UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DETECTING EAVESDROPPING ACTIVITY IN FIBER OPTIC NETWORKS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

GREGORY G. MACDONALD
Norman, Oklahoma
2012

DETECTING EAVESDROPPING ACTIVITY IN FIBER OPTIC NETWORKS

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF ENGINEERING

BY

_____
Dr. James J. Sluss Jr., Chair

_____
Dr. Samuel Cheng

_____
Dr. Hillel J. Kumin

_____
Dr. William O. Ray

_____
Dr. Pramode K. Verma

This dissertation is dedicated to my father and mother, Gerald and Joy MacDonald, who taught me the important things in life.

Acknowledgements

I would like to thank each of my committee members, whose classes in fiber optics, computer security, stochastic processes and information theory interested me in the world of uncertainty, learning algorithms and machine learning.

Table of Contents

List of Tables

List of Figures

Abstract

DETECTING EAVESDROPPING ACTIVITY IN FIBER OPTIC NETWORKS

Gregory G. MacDonald
University of Oklahoma, 2012

Advisor: James J. Sluss, Jr.

The secure transmission of data is critical to governments, military organizations, financial institutions, health care providers and other enterprises. The primary method of securing in-transit data is though data encryption. A number of encryption methods exist but the fundamental approach is to assume an eavesdropper has access to the encrypted message but does not have the computing capability to decrypt the message in a timely fashion. Essentially, the strength of security depends on the complexity of the encryption method and the resources available to the eavesdropper. The development of future technologies, most notably quantum computers and quantum computing, is often cited as a direct threat to traditional encryption schemes. It seems reasonable that additional effort should be placed on *prohibiting* the eavesdropper from coming into possession of the encrypted message in the first place.

One strategy for denying possession of the encrypted message is to secure the physical layer of the communications path. Because the majority of transmitted information is over fiber-optic networks, it seems appropriate to consider ways of enhancing the integrity and security of the fiber-based physical layer.

The purpose of this research is to investigate the properties of light, as they are manifested in single mode fiber, as a means of insuring the integrity and security of the physical layer of a fiber-optic based communication link. Specifically, the approach focuses on the behavior of polarization in single mode fiber, as it is shown to be especially sensitive to fiber geometry. Fiber geometry is necessarily modified during the placement of optical taps.

The problem of detecting activity associated with the placement of an optical tap is herein approached as a supervised machine learning anomaly identification task. The inputs include raw polarization measurements along with additional features derived from various visualizations of the raw data (the inputs are collectively referred to as "features"). Extreme Value Theory (EVT) is proposed as a means of characterizing normal polarization fluctuations in optical fiber. New uses (as anomaly detectors) are proposed for some long-time statistics (Ripley's ***K*** function, its variant the ***L*** function, and the Hopkins statistic). These metrics are shown to have good discriminating qualities when identifying anomalous polarization measurements. The metrics have such good performance only simple algorithms are necessary for identifying modifications to fiber geometry.

Chapter 1: Introduction

The secure transmission of data is critical to governments, military organizations, financial institutions, health care providers and other enterprises. Traditionally, the emphasis in security has focused on protecting the message to be transmitted by encrypting it at the sender and decrypting it at the receiver. The encryption/decryption process is usually performed by software at the presentation layer of the OSI reference model. The primary goal of the encryption/decryption process is to secure the message for as long as it has some inherent value to an eavesdropper[1]. This process usually assumes an eavesdropper has access to the encrypted message but does not have the computing capabilities to decrypt the message.

The future development of quantum computers and quantum computing is frequently cited as a threat to present day software-based encryption methods. Even *simulations* of quantum computers may become a looming threat to security. In 2010, the Institute for Advanced Simulation at the Julich Supercomputing Centre reported the successful 42-qubit *simulation* of Shor's integer factorization algorithm [2] using the

[1] Absolute security can be achieved by use of the one-time pad; a non-repeating random key that is as long as the message itself [1].

JUGENE supercomputer [3]. Given this looming threat to software-based encryption methods, protection at other levels of the OSI reference model are likely to become more relevant and important. For example, securing the transmitted message at the physical layer has been previously investigated and includes methods such as optical encryption and optical steneganography [4].

It is safe to say that most transmitted information traverses a fiber-optic link at some point in its journey from the sender to the receiver. Once thought to be highly secure, fiber-optic links are easily tapped as demonstrated and described by numerous YouTube videos [5-7]. The tapping devices described in these videos are available at a moderate cost, usually well under $1000. ***In all cases, fiber optic taps require direct access to the fiber optic cable (outer jacket removed).***

Specific instances of fiber-optic tapping are understandably difficult to verify. Organizations that experience incidents of unwanted optical tapping are not usually interested in publicizing the details of such events. One popularly cited incident is described in a 2003 Black Hat Federal Briefing by iDefense and Opterna [8]. In this case, authorities discovered an optical fiber tap (similar to those described in the aforementioned YouTube videos) on a portion of Verizon's network leading to a mutual fund company. The timing of the discovery coincided with the release of quarterly financial numbers.

The desire to intercept in-transit data is not limited to criminal activity. The press has reported that the United States Government recently spent $1 billion dollars to outfit the U.S.S. Jimmy Carter (a Seawolf class nuclear attack submarine) with the capability of tapping undersea fiber-optic links (2005). Conspiracy theories abound as

to the cause of numerous undersea fiber-optic cable interruptions in the Middle East in 2008.

It is clear the U.S. Federal Government considers fiber tapping a potential threat to their own networks and has specified certain measures that must be taken to secure sensitive networks (NSTISSI-7003, AFMAN33-221). Some of these measures include: encasing the fiber-optic cable in cement, installing the fiber-optic cable in a pressurized conduit, and continuous surveillance of the entire fiber-optic link. The first method assumes an eavesdropper will not be able to gain access to the fiber-optic cable (protection but no monitoring). The second method assumes pressure fluctuations will signal attempts to gain access to the fiber-optic cable (monitoring but minimal to no protection). The last method assumes any attempt to gain access to the fiber-optic cable will be directly observed so that appropriate measures can be taken to reduce the possibility of compromising sensitive information.

Over the last few years a market has developed for products capable of detecting fiber-optic taps. A few of the companies that make fiber-optic tap monitors include Eigenlight, Oyster Optics and Optnera. Detection approaches include power analysis for single mode fiber, and speckle analysis for multimode fiber. There are no known papers or patents proposing the use of polarized light for detecting fiber-optic based network tapping activity.

The purpose of this research is to investigate the properties of light, as they are manifested in single mode fiber, as a means of insuring the integrity and security of the physical layer of a fiber-optic based communication link. Specifically, the approach is to examine the behavior of certain properties of light in single mode fiber when the

geometry of the fiber is perturbed. Fiber geometry is necessarily modified during the placement of optical taps. The focus is on the polarization properties of light as they are found to be particularly sensitive to fiber geometry.

This dissertation is organized as follows: Chapter 2 contains a brief overview describing the nature of polarized light, how it is experimentally measured, and its representation in Stokes space. Chapter 3 contains a description of the birefringent properties of single mode fiber, along a description of the physical phenomena that give rise to fiber birefringence. Chapter 4 discusses optical-fiber tapping mechanisms. Chapter 5 documents the experimental configuration used for this dissertation, along with a description of the measured data and various representations of the measured data. These representations lead to other methods and metrics that are helpful in identifying activity leading to the placement of optical taps. This chapter concludes by characterizing the task of detecting optical tapping events as a two-class (no tapping event / tapping event) supervised machine learning anomaly identification problem. Chapter 6 discusses the method for choosing the "best" subset of features for optical tapping identification. Chapter 7 describes the three distributions behind Extreme Value Theory (EVT) and proposes using predictions from EVT to characterize the "natural" state of the fiber. Chapter 8 discusses the results of a simple tree based algorithm on observed and on synthetically produced data. Chapter 9 concludes this dissertation with a discussion of future work and a few summary remarks.

The original contributions of this dissertation are:

- The use of Extreme Value Theory to characterize the "natural" state of optical fiber.
- The use of Extreme Value Theory as an automated domain expert for supervised learning problems.
- The use of Ripley's ***K*** function and specifically its variant, the ***L*** function, as an anomaly detector.
- The use of the Hopkins test statistic as an anomaly detector.
- The use of polarized light in single mode fiber as an ultra-sensitive fiber-optic tapping sensor.

Chapter 2: Fundamentals of Polarized Light

This chapter contains a brief description of polarized light and how it is experimentally measured and represented. Essentially, the state of polarization is determined by the path traced out by the tip of the electric field vector over time. The state of polarization is described in terms of the Stokes parameters. All states of polarization can be conveniently represented with the Stokes parameters mapped onto the Poincare sphere. Every location *on* the Poincare sphere represents a specific state of *fully* polarized light. Every location *inside* the sphere represents some state of *partially* polarized light.

2.1 Fundamental Principles of Polarized Light

Electromagnetic radiation consists of oscillating electric and magnetic fields. The electric field $\boldsymbol{E}$ and magnetic field $\boldsymbol{H}$ vectors are mutually perpendicular and are both orthogonal to the direction of propagation $\boldsymbol{S}$ (sometimes called the *Poynting vector).* The state of polarization is determined by the path traced by the "tip" of the electric field vector $\boldsymbol{E}$ over time**.**

Assuming Cartesian coordinates, and that the direction of propagation ***S*** is along the positive z-axis (outward from the page), we can write the electric field ***E*** vector (for a quasi-monochromatic source) as the sum of two orthogonal components [9].

$$\boldsymbol{E_0} = \hat{\imath}\boldsymbol{E_{0x}} + \hat{\jmath}\boldsymbol{E_{0y}}$$

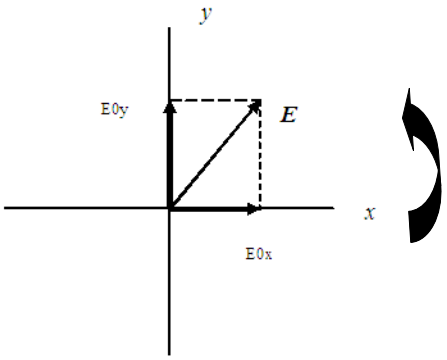


Figure 1 Electric field vector *E* and orthogonal components E_{0x} and E_{0y}.

*This figure shows the electric field vector **E** and the two orthogonal components. The relative amplitudes and phases of the two components determine the direction and path traced by the tip of the electric field vector over time – giving rise to various states of polarization.*

The corresponding wave function is:

$$\boldsymbol{E} = \boldsymbol{E_0}\exp i(\boldsymbol{kz}\text{-}wt)$$

where ***k*** is the propagation constant, ***z*** is the distance along the direction of propagation, *w* is the angular frequency and *t* is time.

If $\boldsymbol{E_0}$ is real, light is linearly polarized. If $\boldsymbol{E_0}$ is complex, light is elliptically polarized. If the real and imaginary parts of $\boldsymbol{E_0}$ are equal, light is circularly polarized.

The relative amplitudes and phases of the two orthogonal components determine the direction and path traced out by the tip of the electric field ***E*** over time. This path is referred to as the state of polarization (***SoP***). If the direction of rotation is clockwise, the ***SoP*** is said to be *right-handed*. If the direction of rotation is counterclockwise, the ***SoP*** is said to be *left-handed*[2].

2.2 Stokes Parameters and the Poincare Sphere

Although the above description of polarization is in terms of the electric field of the light wave, it is not convenient to measure polarization in such a manner. Fortunately, it is possible to measure the optical power of the light wave and *derive* the ***SoP*** in terms of the normalized Stokes polarization parameters (S_1, S_2, and S_3). The normalized Stokes parameters are determined from a set of intensity measurements taken after light is passed through various types of optical elements [10]. Stokes parameters can then be mapped onto the Poincare sphere (figure 2) providing a convenient way to describe any ***SoP***.

Each point on the sphere represents a unique ***SoP***. The region of the sphere where $S_3 = 0$ (the "equator") describes various orientations of linearly polarized light. Areas where $S_3 > 0$ (the "northern hemisphere") represent right-handed elliptically

[2] Various conventions exist for determining right-handedness or left-handedness. As a result, the literature contains conflicting definitions of right and left-handedness. The interpretation appears to depend on whether the person defining polarization is a physicist or electrical engineer.

polarized light, while areas where $S_3 < 0$ (the "southern hemisphere") represent left-handed elliptically polarized light.

The eccentricity of the ellipse depends on its "latitude" while the orientation of the ellipse depends on the "longitude." As we move from the "equator" to the "poles" the polarization ellipticity decreases from 1 to 0. The upper most point on the sphere (the "north pole") represents right handed circularly polarized light, and the lower most point on the sphere (the "south pole") represents left handed circularly polarized light. It should be noted that antipodal points on the Poincare sphere represent mutually orthogonal states of polarization. Points inside the sphere represent *partially* polarized light. Points on the sphere represent *fully* polarized light. The origin of the sphere represents depolarized light.

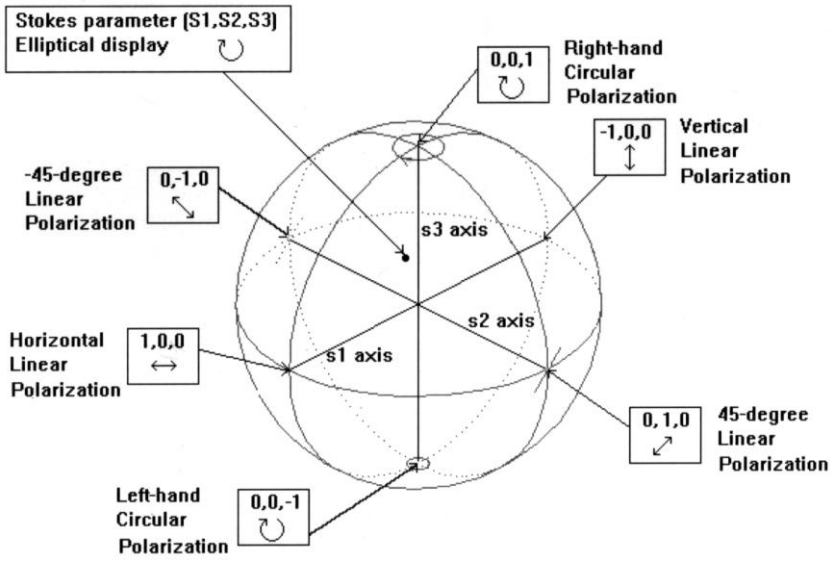


Figure 2 Poincare sphere.

