be mapped to a 10×10 grid as shown in Figure 3.3. Each star represents a location in the visual field of the right eye. Two small stars represent the physiological blind spot.

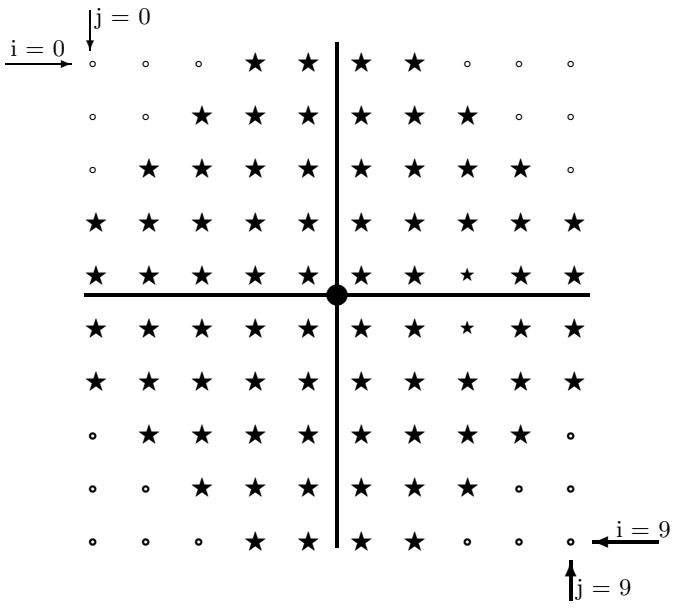


Figure 3.3: A 10×10 grid mapping a visual field. Each star represents a location in the visual field of right eye; two small stars represent the blind spot.

The factor is calculated as

$$(\sqrt{(}4.5 - i)^2 + \sqrt{(}4.5 - j)^2)/6.5 \tag{3.11}$$

where i and j are the coordinates of the location in the 10×10 grid (see (Spry et al., 2000, p. 3)). The upper left corner is $i = 0$, $j = 0$. Values in the outer part of the visual field are more noisy than those closer to the center. This quantity is calculated and added to SD_s, thus increasing the range of possible simulated threshold values.

Noise represented by a function. Noise (including short- and long-term fluctuation) in normal subjects and glaucomatous patients was estimated by Henson et al.

(2000) as a Gaussian probability density function G(μ, SD), where $\mu = 0$ is the mean value and SD is the Standard Deviation.

Henson et al. (2000) gave an estimate for the SD, dependent on the true threshold. The SD was represented by the function:

$$log_e(SD) = A \times threshold(dB) + B \tag{3.12}$$

where the constants A and B are -0.081 and 3.27, respectively. At each location, therefore, the noise was determined by independent random samples from the normal distribution $N(\mu, (SD)^2)$ with the mean value at the true threshold value, and the standard deviation (SD) derived from the Equation (3.12). For example, if the true threshold value is 26dB, when noise is added the measured threshold value is drawn from the distribution $N(26, 3.2^2)$ This procedure was repeated for each value of each sequence. Pseudo code describing the whole procedure follows.

Algorithm 2. *Simulation algorithm.*

Input: a and b;

Output: a sequence;

Begin

Choose a function $f(linear, bilinear, concave, convex, episodic, stable)$

Choose test frequency (6 months or 1 year)

Interpolate n values between a and b using f

Add age-decline (0.05 dB per 6 months or 0.1 dB per year)

Add noise (using either SD_s + SD_l or $exp(-0.081t + 3.27)$

For each produced value,

if the value > 39, then the value $= 39$

else if the value < 0, then the value $= 0$

End

3.3 Experimental Data

In Chapter 4, 5, and 6, we used simulated sequences to evaluate the effectiveness of methods. In Chapter 4 (EA methods) and Chapter 5 (sequence matching techniques), 3330 progressive sequences fell into three classes (types of pattern): 1110 each of linear, bi-linear and convex curvilinear degradation. In each case the starting value of the sequence was randomly chosen from 15 normal visual fields in the stable database described in more detail in Chapter 4. In each case, 370 of the 1110 sequences had final values that were 10 dB less than the initial values of the sequences, 370 had final values that were 15 dB less, and 370 that were 20 dB less. This gives an average decrease of 2 dB, 3 dB or 4 dB per year for 5 years. Note that in sequences with values that dropped below zero, all negative values were replaced by zeros (equivalent to total blindness for that location).

In Chpater 6 (linear regression analysis), we used 15 fields from normal subjects to simulate *3330 (= 15 (fields) × 74 (locations/per field) × 3 (final values/per initial value)* progressing sequences which only fell into linear class. Fields from normal subjects in the database serve as the initial values in the simulation program. Three final values are obtained from each initial value, by subtracting 10, 15 or 20 dB, with a minimum possible value of 0 dB. For example, if the initial value is 28 dB, then the 3 final values obtained are 18 dB, 13 dB and 8 dB. That is, for the initial value 28, the three pair sets of initial and final values are: (28, 18), (28, 13), and (28, 8).

Note that we did not use concave and episodic functions to generate sequences. This is because (1) there is no standard definition of progression; (2) in Figure 3.2, true values indicated by small circles in both concave and episodic curves do not change too much from the $7th$ to $10th$ tests. The changing trend in both cases seems to be stable in the late tests.

To simulate stable sequences, the first visual field from each of the 50 real subjects is used as both the initial and the final values in the simulation. The middle 7, or 9, or 10 values are generated with age-related decline and noise which are described in detail in Chapter 4, 5 and 6.

In total, therefore, we simulated 3330 progressing sequences (locations) and 50(fields) × 74(locations) = 3700 stable sequences (locations). These, together with

the real data sets, are used in the experiments.

3.4 Discussion

In the chapter, we have introduced stable, linear, bilinear, episodic, and one new progressive model − curvilinear, which includes convex and concave cases. For a linear or bilinear progression, the degradation occurs along a straight line or two straight lines. The decline can easily be detected using established methods. Similarly, for convex progression, the rate of change goes from slow to quick. This represents a deteriorating location starting slowly so that it is not noticed by the patient. In contrast, in the case of concave progression, the rate of change goes from quick to slow. This represents a deteriorating location that is gradually controlled with treatment. In the case of the episodic function, the true thresholds at each segment do not change. The changes only occur at the episode points.

When we used Equation 3.12 as the standard deviation of a normal distribution to simulate noise, the *SD* value is very large when the true threshold is below 15. Furthermore, if the true threshold is below 15, the number of subjects used by Henson et al. (2000) is too small to reliably estimate the *SD*. Therefore, we cap the *SD* value at 9dB, instead of the *SD* produced by Equation 3.12. This is because the value of "mean (threshold) ± 1.96 *times* 9 dB" ≈ 0, 1, . . . , 39.

The data produced by using Equation 3.12 is noisier than that produced by the method of Spry et al. (2000). For example, if a true threshold is 20, assuming 45 year old patient, the standard deviation *SD* obtained by using Function 3.12 is 5.207. The simulated threshold value (for the true threshold 20) is probably in the range [9.794, 30.206] [1].

If using the method of Spry et al. (2000), the standard deviations SD_s and SD_l for short- and long-term fluctuation are calculated as follows. SD_s = 1 + *Eccentricity-related fluctuation* + *Defect-related fluctuation*, where 1 is the standard deviation of the standard normal distribution. Eccentricity-related fluctuation calculated using Equation 3.11 is less than 1 for any location. Defect-related fluctuation is $|t-n| \times 0.08$,

[1] 95% interval can be calculated by Mean ± 1.96 $\times SD$

where t is the measurement, and n is the age-matched threshold. The maximum value n in the 45 years old normal field is 35 dB read from our database. Therefore, $|t-n| \times 0.08 = 0.08 \times |20-35| = 1.2$. Thus $SD_s < 1+1+1.2 = 3.2$. According to Spry et al. (2000) and Vesti et al. (2002), the constant SD_l was set to 1 dB (or 2). Assuming short- and long-term fluctuation are independent, the standard deviation of the random variable ε in Equation 3.10 is $\sqrt{(SD_s)^2 + (SD_l)^2} < \sqrt{3.2^2 + 1^2} = 3.353$ (or $< \sqrt{3.2^2 + 2^2} = 3.499$). The simulated threshold value (for the true threshold 20) by using short- and long-term fluctuation is in the range [13.428, 26.572] (or [13.142, 26.858]). Therefore, fluctuation in the data simulated by using short- and long-term noise is smaller than by using Equation 3.12. Therefore, the accuracy of classification in the datasets obtained by using short- and long-term fluctuation is expected to be higher than in the dataset obtained by using Equation 3.12. This will be explored further in Chapter 6.

3.5 Summary

In this chapter, we described an algorithm for simulating sequences of threshold measurements for independent visual field locations. The program not only provides linear, bilinear, episodic, and stable functions, but also provides convex and concave functions to closely model real visual field data. The frequency of tests per year, total number of tests, age-related decline, and short- and long- term fluctuation can all be controlled to suit the needs of the experiment.

Chapter 4

Bias Analysis

4.1 Introduction

In the previous chapter, we generated datasets in which each sequence is a simulation of measurements at a single location in a visual field. In this chapter, we aim to quickly and reliably identify each sequence as progressive or stable.

Early and reliable identification of glaucomatous progression is an important part of the management of glaucoma. Measured values are very noisy, especially at locations with progression. The noise makes it difficult to identify true progression, and many methods have been proposed to address this problem. One group of methods is the bias analysis. Bias is a deviation of a value from a reference value. That is, in this chapter, repeated real normal and stable data will be used as reference data to classify a given location in a visual field as progressing or stable. Bias analyses include (1) probability analysis based on all possible responses (the rate of the false positive and the false negative response) and frequency-of-seeing curve (Turpin and McKendrick, 2005), (2) paired t-test to determine whether significant differences are present between one test result and another (Spry et al., 2002), and (3) the well known Glaucoma Change Probability (GCP) method, described in Section 2.2.4.

In this chapter, we focus our discussion on the GCP method. As mentioned previously, in order to avoid confusion with the GCP software, we refer to the GCP method used here as Event Analysis (EA).

EA method chooses a pre-determined confidence interval for each visual field loca-

tion, and flags a location as changed if its threshold value falls outside that confidence interval. Confidence intervals are selected according to the *baseline defect* of a location; that is, the average of the first two thresholds measured at that location, less an age matched normal value. The steps of using the EA method to identify a single location are as follows.

1. Calculate the baseline defect. Let x_1 and x_2 be the baseline values for one location, and n be an age-matched normal value at that location. The baseline defect $= [(x_1 + x_2)/2] - n$.

2. Select a 95% or 99% confidence interval (CI - see below) according to the baseline defect.

3. Calculate the difference $(x_1 + x_2)/2 - x_n$, where x_n is a value measured after the baseline values.

4. Identify whether the difference $(x_1 + x_2)/2 - x_n$ falls outside the selected CI.

The upper and lower bounds of the confidence interval are established from a database of repeated SAP tests of stable visual fields. There are two components of the algorithm used to calculate EA confidence intervals. The first component is the *classification* of a point into a confidence interval class. A point can be classified based on its total deviation value, its eccentricity (say, the distance between a location and the central point of the visual field), its baseline value alone, or a combination of these. Once a confidence interval class is chosen, the second component is the *differencing* technique used to obtain the confidence interval from the database of stable fields. Typically test-retest values are used to derive a distribution of likely values.

When using a method for determining whether a visual field is stable or progressing, the specificity (proportion of stable cases correctly classified) and sensitivity (proportion of progressive cases correctly classified) both have to be taken into account. The accuracy of the EA method depends on the underlying confidence interval. If the confidence interval is wide, the EA method labels a small visual field loss as stable, reducing sensitivity. If the confidence interval is narrow, the EA method labels a stable case with high noise as progressive, reducing specificity. Establishing 95% or

99% confidence intervals is therefore a key issue for the EA method. In this chapter, we aim to explore this issue. In addition, we propose an alternative method, called "baseline-less-follow-up" differences, to compute the interval and examine its performance against test-retest values.

The chapter is structured as follows. Section 4.2 describes two datasets: one is the real dataset which is used to build confidence intervals; the other is the simulated dataset which is used to examine the performance of the built confidence intervals. Section 4.3 describes the algorithms for building confidence intervals, and the statistical method, *Cochran's Q-test*, which assesses whether any difference between the methods is due to chance. Section 4.4 shows the experimental results followed by the discussion in Section 4.5, and Summary in Section 4.6.

4.2 Datasets

We used two data sets to examine the effectiveness of four ways of deriving EA confidence intervals. The first is real patient data, consisting of five follow-up visual fields, which is used for deriving EA confidence intervals. The effectiveness of these intervals in classifying stable and changing data is measured using a second set of data, which consists of synthetically generated sequences of 12 thresholds. Both sets are described in detail in this section. Figure 4.2 shows a schematic overview of this study.

4.2.1 Patient Data

Confidence intervals were built from stable visual fields collected as part of a study at the Nova Scotia Eye Centre, which is fully described elsewhere (Chauhan and Johnson, 1999). Fifteen subjects with normal eyes, and thirty-five subjects with stable glaucoma, repeatedly performed the 30-2 program of the Humphrey Field Analyzer (Carl Zeiss, Meditech, Dublin CA). The examination interval for subjects with normal eyes was six months. Of the glaucoma patients, 90% were followed up within a one week period between consecutive tests. The subjects were aged from 34 to 82 years (average 60.9 years). The first five follow-up visual fields for each subject were used, even in cases where subjects had more than five SAP tests. Thus a total of $50 \times 5 = 250$

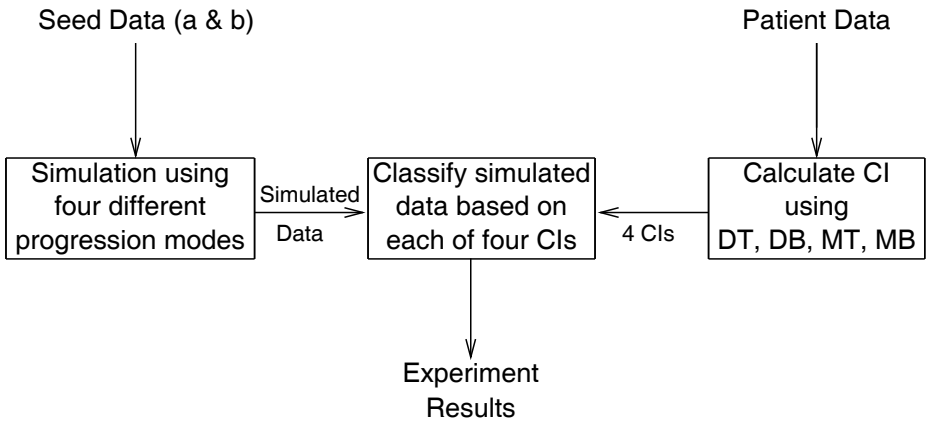


Figure 4.1: Schematic overview of this study.

visual fields were used, providing $50 \times 74 = 3700$ locations, each with a sequence of five measurements. This data set was used to derive the various confidence intervals in the experiments that follow.

Figure 4.2(a) shows a histogram of the mean threshold at individual locations in the real patient data over the five tests for both normal and stable glaucomatous subjects. Figure 4.2(b) shows the distributions of defects (see Building Confidence Intervals below) for the inner, middle, and outer zones of the visual field. Note that in the figures some mean threshold values and some defect values are based on very little data, and so deriving confidence intervals from them is difficult. Section 4.3.4 describes the solution to this problem.

4.2.2 Simulated Data

To test the effectiveness of the EA method with different confidence intervals, 7030 sequences (3330 progressive and 3700 stable) were generated using the simulation program described in detail in Chapter 3. Given the initial and final thresholds for a single location in the visual field, the program generates a sequence of thresholds in between these seed values.

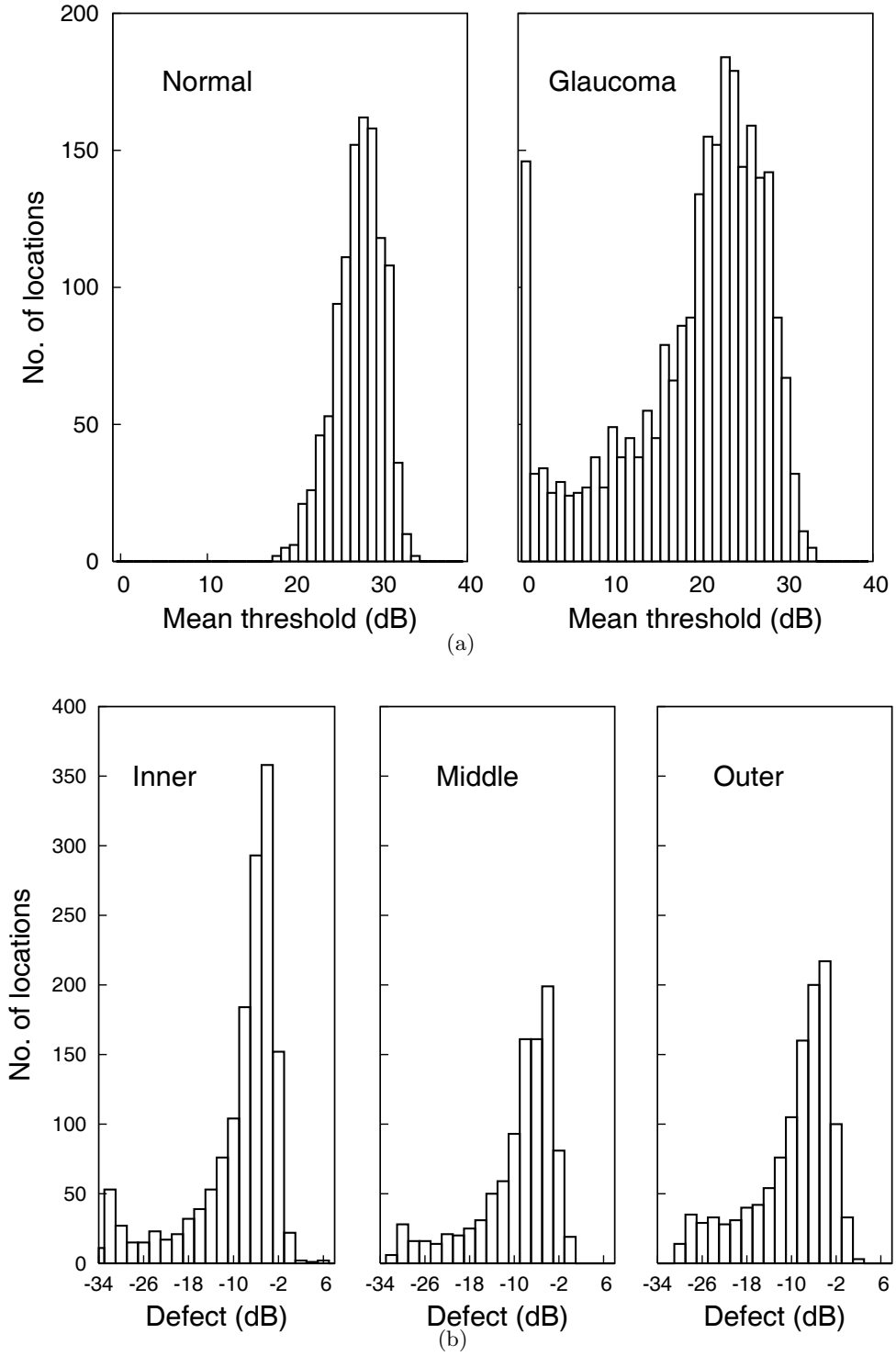


Figure 4.2: Distribution of visual field data in the reference database. (a) Histogram of the mean threshold over five visual field tests for the patient data. (b) The distributions of defects in the inner, middle, and outer zones of the visual field.

For example, given an initial threshold a and final threshold b, the sequence a, x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, b was obtained by using the program to interpolate the middle 10 values. The first interpolated value x_0 is the same as the initial value a (until noise is added). This is because the EA method requires the baseline measurement. The frequency of follow-up for subsequent values was taken to be twice per year (6-month intervals). An age-related decline of 0.1 dB per year was then added to each sequence of measurements. Finally short- and long-term fluctuation were separately added.

We generated 3330 progressive sequences of 12 values. The sequences fell into three classes (types of pattern): 1110 each of linear, bi-linear and convex curvilinear degradation. In each case the starting value of the sequence was randomly chosen from 15 normal visual fields in the stable database described in Section 4.2.1. In each case, 370 of the 1110 sequences had final values that were 10 dB less than the initial values of the sequences, 370 had final values that were 15 dB less, and 370 that were 20 dB less. This gives an average decrease of 2 dB, 3 dB or 4 dB per year for 5 years. Note that in sequences with values that dropped below zero, all negative values were replaced by zeros (equivalent to total blindness for that location).

To simulate stable sequences, the first visual field from each of the 50 real patients was used as both the initial and the final values in the simulation. The middle 10 values were generated as for the linear case in the progression simulation, with age-related decline and noise.

In total, therefore, we simulated 3330 progressing locations and $50 \times 74 = 3700$ stable locations, each with 12 threshold measurements (including the initial and the final seed values from which the other 10 measurements were derived) as summarised in Table 4.1).

Table 4.1: The summary of the simulated sequences.

Classes	Decline Per year	No.	Total
Progressing	2dB	1110	3330
	3dB	1110	
	4dB	1110	
Stable	0dB	3700	3700

4.3 Methods

4.3.1 Building Confidence Intervals

The EA method calculates the difference between a threshold and a baseline on a point-by-point basis, and then determines whether the difference falls inside or outside a 95% or 99% confidence interval established using a database of stable glaucomatous visual fields (Anderson and Patella, 1999). The stable database described above (Section 4.2.1) was used to build the confidence intervals, according to four different methods. The first of these methods is similar to that described by Heijl et al. (1989), and the second arose out of discussions with Chris Johnson around the 2002 review article he co-authored (Spry et al., 2002). The third and fourth methods use the classification component of Heijl et al and Spry et al respectively, but alter the derivation component.

Defect-based, Test-retest (DT). Given a sequence of five threshold values at a single location $y_1, y_2, \ldots, y_5$, the difference is calculated between each value y_i and the corresponding age-matched normal threshold value n. The quantity $y_i - n$ is the *deviation*. The average of the first two deviations in the sequence is taken as the *defect* of the sequence. All sequences are classified according to defect and eccentricity. Eccentricity can be one of: "inner", the 30 central points of the field excluding the two blind spots ($15°, \pm 3°$); "middle", the 20 points directly bordering the inner field; or "outer", the 24 points on the edge of the 30-2 pattern (see Figure 2.15). We call this classification scheme Defect-based (denoted D). For each class (or group), the test-retest differences in each sequence, $y_i - y_{i+1}$, $1 \leq i \leq 4$ are calculated and sorted from the smallest to the largest. We call this method of selecting differences T, so juxtaposing the labels of the classification and differencing methods we have method DT. Once classification and differencing have been performed, the 2.5th and 97.5th percentiles are computed for each group to form a 95% confidence interval about the defect for that group.

Mean-based, Test-retest (MT). The mean of all five values in a sequence for one location is calculated: $\bar{y} = (y_1 + y_2 + \cdots + y_5)/5$. The mean is rounded to the nearest integer, which is used to classify the sequence. We call this classification

scheme Mean-based (denoted M). The calculation of differences now proceeds as for the first method (that is, T), yielding method MT.

We propose an alternative method for building confidence intervals that calculates the difference between each measurement in a sequence and the baseline of that sequence, rather than test-retest differences. The baseline is defined as the average of the two initial measurements. Combining this differencing method with the classification methods D and M described above yields two new methods.

Defect-based, Baseline-less-follow-up (DB). Classify sequences according to method D, but calculate differences as baseline-less-follow-up: $(y_1 + y_2)/2 - y_i$, $3 \leq i \leq 5$. This is method DB.

Mean-based, Baseline-less-follow-up (MB). Classify sequences according to method M, but calculate differences as baseline-less-follow-up: $(y_1 + y_2)/2 - y_i$, $3 \leq i \leq 5$. This is method MB.

When building a confidence interval, DT and DB use the same method of classification according to defect and eccentricity. DT and DB have the same groups. However, for each group, DT uses test-retest differences: $y_i - y_{i+1}$, $1 \leq i \leq 4$; in contrast, DB uses baseline-less-follow-up differences: $(y_1 + y_2)/2 - y_i$, $3 \leq i \leq 5$. Thus confidence intervals between DT and DB may be not the same. These methods give different results. Similarly, MT and MB have the same groups, but maybe do not have the same confidence interval. For example, consider a sequence of threshold values at a single location: 25, 31, 28, 32, 24. The test-retest differences (DT and MT used) for this sequence are -6, 3, -4, 8. In contrast, the baseline is $(25 + 31)/2 = 28$ and the baseline-less-follow-up differences (DB and MB used) are 0, -4, 4. Figure 4.3 gives a summary of all four methods of building a confidence interval.

4.3.2 Classifying the Simulated Test Sequences

Consider a test sequence $X = \{a', x'_0, x'_1, \ldots, x'_k\}$ generated using the simulation procedure. To classify X using either DT or DB it is first necessary to calculate the defect of X, that is the $((a' - n) + (x'_0 - n))/2$, where n is the corresponding age-matched

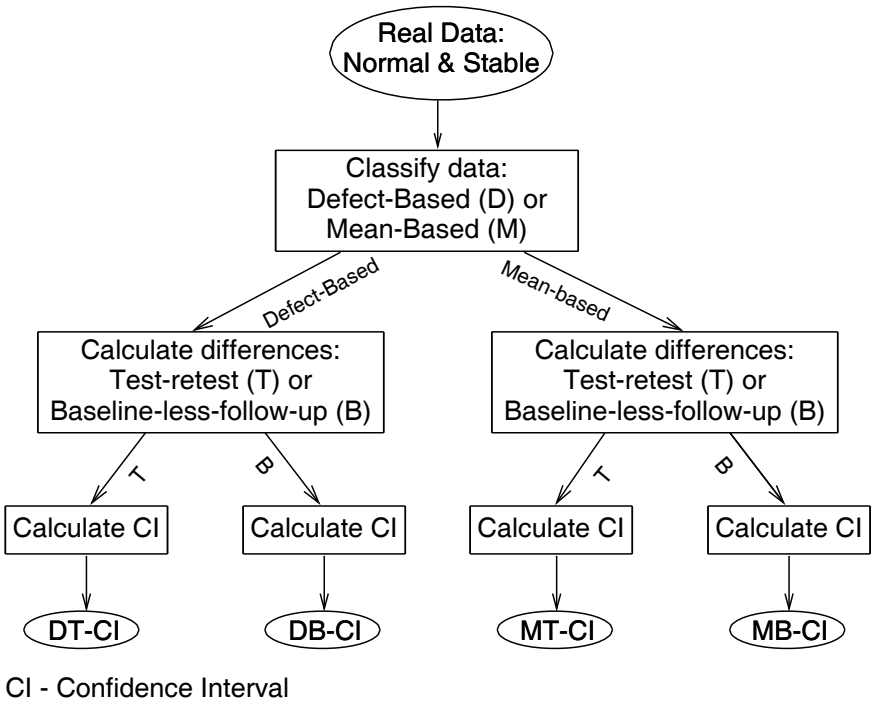


Figure 4.3: Methods for building confidence intervals.

normal threshold value. The defect of the sequence, together with the eccentricity of the location, is used to select an appropriate confidence interval calculated from the reference database. Next, the difference between the baseline $((a' + x'_0)/2)$ and the current value x'_k is calculated. The final step is to evaluate the difference value to determine whether it lies inside (stable) or outside (progressive) the 95% confidence interval.

To classify X using either MT or MB the procedure is similar to that described above, except that the baseline of the sequence is calculated instead of the defect, and this is used to select an appropriate confidence interval from the reference dataset.

4.3.3 Modification of Sparse Patient Data

Confidence limits will only be meaningful if derived from a large enough sample of locations that get classified together. In several of the classes, derived either with method M or D, the number of locations falling in that class from the patient data was small. Rather than simply discarding precious data, I altered sequences at locations

that fell into a sparse class by a constant amount so that they fell into a well populated class.

Classification Using Method *M*. Figure 4.2 (a) shows a histogram of the mean threshold at individual locations in the real patient database over the five tests. There is little data in the groups of mean threshold values larger than 32, hence the confidence limits derived for these classes will be unreliable. Rather than simply discarding these values, we subtracted a constant from all values at any such location so that the mean of the values fell below 32 (or equal to). For example, a sequence such as 33, 33, 30, 32, 29, 31, 32, 30, 29, 28, 33, 32 would have a baseline of (33 + 33)/2 = 33. We therefore modified this sequence by computing baseline - 32 = 33 - 32 = 1. This value (1) was then subtracted from all values in the sequence. The modified sequence (32, 32, 29, 31, 28, 30, 31, 29, 28, 32, 31) has a baseline that equal to 32 and so can now be matched using the patient data. The final simulated data set therefore had all mean thresholds less than 32.

Classification Using Method *D*. The distributions of defects for the inner, middle, and outer zones of the visual field are shown in Figure 4.2 (b). The data for deriving confidence intervals for defects greater than zero is sparse, so simulated locations with defects in this range were adjusted by a constant to make the baseline defect zero. The final simulated data set therefore had all baseline defects less than or equal to zero.

4.3.4 Statistical Methods

Each of the four confidence intervals obtained from the real patient data was used to classify each simulated location using EA as outlined in Figure 4.2. We evaluated all four simulated groups: 10dB, 15dB and 20dB decrease, and stable. For each test sequence, each method produced a binary outcome: classified correctly or not.

Cochran's Q-test (Cochran, 1950) was used to asses whether any difference between the methods was due to chance. This test is essentially an ANOVA for binary (also referred to as dichotomous) data. If the Q-test indicated a difference between the methods ($\alpha < 0.05$), then a non-parametric pairwise comparison technique outlined

by (Sheskin, 2000, p. 689) was applied. This takes into account a correction for multiple comparisons and binary data.

For example, suppose that three methods M_1, M_2, and M_3 are employed, and the number of subjects employed in the experiment is $n = 12$ (see Table 4.2. Each method classifies each subject as "yes" or "no". The null hypothesis to be tested is whether the proportion of **yes** responses in the population represented by the method M_1 equals the proportion of **yes** responses in the population represented by the method M_2, and equals the proportion of **yes** responses in the population represented by the method M_3.

The Cochran's Q statistic is defined as

$$Q = \frac{(k-1)[(k)(C) - (T)^2]}{(k)(T) - R} \tag{4.1}$$

Where k is the number of the methods (here $k = 3$). $C = \sum(\sum C_j)^2$ where C_j is the number of **yes** responses for method M_j. $T = \sum R_i$ where R_i represents the sum of **yes** responses of the k (here $k = 3$) methods for the i^{th} subject. $R = \sum R^2$.

Table 4.2: An example of calculating Q-test.

	M_1	M_2	M_3	R_i	R_i^2
Subject 1	0	0	0	0	0
Subject 2	0	0	1	1	1
subject 3	1	1	0	2	4
...	...	...	...	...	...
Subject 12	1	1	1	3	9
	$\sum C_1 = 3$	$\sum C_2 = 9$	$\sum C_3 = 3$	$\sum R_i = 15$	$\sum R_i^2 = 27$
$p_1 = \frac{\sum C_1}{n} = \frac{3}{12} = 0.25$		$p_2 = \frac{\sum C_2}{n} = 0.75$		$p_3 = \frac{\sum C_3}{n} = 0.25$	

Therefore $C = (\sum C_1)^2 + (\sum C_2)^2 + (\sum C_3)^2) = 99$, $T = 15$, and $R = 27$. Q = 8 for this example computed by Equation 4.1. In order to reject the null hypothesis, the calculated Q value must be greater than or equal to the critical value obtained from the table of chi-square values with $k - 1$ degree of freedom at the pre-specified level of significance. If the null hypothesis is rejected, then the proportion of **yes**

responses in the population represented by method M_1 is not equal to the proportion of **yes** responses in the population represented by method M_2, or it is not equal to the proportion of yes responses in the population represented by method M_3.

The statistic

$$CD_c = z_{adj}\sqrt{2\frac{(k)(T)-R}{(n^2)(k)(k-1)}} \tag{4.2}$$

is used to compare the proportions of two methods (i.e. that if $p_i = p_j, i \neq j$) with the normal distribution, where z_{adj} is the critical value obtained from the table of the normal distribution with probability level $(\alpha/(k(k-1)))$ which can be assured that any pairwise comparison will not exceed α.

Consider Equation 4.2 above. For example, given $\alpha = 0.05$, $z_{adj} = 2.39$ and $CD_C = 0.49$. The obtained value $CD_C = 0.49$ indicates that any difference between two proportions is significant if the difference is greater than or equal to $CD_C = 0.49$. Therefore we can conclude that the differences between M_1 and M_2, and M_2 and M_3 are significant, but the difference between M_1 and M_3 is not significant, since $|p1 - p2| = 0.5$, $|p2 - p3| = 0.5$, and $|p1 - p3| = 0$.

4.4 Results

EA was used to assess all threshold measurements in each simulated sequence, using each confidence interval method. Figure 4.4 shows the accuracy of classification for progressing sequences. The horizontal axis shows the number of visits following the first two at which the baseline measurements were taken. The sensitivity of the methods generally increases with the number of visits. This is because the difference between the baseline (the average of the first two measurements) and each subsequent visit increases with time for progressing fields, and so the chance of a value falling outside the confidence limit increases with time. Figure 4.5 shows the accuracy of classification for stable sequences. The specificity of each method is approximately constant across the ten visits. (Note that the vertical scale is different in the two Figures.) Figure 4.4 shows that before the 4th visit (2 years), the probability of correct classification for progressive sequences is lower than 50% for all four methods (except *MB* on visit 4, which is slightly better than chance). Therefore, the analysis of methods starts at the 4th visit in the remainder of this section.

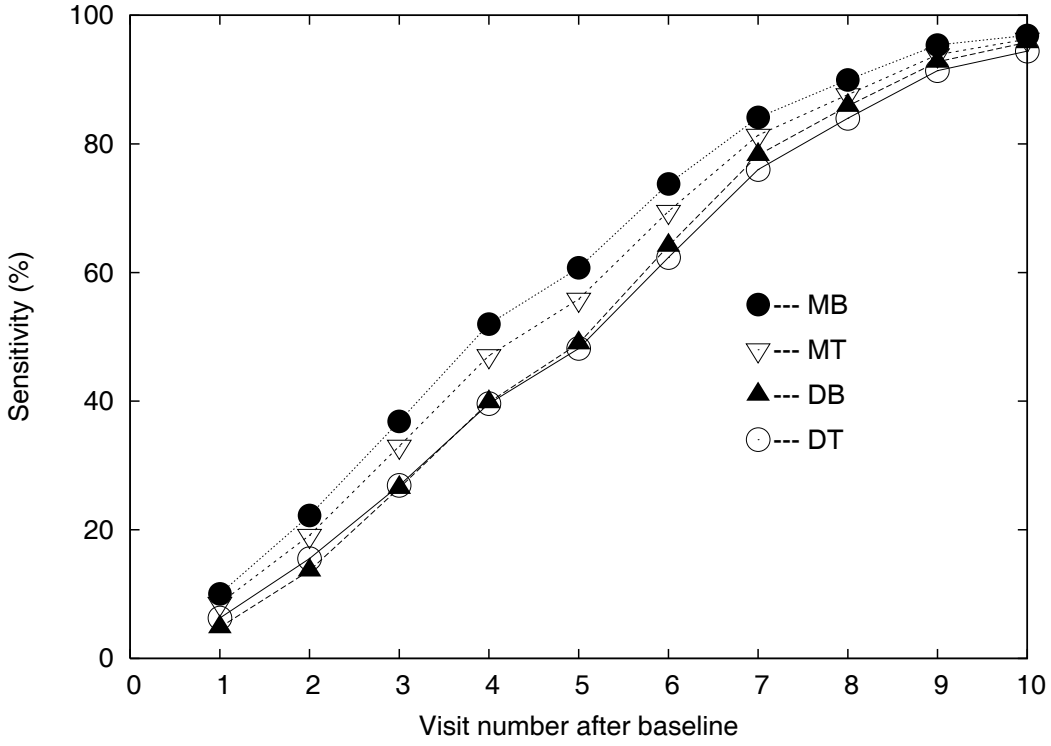


Figure 4.4: Accuracy of EA classification for progressing data (sensitivity). Four different confidence intervals are used on all ten threshold measurements in each sequence.

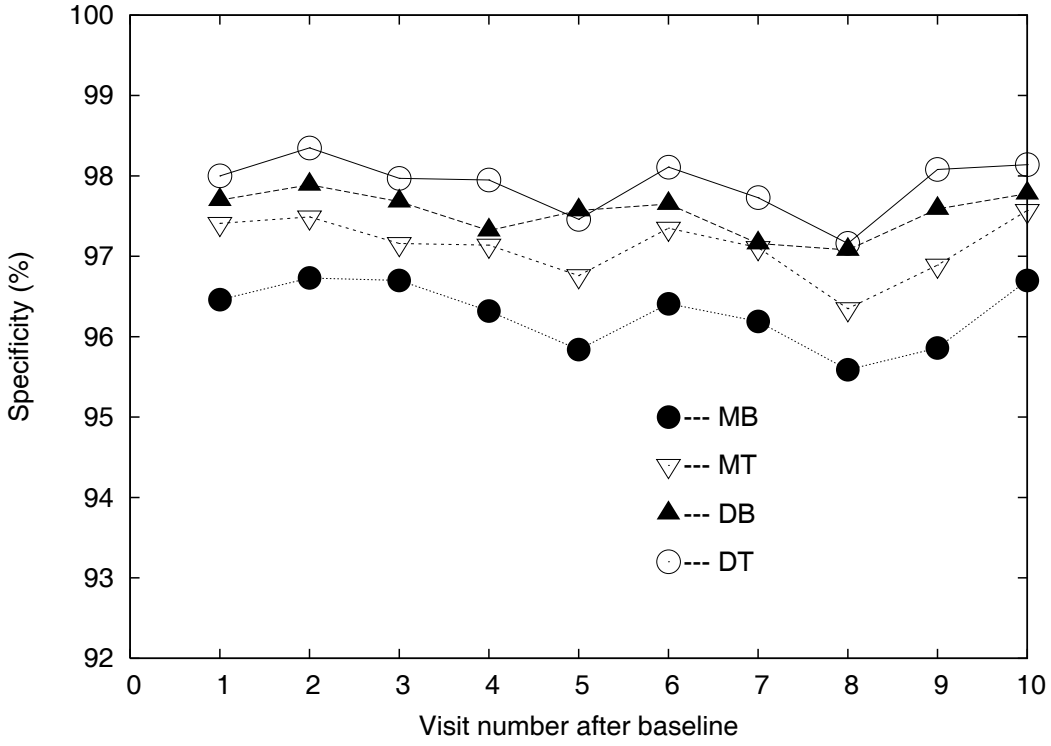


Figure 4.5: Accuracy of EA classification for stable data (specificity). Four different confidence intervals are used on all ten threshold measurements in each sequence.

Method DT attains the highest specificity (between 97.2% and 98.1%) for assessing the 4th to the 10th measurements in a stable sequence of 6-monthly measurements. However, it sacrifices sensitivity, particularly when the change is small that will be seen in -2dB rows Table 4.3. In contrast, method MB attains the highest sensitivity for assessing the 4th to the 10th measurements in a progressive sequence, but has lower specificity (between 95.6% and 96.7%). The discrepancies between MB and DT in correctly classifying progressive sequences are between 2.5% and 12.4%. The discrepancy between MB and DT is smallest at the 10th visit. This is because the change between the baseline and the 10th visit is very large (10, 15, or 20 dB without noise) so that all methods can correctly classify most progressive sequences.

Cochran's Q-test was used to evaluate differences between the methods. The test revealed that there were statistically significant differences ($\alpha < 0.05$), and so pairwise comparisons were performed as shown in Table 4.3.

The first column of differences in Table 4.3 ($DB - DT$) indicates that there are some significant differences between DB and DT from the 6th visit onwards when classifying progressive sequences, particularly in the group showing 2 dB decline per year. For classifying stable threshold measurements, there is no consistent statistically significant difference.

The second and third columns of differences indicate that MT and MB have higher sensitivity than DT, but offer decreased specificity. The MB method offers approximately 2-14% increase in sensitivity, particularly for small change (2 dB decline per year), and particularly for analyzing the 4th-6th visits in a sequence. However, the cost is about 2.2% decrease in specificity. The difference between the second and third columns shows that the sensitivity gains of the MB method over DT are partly due to the classification method (M compared with D), and partly due to the differencing technique (B compared with T).

When the classification method is held constant as M in the final column, there is 2-5.6% increase in sensitivity (except for the 9th and 10th visits, and 4 dB decline per year) due to using the baseline-less-follow-up differencing method over the test-retest method.

Table 4.3: Differences in percentage of correct classifications between the four confidence interval methods (*MT, MB, DT, DB*). Statistically significant change is shown in bold ($\alpha < 0.05$), and using an asterisk $*$ ($\alpha < 0.01$).

	Sequence	$DB-DT$	$MT-DT$	$MB-DT$	$MT-DB$	$MB-DB$	$MB-MT$
	-2dB/yr	+0.1	**+ 6.8***	**+11.5***	**+7.8***	**+12.5***	**+4.7***
4th visit	-3dB/yr	+0.5	**+8.6***	**+14.1***	**+8.1***	**+13.6***	**+5.5***
(2 years)	-4dB/yr	+1.7	**+7.2***	**+11.9***	**+5.3***	**+10.0***	**+4.7***
	Stable	−0.4	**−0.8***	**− 1.7***	−0.4	**− 1.2***	**−0.8***
	-2dB/yr	+0.1	**+7.0***	**+12.6***	**+7.8***	**+13.4***	**+5.6***
5th visit	-3dB/yr	+0.9	**+8.8***	**+13.9***	**+7.4***	**+12.4***	**+5.0***
(2.5 years)	-4dB/yr	+1.5	**+7.7***	**+11.6***	**+5.8***	**+ 9.7***	**+4.0***
	Stable	−0.1	**−0.8**	**− 1.7***	**−0.6**	**− 1.5***	**−0.9***
	-2dB/yr	+1.0	**+8.6***	**+14.1***	**+7.7***	**+13.1***	**+5.4**
6th visit	-3dB/yr	+2.1	**+7.1***	**+12.6***	**+5.0***	**+10.5***	**+5.5***
(3 years)	-4dB/yr	**+2.1***	**+5.8***	**+ 7.7***	**+3.7***	**+ 5.6***	**+1.9**
	Stable	−0.5	**−0.8***	**− 1.8***	−0.4	**− 1.3***	**−0.9***
	-2dB/yr	**+3.2***	**+8.6***	**+13.2***	**+5.4***	**+10.1***	**+4.7***
7th visit	-3dB/yr	**+2.4***	**+5.1***	**+ 7.5***	**+2.7***	**+ 5.0***	**+2.3***
(3.5 years)	-4dB/yr	**+1.3**	**+2.3***	**+ 3.7***	+1.1	**+ 2.4***	**+1.4**
	Stable	**−0.5**	**−0.7***	**− 1.6***	−0.1	**− 1.0***	**−0.9***
	-2dB/yr	**+2.3**	**+6.0***	**+10.7***	**+3.7***	**+8.4***	**+4.7***
8th visit	-3dB/yr	**+2.1***	**+2.9***	**+ 4.9***	+0.8	**+2.8***	**+2.0***
(4 years)	-4dB/yr	**+1.1**	**+2.0***	**+ 2.3***	+0.9	**+1.2***	+0.3
	Stable	−0.1	**−0.8***	**− 1.5***	**−0.7**	**−1.4***	**−0.8***
	-2dB/yr	**+3.0***	**+6.6***	**+10.1***	**+3.6***	**+7.1***	**+3.5***
9th visit	-3dB/yr	+0.8	+0.8	**+ 1.7***	+0.0	**+0.9**	**+0.9**
(4.5 years)	-4dB/yr	+0.2	+0.3	+ 0.4	+0.1	+0.2	+0.1
	Stable	−0.5	**−1.1***	**− 2.2***	**−0.7**	**−1.7***	**−1.0***
	-2dB/yr	**+2.7***	**+4.5***	**+5.8***	**+1.8***	**+3.1***	+1.3
10th visit	-3dB/yr	**+1.1***	**+0.9**	**+1.4***	−0.2	+0.3	+0.5
(5 years)	-4dB/yr	+0.2	+0.1	+0.2	−0.1	+0.0	+0.1
	Stable	−0.4	**−0.6**	**−1.5***	−0.2	**−1.0***	**−0.9***

When the differencing method is held constant in the penultimate column, there is 1-13.6% increase in sensitivity (except for the 9th and 10th visits) due to classifying sequences by mean threshold, rather than defect and eccentricity.

At the 9th and the 10th visits, there are no significant differences between methods for classifying progressive measurements except for 2 dB decline per year. For 4 dB decline per year, the differences between methods are very small due to the large difference between the baseline and the current threshold value.

4.5 Discussion

When using a method for determining whether a visual field is stable or progressing, the specificity and sensitivity both have to be taken into account. The accuracy of the EA method depends on the underlying confidence interval. If the confidence interval is wide, the EA method labels a small visual field loss as stable, reducing sensitivity. If the confidence interval is narrow, the EA method labels a stable case with high noise as progressive, reducing specificity.

In this chapter we examined the EA method with different confidence intervals. We found that building a confidence interval using test-retest differences and classifying locations by eccentricity and defect (method DT) gives wider confidence intervals than calculating the difference between the baseline and follow-up values, or stratifying by mean threshold (methods DB, MT, and MB). Hence method DT has higher specificity than the other methods, but decreased sensitivity. To confirm this finding, we applied the EA method with the four different confidence intervals to the original patient data used for building the 95% confidence intervals. In theory, the accuracy of classifying the original data should be close to 95%. The results are shown in Figure 4.6.

The two B methods for calculating the confidence interval give a specificity of about 95%. In contrast, the two T methods relying on test-retest data have a higher specificity, due to a broader confidence interval. This result is to be expected because, for methods DB and MB, the value used is the difference between the current value and the baseline value, which is the same as the quantity being compared to the confidence limit using the EA method. It seems reasonable, therefore, that a confidence interval

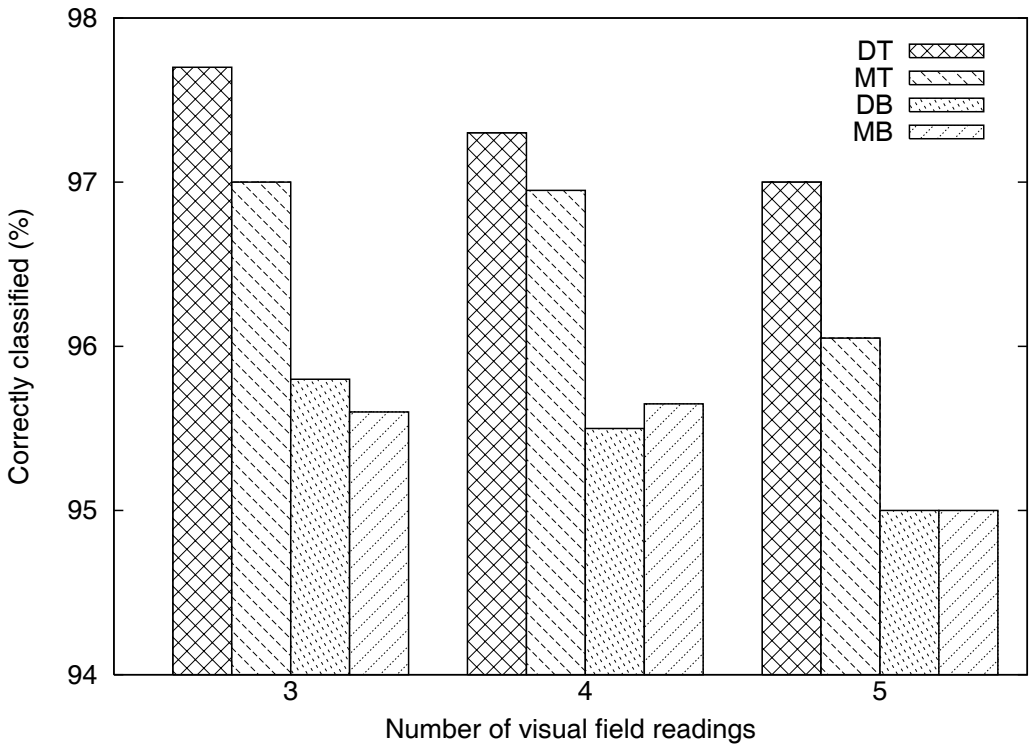


Figure 4.6: Results of classifying the original stable data using EA with various confidence intervals.

built on such values would perform closer to the 95% limit than one built on test-retest values.

The M method of classifying a point into a confidence interval class gives a higher sensitivity than the D method. This is because the first and sometimes second SAP tests carried out by a patient typically show lower thresholds than subsequent fields due to absence of learning effects (Amjad et al., 2002; Fujimoto et al., 2002). This has more effect on the DT and DB methods. In contrast, the best evaluation at each location is the mean of measurements taken in a short period. Therefore, stratifying fields based on mean threshold is more accurate.

It is difficult to compare the sensitivity and specificity of the DT method with other EA studies in the literature. Most studies use EA to classify full fields rather than single locations, and rely on a number of follow-up examinations before a classification (as progressing or stable) is made. A recent simulation study of EA has one scenario with a sensitivity of about 60% and a specificity of 99% (reading from Figure 2 of Vesti et al. (2003)). For comparison, we used the two criteria: (1) "a point is flagged as progressive at the less than 5% level if it occurs in three of three (denoted by 3of3) consecutive measurements"; and (2) "a point is flagged as progressive at the less than 5% level if it occurs in two of three (denoted by 2of3) consecutive measurements" (Vesti et al., 2003, p. 2). In other words, a point is flagged as progressive if it is outside the 95% confidence interval on three consecutive occasions (3of3) or on two of three consecutive occasions (2of3). The percentage of correct classification in a sequence over 4 years is shown in Table 4.4, which concurs with the results in Vesti et al. (2003). Again, however, their simulation applied EA to full fields and multiple locations, so the comparison is somewhat tenuous.

Table 4.4 shows that MB offers about 8.6% increase in sensitivity compared with DT , but with a 0.6% decrease in specificity when the criterion 2of3 is used. When the criterion 3of3 is used, MB increases about 12% in sensitivity compared with DT , but decreases about 0.1% in specificity. In Figure 4.5 and Table 4.4, the difference between DT and MB for specificity decreases from about 2.2% to 0.6% when 2of3 is used, and to 0.1% when 3of3 is used. This is because a point is more likely to be flagged as progressive as the number of measurements increases.

Table 4.4: Results of correct classification using two criteria. (1) A point is flagged as progressive at the 5% level if it occurs outside the 95% confidence interval in three of three consecutive measurements in a sequence (3of3); and (2) a point is flagged as progressive at the 5% level if it occurs outside the 95% confidence interval in two of three consecutive measurements in a sequence (2of3)

Criteria		*DT*	*DB*	*MT*	*MB*
3of3	Progressive (sensitivity)	54.54	56.55	62.61	67.00
	Stable (specificity)	99.31	99.18	99.34	99.20
2of3	Progressive (sensitivity)	77.42	80.09	82.88	86.07
	Stable (specificity)	98.76	98.60	98.46	98.14

When analysing whole visual fields, most methods use more than one location (two or three or four locations) to determine whole visual field progression (Vesti et al., 2003; Membrey et al., 2000; Gardiner and Crabb, 2002b,a). Therefore, if the probability of each point being falsely flagged as progression is 3%, the probability of the whole field (76 locations) being falsely flagged is not 89.5% as Gardiner pointed out. The probability of the whole field being falsely flagged can be calculated as follows.

Suppose that we use the criteria that the whole field is identified as progressive only if at least two locations are identified as progressive. Let B denote the event that the whole field is falsely flagged as progressive. Let A_0 denote the event of all locations being stable, Let A_1 denote the event of only one location being progressive. Therefore a patient with a stable visual field is "all locations are stable" or "only one location is progressive" $(A_0 \bigcup A_1)$. Assuming event A_0 and A_1 to be independent and equal in probability. Thus, $p(A_0) = p(A_1) = 0.5$.

The probability of the whole field (76 locations) being falsely flagged is $p(B)$.

$$p(B) = p(B(A_0 \bigcup A_1)) = p(BA_0) + p(BA_1) = p(A_0)p(B/A_0) + p(A_1)p(B/A_1)$$

$$
\begin{aligned}
p(B/A_0) &= p(\text{at least two locations falsely flagged as progressive}/A_0) \\
&= 1 - p(\text{at most one location being falsely flagged as progressive}/A_0) \\
&= 1 - [p(\text{no location being falsely flagged}/A_0) \\
&\quad + p(\text{only one location being falsely flagged}/A_0)] \\
&= 1 - [(1-0.03)^{74} + 74 \times (1-0.03)^{73} \times 0.03] \\
&= 1 - (0.105 + 0.240) \\
&= 0.655
\end{aligned}
$$

$$
\begin{aligned}
p(B/A_1) &= p(\text{at least one location is falsely flagged as progressive}/A_1) \\
&= 1 - p(\text{no location being falsely flagged}/A_1) \\
&= 1 - (1-0.03)^{73} \\
&= 0.892
\end{aligned}
$$

$$P(B) = 0.5 \times 0.655 + 0.5 \times 0.892 = 77.3\%$$

The probability of the whole field (76 locations) being falsely flagged is reduced to 77.3% (from 89.5%) when the criterion "at least two locations" is used. If using three or more progressive locations to determine whole field progression, the probability of a false flag can be further reduced.

4.6 Conclusion

Given that the EA method treats locations independently, and does not exploit any spatial correlation between locations, it seems reasonable to examine the performance of EA on individual sequences of measurements, rather than on whole visual fields. The study presented in this chapter appears to be the first to use this approach to rigorously examine the sensitivity and specificity of the EA method.

The study has demonstrated that using confidence intervals stratified by mean threshold, and derived from baseline-less-follow-up differences rather than test-retest

differences, yields a test for progression with approximately 2.5-12.4% higher sensitivity than other EA methods in the literature. The cost is a reduction in specificity by about 2.2%.

Chapter 5

Matching Techniques

Chapter 2 reviewed current techniques for classification of glaucomatous visual field loss, including Global indices, scoring systems, event analysis, trend analysis, and machine learning classifiers. In this chapter, we investigate matching techniques applied to glaucomatous visual fields, and then classify visual field loss.