*This figure shows the mapping of the Stokes parameters onto the Poincare sphere. Every **SoP** can be represented by a point on the surface of the sphere (fully polarized light) or inside the surface of the sphere (partially polarized light).*

In terms of the Poincare sphere, the four (non-normalized) Stokes parameters are defined as [11]:

s_0 = total power (polarized + unpolarized)

$s_1 = s_0 \cos(2\gamma)\cos(2\beta)$

$s_2 = s_0 \cos(2\gamma)\sin(2\beta)$

$s_3 = s_0 \sin(2\gamma)$

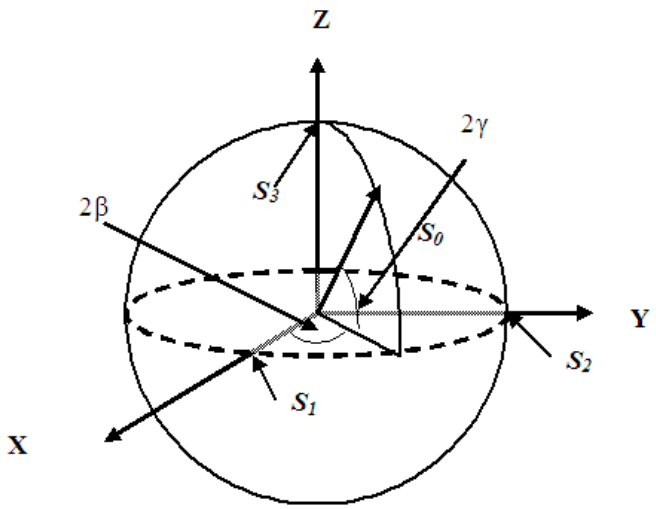


Figure 3 Poincare sphere geometry.

The "normalized" Stokes parameters are given by:

$S_1 = s_1/s_0$

$S_2 = s_2/s_0$

$S_3 = s_3/s_0$

For fully polarized light,

$$s_0 = \{(s_1)^2 + (s_2)^2 + (s_3)^2\}^{1/2}$$

The degree of polarization (***DoP***) is given by:

$$D_{pol} = \frac{\sqrt{S_1^2 + S_2^2 + S_3^2}}{S_0}$$

Example states of polarization along with their respective normalized Stokes parameters are illustrated in figure 4.

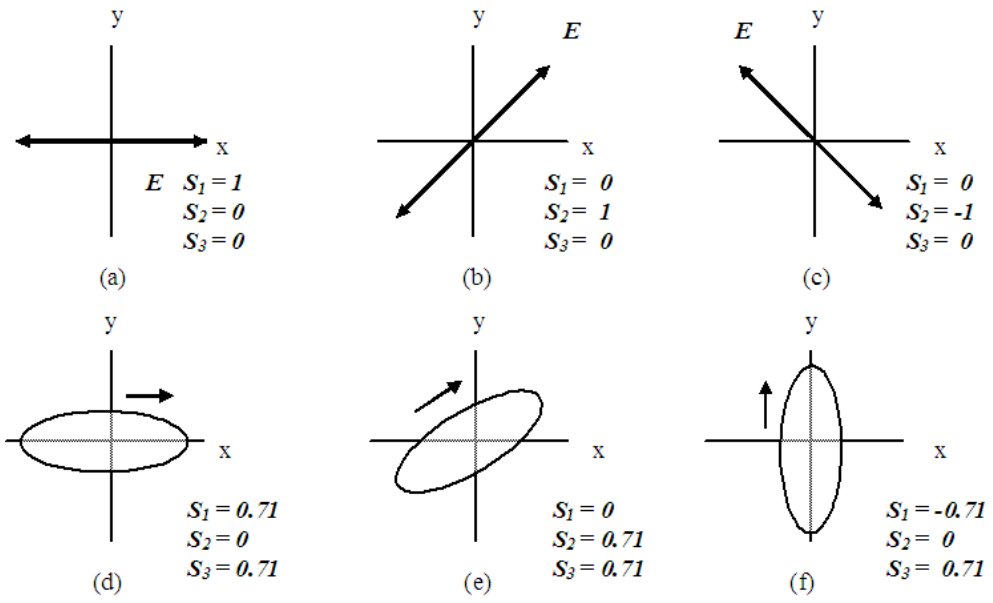


Figure 4 States of polarization and their respective Stokes parameters.

The above figure shows various states of polarization and their respective Stokes parameter values. Figure 4(a) shows linearly polarized light in the horizontal direction (LHP), (b) shows linearly polarized light at 45°, (c) shows linearly polarized light at 135°, (d) – (f) show elliptically polarized light of various orientations.

Some of the examples in figure 4 show Stokes parameters having negative values. How is it possible to have negative values for the Stokes parameters if, as mentioned above, they are derived from intensity measurements? The intensity measures are made using a device called a polarimeter (figure 5). The Stokes parameters are derived using sum and differences of the measured intensities.

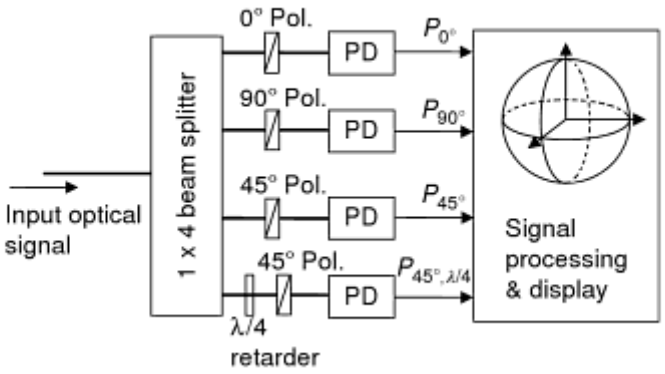


Figure 5 Polarimeter for measuring polarization.

This figure illustrates the typical design of a polarimeter. The optical elements facilitate the measurement of the intensity present in various components of the incoming signal.

The incoming light wave is split into four components of equal intensity using a 4-way beam splitter. The light from each beam splitter is then passed through various optical elements (polarizers and waveplates). Photodiodes (PD) measure the 4 different intensities ($\boldsymbol{P_0}$, $\boldsymbol{P_{90}}$, $\boldsymbol{P_{45}}$ and $\boldsymbol{P_{45,\lambda/4}}$).

The following sums and differences are calculated to arrive at the non-normalized Stokes parameters:

$$s_0 = \boldsymbol{P_0} + \boldsymbol{P_{90}}$$

$$s_1 = \boldsymbol{P_0} - \boldsymbol{P_{90}}$$

$$s_2 = \boldsymbol{2P_{+45}} - s_0$$

$$s_3 = \boldsymbol{2P_{45,\lambda/4}} - s_0$$

Clearly, if the number of photons reaching the $\boldsymbol{P_{90}}$ photodiode is greater than the number of photons reaching the $\boldsymbol{P_0}$ photodiode, the quantity $s_1 < 0$, hence $S_1 < 0$ (recall $S_1 = s_1/s_0$).

Chapter 3: Birefringent Properties of Single Mode Fiber

Fiber birefringence is inherent to all fiber and arises from internal and external sources. It is the birefringent properties of optical-fiber that make tapping activity detectable. The behavior of polarization in optical fiber has been studied from various perspectives. For example, there is an abundance of literature on polarization mode dispersion (PMD) and polarization dependent loss (PDL) as both of these tend to have detrimental effects on the quality of the communication channel. A number of studies have attempted to model fiber birefringence and a few of these are mentioned in passing[3]. The primary importance of this chapter is the studies that characterize the polarization fluctuations in fiber installed in a variety of environments. These fluctuations arise from the birefringent properties of fiber.

3.1 Internal Sources of Birefringence

In single mode fiber above a specific wavelength only one mode of light propagates ($\boldsymbol{HE_{11}}$ mode). This fundamental mode actually consists of two degenerate orthogonal sub-modes ($\boldsymbol{HE_{11(x)}}$ *and* $\boldsymbol{HE_{11(y)}}$). Because of the birefringent properties of

[3] This work treats the effects of birefringence on the properties of light as a stochastic process with minimal reliance on first principles.

single mode fiber these two sub-modes propagate at different velocities, resulting in the random coupling of energy between the two sub-modes. The result is random fluctuations in the state of polarization as the light propagates through the fiber. Randomness aside, there is a characteristic distance along the fiber at which the state of polarization repeats. This distance is known as the beat length and is on the order of 20m – 100m for single mode fiber in a laboratory environment [12].

If a "perfectly" isotropic fiber existed, the received state of polarization would remain unchanged from the transmitted state of polarization. Intrinsic conditions that result in departure from perfectly isotropic conditions occur during the manufacturing process and become a permanent characteristic of the fiber (see section 4.1 for a discussion on fiber construction). They include a non-circular core (geometric birefringence), non-symmetrical stress fields in the cladding (stress birefringence), micro bubbles and other imperfections [13]. Together, these produce anisotropic variations in the indices of refraction along the fiber. These variations in the index of refraction lead to different propagation velocities for the two orthogonal sub-modes.

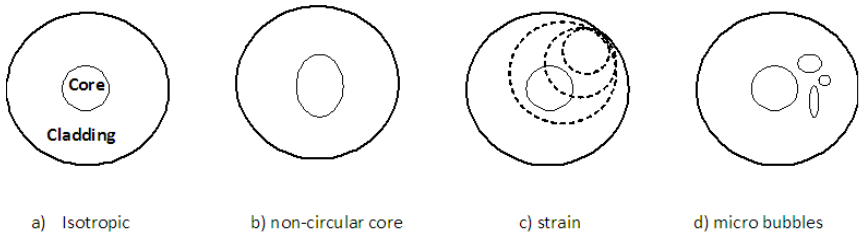


Figure 6 Internal sources of fiber birefringence.

This figure shows some of the intrinsic sources of fiber birefringence, largely attributable to the manufacturing process. These sources give rise to anisotropic conditions causing changing indices of refraction and different propagation velocities.

3.2 External Sources of Birefringence

External sources of birefringence include lateral stresses, bends, temperature variations [14] and external magnetic fields [15-16]. Birefringence is also a function of the wavelength of the light source and so any drift in the source wavelength will add to the overall birefringence properties of the fiber[4]. An analysis of birefringence contributions (produced separately) by asymmetrical lateral stress, bends, twists, and electric/magnetic fields are given in [17]. An analysis of the combined effects is discussed in [18].

3.3 Models of Birefringence

This dissertation treats ***SoP*** fluctuations as a stochastic process with minimal reliance on first principles. However, birefringence modeling efforts include those based on random coupling theory [19-20], probability density functions of the Jones matrix [21], and other statistical characterizations [22-25]. The Jones matrix describes the rotation of the ***SoP*** of light as it propagates through an optical device.

3.4 Time Evolving State of Polarization in Fiber

A number of studies have been performed to characterize the magnitude and nature of evolving ***SoP*** in optical fiber. Polarization fluctuations (in the form of PMD) have been studied for aerial cables, buried cables and for spooled cables in an environmentally controlled setting [26]. It was shown that a variation in fiber

[4] This phenomenon was easily observed in the laboratory when switching the light source from 1310 nm to 1550 nm.

temperature leads to PMD fluctuations and that the rate of fluctuation follows the rate of temperature change. Aerial cables were more susceptible to PMD fluctuations. This was attributed to the frequency of temperature changes. Polarization fluctuations on long terrestrial links also confirmed strong dependence of polarization fluctuations with temperature change [27].

One of the first studies performed on buried fiber found polarization fluctuations to be slow, on the order of hours, with daily fluctuations generally between 2° - 10° as measured on the Poincare sphere [28-29]. This agrees well polarization behavior observed for this dissertation.

Polarization fluctuations for underground fiber and submarine fiber found polarization fluctuations for submarine cable to be significantly larger than the fluctuations for underground fiber [30]. This was presumed to result from physical disturbances produced by wave action, and also to the fact a portion of the cable was exposed on the shoreline, and thus subject to greater temperature variations.

Polarization fluctuations studies were performed in a long-haul terrestrial (1800 km) AT&T WDM link between Florida and Louisiana [31]. Polarization behavior was characterized as "elastic" (a tendency to return to the original state of polarization) or "inelastic" (no such tendency to return to the original state of polarization), with a "transient" event defined as a greater than 10° rotation on the Poincare sphere within 100 msec. Many of the transients were attributed to human activity and to maintenance activity on adjacent fiber cables.

PMD studies of two fibers installed in the *same* physical cable found that polarization drift is remarkably similar between the two fibers [32][5]. This is not to say the ***SoP*** was the same in the two fibers, but rather the magnitude and rate of change of the ***SoP*** in the two fibers were similar.

[5] This observation can be used to provide further discrimination between non-threatening fiber perturbations (maintenance activity) and suspicious fiber perturbations (optical tapping).

Chapter 4: Fiber Optic Tapping Methods

Unlike copper cables, fiber optical cables produce no unwanted or compromising emissions. Interception of the data stream requires access to the fiber-optic cable. The general technique for intercepting data in fiber optic networks involve coupling a portion of the light pulse from the network fiber to an eavesdropping fiber or eavesdropping photodetector using one of the following mechanisms: fiber bending, optical splitting, evanescent coupling, or scattering [33]. *All techniques involve removing some portion of the fiber cable jacket*. Previously mentioned YouTube video references of fiber optic tapping demonstrations all begin with the fiber jacket already removed. Jacket removal is required prior activity. It is this activity that is detectable through polarization monitoring. This activity is detectable *prio*r to the actual placement of the optical tapping device.

4.1 Fiber Construction

The figure below illustrates the most important components of the fiber cable. In single mode fiber (SMF), the purpose of the cladding is to keep the information (light) confined to the core. It is able to do this because the index of refraction of the

cladding is less than that of the core resulting in the confinement of light to the core by way of total internal reflection (Snell's Law).

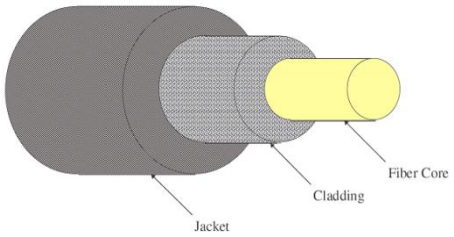


Figure 7 Basic construction of fiber optic cable

In order to "tap" transmissions, light must be coupled out of the core to another device. What follows is a description of typical methods used to couple light out of the fiber core.

4.2 Tapping Methods

In-transit data propagates in the core of the fiber cable. In order to gain access to the transmitted information it is necessary to remove the outer jacket followed by removal of the cladding by a polishing process.

4.2.1 Fiber Bending

Light is coupled out of the core when the bend radius reaches a critical angle. At this radius, total internal reflection is compromised and some portion of the light is emitted from the core. The emitted light can then be detected by a suitably placed optical detector.

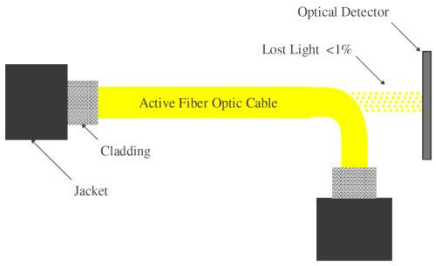


Figure 8 Optical tapping via fiber bending.

4.2.2 Optical Splitting / Evanescent Coupling

These two techniques are essentially the same. In the case of optical splitting, the original path is cut and a pre-manufactured device (splitter) is inserted into the path. Clearly, the act of cutting the path is easily detectable as no flow of information is possible immediately after the cut occurs.