5.1 Introduction

Matching techniques have been widely used in applications such as text retrieval, DNA analysis, signal processing, anomaly detection for computer security, and music matching (Lane and Brodley, 1997; Navarro, 2001). Two well known applications are genome processing and object matching/recognition. In the case of genome sequence processing, the data is in the form of sequences of the letters C, G, T, and A, representing the four nucleotide bases found in DNA. Some applications such as shape matching involve approximate string matching. In shape matching, to retrieve similar shapes (or images) from a database, the user can either input a shape and ask the system to find all shapes similar to it, or sketch the boundary of the desired object. After computation and comparison, the system assigns a matching value to every shape in the database which is similar to the input, and shows the first n matches. The candidates are those shapes that fall within a certain range of the input one (Mokhtarian et al., 1997). In general, for a given query pattern, the following are true for any matching technique.

- A database in that domain is absolutely necessary for searching.

- The matching approach can be exact or inexact for query processing. For sequence matching of DNA letters, for example, matching includes *insert*, *delete*, and *replace* operations to calculate the "errors".
- All matches are found that have up to "*k errors*" permitted. A scoring approach is used to measure how similar patterns are.

The general approach we adopt, therefore, is to have a database of pre-classified sequences of visual field measurements. We then classify a test sequence in the same way as the closest matching sequences in the database.

We have attempted to investigate matching techniques applied to glaucomatous visual fields. However, this problem is complex. First, the data obtained by using Standard Automated Perimetry is noisy, which makes the correct diagnosis of glaucoma and the detection of progression difficult. Second, there is no universally accepted standard for detecting progression. Different clinical trials use different definitions of "progression". The rate of agreement of detecting visual field progression between methods is not very high, 22.4% and 35.5%, as Vesti et al. (2003) found. Hence, we decided that (1) the reference database used for matching techniques excludes progressive sequences; (2) sequences collected for constructing the database are normal visual fields, and stable glaucomatous visual fields tested in a short period (test-retest). That is, the database only contains stable sequences, and no declining (natural or diseased) sequences.

We propose two different matching methods, both of which utilise *cut-off* scores to optimise sensitivity and specificity of classification. One is the weighted Sequence Matching (SM) method which uses weighted distance functions to find the best matches. The other one is the Baseline Matching Stable Sequence (BMS) method which uses the baseline in a query sequence to match the means of sequences in the database. The best matches then are used to classify the query sequence.

This chapter is organized as follows. Section 5.2 briefly describes the datasets used in sequence matching. Section 5.3 describes sequence matching methods that we have explored. The experiments and results are described in Section 5.4, and the discussion and summary are presented in Section 5.5 and 5.6 respectively.

5.2 Datasets

To examine the effectiveness of proposed sequence matching methods, we need a reference database for matching, and a test dataset for measuring the effectiveness of the methods in classifying stable and changing data. The first one is real patient data, described in Section 4.2.1. All visual fields are stable. Values at each location of the visual field are taken as a sequence over five follow-up measurements. All sequences ($50 \times 74 = 3700$) are combined to form a reference database R.

The other dataset is synthetic and contains not only progressive sequences, but also stable sequences. This synthetic dataset is generated by using the program described in Chapter 3 which is similar to that used by Spry et al. (2000). This provides sequences with stable, linear, bilinear, and convex exponential degradation over time. Each sequence in the synthetic dataset is used as a query sequence to look for matches in the database R. The matches are then used to classify whether a given query sequence is stable or progressive.

Because we will discuss sequence matching methods for detecting change at individual locations, the number of elements in a query sequence is the same as in stable sequences in R. Given an initial threshold a and a final threshold b, five middle values were interpolated by the simulation program to obtain the test sequence a, x_0, x_1, x_2, x_3, x_4, x_5, b. The frequency of follow-up for measurements was once per year. An age-related decline of 0.1 dB per year was added to each sequence of measurements. Finally noise (short- and long-term fluctuation) was added to each interpolated value. x_0 is equal to a before noise is added. Noise was not added into the initial a and final b because values a and b were removed from each sequence.

As described in Chapter 3 and 4, 3330 progressive and 3700 stable, sequences are used as query sequences to evaluate the matching methods.

5.3 Methods

The proposed matching methods are described in detail in this section. Figure 5.1 shows a schematic overview of the sequence matching techniques.

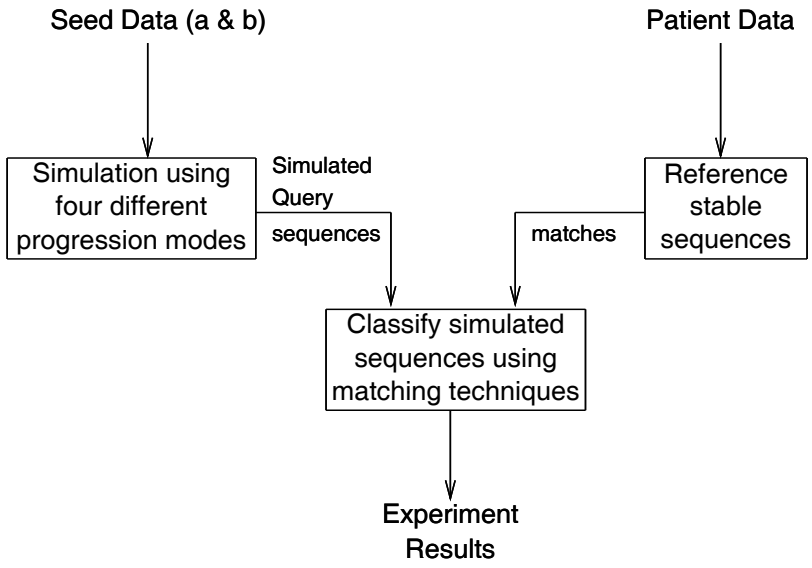


Figure 5.1: Schematic overview of sequence matching techniques.

5.3.1 Weighted Sequence Matching (*SM*)

Let $X = \{x_1, x_2, \ldots, x_N\}$ be a sequence of measurements for one location, where x_i is the i^{th} measurement, and N is the number of observations. For one query sequence, we use a sequence matching method to choose the most closely matched sequences in the database R. If the query sequence is stable, the i^{th} measurement x_i should fall within the interval calculated using all of the i^{th} measurement values in the matched sequences.

Let $R = \{S_1, S_2, \ldots, S_n\}$ denote the database R where $n = 3700$. $S = \{y_1, y_2, \ldots, y_N\}$ is a sequence of threshold values for one location of either a normal or stable glaucomatous eye. Let $Q = \{x_1, x_2, \ldots, x_N\}$ be a given query sequence.

A distance function is an important tool for choosing which sequence in R is close to a given query sequence. There are many distance functions, such as Euclidean Distance and Manhattan Distance.

Euclidean Distance is a simple measure of distance between two points

$X = \{x_1, x_2, \ldots, x_N\}$ and $Y = \{y_1, y_2, \ldots, y_N\}$, and is given by

$$d_{Euclidean}(Q, S) = \sqrt{\sum_{i=1}^{N-1} (x_i - y_i)^2} \tag{5.1}$$

The square root is often not computed in practice to save computation time, because measurement ordering is preserved.

Manhattan Distance is given by

$$d_{Manhattan}(Q, S) = \sum_{i=1}^{N-1} |x_i - y_i| \tag{5.2}$$

Manhattan distance requires less computation than the $d_{Euclidean}$.

In practice, weighted distance functions may be more effective for selecting matches when some information is emphasized.

The weighted Euclidean Distance function is given by

$$d_{Euclidean}(Q, S) = \sqrt{\sum_{i=1}^{N-1} w_i \times (x_i - y_i)^2} \tag{5.3}$$

The weighted Manhattan Distance function is given by

$$d_{Manhattan}(Q, S) = \sum_{i=1}^{N-1} w_i \times |x_i - y_i| \tag{5.4}$$

where the sum of all weights in the Manhattan or Euclidean Distance function equals 1.

It is difficult to find exact matches for a given query sequence in glaucomatous data. The distance function is therefore used to find the closest matching sequences S_{Q_1}, S_{Q_2}, ..., S_{Q_k}. For example, suppose a stable location has a baseline value of 20. The 95% confidence interval is [14, 24] read from Turpin and McKendrick, 2005, p. 3283. If this location is measured five times and each measured value is between 14 and 24, then x_i $(i = 1, 2, 3, 4, 5)$ can be one of 10 possible values. If the first value of a sequence is 20 dB, there are 4^{10} (= 1,048,576) possible stable sequences for this location. The measured sequence x_1, x_2, x_3, x_4, x_5 is one of these 1048576 different stable sequences. Clearly, the database R is not large enough to contain exact matches for all possible query sequences. We therefore used the distance function to select the most closely matched sequences for a given query sequence.

For a given query sequence, we used an approximate matching method using a fixed distance cut-off. For each matched sequence $S_{Q_i} = \{y_1^i, y_2^i, \ldots, y_N^i\}\ \forall i \in \{1, 2, \ldots, k\}$, we extracted the last value y_N^i from each S_{Q_i}. All $y_N^1, y_N^2, \ldots, y_N^k$ were then sorted in ascending order. Finally, the 2.5% and 97.5% percentiles in this range were calculated to form a 95% interval. If the Nth measurement value x_N in Q is equal to or greater than the lower limit of the interval, the query sequence Q is said to be stable. Otherwise Q is said to be progressive. This method, weighted sequence matching, is referred to as *SM* in this thesis.

Pre-processing Simulated Sequences

To examine the sequence matching method described above, query sequences are pre-processed. This is because (1) each sequence in R is known to be normal or stable, and a given query sequence Q can be progressive or stable; (2) in the *SM* method, the distance cut-off between the query and reference sequence in R is used to choose the closest sequences. The distance between a stable query sequence and reference sequences is smaller than the distance between a progressive query sequence and reference sequences. Therefore query sequences are pre-processed using a linear regression function. Note that in linear regression analysis, a sequence with a slope $<$ -1/year is considered progressive, provided that it is significant ($\alpha < 0.05$ or 0.01). Here only the slope of a sequence is used. The steps are as follows

1. Calculate the slope of each sequence of measurements in the simulated visual fields using univariable linear regression, regardless of significance level α.

2. If the slope of a query sequence $Q = \{x_1, x_2, \ldots, x_N\}$ is less than -1 (that is, slope $\leq$ -1/year), the values x_i $i = 1, 2, \ldots, N$ were transformed by a rotation and translation of coordinates (see Figure 5.2 (a) and (b)). The rotating angle of the coordinate system was fixed by the slope. That is, the axes were rotated clockwise, such that the slope became horizontal. The corresponding values of x_i' $i = 1, 2, \ldots, N$ in the new coordinates were calculated by the following formula:

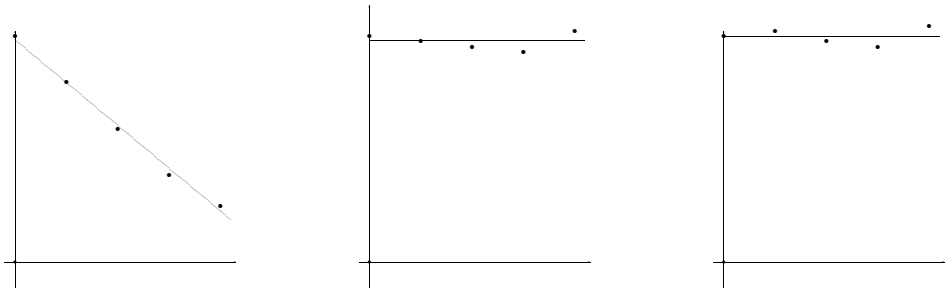
$$x_i' = \frac{1}{\sqrt{1 + slope^2}} \times ((i-1) \times (-slope) + x_i) \qquad (5.5)$$

$$\text{i} = 1,\ 2,\ \ldots,\ \text{N}$$

To avoid altering the initial value, the formula used to compute the values was replaced by

$$x'_i = (i-1) \times (-slope) + x_i \tag{5.6}$$

$$\text{i} = 1, 2, \ldots, \text{N}$$



(a) (b) (c)

Figure 5.2: Rotation of axes. (a) A decreasing sequence in the original coordinates. (b) The decreasing sequence transformed into the new coordinates. (c) A stable sequence which remains the same in both the original coordinates and the new coordinates

Let $Q' = \{x'_1, x'_2, \ldots, x'_N\}$ denote the pre-processed sequence Q. As an example, assume that the query sequence Q is 18, 16, 14, 12, 11 with slope -1.8, then the corresponding pre-processed sequence Q' will be 18, 17.8, 17.6 17.4, 18.2. If the slope of a sequence is greater than -1, the query sequence Q is unchanged. That is $Q' = Q$ (see Figure 5.2 (c)).

3. The sequence Q' was then used as the pattern to look for matches, as described before Subsection of "Pre-processing Simulated Sequences". The 95% interval $[CI_l, CI_u]$ was formed by calculating matched sequences for Q' and was used to classify the query sequence Q. If x_N which is in Q is less than CI_l, the query sequence Q is progressive. Otherwise it is stable.

5.3.2 Baseline Matching Stable Sequences (*BMS*)

In this section, we introduce the baseline matching stable sequences (*BMS*) method. The method uses the baseline of a query sequence to match stable sequences in the reference database R. A similarity measure D is calculated as follows:

$$D_{distance} = |baseline - \bar{S}_i| \leq \text{cut-off} \tag{5.7}$$

where $\bar{S}_i$ is the average of five visual field measurements from the stable sequence $S_i = \{y_1, y_2, \ldots, y_N\}$, and the *baseline* is from Q. This function is used to choose the best matching reference sequences in R for the query sequence Q.

The rationale behind this method is that the baseline value for each query sequence Q is an observation of the initial condition, and is the comparison condition for future follow-up examinations (Anderson and Patella, 1999). For a stable sequence $S_i = \{y_{i1}, y_{i2}, \ldots, y_{i5}\}$, $\bar{S}_i$ is an unbiased estimate of the population mean at the corresponding location. Therefore, $|baseline - \bar{S}_i|$ indicates that, for a given query sequence Q, we selected some stable sequences which are closest to the baseline of Q. The degree of similarity depends on the cut-off.

For a given cut-off and query Q, we collected all measurements from matched sequences and sorted them from the smallest to the largest. The 2.5% and 97.5% percentiles in this range were calculated to form a 95% interval. If the N^{th} value x_N in Q fell into the 95% interval, the query sequence was said to be stable. If x_N was less than the lower limit of the interval, the query sequence was said to be progressing. Otherwise, the sequence was improving. Improving and stable sequences were classified together as non-progressing.

Note that the confidence interval built using propsed method *BMS* differs from the confidence interval MB in Chapter 4. Here we only used measurement values in matches (that is test-retest sequences) to build a confidence interval. In Chapter 4, the differences in test-retest sequences were used to build the confidence interval.

5.4 Experiments and Results

5.4.1 Experiments and Results using *SM*

In *SM*, many kinds of weights can be used. As the baseline visual field examination(s) is the condition established at the beginning of the follow-up period, we emphasized the beginning measurements as the more important information. Thus, we used the weights based on the geometric series: $\frac{q^{N-1}}{d}, \frac{q^{N-2}}{d}, \ldots, \frac{q^0}{d}$, $(d = \sum_{i=0}^{N-1} q^i,\ q > 0)$ and arithmetic series: $\frac{N}{d}, \frac{N-1}{d}, \ldots, \frac{1}{d}$ $(d = \sum_{i=1}^{N} i,\ )$ in formula (5.4) and (5.3). We found that when the "distance" (or cut-off) was fixed, the accuracy of classifying stable visual fields gradually increased as the base of the geometric series increased. If the base of the geometric series was fixed, the accuracy of the classification of the stable visual fields increased progressively as the distance cut-off gradually increased. The gradual increase of distance cut-off led to a greater number of matched sequences. Increased sequences gave a wide 95% interval described in Section 6.3.1. When a wide interval is used, accuracy decreases for classifying progressive sequences, but increases for classifying stable sequences. For the balance of sensitivity and specificity, the distance cut-off was taken from 1.5 to 3.0. Figure 5.3 and 5.4 show the results of classifying both the progressive and stable sequences by using the Euclidean Distance function, based on the different weights.

When we used the weighted Manhattan Distance, the distance cut-off is not the same as the one used for Euclidean Distance. This is because the relationship between the weighted Manhattan and the weighted Euclidean functions is

$$\sum_{i=1}^{n}(w_i|a_i|) \leq \sqrt{\sum_{i=1}^{n} w_i \times (a_i)^2} \tag{5.8}$$

where $\sum_{i=1}^{n} w_i = 1$. The accuracy of classifying stable and progressive sequences by using Manhattan Distance is not higher than that using Euclidean Distance. Table 5.1 shows the results when weighted functions are based on the arithmetic series, with distance cut-off being 2.1 and 3.0 for Manhattan and Euclidean respectively. There is little difference between these two distance functions. Therefore we focus on Euclidean Distance function from this point onwards.

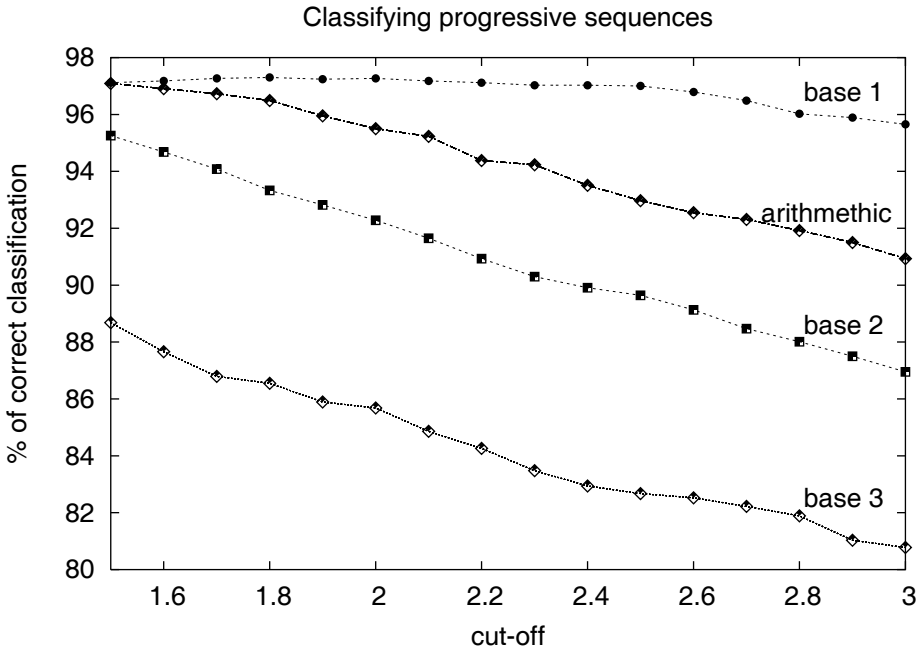


Figure 5.3: The performance of the Euclidean Distance function in sequence matching to classify the 5^{th} measurement in a progressive sequence. The arithmetic and geometric series used bases 1, 2, 3. The selection of the distance cut-off was from 1.5 to 3.0.

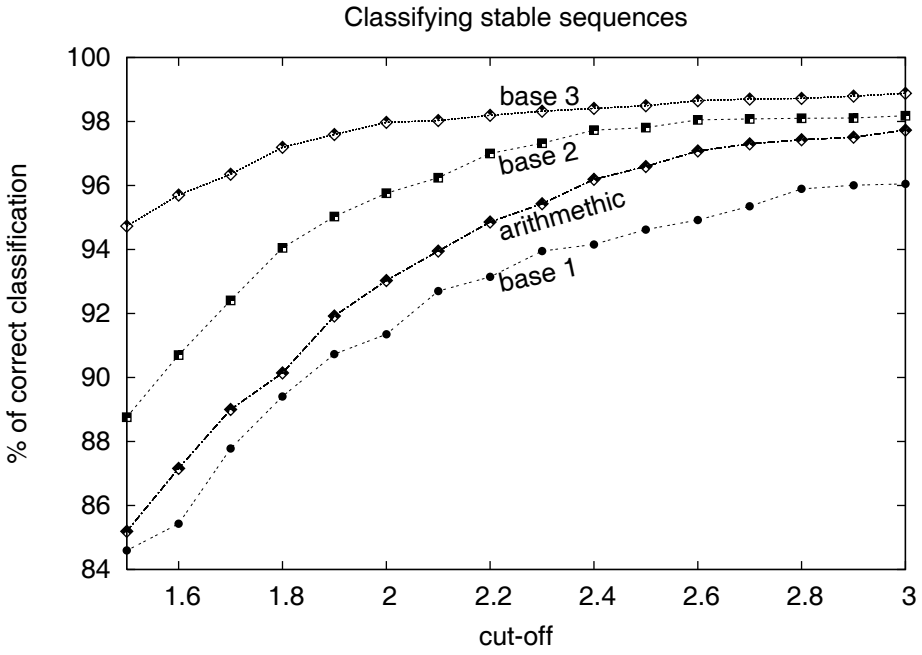


Figure 5.4: The performance of the Euclidean Distance function in sequence matching to detect 5th measurement in a stable sequence. The arithmetic and geometric series used bases 1, 2, 3. The selection of the distance cut-off was from 1.5 to 3.0.

Table 5.1: Comparison between Manhattan and Euclidean metrics at specific distance cut-off points.

Function	Class	3rd VF(%)	4th VF(%)	5th VF(%)
Manhattan	Sensitivity	58.83	81.17	89.43
Manhattan	Specificity	95.95	96.00	97.70
Euclidean	Sensitivity	58.17	81.68	90.93
Euclidean	Specificity	96.19	96.65	97.89

We took the Euclidean Distance function with a cut-off 3.0, and compared it with the EA method (MB from Section 4.3.1 in Chapter 4). As a comparison for each location, we also used the following EA criteria: (a) for a given location, if progression was identified in two of the three consecutive fields (denoted by 2of3), the location was said to be progressing; or (b) in three of the three consecutive fields (denoted by 3of3) (Vesti et al., 2003). After five measurements were used, the results of using the EA criteria with confidence interval MB described in Chapter 4 are shown in Figure 5.5.

Figure 5.5 shows that the sequence matching method increases accuracy, both in sensitivity and specificity. The only exception is the 3rd measurement $SM(3)$ where a reduced specificity was obtained when compared with EA (EA(3)). When criteria 2of3 and 3of3 were used, the matching method SM improves classification for both stable and progressive sequences. Cochran's Q-test (described in Chapter 4) revealed a significant difference between the methods ($\alpha < 0.05$), and so comparisons using Q-tests were conducted and the results are given in Table 5.2.

In Table 5.2, results in the first column show the difference between percentages of correct classification for $2dB/year$ decline. The second column show the difference between percentages of correct classification for $3dB/year$ decline and so on. Results show that the method SM achieves statistical significance for all progressive sequences, and all stable sequences except for the $3rd$ row (last column).

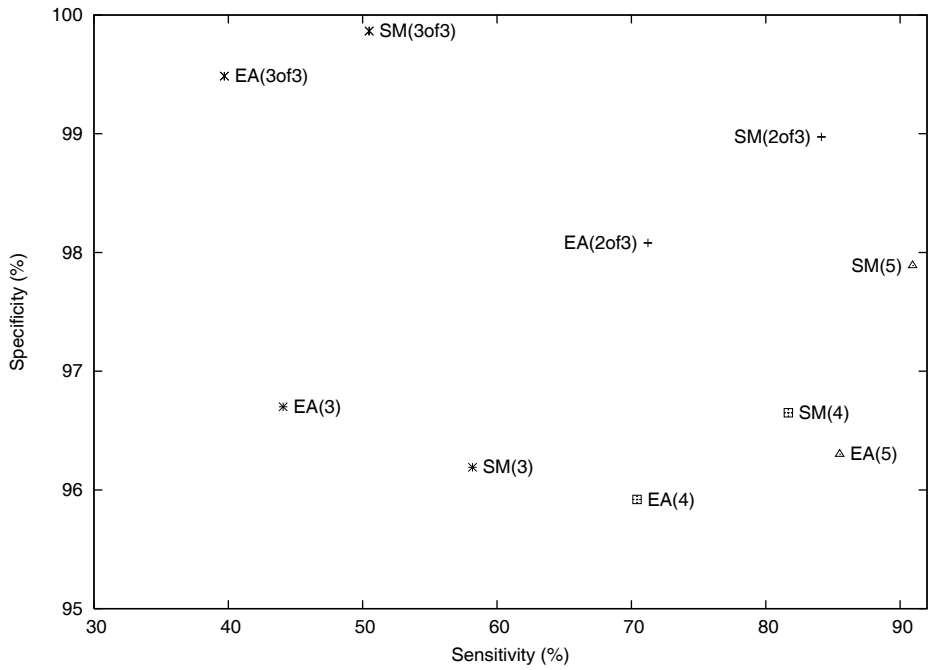


Figure 5.5: Sensitivity and specificity of the EA method and the sequence matching method with arithmetic series and with distance cut-off 3.0. EA(i) indicates that EA was only applied on the i^{th} measurement ($i = 3, 4, 5$). SM(i) represents that sequence matching was only applied on the i^{th} measurement ($i = 3, 4, 5$). EA(2of3) or EA(3of3) and SM(2of3) or SM(3of3), indicate that the criterion 2of3 or 3of3 was used.

5.4.2 Experiments and Results using *BMS*

Experiments

In the experiments of baseline matching stable sequences, the first interpolated field was taken as the baseline for each query. That is, for a query sequence $Q = \{x_1, x_2, \ldots, x_N\}$, x_1 was the baseline. We restricted the baseline values to lie between 16 and 31 dB, because the stable database included only a small number of sequences outside this range. Figure 5.6 shows the distribution of means of sequences in the database R. This represents the combined normal and test-retest datasets, shown in Figures 4.2(a) in Chapter 4.

The evaluation of an interval requires an adequate sample size; otherwise the classification method may not be reliable (Hughes and Grawoig, 1971). The formula for calculating the required sample size n is:

Table 5.2: Percentage difference in correct classification between methods. Statistically significant change is shown in bold ($\alpha < 0.05$). *SM* is the weighted matching method. The EA method with confidence interval *MB* is described in Chapter 4.

SM − MB				
Conditions	-2 dB	-3 dB	-4 dB	Stable
2of3	**+14.4**	**+14.1**	**+10.3**	**+0.9**
3of3	**+2.6**	**+11.2**	**+18.6**	**+0.4**
3*rd*	**+10.5**	**+13.9**	**+17.9**	-0.5
4*th*	**+12.3**	**+11.7**	**+9.8**	**+0.7**
5*th*	**+7.4**	**+5.1**	**+3.8**	**+1.6**

$$n = Z^2 \times \sigma^2 / E^2 \tag{5.9}$$

where E is the maximum allowable error (the difference between the population mean and the sample mean). Z is obtained by using the given confidence interval coefficient α. σ is the population standard deviation. For this study, $E = 1$, and $Z = 1.96$ computed by using $\alpha = 0.95$. σ is computed using Henson et al. (2000) standard deviation

$$log_e(SD) = -0.081 \times dB + 3.27$$

In Equation 5.7, we used values between 0.1 and 1.0 as the cut-off to select matches in R.

Results for Simulated Dataset

Figures 5.7 and 5.8 show the results obtained using the new method *BMS* on the simulated datasets for classifying the 3*rd*, 4*th* and 5*th* measurements at different cut-offs. As the cut-off increases, the accuracy of classification also increases for stable sequences, while it decreases for progressing sequences. This is because when the cut-off is less than 0.4, some query sequences cannot get sufficient matched values in R as required by Equation 5.9. For these sequences, classification is not reliable. On the other hand, when the cut-off is greater than 0.6, query sequences may obtain enough values from the matched stable sequences. However, the means of some matched stable sequences are closer to the integer "baseline + 1" or "baseline - 1". [1] When

[1] $|baseline - mean| \leq$ cut-off $\Rightarrow$ baseline - (cut-off) $\leq$ mean $\leq$ baseline + (cut-off)

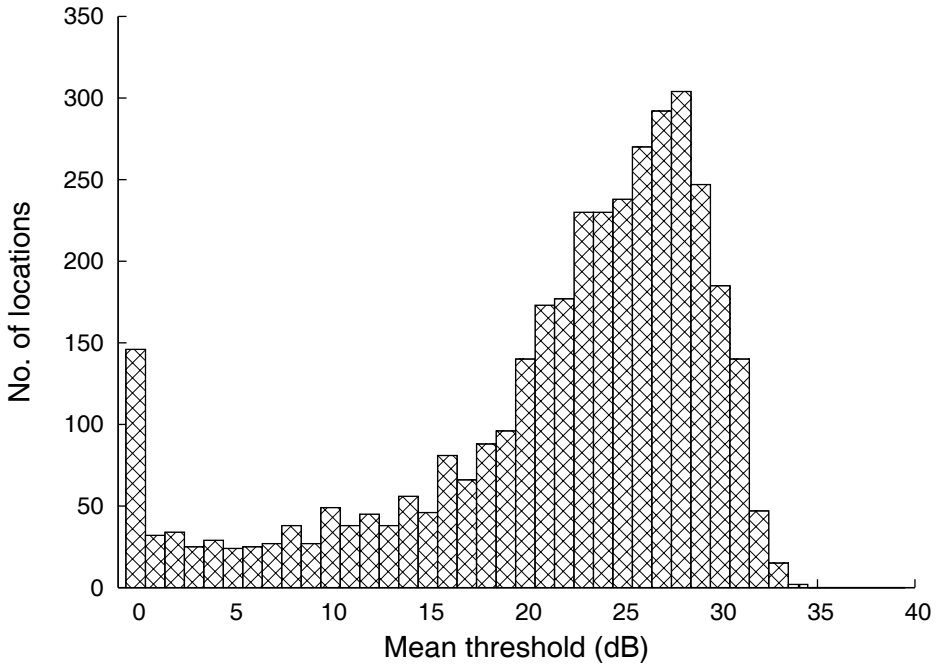


Figure 5.6: The distribution of means of sequences in the database R.

the cut-off is greater than 0.6, the interval formed by taking all values from matched stable sequences is too wide. When the cut-off is 0.4, the minimum number of matches to satisfy Equation 5.9 is 120 (the minimum sample size). Figures 5.7 and 5.8 show that the percentages of correct classification are similar for cut-offs between 0.4 and 0.6, for each of the 3rd, 4th and 5th measurements.

We used a cut-off of 0.5 to compare the EA methods with confidence intervals MB described in Chapter 4. We also used the EA criteria described Section 6.4.1. The results are shown in Table 5.3 using five measurements.

Table 5.3: Percentage of correct classification for baselines between 16 and 31dB. MB is confidence intervals described in Chapter 4. *BMS* is the proposed baseline matching stable sequences method.

Criterion	MB		BMS	
	P	S	P	S
2of3	74.10	98.95	82.72	97.11
3of3	41.52	99.79	52.99	99.30

In Table 5.3, "P" indicates the percentage of correct classification for progressive

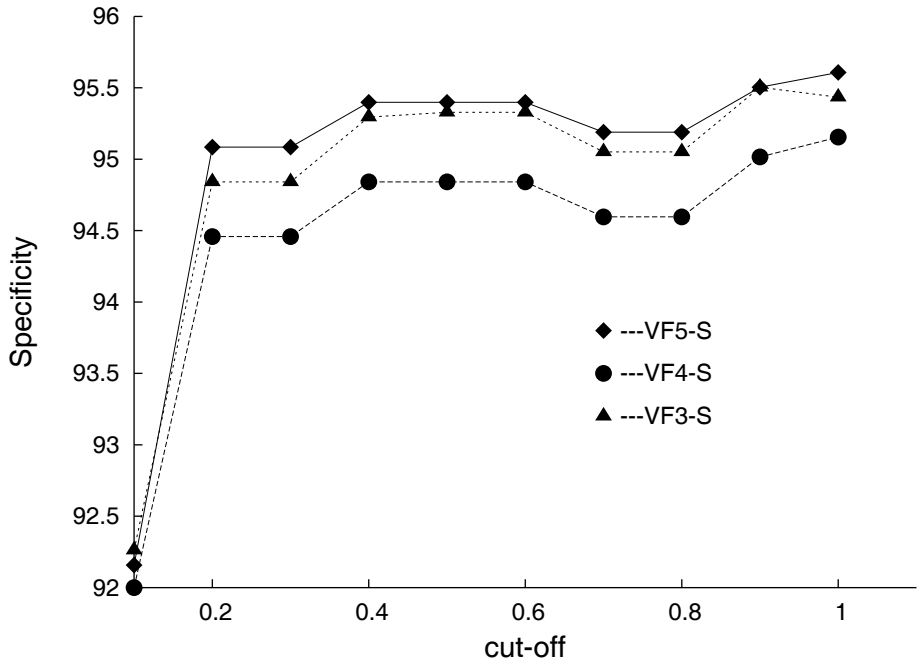


Figure 5.7: Percentage of correct classification for stable sequences (S) with baseline values between 16 and 31dB at different cut-offs, using the $3rd$, $4th$ and $5th$ measurements.

sequences, "S" is similar for stable sequences. The new method offers higher accuracy for classifying progressive sequences, while accuracy is lower for stable sequences. Cochran's Q-test (confidence level of 0.05) indicated a statistically significant difference between the new method SH and EA method. Hence, pair-wise comparisons were performed using Q-test as shown in Table 5.4.

Table 5.3 and 5.4 show that the new method BMS is better than the EA method for correctly classifying progressive sequences. When the EA criterion 2of3 was used, BMS increases accuracy by 8%, with less than 2% decrease for stable sequences. When the EA criterion 3of3 was used, BMS offers about 11% increase for progressing sequences, with about 0.5% decrease for stable sequences.

Results for Real Dataset

The proposed method BMS and the EA method using confidence and MB were also evaluated using the real dataset. The sequences in the real dataset contained 8 values instead of the 5 used for the simulated dataset. The real dataset consists of 60 progressive and 62 stable patients each with 8 visual fields. Following Vesti

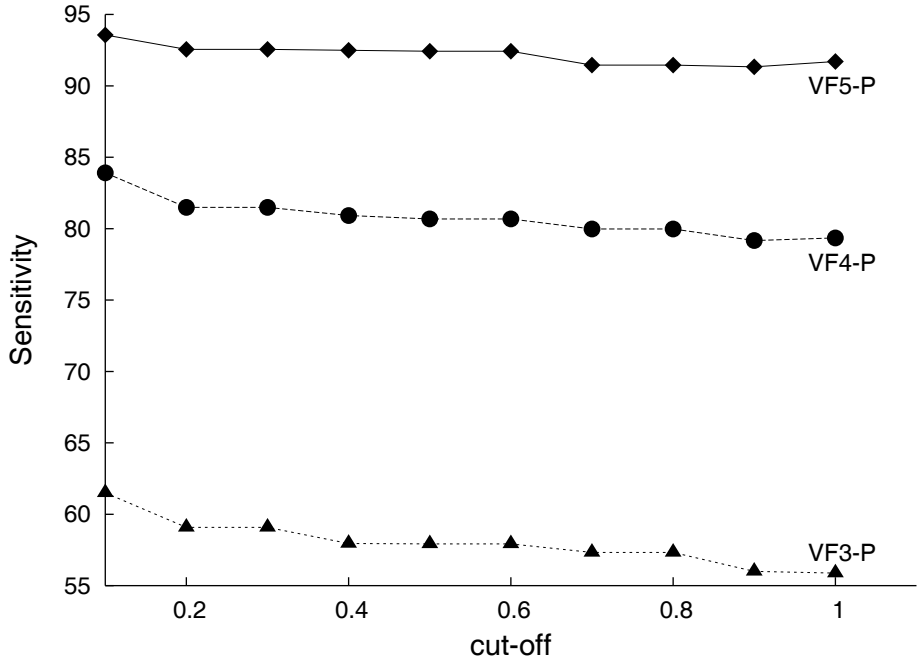


Figure 5.8: Percentage of correct classification for progressive sequences (P) with baseline values between 16 and 31dB at different cut-offs, using the $3rd$, $4th$, and $5th$ measurements.

et al. (2003), we used similar EA criteria to classify the real dataset: (i) for a given patient, if progression at four or more locations is found in two of three consecutive fields (denoted by (4, 2of3)), this patient was said to be progressing; or (ii) similarly in three of three consecutive fields (denoted by (4, 3of3)). The results are shown in Table 5.5 using the first five visual field measurements.

Tables 5.5 shows that the EA method using confidence interval MB is more specific, and BMS is more sensitive. When using the criterion (4, 2of3), BMS offers a 15% increase for correctly identifying progressive visual fields. This is accompanied by a drop of 8% for correctly identifying stable visual fields when compared with the EA method using MB. When the criterion (4, 3of3) is used, BMS offers an 6% rise in correctly detecting progressive visual fields, and a 2% decrease for stable visual fields. When compared with the results obtained from the simulated dataset, the real dataset results show a decrease in accuracy for both the stable and the progressive sequences. This, we believe, is due to the significantly larger variability in visual field sequences extracted from the real data, and the difference between classification methods based on the whole visual field versus a single location.

Table 5.4: Percentage difference (of correct classification) between methods. Statistically significant differences are shown in bold ($\alpha < 0.05$).

Criterion	Group	$BMS-MB$
2 of 3	12dB	**+13.9**
	18dB	**+ 7.2**
	24dB	**+ 3.6**
	Stable	**-1.4**
3 of 3	12dB	**+41.6**
	18dB	**+42.6**
	24dB	**+33.9**
	Stable	**- 2.1**

Table 5.5: Percentage of correct classification. MB is confidence interval described in Chapter 4. BMS is the proposed matching method.

Criterion	MB		BMS	
	P	S	P	S
(4, 2of3)	53.33	79.03	68.33	70.97
(4, 3of3)	20.00	96.77	26.67	95.16

5.5 Discussion

In this chapter, we proposed two sequence matching methods for application on glaucomatous visual field data. The method SM used weighted distance functions to select close matches. In experiments (not reported here), we tried rotating all query sequences and reference sequences in R to horizontal lines. The accuracy of correct classification is less than that by rotating only the query sequences with a slope of less or equal to -1dB/year. This is probably because if all sequences are rotated, some query sequences which are decrease may match some reference sequences which are increase. For example, given a query sequence $Q = \{25, 24, 23\}$ which is rotated to $Q' = \{25, 25, 25\}$, the reference sequence $S = \{25, 26, 27\}$ in R can be chosen because $S' = \{25, 25, 25\}$ is the rotated sequence S.

As can be seen, if less than six measurements are available, SM achieves better

classification for both progressive and stable sequences compared with the EA method. We also compared the two new methods *SM* and *BMS* on the simulated datasets, in which the baseline values are from 16 dB to 31 dB. The results are shown in Table 5.6.

Table 5.6: Comparison of the two new methods *SM* and *BMS*. "P" indicates the percentage of correct classification for progressive sequences, "S" is similar for stable sequences.

Tests	*SM*		*BMS*	
	P	*S*	*P*	*S*
(3)	63.95	95.40	57.92	95.33
(4)	84.57	97.14	80.67	94.84
(5)	94.94	97.70	92.43	95.40

The method of *SM* is not without its limitations. While *SM* offers higher accuracy for classifing both stable and progressive sequence (see table 5.6), its use is limited in that for more than five measurements in a query sequence. *SM* is not effective because the length of the reference sequences in R is only five measurements (five test-retest normal or stable visual field measurements).

The method *BMS* overcomes the limitation of *SM*. *BMS* does not depend on the number of measurements in a query sequence Q. *BMS* only depends on the baseline of a given query sequence for choosing matches. Note that when the cut-off is 0.5, the 95% confidence interval created using *BMS* differs from those created using MB as described in Chapter 4. For example, if the baseline of a given query sequence Q is 25 dB, suppose a matched sequence is $S = \{22, 26, 28, 24, 25\}$, and its mean is 25 dB. the 95% interval created using *BMS* consists of 22, 26, 28, 24, 25. That is, the 95% confidence interval is [22, 28]. Note that the differences between the minimum and 25, and the maximum and 25 are -3 and 3 dB. When building 95% interval MB (1) the mean of five measurements in a reference in R is rounded to the nearest integer; (2) all sequences in R are grouped based on the nearest integers of the means; (3) differences between baseline and follow-up measurements $(y_1 + y_2)/2 - y_i$, $3 \leq i \leq 5$ are calculated for MB. Therefore MB consists of -4, 0, -1. That is [-4, 0].

5.6 Summary

We have described applications of sequence matching to the problem of classifying change in visual field measurements. It is difficult to establish a set of reference sequences that includes progressive sequences because different classification techniques can give different results for a patient with progressive glaucoma. We have therefore focused on the use of a baseline for a given query sequence, regardless of the method used (i.e. *BMS* directly uses a baseline of a query sequence to select matches, and *SM* gives the highest weights for the first measurements). The sequence matching methods were tested with both synthetic and real datasets. The results indicate that the new method *SM* is better than the new method *BMS*. However *SM* is less effective for more than five visual field measurements. The new method *BMS* overcomes the limitation of *SM*. *BMS* can significantly improve the accuracy of identifying progressing sequences (increased sensitivity) compared with EA method using confidence interval MB, though there is a small penalty for correctly identifying stable sequences (decreased specificity).

Chapter 6

Linear Regression Analyses

6.1 Introduction

The previous chapter discussed the Event Analysis and Sequence Matching methods for determining visual field change on a point-by-point basis. In this chapter, we aim to recognise visual field loss using point-wise linear regression (PLR) analysis.

Linear regression analysis is a clinically useful tool for determining glaucomatous visual field change, including detecting the status of each location (progressing, stable or improving) and determining the rate of any change. Although linear regression is a useful tool for analysing longitudinal visual field data, there is no consensus about what value of the regression slope, or significance level α, constitutes progression. The most widely used slope and significance level α are shown in Table 2.4 in Chapter 2. For instance, Noureddin et al. (1991) required significant ($\alpha < 0.05$) slopes of -2.4 dB / per year for it to be considered progressive. Birch et al. (1995), Viswanathan et al. (1997) and M et al. (1997) used the criterion of significant ($\alpha < 0.05$) slope worse than -1.0 dB/year. A number of other investigations employed the criterion of slope less than -1.0 dB/year, together with a significance level of $\alpha < 0.01$ (Katz et al., 1997; Gardiner and Crabb, 2002a; Nouri-Mahdavi et al., 1997; Wild et al., 1997; Spry et al., 2002). Furthermore, for a fixed criterion of slope and significance level, different numbers of consecutive field measurements have been used to determine progression. For example, Katz et al. (1997), Nouri-Mahdavi et al. (1997), Wild et al. (1997) and Spry et al. (2002) used the criteria of slope less than -1.0 dB / per year and significance level of $\alpha < 0.01$ occuring once. Hitchings et al. (1994) used the criteria of slope less

than -1.0 dB / per year and significance level of $\alpha < 0.01$ occuring in two of two consecutive visual field tests for a location. Some researchers determined progression by deleting one measurement and adding a new measurement into the calculation of the linear regression function:

$$x = \beta_1 + \beta_2 t + \varepsilon \tag{6.1}$$

where β_1 and β_2 are the intercept and the slope respectively, both of which are unknown constants; ε is a random error component.

In this study we propose two PLR methods for classification of glaucomatous visual field measurements. The purposes of the methods are (1) to explore whether omitting values at different positions in a sequence produces different results; (2) to investigate the effect of omitting the value that represents the maximum deviation from a fitted line.

The rest of this chapter is organized as follows. Section 6.2 briefly describes the simulated data sets, which differ from the datasets used in Chapter 4. Section 6.3 describes the new methods and discusses the techniques that we have explored in experiments. The results are summarised in Section 6.4. The discussion and summary are presented in Section 6.5 and 6.6 respectively.

6.2 Datasets

The data generated is similar to that used by Gardiner and Crabb (2002a), who simulated vision loss and grouped visual fields as a pool to be analysed. We group simulated sequences as a pool to be analysed. Like Gardiner and Crabb (2002a) used, we adopt Function 3.12 in Chapter 3 as standard deviation in a normal distribution to randomly generate sequences.

Given the initial and final thresholds a and b for a single location in the visual field, the simulation program described in Chapter 3 was used to interpolate the middle 9 values. For convenience of explanation, we use the series x_1, x_2, ..., x_{10}, x_{11} to describe the input seeds and interpolated values before age-related decline and noise were added. (Here $x_1 = a$ and $x_{11} = b$). Note that $x_1 \neq x_2$ for a progressive

sequence, because the linear regression method does not require the baseline value. The frequency of follow-up for subsequent tests was twice per year. An age-related decline of 0.1 dB per year was added to each series of measurements. That is $x_i + 0.05 \times (i-1)$. Finally, noise was added to each value ($x_i + 0.05 \times (i-1)$) by using a normal distribution function in which the standard deviation is calculated as follows:

$$log_e(SD) = A \times threshold(dB) + B = A \times (x_i + 0.05 \times (i-1)) + B \qquad (6.2)$$

where the constants A and B are -0.081 and 3.27, respectively. Note that Equation 6.2 is a decreasing function as the threshold increases. When the threshold is less than 14, the SD is larger than 9 dB (when the threshold is 13 dB, $SD = 9.18$). In the experiment, if a threshold was below 14 dB, 9 dB was used as the standard deviation SD instead of the value calculated using Equation 6.2 as discussed in Chapter 3.

The process of simulation was similar to that described in Chapter 4. We generated 3330 progressive sequences of 11 values. The starting (or initial) value of each sequence was chosen from 15 normal visual fields in the stable database. Three final values are obtained from each starting value, by subtracting 10, 15 or 20 dB, with a minimum possible value of 0 dB. This gives an average decrease of 2 dB, 3 dB or 4 dB per year for 5 years. For example, if the starting value is 28 dB, then the 3 final values obtained are 18 dB, 13 dB and 8 dB respectively. That is, for the starting value of 28, the three pairs of initial and final values are: (28, 18), (28, 13), and (28, 8) which are used to generate three progressive sequences.

In each case, the 1110 (15×74 locations) sequences had final values that were 10 dB less than the initial values of the sequences, 1110 had final values that were 15 dB less, and 1110 that were 20 dB less. This gives an average decrease of 2 dB, 3 dB or 4 dB per year for 5 years. Note that in sequences with values that dropped below zero, all negative values were replaced by zeros (equivalent to total blindness for that location).

To simulate stable sequences, the first visual field from each of the 50 real patients was used as both the initial and the final values in the simulation. The middle 9 values were generated as for the linear case in the progression simulation, with age-related decline and noise.

In total, we simulated 3330 progressive locations and 3700 stable locations, each with 11 threshold measurements (including the initial and the final seed values from which the other 9 measurements were derived).

6.3 Methods

6.3.1 Point-wise Linear Regression Analysis

This section describes six different point-wise linear regression (PLR) analyses. Let pairs $(t_1,\ x_1)$, $(t_2,\ x_2)$, $\ldots$, $(t_n,\ x_n)$ be a sequence of measurements against time for a location X in an eye, where n is the number of measurements, and x_i is the measurement value at time t_i.

A linear regression model $x = \beta_0 + \beta_1 t + \varepsilon$, is estimated from $(t_1,\ x_1)$, $(t_2,\ x_2)$, $\ldots$, $(t_n,\ x_n)$ as described by Equations 2.6

$$\hat{\beta}_0 = \bar{x} - \hat{\beta}_1 \bar{t}$$

and 2.7

$$\hat{\beta}_1 = \frac{S_{tx}}{S_{tt}} \tag{6.3}$$

in Chapter 2.

When a significance level α is given, the slope $\hat{\beta}_1$ of the fitted line is tested to see whether it is statistically significant (in other words, whether there is a strong correlation between t_i and x_i) by using the t-distribution with (n-2) degree of freedom. The value of t_v is calculated as

$$t_v = \frac{\hat{\beta}_2}{Q}\sqrt{(n-2)S_{xx}} \tag{6.4}$$

where $Q = \sqrt{S_{yy} - \hat{b}S_{xy}}$ and $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$. A critical value $(t_{\alpha/2,n-2})$ is obtained from a t-distribution with $n-2$ degrees of freedom. If $|t_v| > t_{\alpha/2,n-2}$, the linear relationship between t and x is statistically significant. Otherwise, the relationship is not statistically significant.

Note that the interval between consecutive tests in glaucoma clinics is usually half a year. Therefore pairs $(t_1,\ x_1)$, $(t_2,\ x_2)$, $\ldots$, $(t_n,\ x_n)$ can be simplified to pairs of (0,

x_1), (0.5, x_2), ..., ($(n-1)\times 0.5$, x_n). In this chapter, we use $X = x_1, x_2, ..., x_n$ instead of (0, x_1), (0.5, x_2), ..., $((n-1)\times 0.5, x_n)$ for simplifying descriptions of methods in the following paragraphs.

Gardiner and Crabb (2002a) compared several PLR methods and the relationships among them were discussed. We describe some of the methods they used, and two new methods below.

- **Basic Criterion** (denoted by *1of1* here): a location X is identified as progressing if the sequence x_1, x_2, ..., x_n satisfies the criterion of the slope being less than -1.0 dB/year at the level $\alpha < 0.01$. Many researchers have used this method to determine the status for a location (Katz et al., 1997; Nouri-Mahdavi et al., 1997; Wild et al., 1997; Spry et al., 2002; Gardiner and Crabb, 2002a).

Suppose a patient had 8 visual field tests. A line fitted by using 8 measurements for one location is shown in Figure 6.1. Given $\alpha < 0.01$, if the slope (decline rate / year) of the fitted line is less than -1.0 dB/year and the statistic $|t_v|$ is greater than the critical value ($t_{\alpha/2,6}$) obtained from a t-distribution with 6 degrees of freedom, the location is progressive. Otherwise, it is stable. This method only uses linear regression once. It is sensitive to noise because one large fluctuation can affect the rate of change.

- **Two of Two** (denoted by *2of2* here): a location X is identified as progressing if the first $n-1$ measurements x_1, x_2, ..., x_{n-1} satisfy the basic criterion, and continue to satisfy it after adding the n^{th} measurement (Hitchings et al., 1994). That is, a line fitted to x_1, x_2, ..., x_n also satisfies the basic criterion.

Suppose a patient has 8 visual field tests. The criterion of *2of2* for identifying a location is that both the first 7 measurements x_1, x_2, ..., x_7 and the 8 measurements x_1, x_2, ..., x_8 satisfy the basic criterion. See Figure 6.2.

- **Two of Three** (denoted by *2of3* here): a location X is identified as progressing if the first $n-2$ measurements x_1, x_2, ..., x_{n-2} satisfy the basic criterion, and continue to satisfy it at least once after adding either the $(n-1)^{th}$ measurement

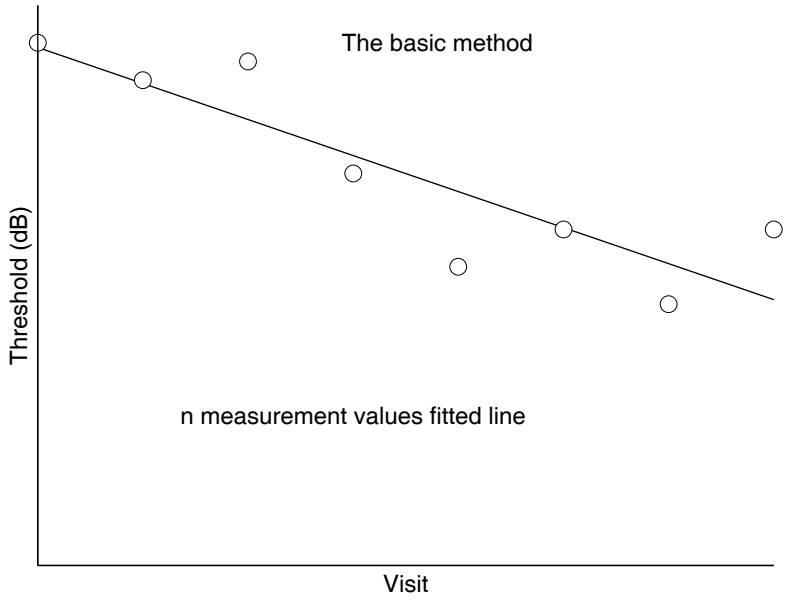


Figure 6.1: Illustration of the *1of1* criterion. The line is fitted by 8 measurement values.

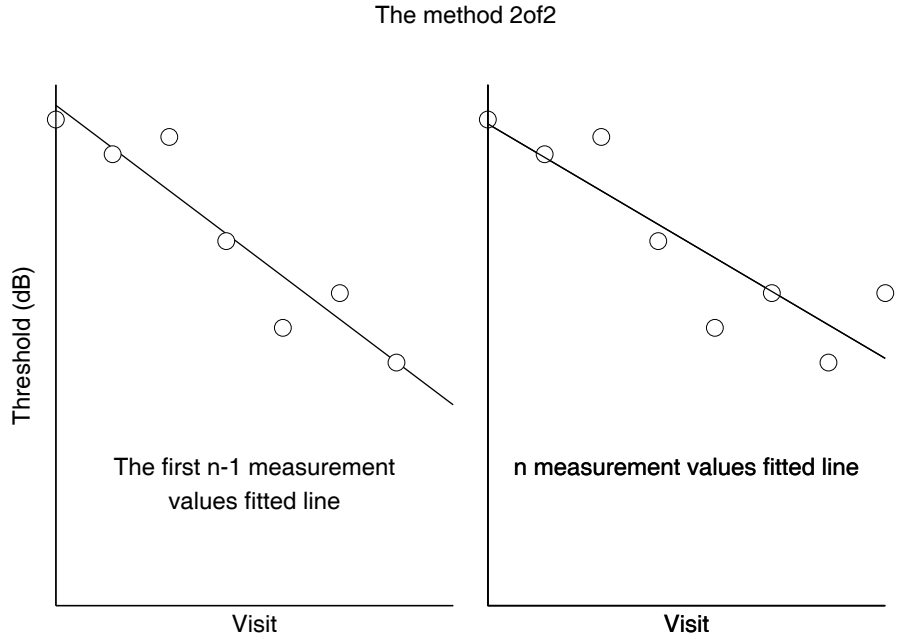


Figure 6.2: Illustration of the *2of2* criterion. The line in the left graph is fitted using the first 7 measurements. The line in the right graph is fitted by adding the 8^{th} measurement to the sequence.

x_{n-1} or the n^{th} measurement x_n. That is, not only do the $n-2$ measurements x_1, x_2, ..., x_{n-2} satisfy the basic criterion, but also at least one of the measurements x_1, x_2, ..., x_{n-2}, x_{n-1} or x_1, x_2, ..., x_{n-2}, x_n also satisfies the basic criterion (Membrey et al., 2000).

Figure 6.3 shows the *2of3* method. The line in the left graph satisfies the basic criterion, and the line in the middle graph or in the right graph or in both satisfies the basic criterion.