In the case of evanescent coupling, the cladding is polished close to the core and the tapping fiber is placed next to the core resulting in partial capture of light.

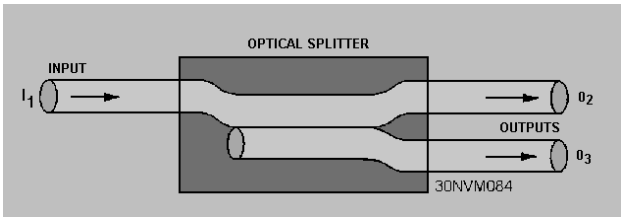


Figure 9 Optical tapping via splitting/coupling.

4.2.3 Scattering

This method involves using a laser to etch a Fiber Bragg grating onto the core of the fiber. Some portion of the light is then scattered out from the core and can then be detected by a suitably placed optical detector.

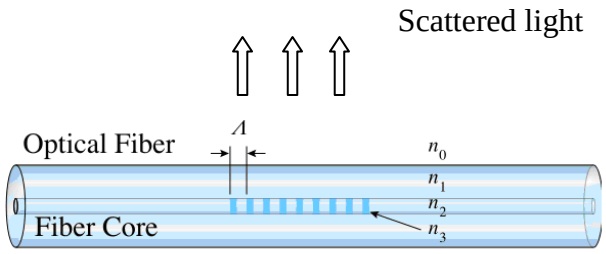


Figure 10 Optical tapping via scattering.

Chapter 5: Fiber Optic Tap Detection – Problem Formulation

This chapter begins by describing the experimental setup used for this dissertation. This is followed by a discussion of a few examples of data obtained from several sets of extended observations. Examples are given for undisturbed fiber as well as for perturbed fiber. ***It will become evident that the attribute of light most sensitive to the environment is the state of polarization.*** This chapter concludes with a discussion of additional representations of the measured data and various features that can be derived from those representations. New uses for some long established metrics are proposed. Some of the metrics are widely used for other purposes in disciplines such as astronomy, botany, ecology, epidemiology and forestry. *The purpose is to construct a set of features that in some way react to the presence of unusual polarization fluctuations.*

5.1 Experimental Configuration

The experimental components for this work consisted of an Agilent 8509C lightwave polarization analyzer, an Agilent 8169A polarization controller (not visible), and a 100m "breakout" cable containing multiple strands of single mode fiber (figure 11). The source laser (λ=1550nm) was located inside of the analyzer. The white

arrows show the path of the light from the polarization analyzer, through the short (orange) cable, into the polarization controller, through the breakout cable, and back to the polarization analyzer.

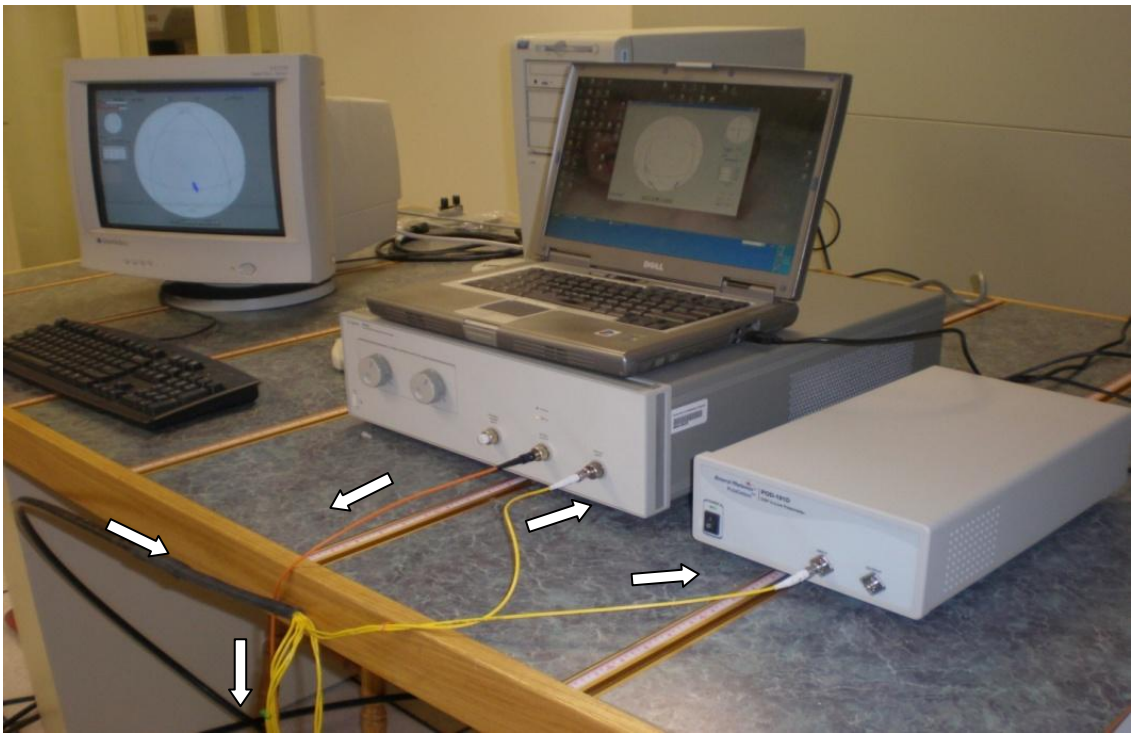


Figure 11 Equipment used for polarization measurements.

The black breakout cable was installed under the raised floor of the TCOM lab. The breakout cable was placed in a cable tray for part of the path and directly on the floor for the remainder of the path. It was not affixed to any structure. The cable ran the length of the raised floor before returning to the instrument area. One of the internal single mode fiber cables was attached to the polarization analyzer. Air handlers produced forced air movement under the raised floor during all data acquisition periods. A nearby air handler produced small but continuous table vibrations.

Measurements were gathered at fifteen (15) second intervals. The following data were recorded at each interval:

1) The power ***P*** of the received signal (this is essentially s_0)
2) The degree of polarization ***DoP*** of the received signal (0%-100%)
3) The state of polarization ***SoP*** of the received signal (expressed in terms of the normalized Stokes parameters - S_1, S_2, S_3)

5.2 Visualization of the Raw Features

The following figures show polarization drift for undisturbed fiber over a period of approximately 68 hours (16,381 measurements taken at 15 second intervals). The polarization drift is relatively slow with little excursion on the Poincare sphere.

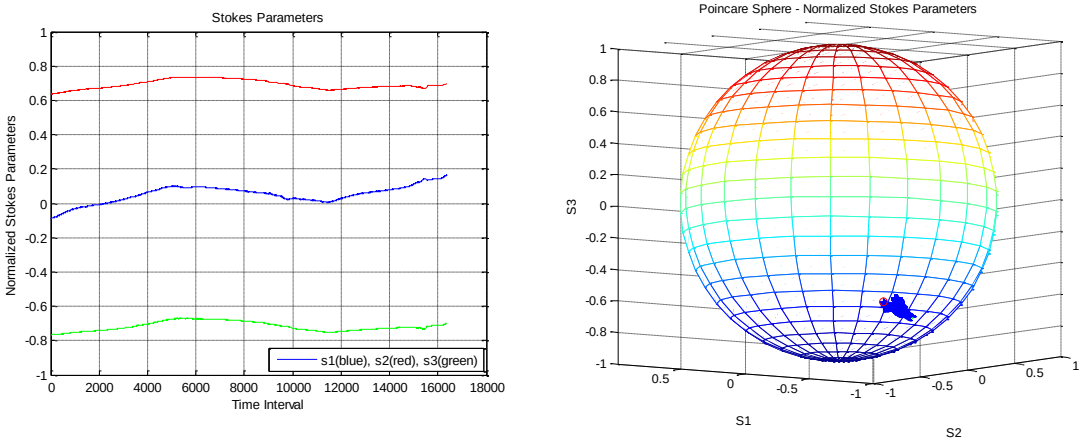


Figure 12 Stokes measurements for unperturbed fiber – case #1.

Another 68 hour collection of data shows much the same behavior. Polarization drift is relatively slow with little excursion on the Poincare sphere.

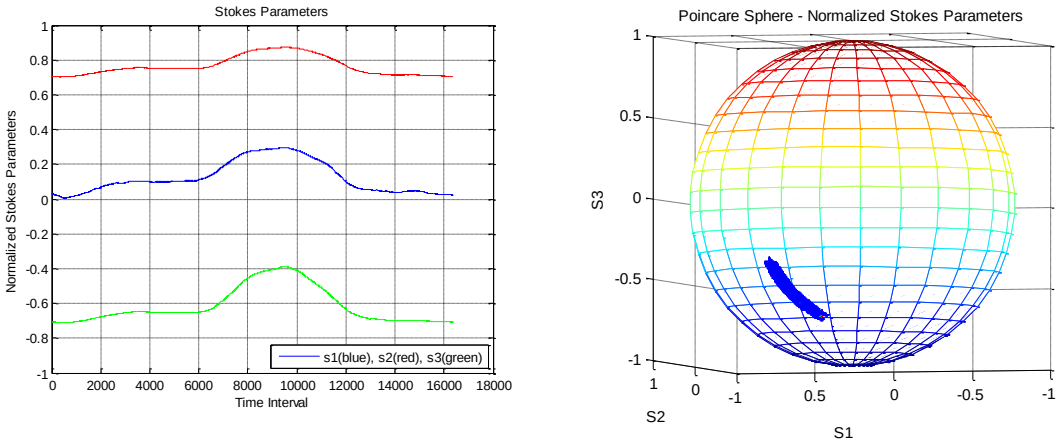


Figure 13 Stokes measurements for unperturbed fiber – case #2.

These following graphs show the degree of polarization ***DoP*** for both cases discussed above (figures 12 and 13). Notice that in both cases the ***DoP*** remains close to 100%. Relative to the ***DoP***, the ***SoP*** is far more sensitive to environmental conditions.

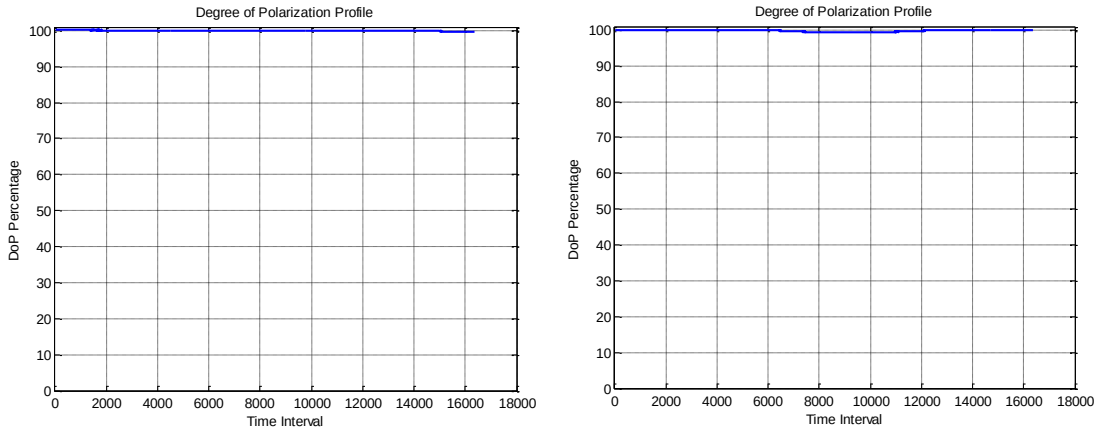


Figure 14 Degree of polarization measurements for unperturbed fiber. (Case #1 [left] and Case #2 [right])

The following graphs show the ***SoP*** and ***DoP*** response to physical perturbations of the fiber. Again, note how much more sensitive the ***SoP*** (figure 15) is to these perturbations as compared to the ***DoP*** (figure 16).

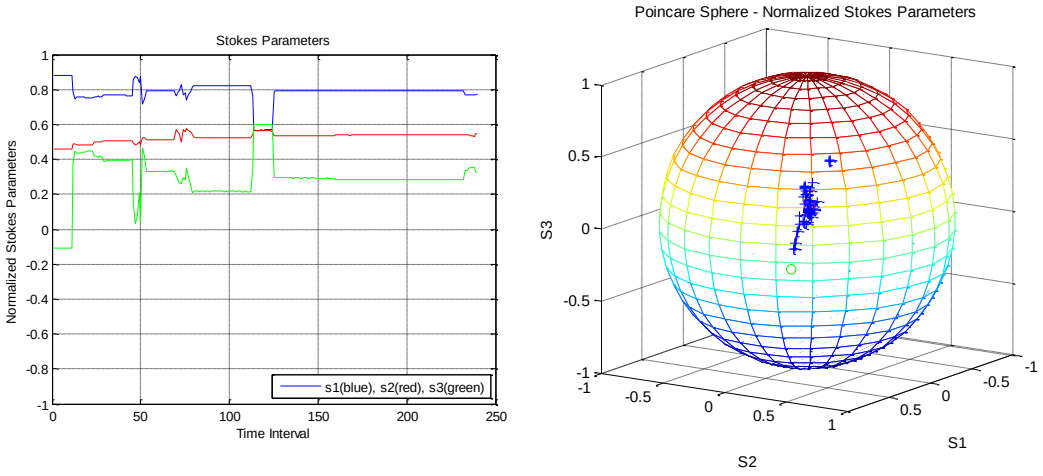


Figure 15 Stokes measurements for perturbed fiber.

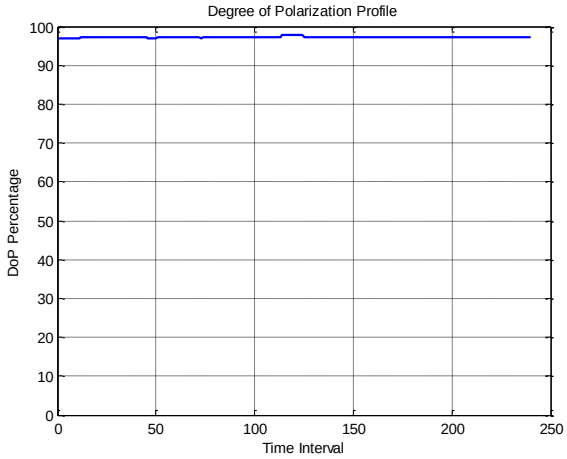


Figure 16 Degree of polarization measurements for perturbed fiber.

This figure shows that the degree of polarization changes very little even when the fiber is perturbed.

The relationship between ***SoP, DoP*** and ***P*** (aka S_0) were discussed in chapter 2. *A key empirical observation is that the degree of polarization* ***DoP*** *remains close to 100% for both cases (undisturbed and perturbed).* Although not shown, the power received ***P*** (S_0) changed very little for the undisturbed and perturbed cases. Essentially, there is little information in either of these attributes.