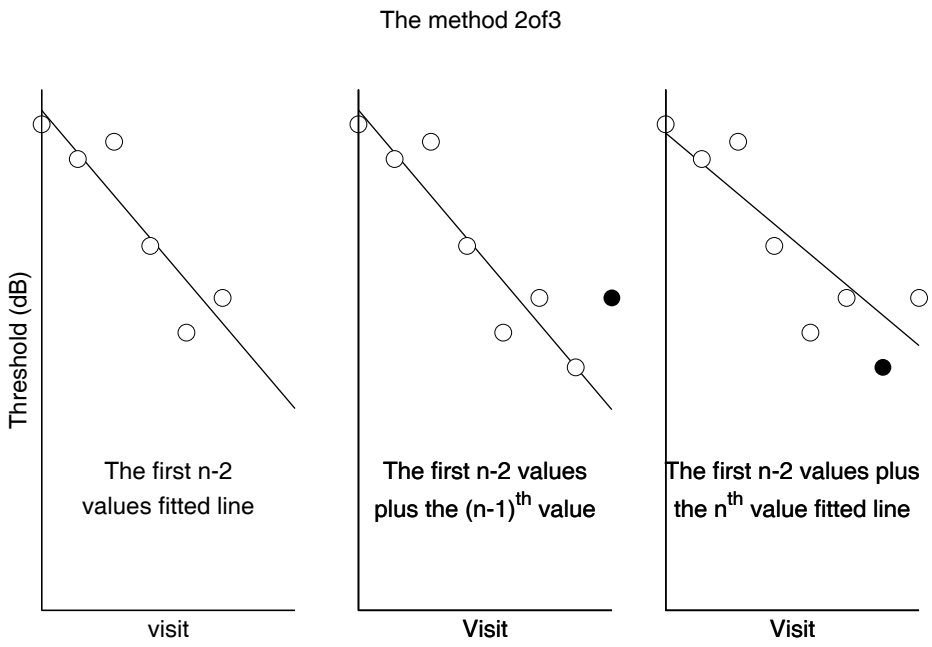


Figure 6.3: Illustration of the *2of3* criterion. The line in the left graph is fitted to the first $n-2$ measurements. The line in the middle is fitted after adding the $(n-1)^{th}$ measurement into the sequence. The line on the right is fitted after adding the n^{th} measurement into the sequence, but excluding the $(n-1)^{th}$ measurement. In each case the filled circle is the measurement excluded for deriving the line.

- **Two-Omitting** (denoted by *(n)to(n-1)* here): a location X is identified as progressing if the first $n-1$ measurements x_1, x_2, ..., x_{n-1} satisfy the basic criterion, and continue to satisfy it after adding the n^{th} measurement x_n and excluding the $(n-1)^{th}$ measurement x_{n-1} (Gardiner and Crabb, 2002a). That

is, a line fitted to x_1, x_2, ..., x_{n-2}, x_n also satisfies the basic criterion. Figure 6.4 shows the (n)to(n-1) method.

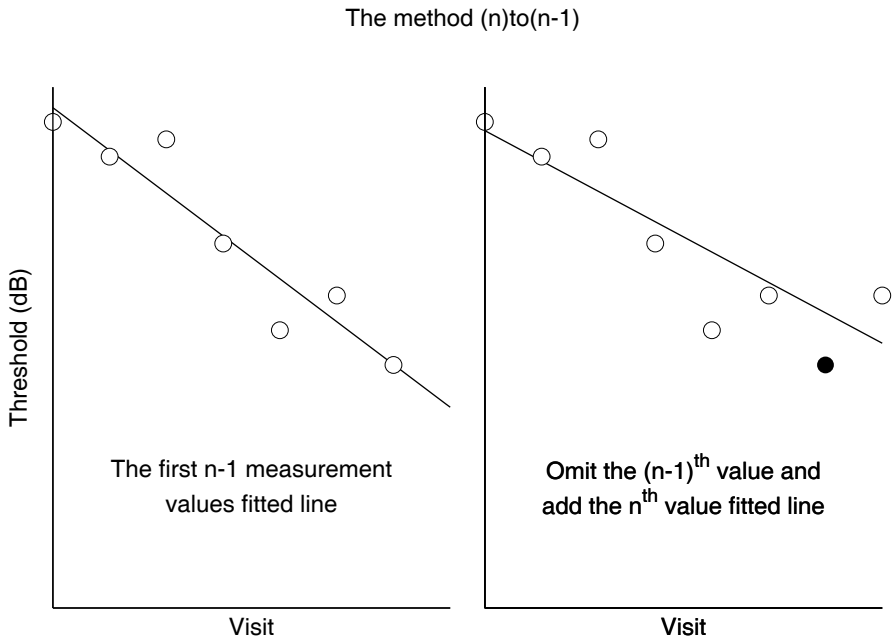


Figure 6.4: Illustration of the *(n)to(n-1)* criterion. The line in the left graph is fitted to the first 7 measurements. The line in the right graph is fitted by omitting the 7^{th} measurement x_7 and adding the 8^{th} measurement x_8. The filled circle is excluded for deriving the line.

Our two new proposed methods are described below.

1. **First-Omitting** (denoted by *(n)to1* here): a location X is identified as progressing if the first $n-1$ measurements x_1, x_2, ..., x_{n-1} satisfy the basic criterion, and continue to satisfy it after adding the n^{th} measurement x_n, but excluding the first measurement x_1. That is, a line fitted to x_2, x_3, ..., x_n also satisfies the basic criterion.

Figure 6.5 shows the *(n)to1* method. When the line fitted by using the first 7 measurements, and the line fitted by deleting the first and adding the 8^{th} measurement, both satisfy the basic criterion, the locaton X is identified as progressive.

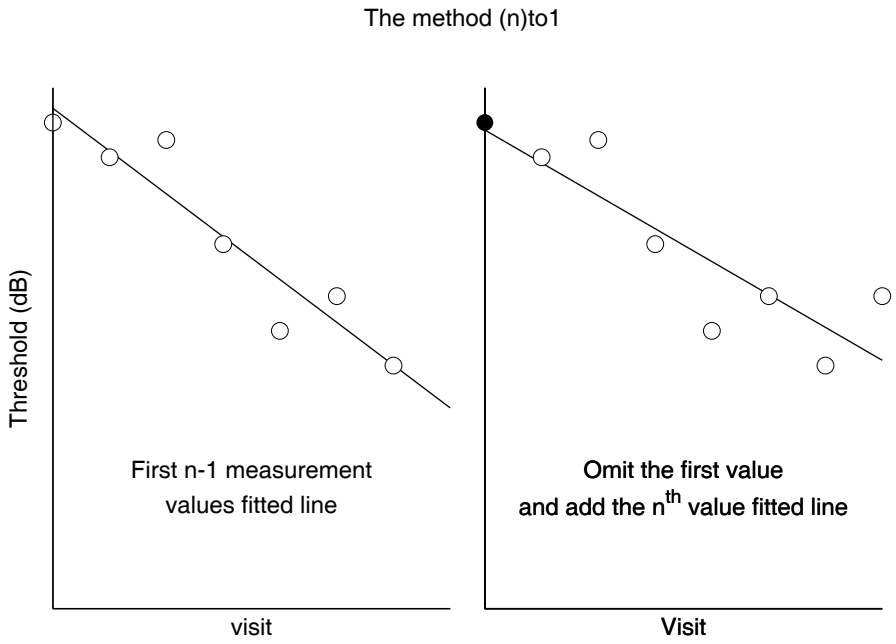


Figure 6.5: Illustration of the *(n)to1* criterion. The line in the left graph is fitted using the first 7 measurements. The line in the right graph is fitted by deleting the first measurement and adding the 8^{th} measurement.

The purpose of proposed method is to examine the fact that the larger the noise (or fluctuation), the harder it is to determine whether a location is progressive or not. This is because for a progressive sequence, the first threshold value is usually greater than the subsequent ones. According to Equation 6.2, a higher threshold has a lower fluctuation.

1. **Omitting-Maximum Difference** (denoted by *(n)to(max)*): a location X is identified as progressing if the first n-1 measurements x_1, x_2, ..., x_{n-1} satisfy the basic criterion, and continue to satisfy it after adding the n^{th} measurement x_n, but excluding the value x_k which represents the maximum absolute deviation between $\hat{x}_k$ and x_k, $1 \leq k \leq n-1$, where $\hat{x}_k$ is calculated by using the x_1, x_2, ..., x_{n-1} linear regression model.

Figure 6.6 shows the *(n)to(max)* method. In the left graph, the maximum $\{|\hat{x}_k - x_k| : k = 1, 2, \ldots, 7\}$ is $|\hat{x}_3 - x_3|$. Therefore the value x_3 is deleted in the right graph, and the line is derived from $x_1, x_2, x_4, \ldots, x_8$.

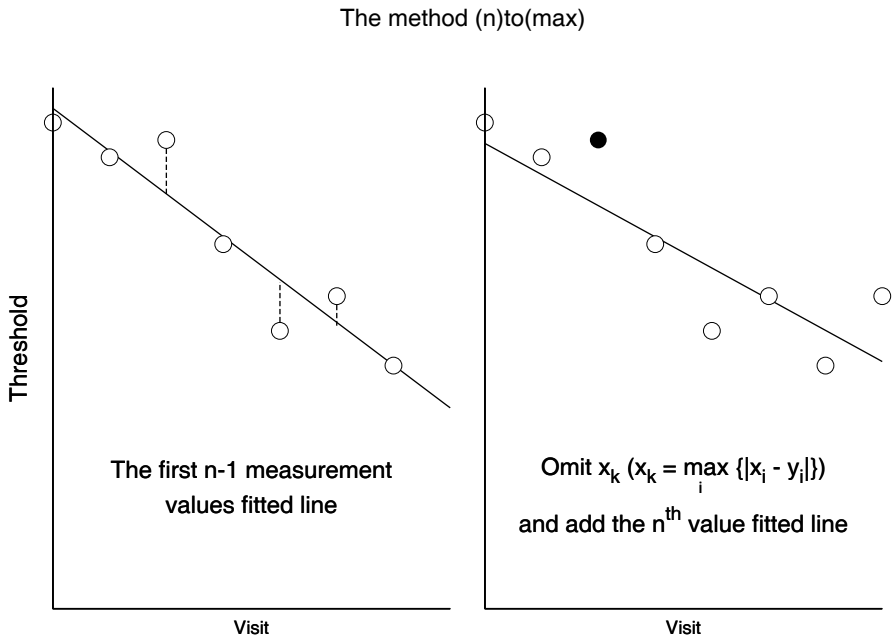


Figure 6.6: Illustration of the *(n)to(max)* criterion. The line in the left graph is fitted by the first 7 measurements. The line in the right graph is fitted by deleting x_3, which corresponds to $\max\{|\hat{x}_k - x_k|\}$, and adding the 8^{th} measurement.

The rationale behind the *(n)to(max)* method is that if a location is identified as satisfying the basic criterion by the first (n - 1) measurements, then deleting the value x_k (x_k= $\max_i\{|\hat{x}_i - x_i|\}$), which may be biased or represent an outlier, may result in the remaining values, together with the n^{th} measurement x_n, fitting a significant line.

6.3.2 The Kappa Statistical Method

The kappa statistic is a measure of agreement between two observers with respect to nominal outcome. It is simply the proportion of agreement corrected for chance. For example, assume that there are two categories A and B for two observers and n subjects. Each observer classifies subjects as shown in Table 6.1.

Kappa is defined as follows:

$$k = \frac{p_o - p_e}{1 - p_e} \tag{6.5}$$

where p_o is the "observe" agreement, and p_e is an expected value by chance alone ("expecte" agreement). The observed agreement is simply the percentage of all clas-

Table 6.1: Results by category, for two observers

		Observer 1		Total Number
		A	B	
Observer 2	A	a	b	a+b
	B	c	d	c+d
Total		a+c	b+d	n

sifications for which the two observers' agree. That is, for the sum of $(a+d)/n$ where $n = a+b+c+d$, the expected agreement according to Cohen (1960) is:

$$p_e = \frac{a+b}{n} \times \frac{a+c}{n} + \frac{c+d}{n} \times \frac{b+d}{n} \tag{6.6}$$

We may also want to know how different the observed agreement (p_o) is from the expected agreement (p_e). The kappa value lies within the range [-1, +1], where 1 is perfect agreement, 0 is exactly what would be expected by chance, and negative values indicate disagreement. Interpretation of kappa values is summarised in Table 6.2 (Viera and Garrett, 2005).

Table 6.2: Interpretation of Kappa

Kappa	Agreement
< 0	Less than chance agreement
0.01 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Almost perfect agreement

6.4 Results

We examined all methods on two datasets, one being the simulated dataset and the other a real dataset described in section 6.4.2. Each criterion for classifying progression was used on each dataset in turn. A location was labeled as progressive if the sequence

of measurements for that location satisfied that particular criterion. The test at which the location was labelled as progressive was recorded. For example, if *2of2* is satisfied by using the first 8 and the first 9 measurements, the location was only labeled as progressive at the $9th$ measurement.

6.4.1 Simulated Data Sets

We tested all 7030 virtual sequences made up of 8, 9, 10 or 11 measurements that were taken during a period of up to five years. It is known that using less than 7 or 8 measurements cannot produce accurate results due to the small sample size (Spry et al., 2002). The results are shown in Table 6.3. Note that

Table 6.3: The percentage of correct identification in the simulated datasets (based on individual test location). "P" indicates progressive sequences; "S" stands for stable sequences. The best results for both progressive and stable sequences are shown in bold.

Criterion	Class	Years of visits			
		3.5 years	4 years	4.5 years	5 years
1of1	P	**23.03**	**30.87**	**40.06**	**49.91**
	S	98.78	98.73	98.27	97.38
2of2	P	10.69	17.48	25.35	34.65
	S	99.49	99.59	99.38	98.78
2of3	P	12.52	20.75	29.73	39.40
	S	99.27	99.08	99.03	98.65
(n)to(n-1)	P	9.16	15.14	22.31	31.02
	S	99.59	99.68	99.51	99.05
(n)to(1)	P	7.48	13.42	20.60	29.64
	S	**99.65**	**99.76**	**99.59**	**99.16**
(n)to(max)	P	10.39	17.39	25.50	35.11
	S	99.46	99.62	99.41	98.78

For each method in Table 6.3, the accuracy of identifying stable sequences was

similar regardless of the number of measurements (compare S values in each row), whereas it gradually increased for progressive sequences as the number of measurements increases. This is because when the number of measurements becomes large, the number of degrees of freedom increases, which results in the critical value in the t-distribution becoming small.

Gardiner and Crabb (2002a) pointed out that the method 1of1 has a high rate of falsely labelling stable sequences as progressive. However, it is the most sensitive method for detecting progressive sequences. The relationships among the methods they used *1of1*, *2of2*, *2of3* and *(n)to(n-1)* is:

$$\textit{(n)to(n-1)} \longrightarrow \textit{2of2} \longrightarrow \textit{2of3} \longrightarrow \textit{1of1}$$
$$\text{(most specific} \longrightarrow \text{most sensitive)}$$

When comparing the new method *(n)to1* to the method *(n)to(n-1)*, which is currently used by Gardiner and Crabb (2002a), Artes et al. (2005), and Nouri-Mahdavi et al. (2005), the results show that the method *(n)to(n-1)* is more sensitive than *(n)to1*. That is, by deleting the first value and adding the n^{th} to the sequence decreases the correct identification for progressive cases. This is because progressive sequences have starting values that are higher than subsequent values.

When comparing the method *(n)to(n-1)* with the new method *(n)to(max)*, the results show that *(n)to(max)* is more sensitive than *(n)to(n-1)* for detecting progressive sequences. The method *(n)to(max)* sacrifices correct identification for stable sequences compared with the method *(n)to(n-1)*.

When comparing the method *2of2* with the new method *(n)to(max)*, the results show that the specificity of the method *(n)to(max)* is higher than that of the method *2of2*, except in the third ("8") column. The sensitivity of the method *(n)to(max)* is higher than that of the method *2of2* in the last two columns. That is, the new method *(n)to(max)* is similar in effectiveness in detecting both stable and progressive sequences when compared with *2of2*.

The results show that the relationship among all methods, including the new methods *(n)to1* and *(n)to(max)*, is:

(n)to1 $\longrightarrow$ *(n)to(n-1)* $\longrightarrow$ {*(n)to(max)*, *2of2*} $\longrightarrow$ 2of3 $\longrightarrow$ 1of1

(The most specific $\longrightarrow$ the most sensitive)

6.4.2 Real Data Sets

The other option to evaluate the new methods is to use real patient data, with progressive and stable sequences grouped in different sets. As previously discussed, it is difficult to separate progressive and stable sequences from whole patient visual fields. This is because there is no consensus on the definition of progressive locations in different study groups (Gardiner and Crabb, 2002a). The status of the whole visual field was known from experts labeling them. Therefore we separately grouped progressive and stable visual fields into different sets in which we examine the new methods.

A real dataset was obtained from 64 patients with progressive eyes and 66 patients with stable eyes. Each patient had at least 8 visual field measurements over a period of at least 3.5 years as a part of their regular ophthalmological examinations. The visual field threshold values were adjusted for age by 1 dB per decade, so in effect all eyes were from 45 year old patients. The definition of visual field progression was that at least one location was confirmed as progressive as Gardiner and Crabb (2002a) used . For example, when the new method *(n)to1* is used, an eye is considered as progressive only if at least one location has been confirmed as progressing as defined by *(n)to1*. Using 8 visual field measurements, the results are shown in Table 6.4.

Table 6.4: The percentage of correct classification for all methods. The criterion of classifying a patient as progressive is based on at least one location being confirmed as progressing.

Criterion	P	S
1of1	**82.81**	54.55
2of2	64.04	75.76
2of3	67.19	69.70
(n)to(n-1)	60.94	78.79
(n)to1	57.81	**84.85**
(n)to(max)	65.62	77.27

On the real dataset, *(n)to(max)* is more sensitive than *(n)to(n-1)* for identifying progressive visual fields. It shows an improvement of 4.5% for progressive visual fields, with a reduction of 1.5% in specificity for stable visual fields. For the new method (n)to1, sensitivity decreases by 3% for progressive visual fields, but increases by 6% for stable visual fields, when comparing with *(n)to(n-1)*. The relationship among all methods remains the same as for the simulated datasets.

We used the kappa statistic to assess the agreement between pairs of methods. The results are shown in Table 6.5. As can be seen, the agreement between *1of1* and the other methods is fair to moderate. Almost perfect agreement exists between 2of2, *2of3*, *(n)to(n-1)*, *(n)to1*, and *(n)to(max)*.

Table 6.5: Pairwise agreement estimated using the kappa statistic.

Methods	*1of1*	*2of2*	*2of3*	*(n)to(n-1)*	*(n)to1*	*(n)to(max)*
1of1	NA	0.53	0.57	0.46	0.34	0.52
2of2		NA	0.89	0.93	0.81	0.96
2of3			NA	0.82	0.71	0.85
(n)to(n-1)				NA	0.85	0.93
(n)to1					NA	0.81
(n)to(max)						NA

6.5 Discussion

Although Pointwise Linear Regression (PLR) is a clinically useful tool for classifying visual field data, there is no consensus on the value of regression slope or significance level α that indicates progression. In this chapter we focused on PLR methods with fixed $\alpha < 0.01$ and slope < -1 dB/year as criteria for classifying visual fields. For a fixed number (n) of measurements, changes in the criteria of slope and α can give rise to different outcomes in classification. For example, suppose the criterion of slope is less than k dB/year for identifying progression when α is fixed, the sensitivity decreases as k decreases. This is because the slope of a sequence of measurements $x_1, x_2, \ldots, x_n$ is the rate of change of the measurements against time. If $k <$ -1,

the criterion slope $< k$ dB/year requires more change to determine progressive. The criterion slope $< k$ dB/year therefore is more specific for classifying stable sequences. On the other hand, if $k >$ -1, the criterion of slope $< k$ dB/year requires less change to determine progressive. The criterion of slope $< k$ dB/year therefore is more sensitive for classifying progressive sequences. For a fixed criterion of slope, as α increases, a smaller critical value must be obtained by the t-distribution. Therefore the accuracy of identification for progressive sequences increases, whereas it decreases for stable sequences. As such, we adopted the most commonly used criteria of slope $<$ -1 dB/year and $\alpha < 0.01$ in this chapter.

We discussed that data generated by using 6.2 is noisier than that by using Spry et al. (2000)'s method in the section 3.4 in Chapter 3. We examined all methods on the simulated datasets, with short- and long-term fluctuations added as noise, as described in Chapter 4. Results are shown in Table 6.6.

Table 6.6: The percentage of correct classification in the simulated datasets described in Chapter 4.

Criterion	Class	3.5 Years	4 Years	4.5 Years	5 Years
1of1	P	40.25	50.43	78.63	87.99
	S	99.49	99.65	99.68	99.73
2of2	P	24.00	43.38	64.38	80.06
	S	99.84	99.84	99.89	99.92
2of3	P	26.08	45.54	66.68	81.39
	S	99.86	99.81	99.70	99.78
(n)to(n-1)	P	22.80	42.80	64.34	79.71
	S	99.87	99.95	99.92	99.92
(n)to(1)	P	19.71	40.04	62.65	78.87
	S	99.89	99.95	99.92	99.92
(n)to(max)	P	23.71	43.50	65.18	80.22
	S	99.86	99.92	99.92	99.92

Table 6.6 shows that the accuracy of classification for both stable and progressive sequences is higher than that the results in Table 6.3. This finding was consistent

with what we have anticipated.

We examined all methods on real datasets. The criterion used Gardiner and Crabb (2002a) for classifying a patient as progressive was that at least one location was confirmed as progressive. However, some investigators required that at least two locations must be progressive to determine whole visual field progression (Vesti et al., 2003). When the criterion of at least two locations confirmed as being progressive was used, the quantitative results changed, as shown in Table 6.7.

Table 6.7: The percentage of correct classification for all methods. The criterion for classifying a patient as progressing is that at least two locations are progressive.

Criterion	P	S
1of1	**71.88**	77.27
2of2	42.19	87.88
2of3	46.88	84.85
(n)to(n-1)	39.06	89.39
(n)to1	31.25	**90.91**
(n)to(max)	42.19	87.88

Table 6.7 shows the results of classifying real patient data using the criterion of at least two locations being progressive. Sensitivity decreases and specificity increases, compared with using the criterion of at least one location being progressive (see Table 6.4). In theory, the accuracy of correctly classifying progressive patients based on at least one location being progressive is higher than that of using two (or more) locations being progressive. The comparable relationship among all methods in Table 6.7 remains the same as on Table 6.4. The pairwise agreements between methods based on Table 6.7 are shown in Table 6.8.

Comparing Tables 6.5 and 6.8, the agreements between pairs of methods differ slightly. Note that perfect agreement occurs between methods *2of2* and *(n)to(max)* in Table 6.8. That means that when using at least two progressive locations to determine whether an eye is progressive or not, the results of classifying each patient using *2of2* and *(n)to(max)* are the same.

Table 6.8: Pairwise agreement estimated by the kappa statistic. Patients are classified as progressing based on at least two progressive location.

Methods	*1of1*	*2of2*	*2of3*	*(n)to(n-1)*	*(n)to1*	*(n)to(max)*
1of1	NA	0.53	0.58	0.48	0.39	0.53
2of2		NA	0.91	0.95	0.85	1.00
2of3			NA	0.86	0.77	0.91
(n)to(n-1)				NA	0.90	0.95
(n)to1					NA	0.85
(n)to(max)						NA

Because this chapter is confined to different point-wise linear regression methods, we also applied the kappa statistic on individual locations. The agreements between methods are shown in Table 6.9. All values in the first row have increased when compared to Table 6.5 and 6.8. The agreement between *1of1* and the other methods are moderate to substantial. Some cell values in rows 2 to 5 in Table 6.9 decreased when compared with Table 6.5 and 6.8. For point-wise linear regression methods, the pairwise agreements in Table 6.9 may show true agreement for classification.

Table 6.9: Pairwise agreements between methods, based on individual locations in the patient dataset.

Methods	*1of1*	*2of2*	*2of3*	*(n)to(n-1)*	*(n)to1*	*(n)to(max)*
1of1	NA	0.63	0.63	0.55	0.48	0.61
2of2		NA	0.93	0.90	0.82	0.98
2of3			NA	0.84	0.75	0.92
(n)to(n-1)				NA	0.88	0.91
(n)to1					NA	0.82
(n)to(max)						NA

Similarly, to assess the agreement between the EA and PLR methods, point-wise agreements may be appropriate. This is because although a patient is identified as progressive using both the EA and PLR methods, the progressive location(s) identified may not be the same. For example, suppose there is a sequence of measurements 20,

18, 20, 15, 9, 20, 9, 8 for one location X_1. The baseline value for that location is 20 dB. If using PLR, slope = -3.21, but the statistic value t_v calculated using Equation 6.4 is -2.6, the absolute value is less than the critical value 3.71. The location therefore is stable. However, if using the EA method, the lower limit for baseline being 20 dB is 14 (Turpin and McKendrick, 2005, p.3). The location is therefore progressive. Similarly one location can be identified as progressing using PLR but not using EA. The result is that eye is identified as progressing by using PLR and EA but the deteriorating location(s) may be in a different area in the eye.

No comparison was made against other methods, such as non-linear or non-pointwise methods. An advantage of PLR methods is that they can provide some spatial information about where disease has occurred. Another advantage is that they are suitable for classifying deepening of an existing defect. However, for an enlarging defect, PLR methods may be not quite as effective (Gardiner and Crabb, 2002a).

6.6 Summary

In this chapter, we proposed two point-wise linear regression methods, *(n)to(max)* and *(n)to1*. We evaluated different PLR methods *1of1*, *2of2*, *2of3* and *(n)to(n-1)* on the simulated and real datasets. Results show the new method *(n)to1* to be the most specific for identifying stable sequences. However it sacrifices accuracy for identification of progressive visual fields. The new method *(n)to(max)* is superior to the method *2of2* for both stable and progressive cases. The method *(n)to(max)* is better than the method *2of3* (or *(n)to(n-1)*) for classifying stable (or progressive) cases although sensitivity (or specificity) is also lost.

Chapter 7

Machine Learning

7.1 Introduction

In previous chapters we discussed bias analysis, point-wise linear regression and sequence matching methods for classification of visual field data, including simulated and real data sets. In this chapter, we will discuss the application of machine learning methods for the classification of visual field data.

Machine learning has been used in the past to classify glaucomatous data by Turpin et al. (2001); Chan et al. (2002); Hothorn and Lausen (2003); Lazarescu and Turpin (2003); Tucker et al. (2004), to determine whether a patient's disease is stable or progressive. The main aim of this chapter is to investigate which machine learning classifier is most effective on glaucomatous data, and to determine whether machine learning approaches are more effective in determining progression than the previously discussed methods.

The chapter is structured as follows: Section 2 describes feature extraction for use in machine learning; Section 3 describes the datasets which will be trained using machine learning; Section 4 briefly introduces the machine learning classifiers that are used in this study; the results are shown in Section 5; the Discussion and Summary are presented in Section 6 and 7 respectively.

7.2 Definition of Features

Feature extraction is an important part of machine learning classification problems. The aim of extracting features from glaucomatous data is to best capture the progression of the disease. Glaucomatous data is numeric, so the most straightforward way to apply machine learning is to use each individual visual field measurement as an input. Excluding the two blind locations, there are $74 \times n \times m$ inputs, where n is the number of measurements and m is the number of patients. However such a feature only shows the response at each location in automated perimetry, and when used individually the locations do not show the changes (if any) that may occur in the overall eye condition. Proceeding in this way also ignores valuable information, such as the differences between measurements, as pointed out by Turpin et al. (2001).

In this study, we used well established background knowledge of classifying glaucomatous visual field loss to choose features. The specific knowledge used was as follows.

- A gradual decrease in overall response indicates that the disease is progressing - it may be deepening or spreading.
- Anomalous response over time may indicate that the disease is progressing.
- The disease usually starts from the periphery of the visual field, and some patients have localized visual field loss that can be located anywhere in fields. Therefore, a consistent decrease in each zone (see Chapter 4) can indicate that the disease is progressing.