Recalling for a moment the expression that relates the Stokes parameters to the degree of polarization;

$$D_{pol} = \frac{\sqrt{S_1^2 + S_2^2 + S_3^2}}{S_0}$$

if ***DoP*** and S_0 change very little, a change in one Stokes parameter requires a change in at least one other Stokes parameter and in all likelihood, a change in the other two Stokes parameters. One special case must be considered and that occurs when a disturbance causes one of the Stokes parameters to change sign but not magnitude. Strictly speaking, neither of the other two parameters is required to change. Random perturbations resulting in such a scenario are thought to be exceedingly rare.

Given any two normalized Stokes parameters, the magnitude of the remaining normalized Stokes parameter is largely determined. For example, linearly polarized light in the horizontal direction is represented by normalized Stokes parameters [S_1, S_2, S_3] = [1 0 0]. Consider an external disturbance that causes S_1 to suddenly change to 0.500. Assuming that either S_2 or S_3 do not *both* change and, recalling that the ***DoP*** remains very near 100%, either S_2 or S_3 must change to ± 0.500. Again, it is far more

probable that both S_2 and S_3 undergo change. For this reason, *an observation will be labeled as anomalous if all three Stokes parameters change by a "substantial" amount.* The notion of "*substantial"* will be developed in Chapter 7 with the help of Extreme Value Theory.

5.3 Visualization of Raw Feature Transformations

Visualizing the data in different forms provide some hints as to what metrics may be useful for detecting tapping activity. This section describes some of the different visualizations and mappings of the raw features. These visualizations suggest the construction and development of additional features. These features come from ordinary statistics, point process statistics, and spatial statistics – although the latter two share many common elements.

5.3.1 First Differences (1D) and Features from Ordinary Statistics

In time series analysis and forecasting, a common preprocessing step is to remove any trend in the data and model the resulting time series. Trends can often be removed by taking first differences. In fact, a stationary stochastic process can often be produced using the first differences of the original data [34]. Stationary stochastic processes are a special type of stochastic process, suggesting that the generating process is in a state of equilibrium.

Below are some examples of first differences of the Stokes parameters using the same data as shown in the examples section 5.2. The vertical scale in figure 17 is

highly magnified to show some amount of minutia[6]. The undisturbed fiber case is case #1 and the perturbed fiber case is case #2.

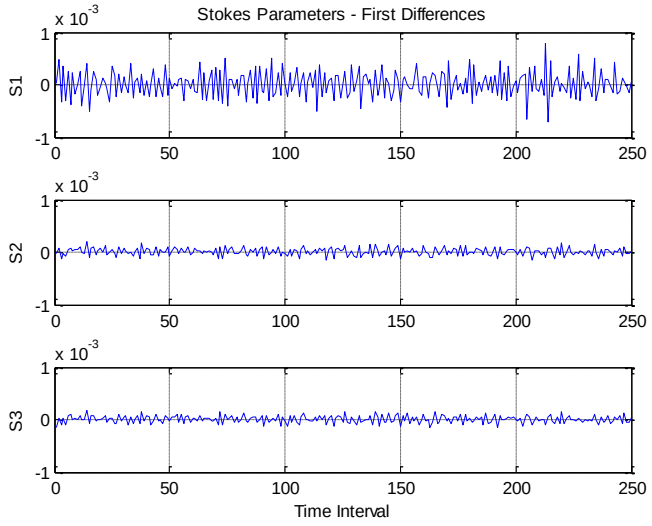


Figure 17 Case #1 - First Differences - undisturbed fiber (highly magnified).

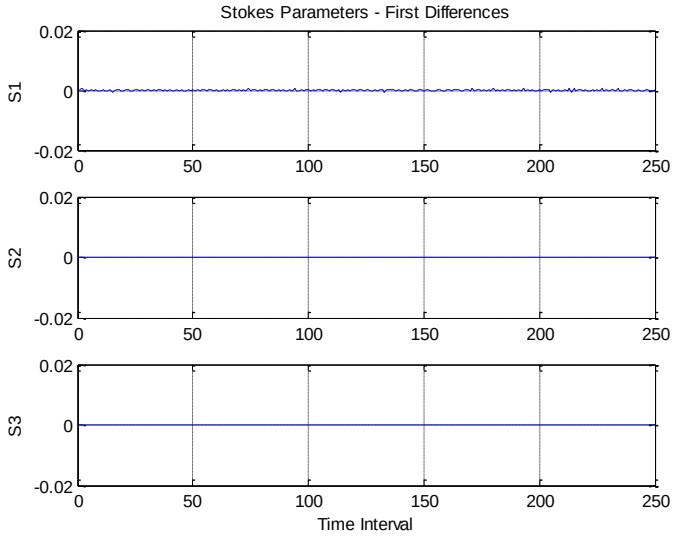


Figure 18 Case #1 - First Differences - undisturbed fiber (slightly magnified).

[6] The "natural" scale for the normalized Stokes parameters is from -1 to 1.

The next two figures show the measures Stokes parameters (figure 19) and the first differences (figure 20). The vertical scale in figure 20 is slightly magnified.

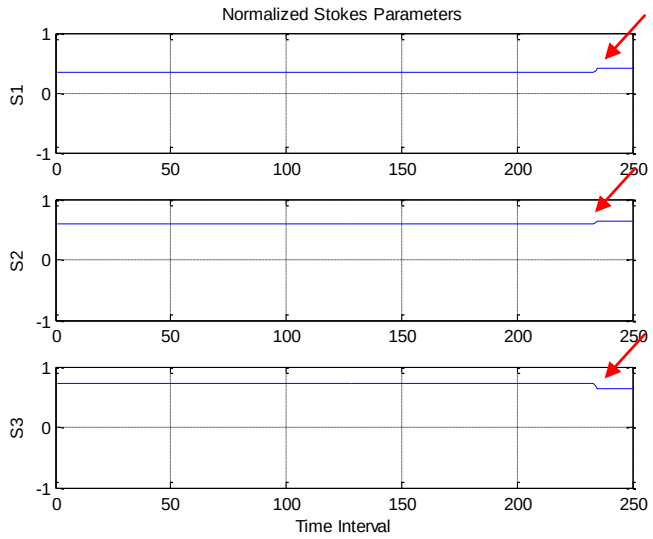


Figure 19 Case #2 - Normalized Stokes parameters - perturbed fiber. (No magnification)

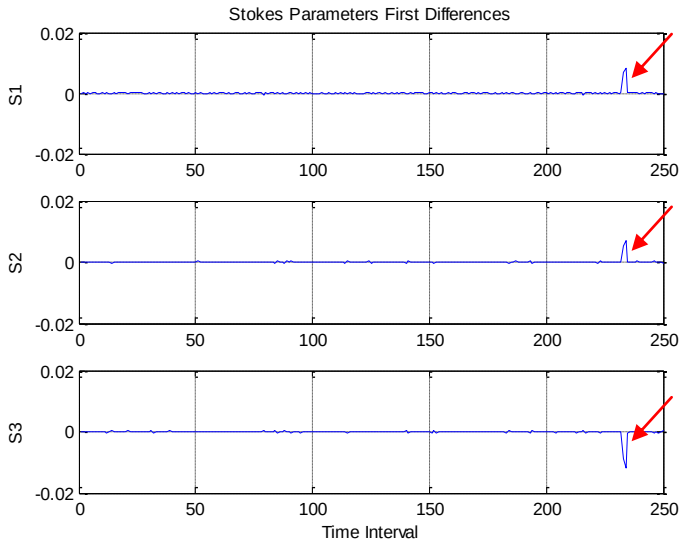


Figure 20 Case #2 - First differences - perturbed fiber. (Slightly magnified)

The polarization measurements obtained can be thought of a sampled points generated from a continuous function. What arouses suspicion of a physical fiber disturbance is the presence of some sort of discontinuity in the measured signal. Discontinuity is well defined for continuous functions but not so well defined for a set of *sampled* data. The definition for discontinuity as it is used here is any "unusually large" change from one interval to the next. As a reminder, the notion of "unusually large" will be developed in Chapter 7 with the assistance of Extreme Value Theory.

Measures from ordinary statistics include the skewness and kurtosis of a distribution. The skewness of a distribution is strongly affected by its symmetry. Skewness may be negative, positive, or undefined. Distributions that are nearly symmetric will have skewness close to zero. The use of skewness to identify anomalies is not new. A recent method called OUTSKEWER identifies outliers based on the skewness of a distribution as extremal values are removed one at a time [35]. Skewness will be close to 0 for a normal distribution.

The kurtosis of a distribution has long been used as an outlier detector [36]. The kurtosis of a distribution is influenced by the tails of the distribution. Kurtosis will be close to 3 for a normal distribution.

Using the data depicted in figure 20 (perturbed fiber) and a moving window with a window size of 50 observations (50 points in the distribution), the graphs of skewness and kurtosis are shown in figures 21 and 22, respectively.

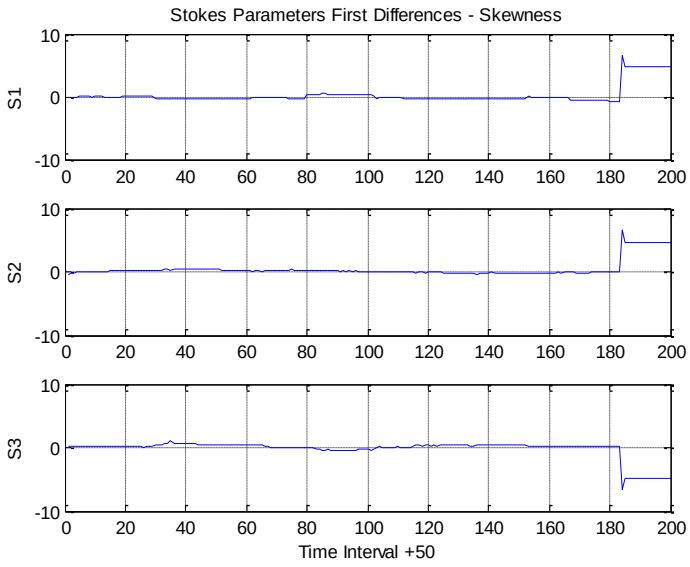


Figure 21 First differences - Skewness (perturbed fiber).

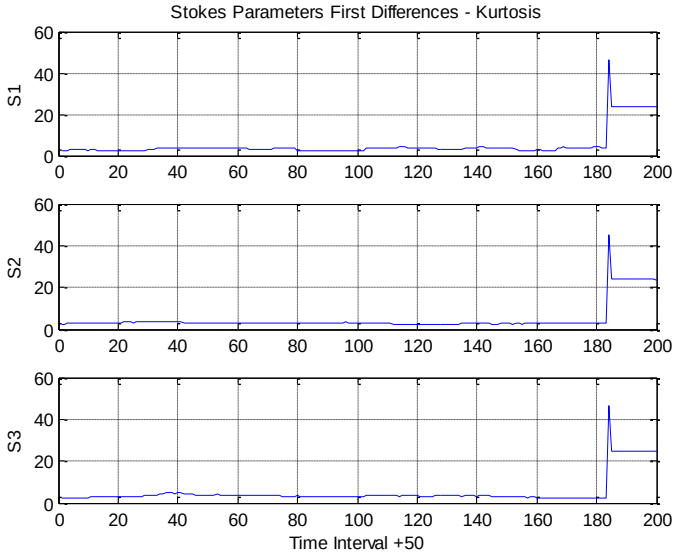


Figure 22 First differences - Kurtosis (perturbed fiber).

The presence of the suspected anomaly as it enters the window of 50 points is clearly visible as a sudden change in the skewness and kurtosis of the distribution. Both the skewness and kurtosis remain elevated as long as this *single* anomaly remains in the window under investigation. It seems obvious that both metrics demonstrate

some potential for identifying the first occurrence of an anomaly in the observation window. The skewness and kurtosis of the window distribution will be added to the list of features for anomaly detection.

5.3.2 First Differences (2D) and Features from Point Process Statistics

Another way to visualize first differences is with scatter plots of the first differences: S_2 vs. S_1, S_3 vs. S_1 and S_3 vs. S_2. The following three graphs show scatter plots of the first differences for undisturbed fiber.

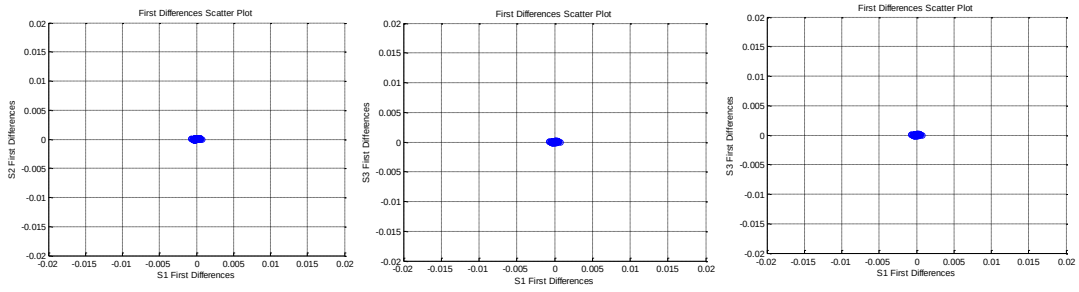


Figure 23 Scatter plot of the first differences - undisturbed fiber.

The following three graphs show scatter plots of the first differences for perturbed fiber.

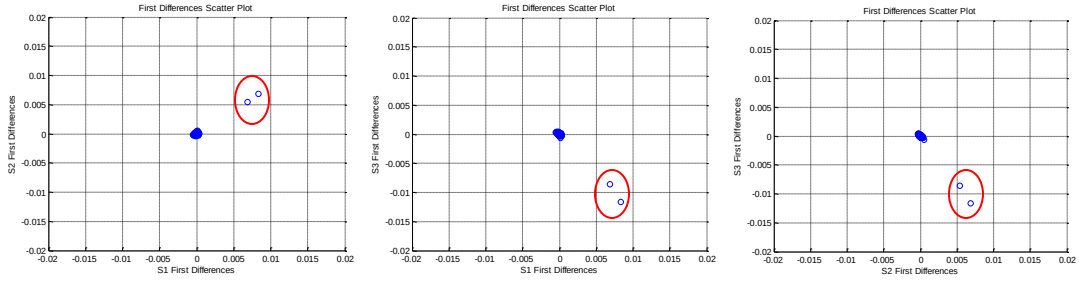


Figure 24 Scatter plot of the first differences - perturbed fiber.

Methods for modeling point processes are used in astronomy [37], ecology [38], epidemiology [39], forestry [40] and other disciplines. There are a plethora of methods, most of which are beyond the scope of this work. As expected, there are the customary characterizations for point processes as there are for ordinary data (e.g., stationarity). A simple test of stationarity for point processes is based on the behavior of the first order intensity λ measure. The standard estimator of intensity λ is defined as the number of points present in a window divided by the size of the window (area for 2D or volume for 3D). Measures involving the degree of dependency amongst events in space are known as spatial autocorrelation measures. Measures of spatial autocorrelation include y include Morans's ***I*** (41), Geary's ***c*** [42], and others.

Two summary characteristics of interest in this dissertation are Ripley's ***K*** function [43], and its variant, the ***L*** function. These two functions are used to detect departure from spatial homogeneity, but actually reveal the same information. The ***L*** function appears to be preferred over the ***K*** function due to some graphical display advantages. The ***K*** function will be discussed first and then the relationship between the ***K*** and ***L*** functions will be described.

Ripley's ***K*** function is a second-order statistic and is commonly used to identify the distance at which clustering occurs. Ripley's ***K*** function characterizes the average number of points found within some distance of a chosen point. A collection of points in space may exhibit clustering, regularity, or complete spatial randomness (CSR). Clustering is easy to visualize because the space appears to consist of one or more groups of points separated away from other points. Regularity is also easy to imagine because the points appear to comprise a grid with regular spacing between the

individual points. Complete spatial randomness is characterized by the homogeneous Poisson point process.

A homogeneous Poisson process ***N*** is characterized by two fundamental properties [44]:

(1) The number of points of ***N*** in any bounded set ***B*** follows a Poisson distribution with mean $\lambda * v(B)$ for some constant λ (λ is known as the intensity or point density and defines the mean number of points in a unit volume ***v***). This is known as the Poisson distribution of point counts.

(2) The numbers of points of ***N*** in ***k*** disjoint sets form ***k*** independent random variables, for arbitrary ***k***. This is known as the completely random property.

The mean number of points to be found in a unit volume is given by:

$$\lambda * \boldsymbol{v}(\boldsymbol{B}) = \mathbf{E}(\boldsymbol{N}(\boldsymbol{B}))$$

The ***K*** function is defined as [45]:

$$\boldsymbol{K}(s) = 1/(\lambda^2 A_r)\sum_{i \neq j}\sum I_s(d_{ij})$$

where A_r is the area of the region of interest, d_{ij} is the distance between the i^{th} and j^{th} events in the region of interest, and $I_s(d_{ij})$ is an indicator function which equals 1 if $d_{ij} \leq s$ and is 0 otherwise. The distance s is usually limited to no more than 0.5 times the length of the shorter side of the rectangular area under study. Ripley's ***K(s)*** has the following interpretations:

If $\boldsymbol{K(s)} = \pi s^2 \rightarrow$ then Poisson process

If $\boldsymbol{K(s)} > \pi s^2 \rightarrow$ then Cluster process

If $\boldsymbol{K(s)} < \pi s^2 \rightarrow$ then Regular process

The result is represented by a graph of $\boldsymbol{K(s)}$ vs. s.

The $\boldsymbol{L}$ function is defined as:

$$\boldsymbol{L(s)} = \sqrt{\boldsymbol{K}(s)/\boldsymbol{\pi}} - s$$

$\boldsymbol{L(s)}$ has the following interpretations:

If $\boldsymbol{L(s)} = 0 \rightarrow$ then Poisson process

If $\boldsymbol{L(s)} > 0 \rightarrow$ then Cluster process

If $\boldsymbol{L(s)} < 0 \rightarrow$ then Regular process

The result is represented as a graph of $\boldsymbol{L(s)}$ vs. s. The graph of the $\boldsymbol{L}$ function has the advantage of showing a Poisson distributed point process as a horizontal line $\boldsymbol{L(s)} = 0$.

Examples of how these functions have been used include the study of neoplasms in humans and dogs in Michigan [46], and the spatial clustering of disease in poultry flocks in Ireland [47]. The first study used the $\boldsymbol{K}$ function to conclude that spatial aggregations of neoplasms in the two species were not independent of one another. The second study used the $\boldsymbol{K}$ function to identify spatial clustering of poultry

flocks affected with the Newcastle disease in Northern Ireland. In practice, it is common to determine the ***K*** function profiles for a control population and diseased population and graph the difference between the two functions as a function of distance *s*.

The behavior of these two functions will be examined using synthetically generated data. The generated data consist of 50 datasets produced by randomly selecting 50 locations in the x-y plane. The randomly generated points are confined to the window bounded by $\{x: 0 \leq x \leq 1\}$ and $\{y: 0 \leq y \leq 1\}$. In this case, the values of *s* will range from 0.01 to 0.50 in increments of 0.01 (in keeping with the practice of limiting the range of *s* to one-half the size of the study region). The goal here is not necessarily to achieve a Poission distributed set of points, but rather to observe the behavior of the two functions when an anomalous data point enters the observation window. Several of the randomly generated datasets are shown in figure 25. The resulting ***K*** and ***L*** function profiles for the 50 randomly generated datasets is shown in figure 26. Both functions tend to suggest a regular distribution of points rather than a tendency to cluster.

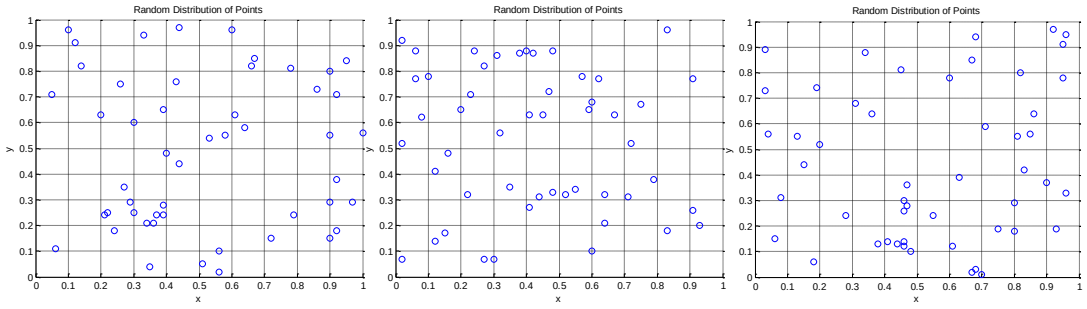


Figure 25 Sample randomly generated datasets.

(No spatial anomaly)

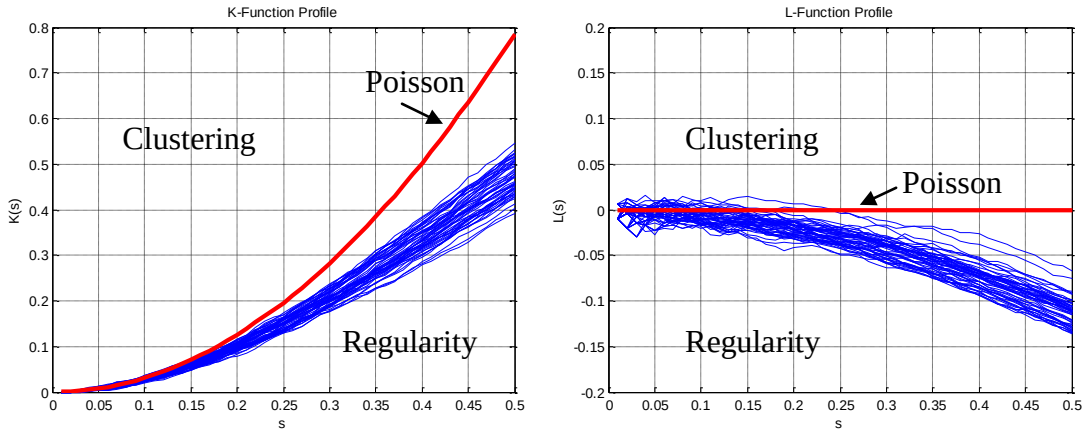
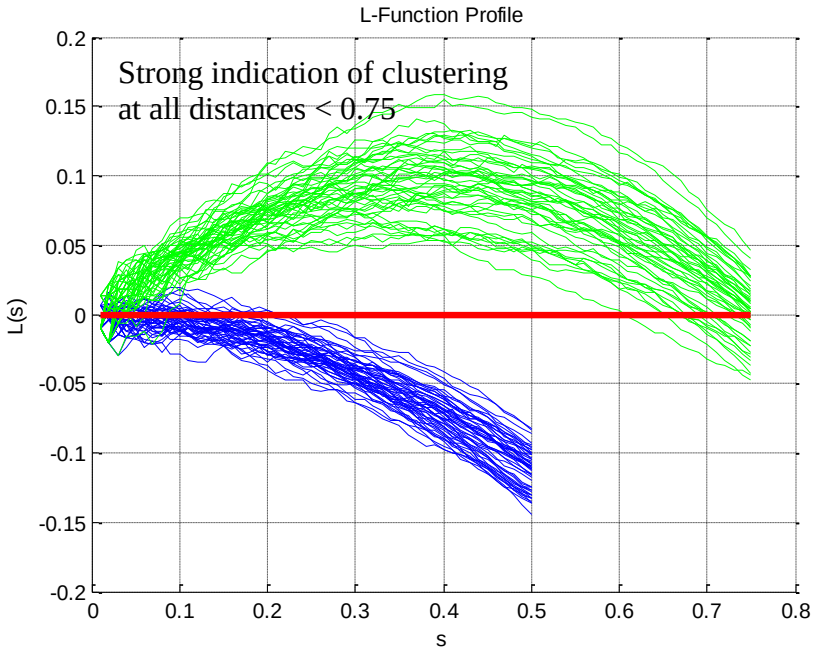


Figure 26 K and L function profiles for randomly distributed data.
(No spatial anomaly)

The following figure shows the ***L*** function profile for the 50 randomly generated datasets and for these same datasets but with one of the data points moved to an "anomalous position" of (x, y) = (1.5, 1.5). The reason why the distance *s* is extended (figure 27) is due to the fact the region of investigation has been expanded because of the location of the anomalous data point.



.

Figure 27 The L-function profile for normal (blue) and abnormal (green) datasets.
(Spatial anomaly present)

The profile for the L function shows strong clustering for datasets containing an anomalous data point.

One modification to the method of calculating the ***K*** and ***L*** profiles is useful for maintaining a constant range of values on the horizontal scale. The modification is to remap all points in the current window to be between {x: -1 ≤ x ≤ 1} and {y: -1 ≤ y ≤ 1}. When no anomalous data points are present in the window, the locations of data points are more or less uniformly spread throughout the bounded region. When an anomalous data point is present in the window, the locations of all other data points are "squeezed" into a much smaller region. This dense region surrounded by space (due to the anomalous data point) suggests the appearance of an emerging cluster of points. In

the presence of an anomalous data point, the expectation is that the function profiles will show evidence of clustering as shown in figure 27.

It seems apparent that both functions demonstrate some potential for identifying the first occurrence of an anomaly in the observation window. Since both functions reveal the same information, only one of these functions is required. The maximum value of the ***L*** function will be added to the list of potential features. The ***L*** function will be calculated after all points in the observation window are mapped into the following space: $\{x: -1 \leq x \leq 1\}$ and $\{y: -1 \leq y \leq 1\}$.

5.3.3 First Differences (2D) and Features from Spatial Statistics

Another way to visualize the data is by using 3D scatter plots of the first differences for S_1, S_2, and S_3. First differences for undisturbed fiber are shown in figure 28. The first differences for perturbed fiber are shown in figure 29.

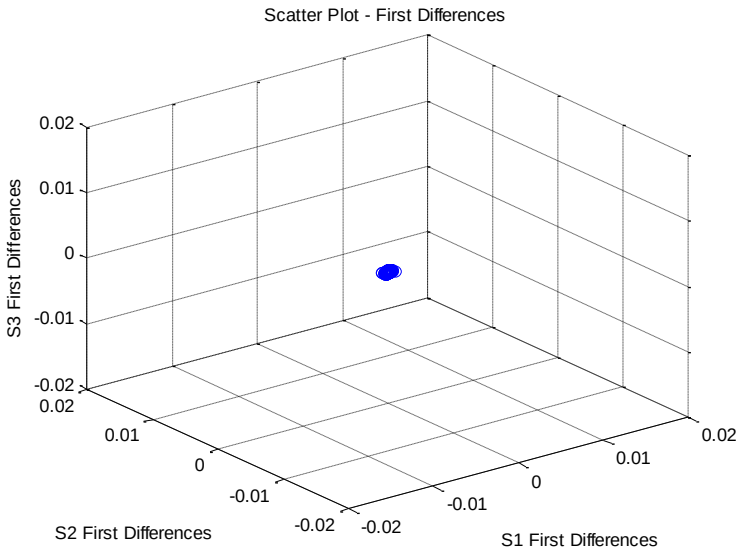


Figure 28 3D first differences – undisturbed fiber.

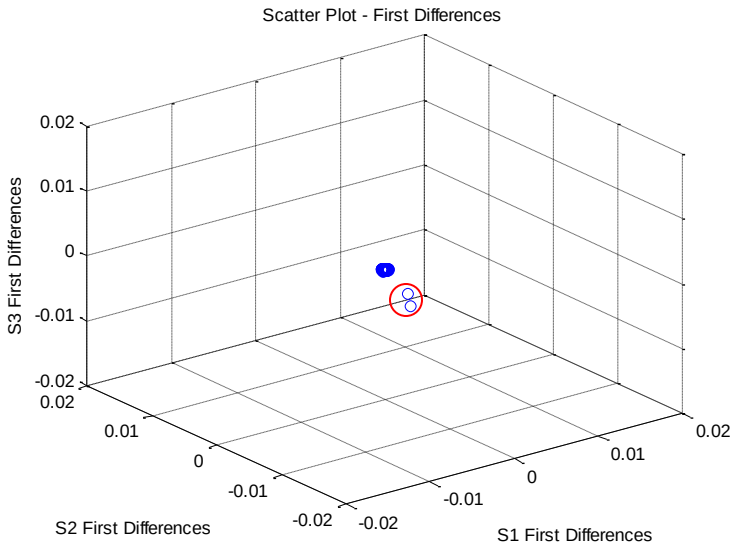


Figure 29 3D first differences - perturbed fiber.

The Hopkins statistic ***h*** [48] is another test statistic for assessing the presence of spatial regularity, spatial randomness, and spatial clustering. The null hypothesis $\boldsymbol{H_0}$ is: the spatial pattern is generated by a Poisson process. The value of the test statistic ***h*** ranges from 0 to 1, with values < 0.50 suggesting spatial regularity, values = 0.50 suggesting randomness (no structure), and values > 0.50 suggesting some measure of clustering. Values close to 1.0 suggest strong clustering.

The Hopkins statistic ***h*** has an *n*-dimensional definition. It is possible to apply the Hopkins statistic ***h*** to the 3D data shown in figures 28 and 29. There are several reasons for choosing multiple 2D analyses instead of a single 3D analysis. The first reason is to permit a direct comparison of the anomaly detection capability of the 2D Hopkins statistic ***h*** against the 2D ***L*** function. If the two are comparable, the ***L*** function would be preferred due to the smaller number of required calculations. The Hopkins method requires multiple simulations in order to determine the value of the Hopkins statistic ***h***.