The details of the features extracted are as follows:

Feature 1. Slope of means. This feature is based on the overall change in the visual field over a series of measurements. A negative slope may indicate the patient's eye gradually deteriorating, because all visual fields are age-corrected. Otherwise the patient's condition is stable. There should not be any change for stable visual fields after the age-correction process, except for noise. For progressive fields, however, deterioration should be obvious. The feature has a single real value calculated by averaging all thresholds in each field, and then computing the slope over time using linear regression.

Feature 2. Consistency in the number of anomalies over the set of observations. This feature is a single, positive integer value that can be obtained by two steps (Lazarescu and Turpin, 2003). First, for each measurement, the mean of its surrounding values is calculated by using a 3-by-3 window (see Figure 7.1). A measured value is considered to be anomalous if it is smaller than the mean by at least 30%. That is, measured value $<$ mean $\times$ (1 - 0.3). Second, the number of anomalies in the field is counted, and is taken as the value of feature 2. For example, consider the measured value 17 (the centre point) and its 8 neighbours highlighted in Figure 7.1. The mean of the neighbouring values is $(29 + 30 + \cdots + 20 + 25)/8 = 27.9$, and $27.9 \times (1 - 0.3) = 19.5$. Therefore the value 17 is anomalous. Note that if a value is located at the edge of a field, for example value 25 at the lower-right corner in Figure 7.1, it has fewer neighbours: 27, 29, 28, and 27. The mean of the four neighbour values is 27.7, and $27.7 \times (1 - 0.30) = 19.4$. Therefore the value 25 is normal.

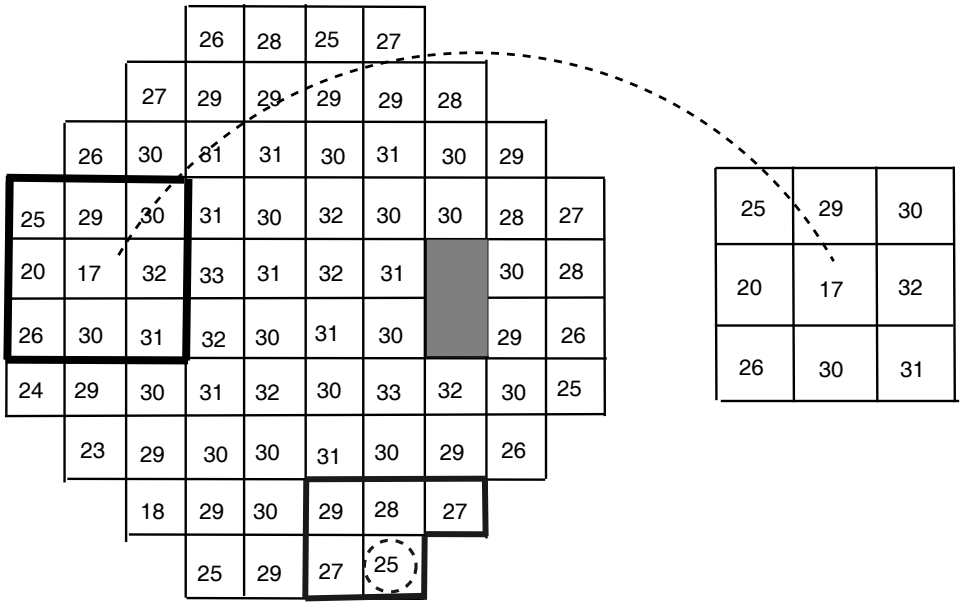


Figure 7.1: The identification of an anomalous location inside a field by using a 3-by-3 window. If a value is located at the edge of a field, fewer neighbours are used to calculate the mean.

Feature 3. Difference between the means of the first and follow-up measurements in the inner zone. We experimented with three different ways of

calculating this feature. (1) Use the slope of linear regression based on means of all thresholds in the inner zone (see definition in Figure 2.15 in Chapter 2). Like Feature 1, this slope can show the change in the inner zone. (2) Use the difference between means of the current and the next measurement (i.e. the change in consecutive measurements). Note that this difference is not obvious for both stable and progressive cases because vision loss usually occurs slowly. (3) Use the difference between means of the first and current measurement. This difference is large, especially for progressive cases when measurements are obtained several tests later. However, for stable cases this difference is small because natural age-related decline is very small (0.1 dB/per year). Methods (1) and (2) did not result in more accurate classification than (3). The experimental results presented in this chapter are based on method (3).

Features 4 and 5. Difference between the means of the first and follow-up measurements in the middle and outer zone. These features are similar to Feature 3, except for the zones (see Figure 2.15 in Chapter 2).

7.3 Feature Data Set

The glaucomatous data consists of five attributes: slope of means, number of anomalies, and difference in inner, middle and outer zones as described in Section 7.2. The target attribute can have values stable or progressive. Table 7.1 shows a set of 130 patients, each represented by the 5 attributes.

A data set of training samples similar to the above was obtained "*from patient charts of glaucoma patients of the Devers Eye Institute, each having at least 8 visual field measurements over at least a 4 year period as part of their regular ophthalmologic examination. Unlike many previous studies, no special efforts were made to ensure patient reliability, nor was the quality of the visual fields evaluated. These are typical patient records from a typical clinical situation. The patients were classed as progressing, or stable, or unknown by an expert (Chris A. Johnson), based on optic disk appearance, and their visual field measurements. The final data set consisted of progressing eyes and stable eyes, each with eight visual field measurements at different*

Table 7.1: Progressive and stable training examples. D stands for Difference, *cont.* for continuous, P for Progressive and S for stable.

Cases	Slope of means	Number of anomalies	D inner	D middle	D outer	Target
1	cont.	cont.	cont.	cont.	cont.	P
2	cont.	cont.	cont.	cont.	cont.	P
...	...	...	...	...	...	...
64	cont.	cont.	cont.	cont.	cont.	P
65	cont.	cont.	cont.	cont.	cont.	S
...	...	...	...	...	...	...
130	cont.	cont.	cont.	cont.	cont.	S

points in time. The visual field threshold values were adjusted for age at measurement by 1 dB per decade, so in effect all eye were from a 45 year old patient. Left eyes were transformed into right eye coordinates" (Turpin et al., 2001, p. 5). These age-corrected data are used in the experiments that follow.

7.4 Machine Learning Classifiers

For this study we used the WEKA package, which provides implementations for Decision Tree, Decision Stump, Naive Bayes, and Bayes Network classifiers, as well as Bagging and Boosting methods for applying the classifiers (Witten and Frank, 2000). In this section, we briefly review these techniques.

Let $S_i = (y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$, (note that $p = 5$ here) denote the i^{th} training sample, where $y_i = \{Stable, Progressive\}$ is the output label (or class), x_{ij} is an input value (or an attribute value) and $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is an instance.

7.4.1 Decision Trees

A decision tree is a tree structure made up of nodes. The top node in the tree is the root node, and each leaf node represents a class. Each internal node is a test, and each branch represents an outcome of test. A basic algorithm for inducing a decision tree from training samples is shown below.

Algorithm 3. *Algorithm for constructing a decision tree.*

Input: each sample S_i, $i = 1, 2, 3, \ldots, n$.

Output: a decision tree.

Method:

(1) create a node (N);

(2) **if** all samples have the same class (C),

then return N as a leaf node labelled with C;

(3) else **if** attributes are empty,

then return N as a leaf node with the most common class;

(4) else select an attribute x that has the highest information gain for node N, and label N with test-attribute x ;

for each possible value y_i (Class) in x_i add a branch as a new tree;

(5) repeat (1) - (4) until all branches have leaf nodes

An example decision tree obtained using the WEKA package is shown in Figure 7.2. Given an instance, this decision tree can classify it as stable or progressive. For example, the instance slope of means = -1/per year, difference in middle = 3.5, number of anomalies = 20 would be sorted down the leftmost branch of this tree, and classified as Progressive.

In general, a decision tree represents a disjunction of conjunctions of constraints on the attribute values of instances (Mitchell, 1997). Therefore, each branch from the root to a leaf in the tree can be converted to a rule. Thus there are 6 rules that can be obtained from the decision tree in Figure 7.2. The leftmost branch in the above tree, for example, can be converted to the rule:

$$\begin{aligned} \text{IF} \quad & ((\text{Slope of means} \leq -0.7) \wedge (\text{Number of anomalies} \leq 29) \\ & \wedge (\text{Difference in middle} \leq 4.35)) \\ \text{THEN} \quad & \text{Progressive.} \end{aligned}$$

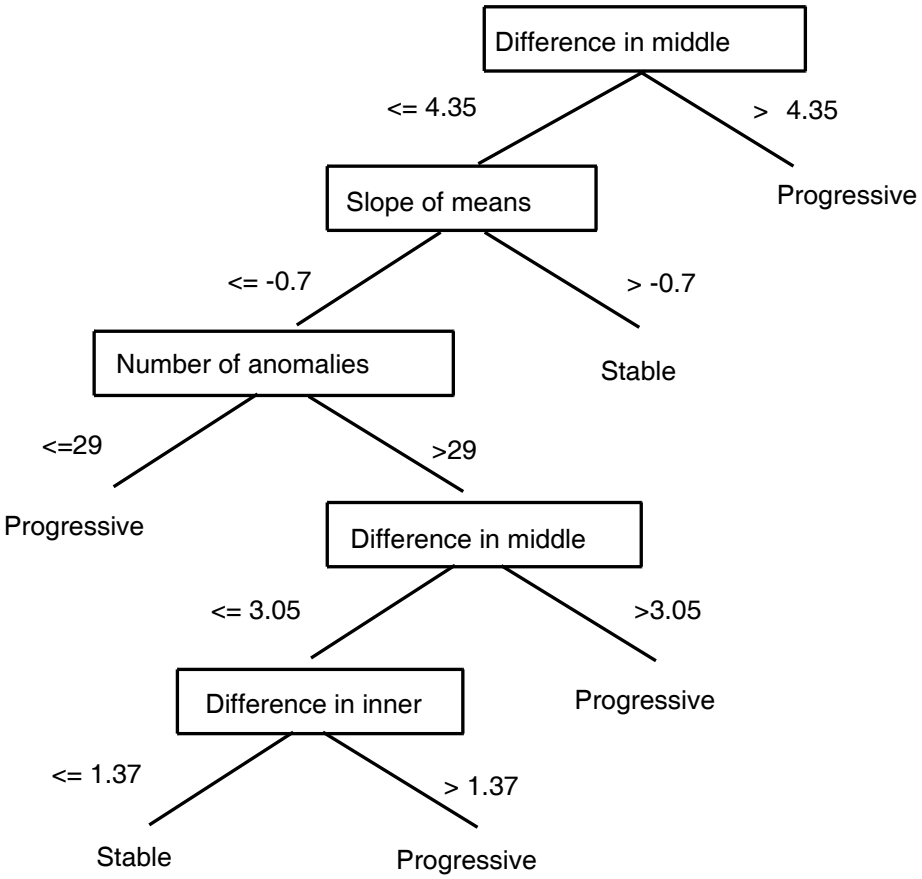


Figure 7.2: A decision tree for the classification of glaucomatous data.

7.4.2 Decision Stumps

A decision stump is a one-level decision tree. It can perform a single test on a single attribute with a threshold value. That is, single feature is used to make the classification decision. An example of a Decision Stump produced by using the WEKA package is shown in Figure 7.3.

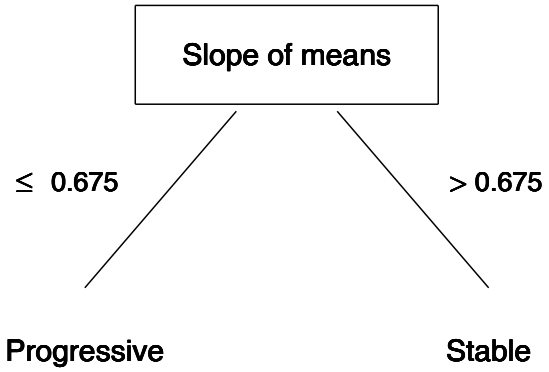


Figure 7.3: A Decision Stump of classification of glaucomatous data.

Given an instance X = (slope of means, number of anomalies, difference in inner, difference in middle, difference in outer), only the *slope of means* is used to make a decision. If the value of the feature slope of means is greater than 0.675, then the instance X is classified as Stable, otherwise it is Progressive. Although decision stumps lack the sophistication of more elaborate decision trees, they are easy to generate and understand, and they are useful in situations where more sophisticated algorithms are proved to be susceptible to over fitting (i.e. the influence of irrelevant "features" and noise).

7.4.3 Naive Bayes

The Naive Bayes approach is one of the most popular data mining tools in the medical domain due to its ease of interpretation over other tools (Lavrac, 1998). The Naive Bayes classifier is applied to learning tasks where each instance X_i is described by a set of attribute values $(x_{i1}, x_{i2} \ldots, x_{ip})$ and the target function f_s can take on any value from some finite set $V = (y_1, y_2, \ldots, y_n)$. Naive Bayes uses Bayesian probabilities calculated from the training data to classify unseen instances.

The Bayesian approach to classifying a new instance is to assign the most probable target value, given the attribute values $(x_{i1}, x_{i2} \ldots, x_{ip})$ that describe the instance. The Naive Bayes classifier is based on the simplifying assumption that the attribute values are conditionally independent given the target value:

$$v_{NB} = argmax P(y_i) \prod_k P(x_{ik}|y_i) \tag{7.1}$$

where v_{NB} denotes the target value output by the Naive Bayes classifier. Each $P(y_i)$ in Equation 7.1 is the probability of the target function f_s being value y_i (called the Prior probability). Each $P(x_{ik}|y_i)$ is the probability of the attribute x_{ik} underlying the target function f_s being value y_i.

Consider again the dataset described in Section 7.3. Each instance consists of five values $(x_1, \ldots, x_5)$ = (*Slope of means, Number of anomalies, Difference in the inner zone, Difference in the middle zone, Difference in the outer zone*). The target function has values of progressive and stable. In Equation 7.1, there are two prior probabilities $P(y_i)$ (i =1, 2). One is for probability $P(Progressive)$ and the other for probability $P(Stable)$. In the dataset there are 130 patients, of whom 64 patients are progressive and 66 patients are stable. Therefore, the probability $P(Progressive)$ = 64/130 = 0.49, and $P(Stable)$ = 66/130 = 0.51.

In Equation 7.1, $P(x_{ik}|y_i)$ is a conditional probability. If an attribute x_{ik} has discrete values, the conditional probability $P(x_{ik}|y_i)$ can be calculated by counting the number of x_{ik} which occur in the progressive patients or stable patients. If an attribute value is continuous, the distribution of the attribute value is estimated to be a normal distribution in progressive patients or stable patients. For example, given that the underlying prior probability $P(Progressive)$ = 0.49 calculated above, the

mean of the normal distribution for the attribute "Slope of means" is estimated to be -1.0079, and standard deviation = 1.4461, in progressive patients. Table 7.2 shows the underlying prior probability, mean and standard deviation for each attribute, using the Naive Bayes classifier provided by the WEKA package.

Table 7.2: Mean and standard deviation of normal distribution for each attribute described in Section 7.3. Each attribute is estimated underlying prior probability P(Probability) or P(Stable). "D" stands for Difference.

Attribute	Prior P(Progressive) = 0.49		Prior P(Stable) = 0.51	
	Normal distribution		Normal distribution	
	Mean	Standard Deviation	Mean	Standard Deviation
Slope of means	-1.01	1.45	0.05	1.08
# of anomalies	29.75	6.26	30.18	7.69
D. inner	1.91	2.98	-0.03	2.14
D. middle	2.02	3.54	-0.36	2.84
D. outer	1.91	3.66	-0.66	3.40

7.4.4 Bayes Network

A Bayesian network is a directed acyclic graph where the nodes represent random variables and the links represent the causal relations among the variables. The links are parameterised by the conditional probabilities of the variables. Assume that $v_1, v_2, \ldots, v_n$ are the variables of a Bayesian network and π_i $(1 < i < n)$ is the set of parents of v_i. The joint probability distribution of $v_1, v_2, \ldots, v_n$ can be defined as:

$$P(v_1, v_2, \ldots, v_n) = \prod_i P(v_i|\pi_i) \tag{7.2}$$

A Bayes Network obtained using the WEKA package is shown in Figure 7.4. Each variable is represented by a node, and the links represent the causal relations among the variables in a dataset. The "Slope of mean" and "Number of anomalies" are associated each other. The "slope of mean" is the immediate parent of node "Difference in outer", and "Number of anomalies" has two children "Difference in inner" and "Difference in middle".

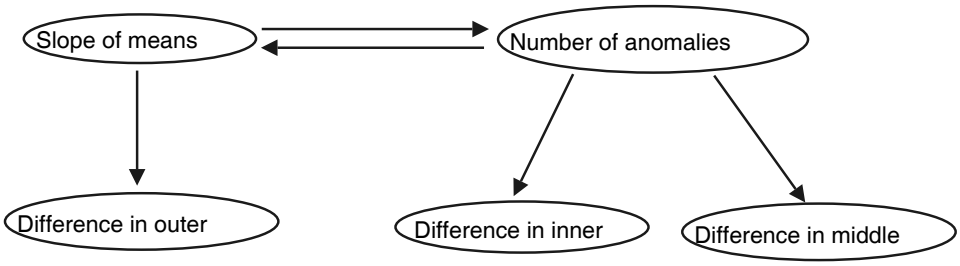


Figure 7.4: A network structure obtained by using the WEKA package. Each variable is presented by a node, and the links represent the causal relations among the variables.

7.4.5 Meta-Learning (Ensemble Learning)

Meta-learning methods use a combination of classifiers to enhance the performance and extend the abilities of learning schemes (Witten and Frank, 2000). There are several methods to combine a set of classifiers, but for experimental purposes in this study we consider only the most well known methods, Bagging and Boosting.

Bagging

The bagging algorithm introduced by Breiman (1996) is a "bootstrap" ensemble method that creates individuals for its ensemble by training each classifier on a random redistribution of the training set. Each classifier's training set is generated by randomly deleting examples, with replacement. This deletion and replacement produces a training set which is equal in size to the original training set. Many of the original examples may be repeated in the new training set while others may be left out. Each individual classifier in the ensemble is generated with a different random sampling of the training set. The WEKA package (Witten and Frank, 2000) automatically provides support for 10-fold cross-validation when training is performed. The bagging algorithm is as follows.

Algorithm 4. *The bagging algorithm.*

Input n instances.

For each of t iterations do

Delete some instances;

Duplicate other instances to form a new sample;

Applying the learning algorithm to the new sample to produce the model;

Store the resulting model;

End for

For each of the t models do

Predict class of instances using model;

End for

Return class that has been predicted most often

Boosting

Boosting is another method to construct an ensemble of classifiers. It is used to boost the performance of a weak learner (Schapire, 1990). A weak learner is a simple classifier whose error is less than 50% on training instances (Ledezma et al., 2001). In this chapter, we used the AdaBoosting algorithm provided by the WEKA package.

Like bagging, boosting generates a set of classifiers on a training set and combines them in the final classifier by using a voting method. However, it differs from the bagging algorithm. In bagging, classifiers are generated independently from each sample. In boosting, classifiers are generated sequentially. First, a classifier is generated by using the original training set. Second, instances misclassified by the classifier are given a larger weight. Third, a new classifier is generated by using the weighted training set. The training error is calculated and a weight is assigned to the classifier in accordance with its performance on the training set. Classifiers with a lower training error have a higher weight. This process is repeated $k-$times until an error rate is reached, and all instances are correctly classified. Finally, the classifier with the highest weight is output (Schapire, 1990; Ledezma et al., 2001).

7.4.6 Stratified 10-Fold Cross-Validation

Stratified 10-fold cross-validation was used to estimate the accuracy of learned concepts. In general for a k-fold cross-validation, the dataset D is randomly divided into k subsets (the folds): $D_1, D_2, \ldots, D_k$ of (approximately) equal size (Bailey and Elkan,

1993). Each one of k subsets is left out for testing and the remaining k-1 subsets are used for training, so that the classifier is trained and tested k times. The cross validation of accuracy is the overall number of correct classifications, divided by the number of instances in the dataset D Kohavi (1995).

The experiments used the Features described in Sections 7.2 and 7.3. The data were classified using the machine learning techniques described in Section 7.4. A number of trials were conducted. The first set involved using only one feature at a time. The second set involved using the features together to determine if the combined features offer a more accurate means of classification.

7.5 Results

All results were obtained by using 10-fold cross-validation and are presented in this section. The abbreviations used in this section are listed below for ease of reference.

Abbreviation	Machine learning classifiers
C4.5	C4.5 Decision Tree
DStump	Decision Stump
NBayes	Naive Bayes
BayesN	Bayesian Network
Bagging	Bagging
Boosting	AdaBoosting

In the following tables, columns show results of using 4 visual field measurements (4VFs) up to 8 measurements (8VFs). Each cell from column 4VFs to 8VFs shows the percentage of correct classification. The best result (s) in each column is (are) highlighted in bold. For each basic classifier, such as C4.5, the results of the sensitivity, specificity and accuracy are shown in the "Basic" rows. For each meta-learning technique, the results of applying it to the basic classifier are shown in the "Bagging" or "Boosting" rows. Each cell in the last column, "Mean", shows the average of all results in the same row.

7.5.1 Classification Based on Slope of Means

In this trial, only feature 1 (the slope of means) was used in classification. The results are shown in Table 7.3.

Table 7.3 shows that the sensitivity (classification accuracy for progressive cases) generally increases with number of measurements. This is because change gradually becomes significant as the number of measurements increases. Boosting based on C4.5 classified 56.3% to 82.8% of visual fields as progressive. On average, boosting based on Decision Stump achieves the best results for classifying progressive visual fields, especially from 6VFs (over 70%). In general, boosting based on each basic classifier performs the same or better than the basic classifier. Comparing bagging and boosting based on each basic classifier, boosting gives better results than bagging. There is an exception for BayesNet in which bagging is better than boosting.

For specificity, the Basic classifier Naive Bayes performs best, given over 80% specificity, except in the 4VFs column. However, on average BayesNet and boosting based on BayesNet give the best results.

For accuracy, all results are over 60%, except for BayesNet and boosting based on BayesNet in column 4VFs. They offer the worst results (49.2%). Although boosting based on Decision Stump offers over 70% accuracy from the 5VFs column, it only achieves 73% for the last column which is lower than bagging based on Decision Stump. In general, boosting has an improvement for each the basic classifier for classification. In contrast, bagging is worse than the basic Naive Bayes. This may is because different classifiers generated based on Naive Bayes misclassify the same instances frequently. Bagging based on C4.5 achieves the best result on average.

7.5.2 Classification Based on Number of Anomalies

In this trial, only feature 2 (the number of anomalies) was used in classification. The results are shown in Table 7.4.

The results in Table 7.4 indicate that feature 2 is less effective than feature 1 for classifying the data. The BayesNet classifier is 100% accurate for classifying stable

Table 7.3: Sensitivity, Specificity and Accuracy based on the slope of means. The best results in each column are shown in bold.

	Classifier		4 VFs	5 VFs	6 VFs	7 VFs	8 VFs	Mean
Sensitivity	C4.5	Basic	59.4	56.3	40.6	**73.4**	75.0	60.95
		Bagging	57.8	56.3	53.1	**73.4**	82.8	64.70
		Boosting	59.4	56.3	60.9	**73.4**	82.8	66.56
	DStump	Basic	59.4	56.3	40.6	54.7	76.6	57.52
		Bagging	**60.9**	56.3	53.1	62.5	**84.4**	63.44
		Boosting	59.4	56.3	**70.3**	**73.4**	76.6	**67.20**
	NBayes	Basic	**60.9**	54.7	54.7	57.8	59.4	57.50
		Bagging	59.4	53.1	56.3	54.7	60.9	56.88
		Boosting	**60.9**	54.7	54.7	57.8	62.5	58.12
	BayesN	Basic	14.1	56.3	40.6	54.7	76.6	48.46
		Bagging	59.4	**57.8**	54.7	62.5	78.1	62.50
		Boosting	14.1	56.3	40.6	54.7	76.6	48.46
Specificity	C4.5	Basic	63.4	**83.3**	**90.9**	75.8	70.0	76.68
		Bagging	66.7	80.3	81.8	78.8	72.7	76.06
		Boosting	63.6	**83.3**	74.2	75.8	66.7	72.72
	DStump	Basic	63.4	**83.3**	**90.9**	77.3	69.7	76.92
		Bagging	66.7	80.3	83.3	78.8	74.2	76.66
		Boosting	63.6	**83.3**	72.7	75.8	69.7	73.02
	NBayes	Basic	72.7	80.3	81.8	**81.8**	**86.4**	80.60
		Bagging	69.7	81.8	81.8	**81.8**	83.3	79.68
		Boosting	72.7	80.3	81.8	**81.8**	84.8	80.28
	BayesN	Basic	**83.3**	**83.3**	**90.9**	77.3	69.7	**80.90**
		Bagging	66.7	78.8	83.3	78.8	74.2	76.36
		Boosting	**83.3**	**83.3**	**90.9**	77.3	69.7	**80.90**
Accuracy	C4.5	Basic	61.5	**70.0**	66.2	74.6	72.3	68.92
		Bagging	62.3	68.5	67.7	**76.2**	77.7	**70.48**
		Boosting	61.5	**70.0**	67.7	74.6	74.6	69.68
	DStump	Basic	61.5	**70.0**	66.2	66.1	73.1	67.38
		Bagging	63.9	68.5	68.5	70.8	**79.2**	70.18
		Boosting	61.5	**70.0**	**71.5**	74.6	73.1	70.14
	NBayes	Basic	**66.9**	67.7	68.5	70.0	73.1	69.24
		Bagging	64.2	67.7	69.2	68.5	72.3	68.38
		Boosting	**66.9**	67.7	68.5	70.0	73.9	69.40
	BayesN	Basic	49.2	**70.0**	66.2	66.1	73.1	64.92
		Bagging	63.1	68.5	69.2	70.8	76.2	69.56
		Boosting	49.2	**70.0**	66.2	66.1	73.1	64.92

Table 7.4: Sensitivity, Specificity and Accuracy based on the number of anomalies calculated using 3-by-3 windows. The best results in each column are shown in bold.

	Classifier		4 VFs	5 VFs	6 VFs	7 VFs	8 VFs	Mean
Sensitivity	C4.5	Basic	7.8	17.2	76.6	0.0	15.6	23.44
		Bagging	68.8	60.9	60.9	32.8	64.1	57.50
		Boosting	7.8	7.8	53.1	0.0	15.6	16.86
	DStump	Basic	**89.1**	**89.1**	**95.3**	**73.4**	**79.7**	**85.32**
		Bagging	79.7	79.7	76.6	57.8	70.3	72.82
		Boosting	71.9	54.7	56.3	35.9	50.0	53.76
	NBayes	Basic	70.3	66.7	62.5	40.6	37.5	55.52
		Bagging	65.6	64.1	56.3	35.9	46.9	53.76
		Boosting	59.4	60.9	60.9	42.2	39.1	52.50
	BayesN	Basic	0.0	0.0	0.0	0.0	0.0	0.00
		Bagging	35.9	32.8	35.9	37.5	35.9	35.60
		Boosting	0.0	0.0	0.0	0.0	0.0	0.00
Specificity	C4.5	Basic	90.9	80.3	25.8	**100**	78.8	75.16
		Bagging	51.5	45.5	43.9	56.1	30.3	45.46
		Boosting	90.9	89.4	42.4	**100**	78.8	80.30
	DStump	Basic	27.3	7.6	15.2	13.6	24.2	17.58
		Bagging	37.9	30.3	33.3	28.8	28.8	31.82
		Boosting	53.0	42.4	62.1	34.8	40.9	46.64
	NBayes	Basic	37.9	45.5	37.9	39.4	57.6	43.66
		Bagging	40.9	42.4	45.5	42.4	59.1	46.06
		Boosting	54.5	51.5	40.9	42.4	57.6	49.38
	BayesN	Basic	**100**	**100**	**100**	**100**	**100**	**100.00**
		Bagging	60.6	59.1	63.6	57.6	57.6	59.70
		Boosting	**100**	**100**	**100**	**100**	**100**	**100.00**
Accuracy	C4.5	Basic	50.0	49.2	50.8	**50.8**	47.7	49.70
		Bagging	60.0	53.9	52.3	44.6	46.9	51.54
		Boosting	50.0	49.2	47.7	**50.8**	47.7	49.08
	DStump	Basic	57.7	47.7	**64.6**	43.1	51.5	**52.92**
		Bagging	58.5	55.4	54.6	43.1	49.2	52.16
		Boosting	**62.3**	48.5	59.2	35.4	45.4	50.16
	NBayes	Basic	53.9	**56.9**	50.0	40.0	47.7	49.70
		Bagging	53.9	53.9	50.8	39.2	53.1	50.18
		Boosting	56.9	**56.9**	50.8	42.3	48.5	51.08
	BayesN	Basic	50.8	50.8	50.8	**50.8**	**55.8**	51.80
		Bagging	48.5	46.2	50.0	47.7	46.9	47.86
		Boosting	50.8	50.8	50.8	**50.8**	50.8	50.80

cases, and the worst for classifying progressive cases (no progressive instance is correctly classified by BayesNet). The Decision Stump classifies progressive data best with all results over 73%. However, for stable data the Decision stump classifier is the worst with all results under 30%.

For accuracy most classification results are lower than 60%. We found that each meta-classifier did not improve the basic classifier. We therefore tried to use the difference between numbers of anomalies at the first and the current visual field. The results are shown in Table 7.5.

As can be seen in Table 7.5, classifiers based on the difference between numbers of anomalies are still less effective for classification. Although there is a little bit of increase for accuracy, most results are below 65%. On average, the maximum accuracy is below 60% achieved by using boosting based on Naive Bayes.

We also tried to use big windows (5-by-5) to calculate the number of anomalies. Figure 7.5 illustrates this approach for the measured value 33 (in bold font) in the area surround by the bold square. Its neighbours are all involved in classification. The method of calculating anomalies is the same as for Feature 2, but with more neighbours. The experimental results based on using 5-by-5 windows are shown in Table 7.6.

Comparing Tables 7.4 and 7.6, most accuracy results increase using 5-by-5 windows, especially in 7VFs and 8VFs columns, where most results are over 60%. On average, although the maximum sensitivity (by Decision Stump) and specificity (by BayesNet) in Table 7.4 are higher than those (by Decision Stump and boosting based on BayesNet) in Table 7.6, the maximum accuracy (by Decision Stump) in Table 7.4 is lower than that in Table 7.6. Boosting based on Naive Bayes using 5-by-5 windows gives the best result. The reason for higher accuracy using 5-by-5 windows is that more neighbours filter out some noise.

7.5.3 Classification Based on Differences in Three Zones

In this trial, features 3, 4, and 5 are used in classification. The classification results are shown in Table 7.7.

Table 7.5: Sensitivity, Specificity and Accuracy of classification, according to the difference between numbers of anomalies at the first and the current visual field. The best results in each column are shown in bold.

	Classifier		4 VFs	5 VFs	6 VFs	7 VFs	8 VFs	Mean
Sensitivity	C4.5	Basic	59.4	21.9	58.8	43.8	35.9	43.96
		Bagging	57.8	45.3	67.2	45.3	**46.9**	52.50
		Boosting	35.9	21.9	**78.1**	43.8	40.6	44.06
	DStump	Basic	**81.3**	50.0	62.5	**46.8**	35.9	**55.30**
		Bagging	71.9	46.9	64.1	45.3	43.8	54.40
		Boosting	48.4	34.4	68.8	43.8	40.6	47.20
	NBayes	Basic	50.0	**56.3**	34.4	43.8	**46.9**	46.28
		Bagging	48.4	53.1	40.6	43.8	45.3	46.24
		Boosting	50.0	**56.3**	43.8	45.3	**46.9**	48.46
	BayesN	Basic	0.0	0.0	0.0	6.3	18.8	5.02
		Bagging	35.9	35.9	34.4	43.8	43.8	38.76
		Boosting	0.0	0.0	0.0	6.3	18.8	5.02
Specificity	C4.5	Basic	42.4	68.2	36.4	86.4	**93.9**	65.46
		Bagging	40.9	42.4	43.9	80.3	84.8	58.46
		Boosting	50.0	68.2	36.4	86.4	84.8	65.16
	DStump	Basic	22.7	47.0	39.4	86.4	93.9	57.88
		Bagging	36.4	50.0	45.5	80.3	80.3	58.50
		Boosting	51.5	59.1	45.5	86.4	84.8	65.46
	NBayes	Basic	56.1	60.6	77.3	78.8	81.8	70.92
		Bagging	54.5	57.6	74.2	78.8	80.3	69.08
		Boosting	56.1	60.6	72.7	74.2	81.8	69.08
	BayesN	Basic	**100**	**100**	**100**	**89.4**	**93.9**	**96.66**
		Bagging	60.6	60.6	57.6	83.3	80.3	68.48
		Boosting	**100**	**100**	**100**	**89.4**	**93.9**	**96.66**
Accuracy	C4.5	Basic	50.8	45.4	52.3	**65.4**	65.4	55.86
		Bagging	49.2	43.9	55.4	63.1	**66.2**	55.56
		Boosting	50.0	45.4	56.9	**65.4**	63.1	56.16
	DStump	Basic	51.5	48.5	50.8	**65.4**	65.4	56.32
		Bagging	**53.8**	48.5	54.6	61.1	62.3	56.06
		Boosting	50.0	46.9	56.9	**65.4**	63.1	56.46
	NBayes	Basic	53.1	**58.5**	56.2	61.5	64.6	58.78
		Bagging	51.5	55.4	57.7	61.5	63.1	57.84
		Boosting	53.1	**58.5**	**58.5**	60.0	64.6	**58.94**
	BayesN	Basic	50.8	50.8	50.8	48.5	56.9	51.56
		Bagging	48.5	48.5	46.2	63.9	62.3	53.88
		Boosting	50.8	50.8	50.8	48.5	56.9	51.56

Table 7.6: Sensitivity, Specificity, and Accuracy according to the number of anomalies calculated by using 5-by-5 windows. The best results in each column are shown in bold.

	Classifier		4 VFs	5 VFs	6 VFs	7 VFs	8 VFs	Mean
Sensitivity	C4.5	Basic	35.9	45.3	81.3	57.8	**75.0**	59.06
		Bagging	51.6	43.8	62.5	56.3	**75.0**	57.84
		Boosting	29.7	15.6	**90.6**	57.8	**75.0**	53.74
	DStump	Basic	57.8	**81.3**	57.8	57.8	62.5	**63.44**
		Bagging	**60.9**	50.0	59.4	56.3	62.5	57.82
		Boosting	34.4	39.1	73.4	57.8	**75.0**	55.94
	NBayes	Basic	53.1	39.1	32.8	46.9	39.1	42.20
		Bagging	54.7	50.0	32.8	51.6	40.6	45.94
		Boosting	53.1	40.6	32.8	**60.9**	56.3	48.74
	BayesN	Basic	0.0	9.4	0.0	14.1	62.5	17.20
		Bagging	42.2	56.3	59.4	56.3	62.5	55.34
		Boosting	0.0	0.0	0.0	14.1	62.5	15.32
Specificity	C4.5	Basic	56.1	56.1	39.4	65.2	65.2	56.40
		Bagging	53.0	60.6	42.4	68.2	65.2	57.88
		Boosting	62.1	**100**	33.3	65.2	65.2	65.16
	DStump	Basic	37.9	27.3	57.6	65.2	68.2	51.24
		Bagging	45.5	56.1	48.5	68.2	68.2	57.30
		Boosting	78.8	60.6	45.5	65.2	65.2	63.06
	NBayes	Basic	65.2	74.2	77.3	**84.8**	**81.8**	76.66
		Bagging	60.6	68.2	80.3	81.8	**81.8**	74.54
		Boosting	65.2	72.7	75.8	75.8	71.2	72.14
	BayesN	Basic	**100**	90.9	**100**	83.3	68.2	88.48
		Bagging	56.1	48.5	50.0	71.2	68.2	58.80
		Boosting	100	98.5	**100**	83.3	68.2	**90.00**
Accuracy	C4.5	Basic	46.2	50.8	60.0	61.5	**70.0**	57.70
		Bagging	52.3	52.3	52.3	62.3	**70.0**	57.84
		Boosting	46.2	53.8	**61.5**	61.5	**70.0**	58.60
	DStump	Basic	47.7	52.3	57.7	61.5	65.4	56.92
		Bagging	53.1	53.1	53.8	62.3	65.4	57.54
		Boosting	55.4	50.0	59.2	61.5	**70.0**	59.22
	NBayes	Basic	**59.2**	56.9	55.4	66.2	60.8	59.70
		Bagging	57.7	**59.2**	56.9	66.9	61.5	60.44
		Boosting	**59.2**	56.9	54.6	**68.5**	63.8	**60.60**
	BayesN	Basic	50.8	50.8	50.8	49.2	65.4	53.40
		Bagging	49.2	52.3	54.6	63.8	65.4	57.06
		Boosting	50.8	50.0	50.8	49.2	65.4	53.24

Table 7.7: Sensitivity, Specificity and Accuracy based on the Differences in Three Zones. The best results in each column are shown in bold.

	Classifier		4 VFs	5 VFs	6 VFs	7 VFs	8 VFs	Mean
Sensitivity	C4.5	Basic	54.7	51.6	**75.0**	75.0	**78.1**	66.88
		Bagging	60.9	54.7	68.6	64.1	76.6	64.98
		Boosting	54.7	54.7	70.3	76.6	**78.1**	66.88
	DStump	Basic	54.7	43.8	**75.0**	75.0	**78.1**	65.32
		Bagging	57.8	56.3	62.5	67.2	**78.1**	64.38
		Boosting	54.7	56.3	71.9	76.6	**78.1**	**67.52**
	NBayes	Basic	**62.5**	50.0	54.7	57.8	62.5	57.50
		Bagging	60.9	53.1	59.4	56.3	64.1	58.76
		Boosting	56.3	50.0	54.7	67.2	62.5	58.14
	BayesN	Basic	39.1	53.1	65.6	**78.1**	67.2	60.62
		Bagging	56.3	**57.8**	65.6	71.9	71.9	64.70
		Boosting	39.1	53.1	65.6	**78.1**	67.2	60.62
Specificity	C4.5	Basic	63.6	63.6	71.2	75.8	72.7	69.38
		Bagging	63.6	72.7	69.7	72.7	71.2	69.98
		Boosting	63.6	62.1	74.2	75.8	72.7	69.68
	DStump	Basic	63.6	77.3	69.7	75.8	72.7	71.82
		Bagging	68.2	77.3	72.7	74.2	75.8	73.64
		Boosting	62.1	66.7	71.2	75.8	72.7	69.70
	NBayes	Basic	63.6	**80.3**	80.3	**81.8**	**84.8**	78.16
		Bagging	60.6	75.8	**81.8**	**81.8**	81.8	76.36
		Boosting	65.2	**80.3**	80.3	80.3	**84.8**	**78.18**
	BayesN	Basic	**71.2**	71.2	74.2	77.3	81.8	75.14
		Bagging	65.2	75.8	75.8	77.3	78.8	74.58
		Boosting	**71.2**	71.2	77.3	77.3	81.8	75.76
Accuracy	C4.5	Basic	59.2	57.7	**73.1**	75.4	75.4	68.16
		Bagging	62.3	63.9	69.2	69.2	73.9	67.70
		Boosting	59.2	58.5	72.3	76.2	75.4	68.32
	DStump	Basic	59.2	60.8	72.3	75.4	75.4	68.62
		Bagging	**63.1**	**66.9**	67.7	70.8	**76.9**	69.08
		Boosting	58.5	61.5	71.5	76.2	75.4	68.62
	NBayes	Basic	**63.1**	65.4	67.7	70.0	73.9	68.02
		Bagging	60.8	64.6	70.0	70.0	73.1	67.70
		Boosting	60.8	65.4	67.7	73.9	73.9	68.34
	BayesN	Basic	55.4	62.3	70.0	**77.7**	74.6	68.00
		Bagging	60.8	**66.9**	70.0	74.6	75.4	**69.54**
		Boosting	55.4	62.3	71.5	**77.7**	74.6	68.30

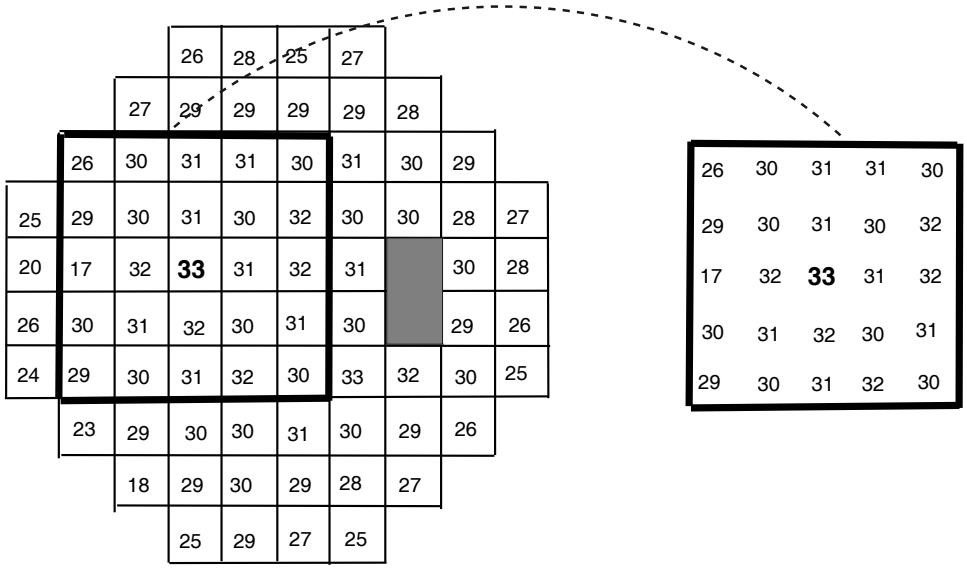


Figure 7.5: An example of calculating the number of anomalies by using a 5-by-5 window.

Table 7.7 shows that boosting improves the performance for classifying progressive visual fields, especially based on Decision Stump that is the most effective on average. For specificity, if over five visual fields are available, Naive Bayes offers the best results in 5VFs, 7VFs and 8VFs. Boosting has an improvement for each basic classifier, except for Decision Stump which specificity is less than that of Naive Bayes.

Decision Stump is the best basic classifier for accuracy on average. If over six visual fields are available, it correctly classifies more than 72% visual fields. In general, boosting achieves an improvement for the basic classifier. Although bagging based on BayesNet offers the best result, bagging has decreases for C4.5 and Naive Bayes.

In Tables 7.3 and 7.7, bagging gives the best accuracy on average. Comparing the best results in Tables 7.3 and 7.7, classification based on "slope of means" is generally better than that based on "the differences in three zones".

7.5.4 Classification Based on Combined Features

In this trial, all five features are used in classification. The purpose is to investigate whether the combination of all features can improve classification. The sensitivity, specificity, and accuracy of classification are shown in Table 7.8.

Table 7.8: Sensitivity, Specificity and Accuracy based on the combined features. The best results in each column are shown in bold.

	Classifier		4 VFs	5 VFs	6 VFs	7 VFs	8 VFs	Mean
Sensitivity	C4.5	Basic	57.8	51.6	67.2	70.3	73.4	64.06
		Bagging	**67.2**	**60.9**	67.2	65.6	76.6	**67.50**
		Boosting	48.4	48.4	**73.4**	67.2	75.0	62.48
	DStump	Basic	53.1	56.3	45.3	54.7	**78.1**	57.50
		Bagging	54.7	56.3	56.3	65.6	**78.1**	62.20
		Boosting	53.1	51.6	65.6	65.6	75.0	62.18
	NBayes	Basic	62.5	56.3	54.7	59.4	60.9	58.76
		Bagging	62.5	59.4	54.7	59.4	62.5	59.70
		Boosting	57.8	**60.9**	60.9	60.9	62.5	60.60
	BayesN	Basic	40.6	**60.9**	65.6	**78.1**	71.9	63.42
		Bagging	62.5	**60.9**	60.9	70.3	76.6	66.24
		Boosting	40.6	**60.9**	65.6	**78.1**	71.9	63.42
Specificity	C4.5	Basic	60.6	71.2	66.7	80.3	66.7	69.10
		Bagging	57.6	66.7	71.2	74.2	72.7	68.48
		Boosting	65.2	74.2	66.7	80.3	72.7	71.82
	DStump	Basic	57.6	**80.3**	**83.3**	78.8	66.7	73.34
		Bagging	62.1	**80.3**	78.8	80.3	69.7	74.24
		Boosting	63.6	**80.3**	72.7	77.3	72.7	73.32
	NBayes	Basic	69.7	**80.3**	78.8	**81.8**	**86.4**	**79.40**
		Bagging	66.7	78.8	80.3	**81.8**	84.8	78.48
		Boosting	**72.7**	75.8	74.2	80.3	83.3	77.26
	BayesN	Basic	65.2	77.3	74.2	77.3	78.8	74.56
		Bagging	63.6	77.3	75.8	77.3	78.8	74.56
		Boosting	65.2	77.3	74.2	77.3	78.8	74.56
Accuracy	C4.5	Basic	59.2	61.5	66.9	75.4	70.0	66.60
		Bagging	62.3	63.9	69.2	70.0	74.6	68.00
		Boosting	56.9	62.3	**70.0**	73.9	73.1	67.24
	DStump	Basic	55.4	68.5	64.6	66.9	72.3	65.54
		Bagging	58.5	68.5	67.7	73.1	73.9	68.34
		Boosting	58.5	66.2	69.2	71.5	73.6	67.80
	NBayes	Basic	**66.2**	68.5	66.9	70.8	73.9	69.26
		Bagging	64.6	69.2	67.7	70.8	73.9	69.24
		Boosting	65.4	68.5	67.7	70.8	73.1	69.10
	BayesN	Basic	53.1	69.2	**70.0**	**77.7**	75.4	69.08
		Bagging	63.1	**69.2**	68.5	73.9	**77.7**	**70.48**
		Boosting	53.1	**69.2**	**70.0**	**77.7**	75.4	69.08

The first observation from Table 7.8 is that C4.5 is the best basic classifier for sensitivity, and Naive Bayes is the best basic classifier for specificity and accuracy. Boosting keeps the same or improves sensitivity of each basic classifier according to single attribute "slope of means" or "difference in three zones". Boosting does not improve sensitivity for C4.5 based on the combination of all attributes. However, bagging achieves improvement of the each basic classifier except for Bayesnet for sensitivity. For accuracy, bagging yields improvements for most basic classifiers (except for Naive bayes).

The second observation is that the combination of all features improves classification of BayesNet, Naive Bayes. However, it influences classification of C4.5 and Decision Stump.

7.5.5 Classification Using EA Method

To evaluate the effectiveness of machine learning classifiers, we used the Glaucoma Change Probability method to classify the data described in Section 7.3. The steps of using the EA method to classify a location as stable or progressive are described in Chapter 4. We used two criteria for identifying a visual field as progressing: (1) EA(2of3) is the presence of four or more overlapping locations at the 95% level (we used 95% confidence interval to classify a location) occurring in two of three consecutive fields; (2) EA(3of3) is the presence of four or more overlapping locations at the 95% level in three consecutive visual fields (Vesti et al., 2003). The 95% confidence interval used is from a large population of test-retest data. The classification results are shown in Table 7.9. To compare machine learning classifiers and EA methods, the results based on best mean accuracy in Table 7.8 are listed in Table 7.9. Bagging based on the basic BayesNet classifier performs best in the mean of all 5 columns, so the sensitivity and specificity of Bagging based on the basic BayesNet classifier are listed.

The sensitivity of the EA methods gradually increases with the number of measurements. This is because the difference between the baseline and current measurements is becoming larger as the number of measurements is increasing, and the lower limit of a confidence interval based on the baseline value does not change. The sensitivity of EA(2of3) is the best for 8VFs (84.4%). However, EA(2of3) offers specificity below

Table 7.9: Percent correct classification for different number visual field measurements according the EA methods. P is for Sensitivity, and S is for Specificity. The best results in each column are shown in bold.

Methods	5 VFs		6 VFs		7 VFs		8 VFs	
	P	S	P	S	P	S	P	S
EA(2of3)	59.4	72.7	62.5	69.7	75.0	66.7	84.4	59.1
EA(3of3)	14.1	95.5	28.1	93.9	34.4	93.9	53.1	93.9
Bagging(BayesN)	60.9	77.3	60.9	75.8	70.3	77.3	76.6	78.8

72.7%. It is less effective for classifying stable visual fields compared with EA(3of3) which is over 93% correct for over 5 visual fields. The method EA(3of3) only correctly classifies 14% of visual fields as progression in the 5VFs column.

Comparing the methods, EA(3of3) is less effective than machine learning classifying progressive visual fields. However, EA(3of3) gives the best results for classifying stable visual fields. The sensitivities of the method EA(2of3) are the best among all classifiers, except for bagging in the 5VFs column. However, EA(2of3) offers the worst specificity, especially in the 8VFs column.

7.6 Discussion

Correctly classifying glaucomatous patients based on visual field measurements is a challenge. Visual field measurements provide a large number of attributes such as:

- Means of thresholds based on the nerve bundle map (see Figure 2.14 of chapter 2), or the difference between the means;
- Means of thresholds based on the ten sectors (see Figure 2.15), or the difference between the means of symmetrical sectors in the upper and lower hemi-field;
- Mean Deviation;
- Pattern Standard Deviation;
- Sum of some slopes, in which each slope is calculated based on each location, and less than 1 dB per year decreasing with $\alpha < 0.01$.

We tried classification by using machine learning classifiers on each attribute above. The results show that none is able to significantly improve classification results compared with feature 1, or 3 or combined all features.

Feature extraction is the key to successful classification in machine learning problems. The results using number of anomalies calculated by 3-by-3 windows show that each classifier is not effective (below 60% for accuracy). The approach of using the difference between the numbers of anomalies at the first and the current visual field also fails to solve the problem. That is, the difference still cannot distinguish progressive and stable instances. The causes probably are that some stable (or progressive) locations are considered as progressive (stable) locations due to noise. Figure 7.6 (a) illustrates a stable location in bold in a bad vision area being considered as progressive one. Measured 4 dB in left window is greater than 70% of the mean of its neighbours so that the location is a normal one, but corresponding location is 0 dB which is less than 70% of the mean of its neighbours. Therefore one more is added into the number of anomalies for a stable location. Similarly, Figure 7.6 (b) shows a progressive location in bold in a normal neighbour area, being classified as stable using 3-by-3 windows. These mixed values for the attribute may lead to poorly trained classifier.

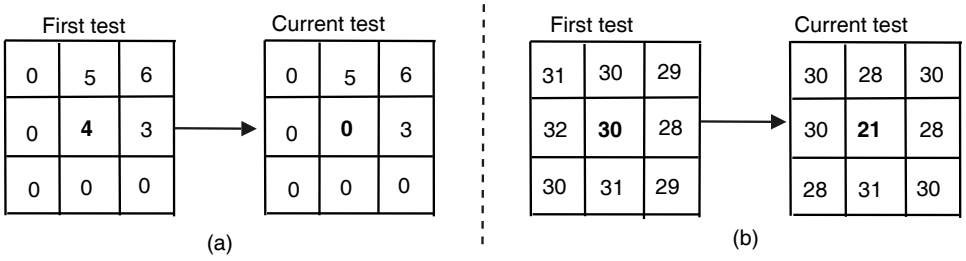


Figure 7.6: An example of misclassifying. (a) A stable location is considered to be a progressive one. (b) A progressive location is considered to be a stable one.

For comparison of results based on a single feature "slope of means" (or "difference in three zones"), we have found that boosting improves each basic classifier. However, for the combination of all features, it does not improve some basic classifiers. This may be because the error on some weighted training sets is not smaller than 50%.

The data used for this chapter was previously classified by Turpin et al. (2001) using linear regression, linear support vector machines, 1R decision stumps, and stumps

based on features which are the differences between the first and the last measurement for each location. There is one scenario with accuracy from 61.5% to 80% for 8 VFs (reading from page 9 of Turpin et al. (2001)). Overall we have found that the accuracy of classification does not exceed 80%. This may be due to noise, or because the classification of the original data is based on more information, such as eye pressure, false response rate, and optic nerve damage. These tests (as features) can be combined as input for a machine learning classifier. Eliminating less useful features from tests may be helpful for analysis by one machine learning classifier.

7.7 Summary

In this chapter we present an application of machine learning techniques to the problem of classifying patients that have either stable or progressive glaucoma. We focus on four different basic machine learning classifiers, and bagging and boosting to classify visual fields. The features are extracted from raw visual field measurements and based on knowledge gleaned from domain experts. Results obtained from classification based on the features show that C4.5 is the best among all basic classifiers for classifying progressive visual fields, and Naive Bayes is the best for classifying stable visual fields. In general, bagging goes to the best improvement to each basic classifier for classification.

Chapter 8

Conclusions

Millions of people worldwide are currently suffering from glaucoma. There is an urgent need for improving the methods of diagnosis to prevent or delay the onset of advanced vision loss. There are a number of methods currently available for the identification of progressive glaucomatous visual field loss. However no single method for identification of visual field progression is ideal, as shown in the review of the literature.

This thesis focuses on point-by-point analysis of longitudinal glaucomatous data and the application of machine learning classifiers to glaucomatous data. In this chapter, we first summarise the work achieved by this thesis, and then present several research directions for future studies.

8.1 Summary

We have investigated the ability of each existing point-by-point methods and machine learning approach for distinguishing visual field change from noise and introduced matching techniques to data from glaucoma patients. In particular, we aimed at answering the following questions:

- How can we use empiric data to classify a given location in a visual field as progressing or not progressing?
- What is the effect of addition or deletion of an observation on the accuracy of a point-wise linear regression method?

- What are the most and least accurate machine learning classifiers and which features improve the accuracy of classification?

To evaluate the effectiveness of these classification methods, this thesis first modified and improved a simulation program to generate sequences at individual locations in a visual field. The program provides rate of natural decline, functions of linear and no-linear deterioration to closely model real data. In non-linear deterioration, the concave and convex functions are drawn through two given points. It also provided different calculation of noise which appears during visual field tests. We have investigated and compared two algorithms for calculating noise. The frequency of tests per year and total number of tests can all be controlled to suit the needs of the experiment. The advantage of computer simulation is to obtain large, longitudinal visual field series with minimal cost. Visual fields simulated are known to be progressing or stable by controlling variability parameters and levels of fluctuation. However, the disadvantage of simulation based on individual location is that data simulated may be only used for point-wise methods.