The Hopkins statistic $\boldsymbol{h}$ is calculated as follows: Let $\boldsymbol{X_n}$ be a collection of $\boldsymbol{n}$ points in 2D space. Let $\boldsymbol{Y_m}$ be a collection of $\boldsymbol{m}$ randomly location points in the same 2D space, with $\boldsymbol{m} << \boldsymbol{n}$[7]. Let $\boldsymbol{u_j}$ be the minimum distance from $\boldsymbol{Y_j}$ to the nearest point in $\boldsymbol{X_n}$, and let $\boldsymbol{w_j}$ be the minimum distance from a randomly selected point in $\boldsymbol{X_n}$ to its nearest point. The Hopkins test $\boldsymbol{h}$ statistic in 2-dimensional space is defined as follows:

$$\boldsymbol{h} = \frac{\sum_{j=1}^{m} u_j}{\sum_{j=1}^{m} (u + w)}$$

In practice, the Hopkins statistic $\boldsymbol{h}$ calculated as the mean value after a number of iterations. Under the null hypothesis (points are generated by a Poisson process) $\boldsymbol{h}$ will be beta distributed with parameters ($\boldsymbol{m}$, $\boldsymbol{m}$).

The notion of extending the usefulness of the Hopkins test statistic beyond assessing the spatial distribution of points includes an estimate of the appropriate *number* of clusters present in spatial patterns that contain multiple sub-clusters [49]. In this dissertation, the usefulness of the Hopkins statistic is redefined to signify the presence of an anomalous event. A point distant from all other points expands the region under investigation making the previously existing points appear to cluster. The expectation is a sudden elevation of the Hopkins statistic $\boldsymbol{h}$ to nearly 1. This is similar to what happens with the $\boldsymbol{K}$ and $\boldsymbol{L}$ functions.

Consider the following random distribution of 50 points in 2D space. Visual inspection of this set of points does not strongly suggest the presence of clustering. The

[7] The usual practice is to choose $\boldsymbol{m} = 0.1\boldsymbol{n}$.

Hopkins test statistic ***h*** for this arrangement of point locations is 0.44 (mean value after 1000 iterations).

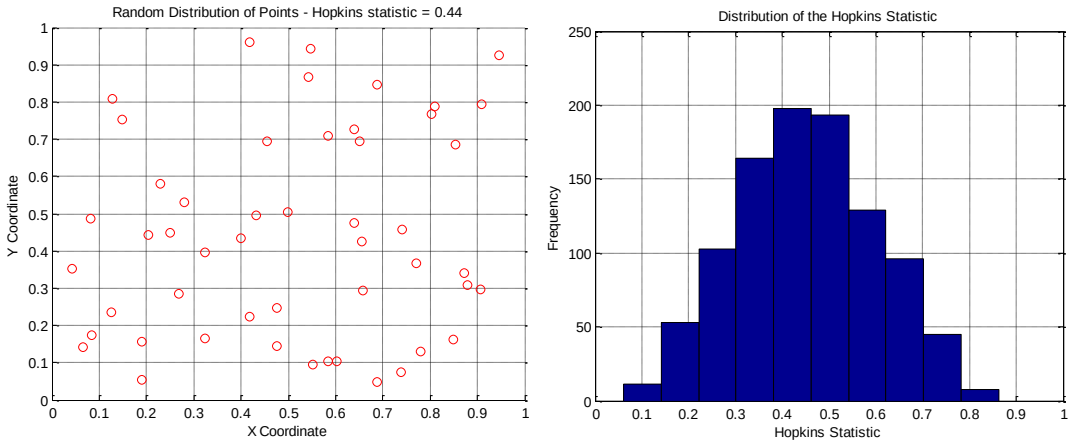


Figure 30 Example of the distribution of the Hopkins statistic. (No spatial anomaly)

Suppose the point located near (x = 0.50, y = 0.50) is moved to (x = 1.2, y = 1.2). Because the observation window has expanded, casual inspection now suggests the presence of one emerging cluster and one data point that seems to be unusually distance from the majority of points (the cluster). The Hopkins statistic for this arrangement of point locations is 0.65 (mean value after 1000 iterations).

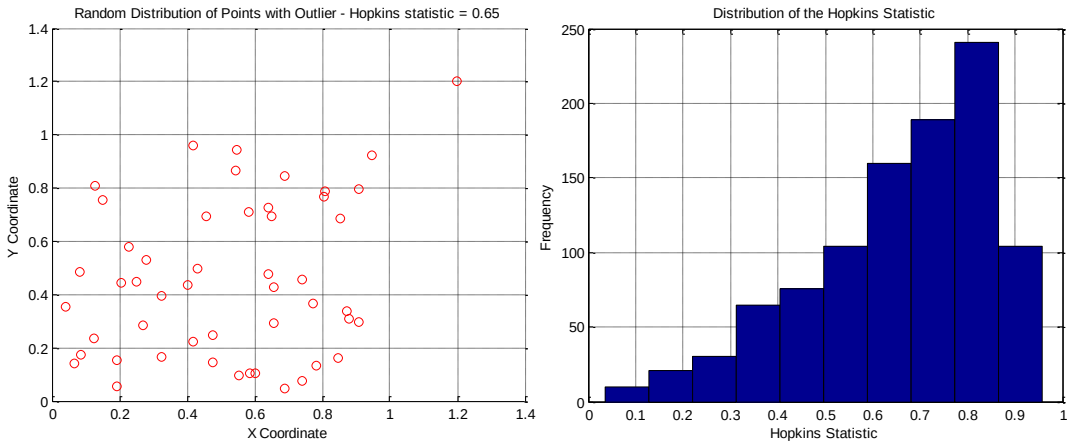


Figure 31 Example of the distribution of the Hopkins statistic. (One moderately anomalous event)

Finally, suppose the point is moved from (x = 1.2, y = 1.2) to (x = 1.5, y = 1.5). This makes the point more distant and consequently "more" anomalous. The Hopkins test statistic for this arrangement of point locations is now 0.83.

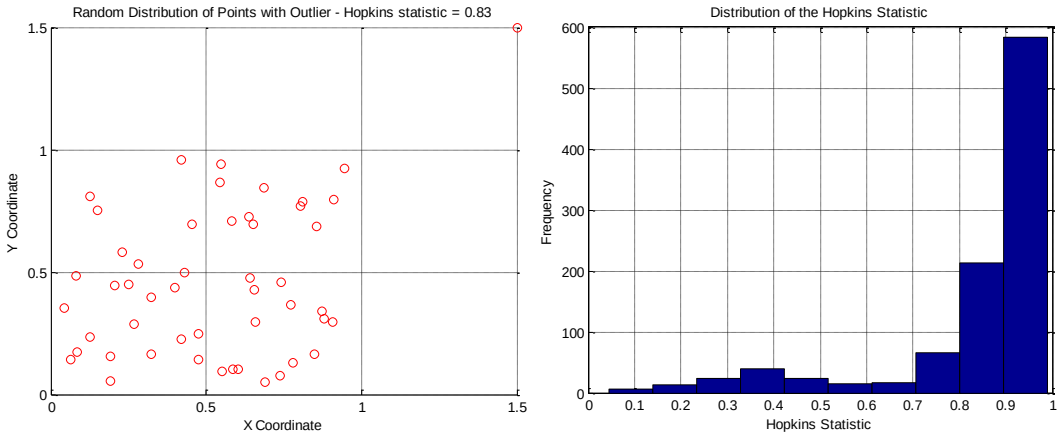


Figure 32 Example of the distribution of the Hopkins statistic. (One severely anomalous event)

Consider the behavior of the Hopkins statistic when applied to a sliding window. Two cases are of interest:

(1) The anomalous set of points persist for a *longer* period of time that is spanned by the sliding window

(2) The anomalous set of points persist for a period of time that is *shorte*r than the time that is spanned by the sliding window

The first case is illustrated using a sequence of measurements shown in figure 33. Let the window size be 50 samples and let the period spanned by the anomalous sequence be 185 intervals, beginning with interval 8586 (figure 33). The Hopkins statistics will be calculated using the spatial arrangement created by a scatter plot of the S_3 first differences vs. the S_1 first differences, taken 50 points at a time.

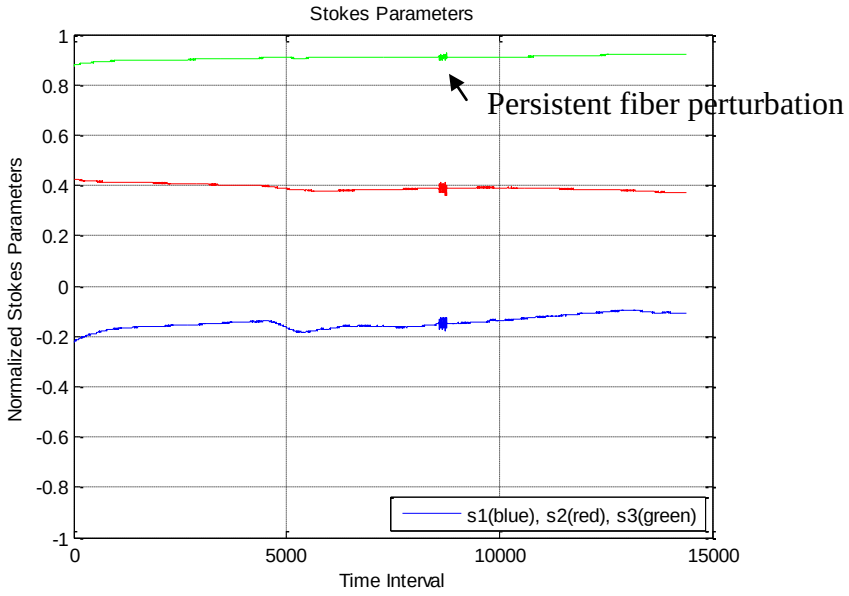


Figure 33 Stokes measurements with synthetically introduced anomalies.

The graph in figure 34 shows the calculated Hopkins statistic prior to the sliding window entering the anomalous sequence ***A***, while the window is fully engulfed by anomalous sequence ***B***, and when the window is exiting the anomalous sequence ***C***. The first anomalous event results in a sudden change in the Hopkins statistic – suggesting the presence of clustering (defined here as an anomaly). During the period when the window is fully engulfed by the anomalous events, the Hopkins statistic eventually returns to values suggesting no clustering present (the exact behavior will depend on the distribution of points). Essentially, a prolonged anomalous period (relative to past behavior) becomes the "new normal" and will diminish the discriminating nature of this statistic. When exiting the anomalous sequence, the Hopkins statistic again rises and then eventually returns to a state that suggests the absence of clustering.

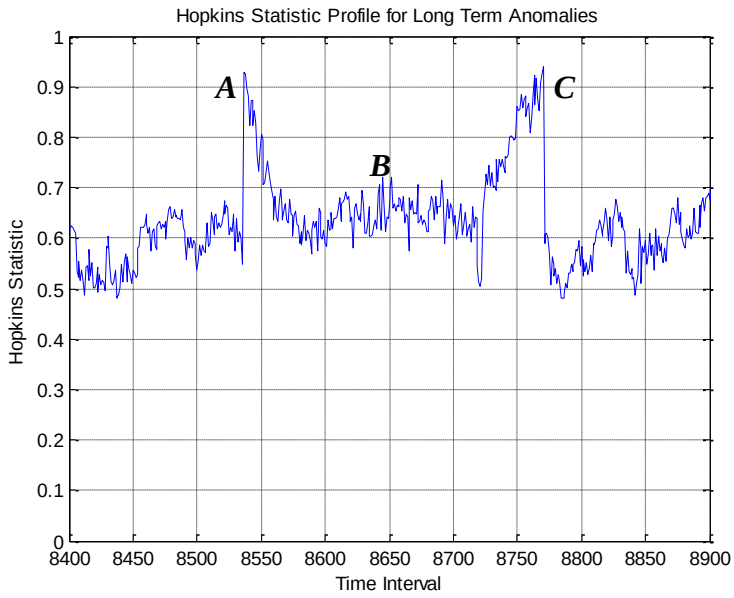


Figure 34 Hopkins statistic profile for long term anomalies.

As another example, consider much smaller perturbations with a shorter duration (shorter than the window size). The perturbations start at interval t = 1280 (figure 35). The first anomalous event results in a sudden change in the Hopkins statistic – suggesting the presence of clustering (figure 36). During the period where the window is larger than the anomalous sequence, the Hopkins statistic decreases but remains somewhat elevated. The Hopkins statistic suddenly decreases as all anomalous events exit the window. The Hopkins statistic tends to maintain its discriminating nature for short term anomalies.

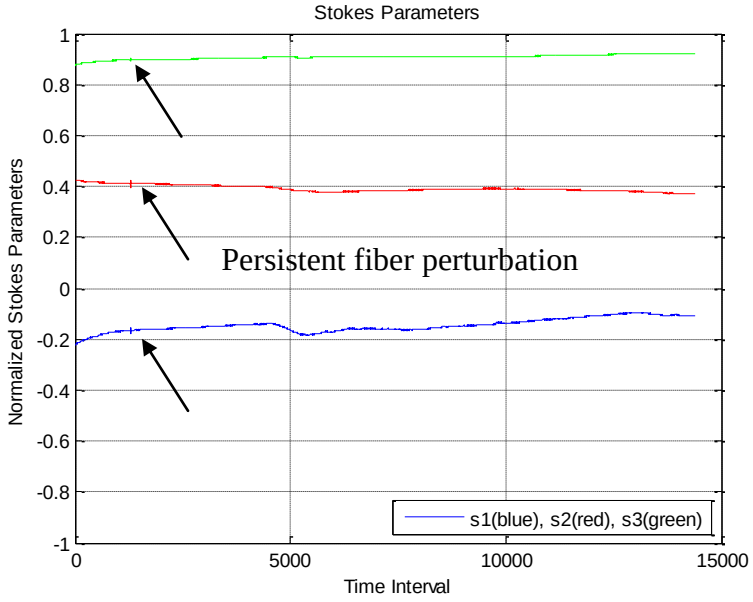


Figure 35 Stokes measurements with synthetically introduced anomalies.

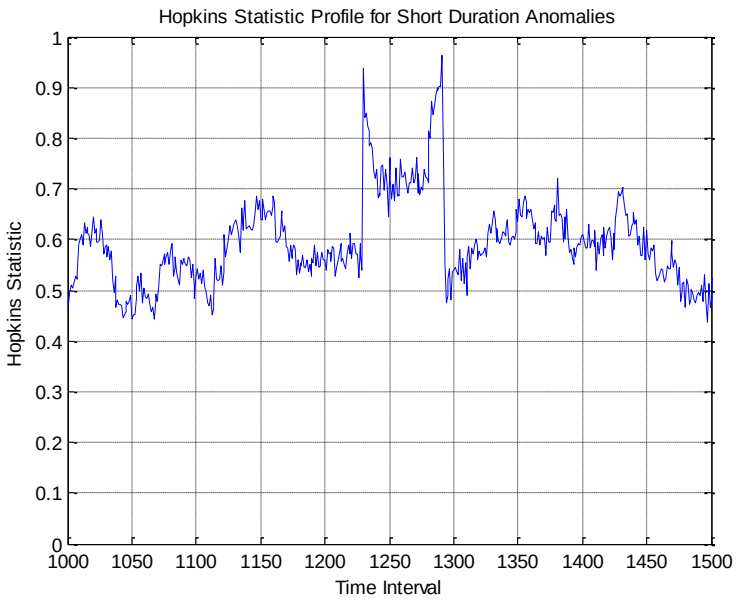


Figure 36 Hopkins statistic profile for short term anomalies.