The first question was explored in Chapter 4 and 5. In Chapter 4, we presented the Event Analysis (EA) method to classify whether or not a location in visual field is progressive. The EA method identifies whether the difference $(x_1+x_2)/2-x_n$ falls outside the selected confidence interval (CI). In this chapter, we proposed an algorithm, called "baseline-less-follow-up", to calculate a confidence interval by using a set of normal and stable data taken in a short period. The "baseline-less-follow-up" calculates the difference between each measurement in a sequence and the baseline of that sequence instead of the difference between two consecutive measurements (test-retest). We not only examined the method on the simulated data set, but also examined the original data set for building 95% CI. The experiments were based on point-by-point basis to rigorously examine the sensitivity and specificity of the EA method. Our experimental results showed that the "baseline-less-follow-up" is more sensitive than test-retest. However, there is small penalty in specificity.

In Chapter 5, we introduced sequence matching techniques to apply on glaucomatous data. We proposed two different matching method, Weighted Sequence Matching (SM) and Baseline Matching Stable Sequences (BMS). For a given query sequence,

SM used the weighted Euclidean or weighted Manhattan Distance function to choose closest matches in a reference database R which consists of normal and stable data. Although SM offers high accuracy in classification, its application is limited for query sequences with more than five measurements. This is because there are only five values in a sequence in R. The method BMS overcomes this limitation of SM, and has less calculation compared with the method SM. We tested the BMS method using both synthetic and real datasets. The results indicate that using the new method BMS can significantly improve the accuracy of identifying progressive sequences compared with the EA method, though there is a small penalty in the correct identification of stable sequences.

The second question was investigated in Chapter 6. In this chapter, We proposed two PLR methods, $(n)to1$ and $(n)to(max)$. We compared these with the PLR methods $1of1$, $2of2$, $2of3$ and $(n)to(n-1)$. The method $(n)to1$ is shown to be the most specific. The $(n)to(max)$ method is superior than the $2of2$ method for both stable and progressive sequences. The method $(n)to(max)$ improves the method $(n)to(n-1)$ which is currently used by some researchers for classifying progressive sequences. An advantage of the PLR methods is that they can provide some spatial information about where disease have occurred. Another advantage is that they are suitable for classifying deepening progression of an existing defect. However, for an enlarging defect, PLR methods may not be as effective.

In Chapter 7, we presented the application of machine learning techniques in classifying patients with stable or progressive glaucoma. We have focused on 4 different basic machine learning classifiers, together with bagging and boosting classifiers. Five features were extracted from raw visual field measurements and based on the knowledge gleaned from domain experts. Each feature and combination of all features have been investigated in classification. The results have shown that classification from the combination of all features does not improve the results from individual features.

We also compared machine learning classifiers and the EA method. Our experimental results indicated that the machine learning classifiers do not offer improvements in classifying both stable and progressive patients.

8.2 Future Directions

Potential areas to be further explored in future studies are presented and discussed below.

- The weighted sequence matching techniques presented in this thesis can be further examined in real visual fields when more measurements in each reference sequence are available.

- Several extensions could be made to improve the sequence matching techniques. Currently, the reference database in sequence matching techniques only consists of normal and stable visual field data. We require a more sophisticated similarity algorithm to accurately process a given query sequence which may be stable or progressive. In addition, capacity to match a set number of closest matches in the reference database for a given query sequence can be investigated.

- More features can be extracted for classification when machine learning classifiers are being used. So far we have focused on the use of visual field data measured in 76 locations. Additional information, such as the patient's age, false response rate, intraocular pressure and specific features of the damaged optic nerve were gathered during the visual field assessment and may be helpful for classification when machine learning classifiers are being used.

- The EA method for identifying whether a location is progressive or not is based on the use of baseline value(s) to choose a suitable confidence interval. The baseline value is established at the beginning of a visual field measurement. A question exists as to how the baseline value may be defined years later, that is, to account for changes occurred in the visual field? For example, a baseline value is 30 dB for a location, and subsequently measured values (twice per year) are 30, 28, 25, 27, 24, 23, 20, 22, 21, 20, 23. As can be seen, during the first three years, this location was progressive. Over the last two years however, this location appeared to be stable. If 30 dB was still used as the baseline value, then this location would be incorrectly identified as progressive over the last two years. Therefore, taken a threshold value measured years earlier (say three year earlier) as baseline to choose a confidence interval may be appropriate.

Bibliography

(2003, September). International Glaucoma Association. http://www.iga.org.uk/.

Amjad, H., F. Shahar, Y. Claudia, F. M. D., T. Uriel, and B. E. Z (2002, December). The learning effect in visual field testing of healthy subjects using frequency doubling technology. *Journal of Glaucoma 11*(6), 511–516.

Anderson, D. R. and V. M. Patella (1999, July). *Automated static perimetry.* (second ed.). Mosby. ISBN: 0815143842.

Artes, P. H., M. T. Nicolela, R. P. LeBlanc, and B. C. Chauhan (2005, December). Visual field progression in glaucoma: Total versus pattern deviation analysis. *Investigative Ophthalmology and Visual Science 46*(12), 4600–4605.

Asman, P. and A. Heijl (1992a, June). Glaucoma hemifield test automated visual field evaluation. *Arch Ophthalmo 110*, 812–819.

Asman, P. and A. Heijl (1992b). Weighting according to location in computer-assisted glaucoma visual field analysis. *Arch Ophthalmo 70*, 671–678.

Asman, P., A. Heijl, J. Olsson, and H. Rootzen (1992). Spatial analyses of glaucomatous visual fields; a comparison with traditional visual field indices. *Arch Ophthalmo 70*, 679–686.

Asman, P., J. Wild, and A. Heijl (2004, September). Appearance of the pattern deviaiton map as a function of change in area of localized field loss. *Investigative Ophthalmology and Visual Science 45*(9), 3099–3106.

Bailey, T. L. and C. Elkan (1993). Estimating the accuracy of learning concepts. In *Proceedings of International Joint conference on Artificial Intelligence*, pp. 895–900.

Birch, M., P. Wishart, and N. ODonnell (1995). Detecting progressive visual field loss in series humphrey visual fields. *Ophthalmology 102*, 1227–1234.

Boden, C., E. Z. Blumenthal, J. Pascual, G. McEwan, R. N. Weinreb, F. Medeiros, and P. A. Sample (2004, December). Patterns of glaucomatous visual field progression identified by three progression criteria. *American Journal of Ophthalmology*, 1029–1036.

Boden, C., P. Sample, A. Bochm, C. Vasile, R. Akinepalli, and R. N. Weinreb (2002). The structure-function relationship in eyes with glaucomatous visual field loss that crosses the horizontal meridian. *Arch Ophthalmol 120*, 907–912.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 123–140.

Brigatti, L., D. Hoffman, and J. Caprioli (1996). Neural networks to identify glaucoma with structural and functional measurements. *American Journal of Ophthalmology*, 511–521.

Cello, K. E., J. M. Nelson-Quigg, and C. J. Johnson (2000). Frequency doubling technology perimetry for detection of glaucomatous visual field loss. *American journal of Ophthalmology 129*, 314–322.

Chan, K., T. Lee, P. A. Sample, M. Goldbaum, R. N. Weinreb, and T. J. Sejnowski (2002, September). Comparison of machine learning and traditional classifier in glaucoma diagnosis. In *IEEE Transactions on Biomedical Engineering*, pp. 963–974.

Chauhan, B., S. Drance, and G. Douglas (1990). The use of visual field indices in detecting changes in the visual field in glaucoma. *Investigative Ophthalmology and Visual Science 31*, 512–520.

Chauhan, B. and P. House (1994). Estimating thresholds in conventional and high pass resolution perimetry using computer simulation. *Journal of Glaucoma 3*, 132–139.

Chauhan, B., P. House, T. McCormick, and R. LeBlanc (1999, January). Comparison of conventional and high-pass resolution perimetry in a prospective study of patients with glaucoma and healthy controls. *Arch Ophthalmol 117*, 24–33.

Chauhan, B. and C. Johnson (1999). Test-retest variability of frequency-doubling perimetry and conventional perimetry in glaucoma patients and normal subjects. *Investigative Ophthalmology and Visual Science 40*, 648–656.

Cochran, W. G. (1950, December). The comparison of percentages in matched samples. *Biometrika 37*(3/4), 256–266.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement 100*(1), 37–46.

Distelhorst, J. S. and G. M. Hughes (2003). Open-angle glaucoma. *American Family Physician 67*, 1937–1944.

Drance, S. M. and D. Anderson (1985). *Automatic perimetry in glaucoma*. Grune & Stration, Inc. ISBN: 0-8089-1705-6.

Draper, N. R. (1981). *Applied regression analysis*. John Wiley & Sones, Inc. ISBN: 0-471-02995-5.

Duggan, C., A. Sommer, C. Auer, and K. Burkhard (1985). Automation differential threshold perimetry for detecting glaucomatous visual field loss. *Am J Ophthalmol. 100*, 420–423.

Fankhauser, F., P. Koch, and A. Roulier (1972). On automation of perimetry. *Albrecht Von Graefes Arch Klin Exp Ophthalmol. 184*, 126–150.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning 2*, 139–172.

Fitzke, F. W., R. A. Hitchings, D. Poinoosawmy, A. I. McNaught, and D. P. Crabb (1996). Analysis of visual field progression in glaucoma. *British Journal of Ophthalmology 80*, 40–48.

Flammer, J., S. M. Drance, and A. Funkhouser (1985). Quantification of glaucomatous visual field defects with automated perimetry. *Investigative Ophthalmology and Visual Science 26*, 176–181.

Flammer, J., S. M. Drance, and M. Schulzer (1984, June). Covariates of the long-term fluctuation of the differential light threshold. *Arch Ophthalmol 102*, 880–882.

Fujimoto, N., K. Minowa, O. Miyauchi, T. Hanawa, and E. Adachi-Usami (2002, February). Learning effect for frequency doubling perimetry in patients with glaucoma. *American Journal of Ophthalmology 133*(2), 269–270.

Gardiner, S. K. and D. P. Crabb (2002a, May). Examination of different pointwise linear regression methods for determining visual field progression. *Investigative Ophthalmology and Visual Science 43*(5), 1400–1407.

Gardiner, S. K. and D. P. Crabb (2002b). Frequency of testing for detecting visual field progression. *British Journal of Ophthalmology 86*, 560–564.

Gillespie, B. W., D. C. Musch, K. E. Guire, R. P. Mills, P. R. Lichter, N. K. Janz, P. A. Wren, and the CIGTS Study Group (2003, June). The collaborative initial glaucoma treatment study: baseline visual field and test-retest variablility. *Investigative Ophthalmology and Visual Science 44*(6), 2613–2620.

Glass, E., M. Schaumberger, and B. J. Lachenmayr (1995, August). Simulations for fastpac and the standard 4-2 db full threshold strategy of the humphrey field analyzer. *Investigative Ophthalmology and Visual Science 36*(9), 1847–1853.

Goldbaum, M. H., P. A. Sample, K. Chan, J. Williams, T. W. Lee, E. Blumenthal, C. A. Girkin, L. M. Zangwill, C. Bowd, T. Sejnowski, and R. N. Weinreb (2002, January). Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry. *Investigative Ophthalmology and Visual Science 43*(1), 162–169.

Goldbaum, M. H., P. A. Sample, H. White, B. Côté, P. Raphaelian, R. D. Fechtner, and R. N. Weinreb (1994, August). Interpretation of automated perimetry for glaucoma by neural network. *Investigative Ophthalmology and Visual Science 35*(9), 3362–3373.

Gupta, D. (2005). *Glaucoma Diagnosis and management.* Lippincott Williams & Wilkins. ISBN: 0-7817-54038.

Han, J. and M. Kamber (2000). *Data mining: concepts and techniques.* Morgan Kaufmann Publishers. ISBN 1558604898.

Harrington, D. O. (1971). *The visual fields* (Third ed.). The C. V. Mosby Company. ISBN: 8016-2056-2.

Heijl, A. and B. Bengtsson (1996). The effect of perimetric experience in patients with glaucoma. *Arch Ophthalmol 114*, 19–22.

Heijl, A., M. C. Leske, B. Bengtsson, B. Bengtsson, M. Hussein, and the EMGT Group (2003). Measuring visual field progression in the early manifest glaucoma trial. *Acta Ophthalmologica Scandinavica 81*, 286–293.

Heijl, A., A. Lindgren, and G. Lindgren (1989, August). Test-retest variability in glaucomatous visual fields. *American Journal of Ophthalmology 108*, 130–135.

Heijl, A., G. Lindgren, and J. Olsson (1986). A package for the statistical analysis of visual fields. *Doc Ophthalmol Proc Ser 49*, 153–168.

Heijl, A., G. Lindgren, J. Olsson, and P. Asman (1989, February). Visual field interpretation with empiric probability maps. *Arch Ophthalmol 107*, 204–208.

Henson, D. B., S. Chaudry, P. H. Artes, E. B. Faragher, and A. Ansons (2000, February). Response variability in the visual field: comparison of optic neuritis, glaucoma, ocular hypertension, and normal eyes. *Investigative Ophthalmology and Visual Science 41*(2), 417–421.

Henson, D. B., S. E. Spenceley, and D. R. Bull (1996). Spatial classification of glaucomatous visual field loss. *British journal of ophthalmol 80*, 526–531.

Henson, D. B., S. E. Spenceley, and D. R. Bull (1997). Artificial neural network analysis of noisy visual field data in glaucoma. *Artifitial intelligence in medicine 10*, 99–113.

Hitchings, R., C. M. R. Wormald, D. Poinooswamy, and F. Fitzke (1994). The primary treatment trial: changes in the visual field analysis by computer-assisted perimetry. *Eye 8*, 117–120.

Hothorn, T. and B. Lausen (2003). bagging tree classifiers for laser scanning images: a data- and simulation-based strategy. *Artificial intelligence in medicine 27*, 65–79.

Hudson, C., J. Wild, and E. O'Neill (1994). Fatigue effects during a single session of automated static threshold perimetry. *Investigative Ophthalmology and Visual Science 35*, 268–280.

Hughes, A. and D. Grawoig (1971). *Statistics: A foundation for analysis.* Addison-Wesley. ISBN: 76-133891.

Hutchings, N., J. M. Wild, M. K. Hussey, J. G. Flanagan, and G. E. Trope (2000, October). The long-term fluctuation of the visual field in stable glaucoma. *Investigative Ophthalmology and Visual Science 41*(11), 3429–3436.

Jansonius, N. M. (2005). Bayes' theorem applied to perimetric progression detection in glaucoma: from specificity to positive predictive value. *Graefe's Arch clin Exp Ophthalmol 243*, 433–437.

Johnson, C., A. J. Adams, and R. A. Lewis (1988). Fatigue effects in automated perimetry. *Appl Optics 27*, 1030–1037.

Johnson, C. A. (1995). Standardizing the measurement of visual fields for clinical research. guidelines from the eye care technology forum. *Ophthalmology 103*, 186–189.

Johnson, C. A., A. J. Adams, E. J. Casson, and J. D. Brandt (1993). Progression of early glaucomatous visual fields loss as detected by blue-on-yellow and standard white-onwhite automated perimetry. *Arch Ophthalmol 11*, 651–656.

Johnson, C. A., S. C. Chauhan, and L. R. Shapiro (1992). Properties of staircase procedures for estimating thresholds in automated perimetry. *Investigative Ophthalmology and Visual Science 33*, 2966–2977.

Katz, J. (1999, February). Scoring systems for measuring progression of visual field loss in clinical trials of glaucoma treatment. *Ophthalmology 106*(2), 391–395.

Katz, J. (2000, April). A comparison of the pattern- and total deviation-based glaucoma change probability programs. *Investigative Ophthalmology and Visual Science 41*(5), 1012–1016.

Katz, J., N. Congdon, and D. S. Friedman (1999, September). Methodological variations in estimating apparent progressive visual field loss in clinical trials of glaucoma treatment. *Arch Ophthalmol 117*(9), 1137–1142.

Katz, J., D. Gibert, H. Quigley, and A. Sommer (1997, June). Estimating progression of visual field loss in glaucoma. *Ophthalmology 104*(6), 10171025.

Katz, J., A. Sommer, D. E. Gaasterland, and D. R. Anderson (1991, December). Comparison of analytic algorithms for detecting glaucomatous visual field loss. *Arch Ophthalmol 109*(12), 1684–1689.

Koch, P., A. Roulier, and F. Fankhauser (1972). Perimetry-the information theoretical basis for its automation. *Vision Res. 12*, 1619–1630.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *The international joint conference on artificial intelligence (IJCAI)*.

Lane, T. and C. E. Brodley (1997). Sequence matching and learning in anomaly detection for computer security. In Fawcett, Haimowitz, Provost, and Stolfo (Eds.), *AI Approaches to Fraud Detection and Risk Management*, pp. 43–49. AAAI Press.

Lavrac, N. (1998). Selected data mining techniques in medicine. In *Proceedings of ECAI-98 workshop on intelligent data analysis in medicine and pharmacology*.

Lazarescu, M. and A. Turpin (2003). Classifying glaucomatous progression using decision trees. In *Proceedings of IASTED International Symposia on Applied Informatics*, Innsbruck, Austria, pp. 205–210.

Ledezma, A., R. Aler, and D. Borrajo (2001). *Data Mining*. Idea Group Publishing.

Lee, A. C., P. A. Sample, E. Z. Blumenthal, C. Berry, L. Zangwill, and R. N. Weinreb (2002). Infrequent confirmation of visual field progression. *Ophthalmology 109*, 1059–1065.

Lin, A., D. Hoffman, D. E. Gaasterland, and J. Caprioli (2003, January). Nerual networks to identify glaucomatous visual field progression. *American Journal of Ophthalmology 135*(1), 49–54.

M, Wall, K. Kutzko, and B. Chauhan (1997). Variability in patients with glaucomatous visual field damage is reduced using size v stimuli. *Investigative Ophthalmology and Visual Science 38*, 426–435.

M. Lazarescu, A. Turpin, S. V. (2002). An application of machine learning techniques for the classification of glaucomatous progression. In *Proceedings of the Syntactic and Structural Pattern Recognition Workshop*, Windsor, Canada, pp. 73–78.

Maddess, T., W. L. Severt, and G. Stange (2001). Comparison of three tests using the frequency doubling illusion to diagnose glaucoma. *Clinical and experimental ophthalmology 29*, 359–367.

Manassakorn, A., K. Nouri-Mahdavi, B. Koucheki, S. K. Law, and J. Caprioli (2006). Pointwise linear regression analysis for detection of visual field progression with absolute versus corrected threshold sensitivities. *Investigative Ophthalmology and Visual Science 47*, 2896–2903.

Membrey, W. L., D. P. Poinoosawmy, C. Bunce, F. W. Fitzke, and R. A. Hitchings (2000). Comparison of visual field progression in patients with normal pressure glaucoma between eye with and without visual field loss that threatens fixation. *British journal of ophthalmolo 84*, 1154–1158.

Mitchell, P., W. Smith, K. Attebo, and P. Healey (1996, October). Prevalence of open-angle glaucoma in Australia. The Blue Mountains Eye Study. *Ophthalmology 103*(10), 1661–1669.

Mitchell, T. M. (1997). *Machine learning.* The MIT Press and The McGraw-Hill Companies, Inc.

Mokhtarian, F., S. Abbasi, and J. K. J (1997). Efficient and robust retrieval by shape content through curvature scale space. *Series on Software Engineering and knowledge Engineering 8*, 51–59.

Montgomery, D. and E. Peck (1992). *Introduction to linear regression analysis* (second ed.). A Wiley-Interscience Publication.

Morgan, R. K., W. J. Feuer, and D. R. Anderson (1991, December). Statpac 2 glaucoma change probability. *Arch Ophthalmol 109*(12), 1690–1692.

Musch, D. C., P. R. Lichter, K. E. Guire, C. L. Standardi, and the CIGTS Study Group (1999, April). The Collaborative Initial Glaucoma Treatment Study. *Ophthalmology 106*(4), 653–662.

Navarro, G. (2001, March). A guided tour to approximate string matching. *ACM Computing Surveys 33*(1), 31–88.

Noureddin, B., D. Poinoosawmy, F. Fietzke, and R. Hitchings (1991). Regression analysis of visual field progression in low tension glaucoma. *British journal of ophthalmol 75*, 493–495.

Nouri-Mahdavi, K., L. Brigatti, M. Weitzman, and J. Caprioli (1997). Comparison of methods to detect visual field progression in glaucoma. *Ophthalmology 104*, 1228–1236.

Nouri-Mahdavi, K., J. Caprioli, A. L. Coleman, D. Hoffman, and D. Gaasterland (2005). Pointwise linear regression for evaluation of visual field outcomes and comparison with the advanced glaucoma intervention study methods. *Arch Ophthalmol 123*, 193–199.

Nouri-Mahdavi, K., D. Hoffma, D. Gaasterland, and J. Caprioli (2004). Predication of visual field progression in glaucoma. *Investigative Ophthalmology and Visual Science 45*, 4346–4351.

Pearson, P., L. B. Baltwin, and T. Smith (1990). The relationship of mean defect to corrected loss variance in glaucoma and ocular hypertension. *Ophthalmologica 200*, 16–21.

Portney, G. and M. A. Krohn (1978). Automated perimetry background, instruments and methods. *Surv Ophthalmol 22*, 271–178.

Quigley, H. A. and A. T. Broman (2006). The number of people with glaucoma worldwide in 2010 and 2020. *British journal of ophthalmology 90*, 262–267.

Quinlan, J. R. (1993). *C4.5 programs for machine learning*. Morgan Kaufmann Publishers, Inc.

Rochtchina, E. and P. Mitchell (2000, June). Projected number of Australians with glaucoma in 2000 and 2030. *Clin Experiment Ophthalmol. 28*(3), 146–148.

Sample, P. A., C. Boden, Z. Zhang, J. Pascual, T. Lee, L. M. Zangwill, R. N. Weinreb, J. G. Crowston, E. M. Hoffmann, F. A. Medeiros, T. Sejnowski, and M. H. Goldbaum (2005, October). Unsupervised learning with independent component analysis to identify areas of progression in glaucomatous visual fields. *Investigative Ophthalmology and Visual Science 46*(10), 3684–3692.

Sample, P. A., K. Chan, C. Boden, T. Lee, E. Z. Blumenthal, R. N. Weinreb, A. Bernd, J. Pascual, J. Hao, T. Sejnowski, and M. H. Goldbaum (2004, August). Using unsupervised learning with variational bayesian mixture of factor analysis to identify patterns of glaucomatous visual field defects. *Investigative Ophthalmology and Visual Science 45*(8), 2596–2605.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning 5*(2), 197–227.

Sheskin, D. J. (2000). *Handbook of parametric and nonparametric statistical procedures.* (second ed.). Chapmanand & Hall/CRC.

Smith, S. D., J. Katz, and H. A. Quigley (1996, June). Analysis of progressive change in automated visual fields in glaucoma. *Investigative Ophthalmology and Visual Science 37*(7), 1419–1429.

Sommer, A., C. Enger, and K. witt (1987). Screening for glaucomatous visual field loss with automated threshold perimetry. *Am J Ophthalmol 103*, 681–684.

Spaeth, G., J. Walt, and J. Keener (2006). Evaluation of quality of life for patients with glaucoma. *American Journal of Ophthalmology 141*, S3–S14.

Spenceley, S. E. and D. B. Henson (1996). Visual field test simulation and error in threshold estimation. *British Journal of Ophthalmol 80*, 304–308.

Spry, P. G., A. B. Bates, C. A. Johnson, and B. C. Chauhan (2000, July). Simulation of longitudinal threshold visual field data. *Investigative Ophthalmology and Visual Science 41*(8), 2192–2200.

Spry, P. G., C. A. Johnson, A. B. Bates, A. Turpin, and B. C. Chauhan (2002, February). Spatial and temporal processing of threshold data for detection of progressive glaucomatous visual field loss. *Arch Ophthalmol 120*(2), 173–180.

Spry, P. G., C. A. Johnson, and B. C. Chauhan (2002, March-April). Identification of progressive glaucomatous visual field loss. *Survey of Ophthalmology 47*(2), 158–173.

Spry, P. G., C. A. Johnson, A. M. McKendrick, and A. Turpin (2001, May). Variability components of standard automated perimetry and frequency-doubling technology perimetry. *Investigative Ophthalmology and Visual Science 42*(6), 1404–1410.

Tan, J., W. A. Franks, and R. A. Hitchings (2002). Interpreting glaucoma progression by white-on-white perimetry. *Grarfe's Arch Clin Exp Ophthalmol 240*, 585–592.

The Advanced Glaucoma Intervention Study Investigators (1994). Advanced Glaucoma Intervention Study, 2: Visual field test scoring and reliability. *Ophthalmology 101*, 1445–1455.

The AGIS Investigators (2000, December). The Advanced Glaucoma Intervention Study, 6: effect of cataract on visual field and visual acuity. *Arch Ophthalmol 118*(12), 1639–1652.

Traquair, H. M. (1944). The nerve fiber bundle defect. *Trans Ophthalmol Soc UK 64*, 1–23.

Tucker, A., V. Vinciotti, X. Liu, and D. Garway-Heath (2004). A spatio-temporal bayesian network classifier for understanding visual field deterioration. *Artifficial intelligence in medicine*.

Turpin, A., E. Frank, M. Hall, I. H. Witten, and C. A. Johnson (2001, April). Detecting progression in glaucoma using data mining techniques. In *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 136–147.

Turpin, A. and A. M. McKendrick (2005). Observer based rather than population based confidence limits for determining probability of change in visual fields. *Vision Research 45*, 3277–3289.

Turpin, A., A. M. McKendrick, C. A. Johnson, and A. J. Vingrys (2002). Development of efficient threshold strategies for frequency doubling technology perimetry using computer simulation. *Investigative Ophthalmology and Visual Science 43*, 322–331.

Vesti, E., C. A. Johnson, and B. C. Chauhan (2003, September). Comparison of different methods for detecting glaucomatous visual field progression. *Investigative Ophthalmology and Visual Science 44*(9), 3873–3879.

Vesti, E., P. G. Spry, B. C. Chauhan, and C. A. Johnson (2002, February). Sensitivity differences between real-patient and computer-stimulated visual fields. *Journal of Glaucoma 11*(1), 35–45.

Viera, A. and J. Garrett (2005, May). Understanding interobserver agreement: The kappa statistic. *Research series 37*(5), 360–363.

Viswanathan, A., F. Fitzke, and R. Hitchings (1997). Early detection of visual field progression in glaucoma: a comparion of progressor and statpac 2. *British Journal of Ophthalmol 81*, 1037–42.

Werner, E., T. Krupin, A. Adelson, and M. Feitl (1990). Effect of patient experience on the results of automated perimetry in glaucoma suspect patients. *Ophthalmology 97*, 838.

Werner, E. B., K. I. bishop, J. Koelle, G. R. Douglas, R. P. LeBlanc, R. P. Mills, B. Schwartz, and W. R. W. J. T. Wilensky (1988). A comparison of experienced clinical observers and statistical tests in detection of progressive visual field loss in glaucoma using automated perimetry. *Arch Ophthalmol 106*, 619–623.

Wikins, M. R., F. W. Fitzke, and P. T. Khaw (2006). Pointwise linear regression criteria and detection of visual field change in a glaucoma trial. *Eye 20*, 98–106.

Wild, J. M., N. Hutchings, M. K. Hussey, J. G. Flanagan, and G. E. Trope (1997, May). Pointwise univariate linear regression of perimetric sensitivity against follow-up time in glaucoma. *Ophthalmology 104*(5), 808–815.

Wilson, M. R. (2002). Progression of visual field loss in untreated glaucoma patients and suspects in st lucia, west indies. *Trans.Am.Ophthalmol. Soc. 100*, 365–410.

Witten, I. H. and E. Frank (2000). *Data mining.* San Francisco, California: Morgan Kaufmann Publishers.