From these examples, it seems apparent that the Hopkins statistic can be used as an indicator of the presence of a spatial anomaly. The Hopkins statistic will be added to the list of potential features.

5.4 Anomaly Detection: Methods and Tools

The definition of an anomaly in this dissertation is the one given by D. Hawkings [50]:

"An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism."

In this case, the generating mechanism is presumed to be activity leading to the placement of a fiber-optic tap.

Anomaly detection[8] techniques fall into two major categories: detection based on supervised learning, and detection based on unsupervised learning. Tools include neural networks (e.g., replicator neural networks [51] and self-organizing maps), one-class support vector machines, tree methods (e.g., isolation forest [52]), and many others. So many methods and approaches exist that there is even a considerable volume of survey literature [53-56].

Methods are often developed for specific application domains or specific classes of problems and may not be perform well in other application domains. Examples of application domains include: hyperspectral imagery for target detection [57], network intrusion detection [58-61], trajectory-based anomaly identification [62], VHF radar anomaly detection for oceanography applications [63], and others.

For some application domains, anomalous data is not considered to be of value but a nuisance, and is removed from the collection as part of a data preprocessing task. These kinds of anomalies are usually referred to as outliers.

[8] Anomaly detection is also referred to as novelty detection or outlier detection.

Chapter 6: Feature Selection

The previous chapter produced a number of features. Each of these appears to have some ability to identify the presence of anomalous behavior. Some features may prove to be better "predictors" than others. This chapter discusses the method for selecting the features that have the greatest potential for producing a high performing anomaly detector. The method of choice is the ReliefF algorithm. Examples of feature selection using the ReliefF algorithm will be given for classification and regression problems.

6.1 Feature Selection – Common Approaches

One of the tasks at hand in learning any concept is to choose the subset of features that appear to be the best "predictors" of the target value, and to discard features that do not appear to be of significant predictive value. One of the benefits of eliminating these so called low-value features is a reduction in computational complexity. This translates to quicker runtimes and results. Elimination of low-value

features can also help the generalization characteristics and overall performance of the resulting algorithm [64-65][9].

There are many methods of performing feature selection. They include myopic and non-myopic methods. Myopic approaches rank feature importance independently of other features. Essentially, feature independence is assumed as far as the target value is concerned. Non-myopic approaches rank feature selection after considering feature interactions. Myopic methods include statistical hypothesis testing, class separability measures (divergence, Bhattacharyya distance), scatter matrices (within-class scatter, between-class scatter) [67], and information theoretic measures (information gain, gain-ratio, minimum description length) [68]. The *J-measure* is sometimes used to assess the importance of a particular feature *value*.

Caution should be exercised when using myopic feature selection methods. A feature is often considered to be of low-value if it is highly correlated with another feature. The common belief is that selecting one feature is sufficient because no further information can be teased out from the highly correlated feature. Several interesting example have been given that counter this and some other widely believed practices regarding feature selection [69]. The examples demonstrate the following:

1) Noise reduction and consequently better class separation may be obtained by adding variables that are presumably redundant.

[9] This appears to be somewhat counterintuitive, especially in context of the information theoretic "data processing inequality theorem" [66] which loosely states: "no clever manipulation of the data can improve the inferences that can be made from the data."

2) Very high feature correlation (or anti-correlation) does not mean the absence of variable complementarity. However, perfectly correlated features are truly redundant.

3) A feature that is completely useless by itself can provide a significant performance improvement when taken with others.

4) Two features that are useless by themselves can be useful together.

Feature selection also includes filter methods and wrapper methods. Filter methods select subsets of variables as a pre-processing step without regard to concept to be learned. The selection criteria are independent of the problem at hand and do not consider the target value. Wrapper methods consider the target value and try different collections of features to see which subset performs best. This can be dependent on the learning algorithm as well as the concept to be learned. Wrapper methods are more complex since feature selection depends on solving the problem at hand many times. The approach used for this dissertation is the non-myopic assessment provided by the ReliefF algorithm.

6.2 The Relief/ReliefF Algorithms

The original Relief algorithm [70] provides a non-myopic method of assessing feature importance for binary classification problems. The Relief algorithm takes into consideration not only the difference in feature values and classes, but also the *distance* between the feature vectors. By taking the distance between feature vectors into account, the contribution of each feature is assessed in a non-myopic manner. The

basic idea is to assign a relative weight to each feature indicating the feature's value to the concept to be learned: the higher the relative weight, the higher value the feature is to learning the concept. Features with relative weights < 0 are considered to be irrelevant to the concept to be learned.

The original Relief algorithm is as follows: Given a random sample of ***m*** feature vectors from a dataset having ***n*** feature vectors and ***a*** features, assign a relative quality measure (weight) $-1 \leq \boldsymbol{W} \leq 1$ to each feature according to the following:

1. assign all features to have a weight ***W*** of 0
2. select one of the feature vectors at random
3. find the nearest feature vector from the same class ***H*** and the nearest feature vector from the other class ***M***
4. for each feature, calculate/update the weight ***W*** of the feature
5. go to step 2 until all ***m*** samples are processed

The weight ***W*** of each feature is calculated and updated according to the following:

$$\boldsymbol{W} = \boldsymbol{W} - \text{diff(feature,instance,H)}/\boldsymbol{m} + \text{diff(feature,instance,M)}/\boldsymbol{m}$$

where, diff(feature, instance, ***H*** or ***M***) has several different definitions depending on whether the feature is discrete or continuous.

The presence of noise (either in classification or in the feature values themselves) can affect the determination of the nearest feature vector. The original Relief algorithm is very sensitive to noise. An extension to the Relief algorithm, known as ReliefF [71], is less sensitive to noise and will even handle cases where some

feature values are missing[10]. The effect of noise is mitigated by taking not just the nearest feature vector from both classes, but the nearest $\boldsymbol{k}$ feature vectors from both classes where $\boldsymbol{k} \in [5, 20]$.

6.2.1 Feature Section for Classification

As an example of ReliefF, consider a feature vector $\boldsymbol{x}$ consisting of 5 features whose values are drawn at random from beta distributions having the following parameters: (4.6, 2.0), (5.3, 2.1), (9.8, 2.8), (2.7, 2.1) and (8.7, 9.8). Values generated from the distributions have range [0, 1]. Assume that class membership is determined according to the unknown rule:

If the value of feature 1 < 0.4360 and the value of feature 4 > 0.8620 → class 1, Otherwise → class 2

Clearly, features 2, 3, and 5 are irrelevant when it comes to determining class membership. The ReliefF algorithm ($\boldsymbol{k}$=5) assessed feature importance according to the following relative weights (figure 37):

[10] A theoretical and empirical comparison of the various Relief algorithms is given in [72].

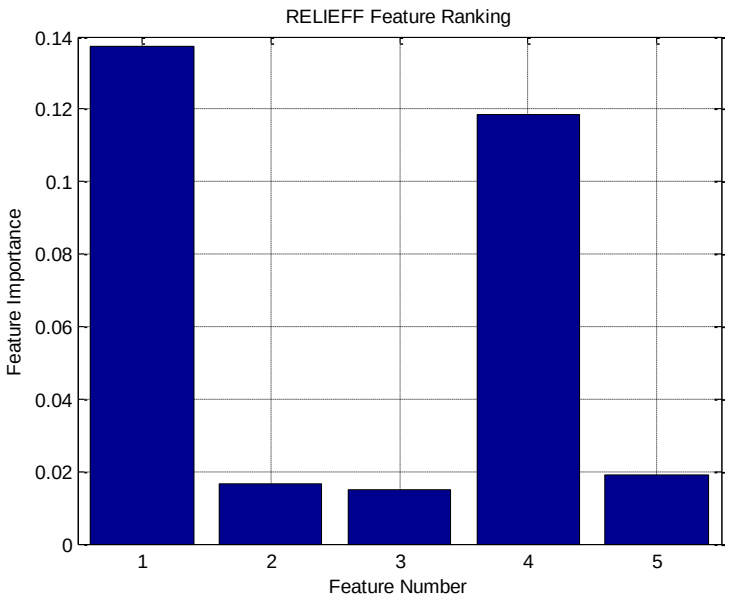


Figure 37 Example feature selection using the ReliefF algorithm – Classification.

Although no concept has been learned, ReliefF correctly determined that features 1 and 4 were more relevant to learning the concept than were features 2, 3, and 5.

6.2.2 Feature Section for Regression

As another example, consider a feature vector ***x*** consisting of 10 features whose values are drawn at random from a uniform distribution (all values are between 0 and 1). Let the unknown concept to be learned be defined as follows:

$$Target = feature(1)^{2.5} + feature(4) - 2*feature(9)$$

Clearly, features 2, 3, 5, 6, 7, 8, and 10 are irrelevant to the concept. The ReliefF algorithm (***k***=5) assessed feature importance according to the following relative weights (figure 38):

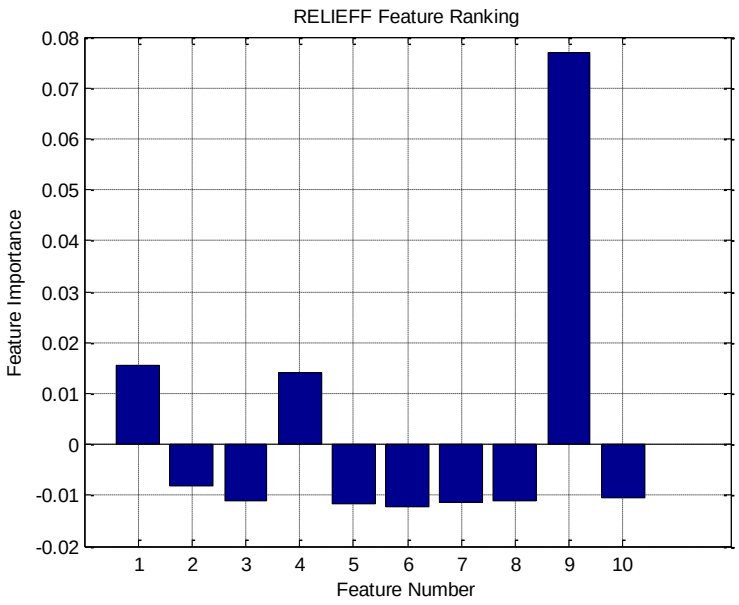


Figure 38 Example feature selection using the ReliefF algorithm – Regression.

Again, although no concept has been learned, ReliefF determined that features 1, 4 and 9 were more relevant to learning the concept than were the remaining features. (Recall that feature importance scores < 0 are considered to come from features that are irrelevant to the concept).

The ReliefF algorithm will be used to assess the relative quality of the features developed in Chapter 5.

Chapter 7: The Domain Expert – Extreme Value Theory

One of the important goals in this dissertation is to identify a small subset of events that are substantially different from the majority of events, without defining what "substantially different" means *a priori*. The approach is to learn the majority behavior from the data so that aberrations in the data are easily identified. For this application it is assumed that aberrations are extremely rare.

The two main learning methods are supervised learning and unsupervised learning. For binary classification tasks, supervised learning involves the labeling of each event (usually performed by a domain expert) as being in member of one class or the other class. Supervised learning is tedious in that the domain expert must manually examine and label each event. Unsupervised learning does not require class assignment by a domain expert, but attempts to classify events based either on measures of similarity or measures of dissimilarity (distance measures, clustering methods, forest methods, and one-class support vector machines). The approach taken here is to treat this problem as a supervised learning problem using Extreme Value Theory (EVT) to characterize the state of the undisturbed fiber and then to transition to the role of the domain expert (automated).

7.1 Classical Extreme Value Theory (EVT)

Extreme value theory [73] is somewhat unique in that it is used to model events that are more extreme than the events observed in the past[11]. EVT is concerned with the study of asymptotic behavior of extreme and consequently, rare events. In its most straight forward form, EVT provides a method of probabilistically predicting the occurrence of events having magnitudes that are less than or greater than some threshold[12]. Generally, the threshold is usually lower or higher than the magnitude of any previously recorded event. EVT is used in the field of hydrology to predict the probability and severity of flooding events [74], and in climatology to predict the probability and severity of rainfall events [75]. It is also used in financial risk management analysis, value at risk analysis and other applications.

EVT focuses on the statistical behavior of,

$$M_i = \max\{X_1, \ldots, X_n\}_i$$

where $X_1, \ldots, X_n$ is sequence of independent random variables having a common distribution function $\boldsymbol{F}$.

In practice, $X_1, \ldots, X_n$ are a sequence of observations (or measurements) taken at regular time intervals, while M_i is the maximum of the i^{th} sequence. A common example would be the daily rainfall amounts ($X_1, \ldots, X_{365}$) for some geographical

[11] Gaussian methods tend to model central tendency of distributions. EVT models the tails of distributions.

[12] EVT is also used to predict probabilities of minimal magnitude events. This is commonly found in applications of structural analysis and failure where the minimal behavior of some structural element is of interest.

location, with M_i being the annual maximum rainfall for the same geographical location. The *Extremal Types Theorem*[13] says the following:

If sequences of constants $a_n > 0$ and b_n exist such that:

$$\Pr\{(M_n - b_n)/a_n \leq z\} \rightarrow G(z) \text{ as } n \rightarrow \infty$$

where G is a non-degenerate distribution function[14], then G belongs to one of the following distributions:

I: $G(z) = \exp\{ -\exp [-(z-b)/a]\}$ for $-\infty < z < \infty$

II: $$G(z) = \begin{cases} 0 & \text{for } z \leq b \\ \exp\{ -[(z-b)/a]^{-\alpha}\} & \text{for } z > b \end{cases}$$

III: $$G(z) = \begin{cases} \exp\{ -[-(z-b)/a]^{-\alpha}\} & \text{for } z < b \\ 1 & \text{for } z \geq b \end{cases}$$

for parameters $a > 0$, and b. In the case of families **II** and **III**, $\alpha > 0$.

In other words, the rescaled sample maxima converge in distribution to a variable having a distribution within one of the above mentioned families. Each of the above distributions (Types **I**, **II**, and **III**) are known as extreme value distributions.

[13] This theorem is also known as the Fisher-Tippett-Gnedenko theorem.

[14] A degenerate distribution is: $\Pr(X=k) = 1$ when $k = x$ and 0 when $k \neq x$. In physics this is a translated Dirac delta function.

Type **I** is known as the Gumbel distribution, type **II** is known as the Frechet distribution, and type **III** is known as the Weibull distribution. *A significant observation is that the three types of extreme value distributions are the only possible limits for the distributions of the maximums regardless of the parent distribution **F***.

These three types of distributions can be combined into a single distribution known as the *Generalized Extreme Value* distribution:

$$G(z) = \exp\{ -[1 + \xi (z - \mu)/\sigma)]^{-1/\xi}\}$$

where,

$$\{z : 1 + \frac{\xi(z-\mu)}{\sigma} > 0\}, -\infty < \mu < \infty \text{ and } -\infty < \xi < \infty$$

The scale parameter σ, shape parameter ξ, and location parameter μ are estimated from the data using maximum likelihood methods. *G(z)* represents a distribution of the maxima (or minima, depending on the problem). It is now possible to ask about the probability of observing a value greater than some extreme value over the next 100 blocks. In this case, the 100 blocks is known as the *return period*.

There are two popular approaches for implementing EVT. The first method is called the peak over threshold (POT) model. The second method is sometimes called the block maxima (or minima) method[15].

[15] The block maxima approach is a special case of the more general approach of modeling the n^{th} largest order statistics in a block.

7.1.1 Peak over Threshold Models

Peak over threshold (POT) models use only values that exceed some predetermined threshold. Consider a model where the rainfall total for some geographic location has been recorded on a daily basis for a period of 20 years. Further assume the block size for the model is 365 and that a threshold has been set at 5 inches of rainfall in any given day. In the case where no daily rainfall exceeds 5 inches, no data for that year will be used in the model. This approach can lead to a waste of potentially useful data.

7.1.2 Block Maxima Models

Block maxima models address this problem by using the maximum occurring data value from *each* block without regard to threshold exceedance. With this approach and the example above, the model will use data from each of the 20 years.

7.2 Block Maxima Models - Example

The following is an example of EVT block maxima predictions for the probability of observing various first difference maxima for normalized Stokes parameter S_1. The magnitude of the first differences for the first 300 measurements of the Stokes parameter S_1 is divided into 10 blocks of 30 measurements per block. Measurements were taken every 15 seconds.

The leftmost graph shows the magnitude of the first differences for the first 300 measurements of S_1 (figure 39). The rightmost graph shows the block maximum for each of the 10 blocks. EVT maximum predictions for various return periods are shown

in figure 40. For example, the probability of observing an event with magnitude difference > 4.0 x 10^{-4} in the next block is 0.01. In this case, the *return period* is 1/0.01 or 100 blocks.

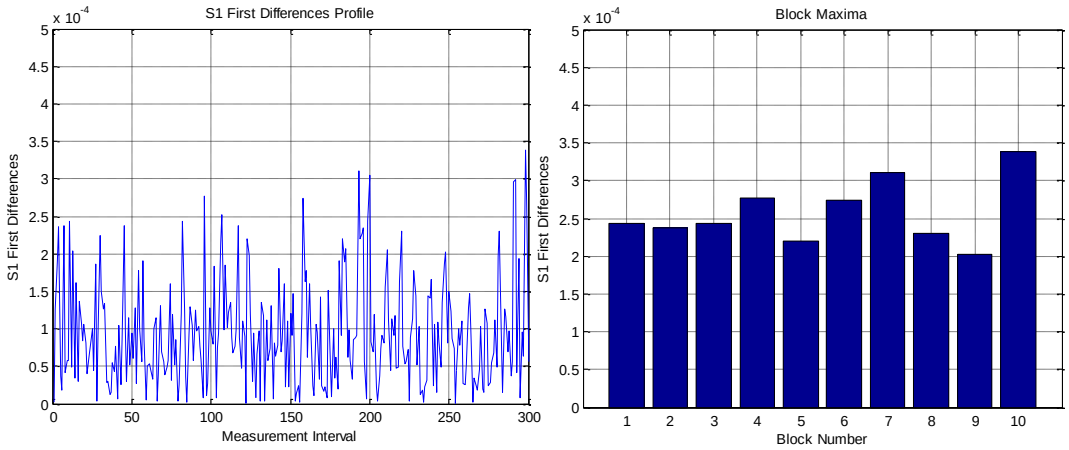


Figure 39 S1 first differences and block maxima.

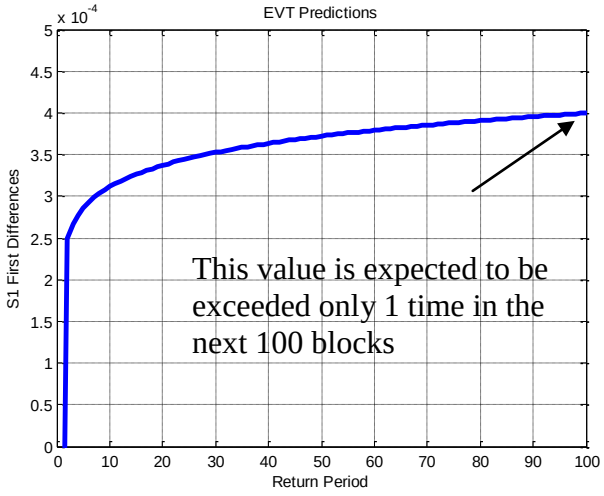


Figure 40 EVT return predictions.

The value of 0.0004 is expected to be exceeded only once in the next 100 blocks, or with probability 0.01 in the next block. Since each block consists of 30 intervals, and each interval represents 15 seconds, the value of 0.0004 is expected to be exceeded only once in the next 45,000 seconds (30 x 15 x 100 = 12.5 hours).

EVT will be used to characterize the "natural" behavior of the fiber (no disturbance). Once EVT characterizes this behavior, it will serve as the domain expert and use EVT block maxima predictions based on a return period of 1000 blocks to label observations as normal or anomalous. Assuming measurements are taken at 15 second intervals, and a block size of 50 is used, 1000 blocks represents approximately 8.68 days. As mentioned earlier, the rule for labeling an observation as anomalous will require the observation to exceed the return value (for a return period of 1000) for all three Stokes parameters.

Chapter 8: Experimental Results

EVT will monitor the natural state of the fiber for the purpose of characterizing the *usual* magnitudes of the first differences of all three Stokes parameters. The monitoring period will consist of 30 blocks, with each block containing 50 observations (no overlapping blocks)[16]. After the characterizing phase is completed, thresholds for *unusual* behavior will be derived and will be based on a return period of 1000 blocks. From this point on, EVT will take on the role as the domain expert, labeling each incoming observation as normal or anomalous. Each incoming observation displaces the oldest observation in the window. All of the features in chapter 5 will be calculated for each observation. Relevant features will be selected by the ReliefF feature selection algorithm. The selected features along with the label assigned by EVT will be used to search for a suitable algorithm for detecting anomaly polarization measurements.

[16] Extreme Value Theory distinguishes between independent and dependent sequences. The non-overlapping block approach more closely reflects a series of independent observations. No block contains information from any other block.

8.1 Example EVT Fiber Characterizations

Measurements from two of the more lengthy data acquisition periods are examined in this section. Both datasets consist of data collected every 15 seconds for over 3 days (over 16,000 observations). For dataset #1, the 1000 block return period thresholds are: S_1 = 0.00093, S_2 = 0.00025 and S_3 = 0.00033. The distribution of observed values for the entire 3 day period is shown in figure 41. The thresholds are illustrated by the red vertical lines for all but S_1, as the threshold for S_1 is beyond the right hand limit of the horizontal scale. The return value for S_1 was not exceeded during the duration of the collection. The return value for S_2 was exceeded during intervals 1572, 15519, and 15525. The return value for S_3 was exceeded during intervals 4327, 4507, and 5625. No exceedance intervals were in common. Therefore, EVT did not flag an anomalous event during the 3 day period.

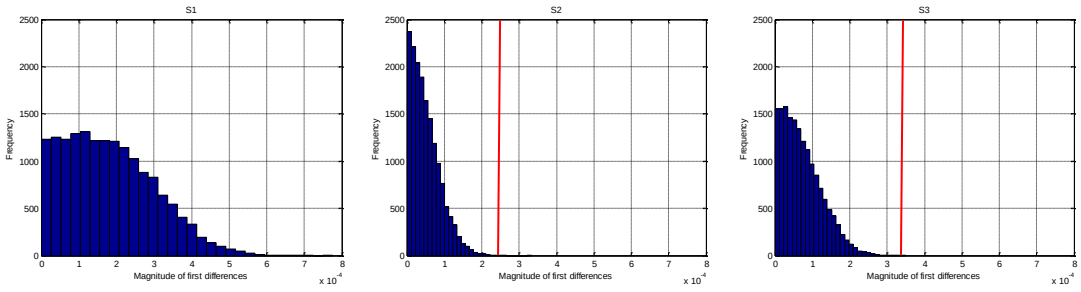


Figure 41 Distribution of first differences for dataset #1.

For dataset #2 the 1000 block return period threshold for S_1 = 0.00052, S_2 = 0.00031, and S_3 = 0.00056. The distribution of observed values for the entire 3 day period is shown in figure 42, along with their respective thresholds.

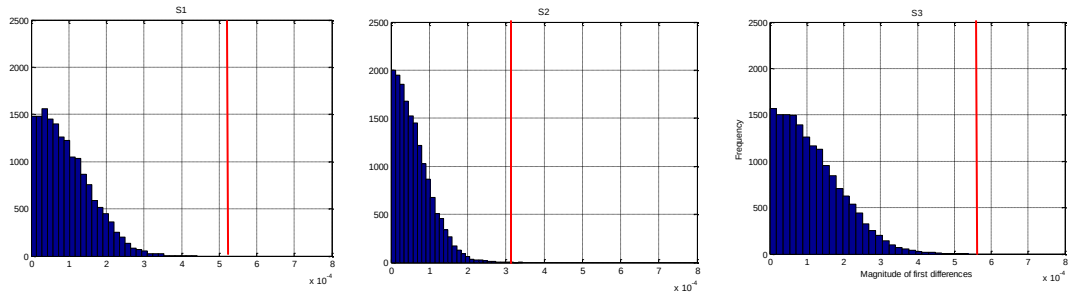


Figure 42 Distribution of first differences for dataset #2.

The return value was not exceeded for any of the three Stokes parameters. Therefore, EVT did not flag any anomalous events during the 3 day period for dataset #2. In fact, observing anomalous events for the undisturbed fiber did not occur. This presents a problem for the ReliefF algorithm as it must have examples from both classes in order to perform feature selection. Either the decision rule on what constitutes anomalous behavior must be changed, or synthetic (or physical) disturbances of various magnitudes must be created. The choice was to introduce synthetic disturbances of various magnitudes. Even extremely small disturbances were effective in producing anomalous cases (according to the EVT anomaly rule), as explained in the next example.

An anomalous period was introduced in dataset #2 beginning at interval 8335 and ending at interval 8612. The Stokes parameters for the original dataset and the

dataset containing the synthetic anomalies are shown in figure 43. The synthetic anomalies are of such small magnitude they cannot be seen.

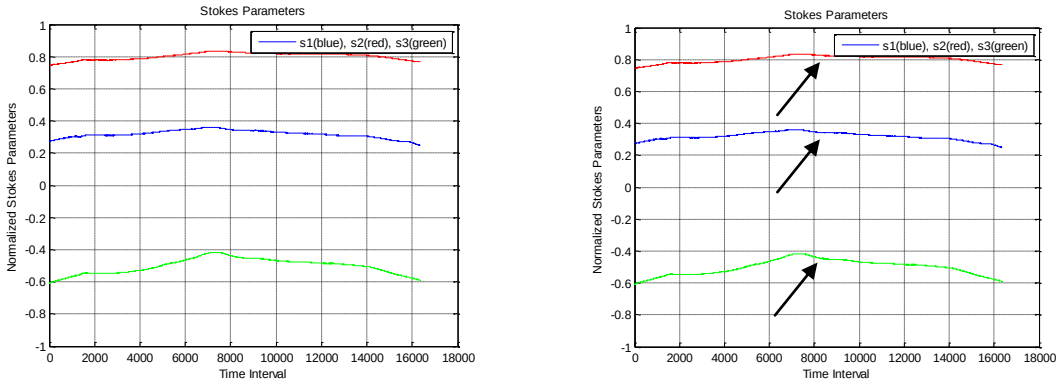


Figure 43 Small magnitude synthetic anomalies for dataset#2.

Original dataset #2 (L), Anomalous dataset #2 (R)
(Fiber perturbations are from interval 8335 to 8612)

The 1000 block return period threshold remains the same as above since the anomalies do not occur until after block 1500 (S_1 = 0.0052, S_2 = 0.0031, and S_3 = 0.0056). The return period threshold was exceeded for all three Stokes parameters for most intervals from 8336 to 8612.

8.2 Feature Selection using ReliefF

Once the characterization phase is complete, EVT begins to classify each observation as normal or anomalous according to the previously discussed rule requiring threshold exceedance for all three Stokes parameters. All of the features in chapter 5 are calculated for each observation. Any given observation will necessarily influence feature values for all windows in which the observation is present.

The relative feature weights for the synthetic anomalies discussed in the previous section are shown in figure 44 (using ReliefF with $\boldsymbol{k}$ = 15). Features 1-3 are the first differences for the three Stokes parameters, features 4-6 are the skewness values, features 7-9 are the kurtosis values, features 10-12 are the L-function values, and features 13-15 are the Hopkins values. Features 1-9 are 1-dimensional features and features 10-15 are 2-dimensional features. Note the most relevant features appear to be the first differences (features 1-3)! ReliefF has simply discovered the EVT parameters used to classify the observations. Ignoring these for the moment, it would appear the most relevant features in descending order are:

14 13 15 12 7 4 11 10 5 6 8 9

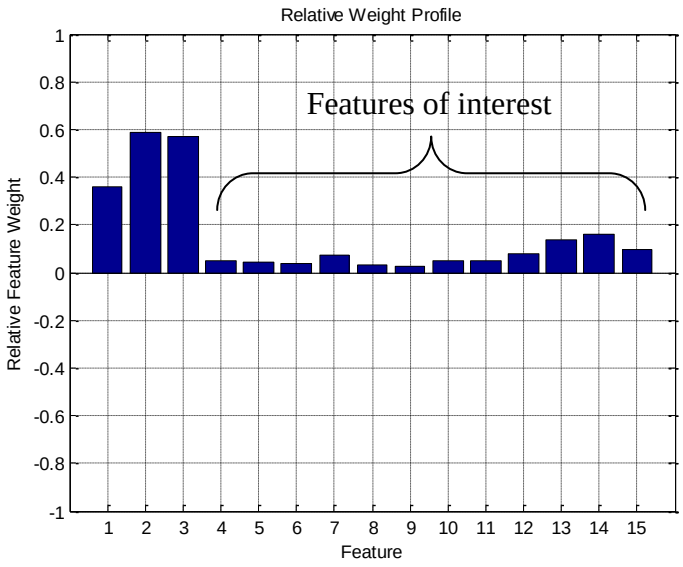


Figure 44 Relative feature weights for small magnitude synthetic anomalies.

Running this analysis again with larger anomalies yields the relative feature weights graph shown in figure 46.

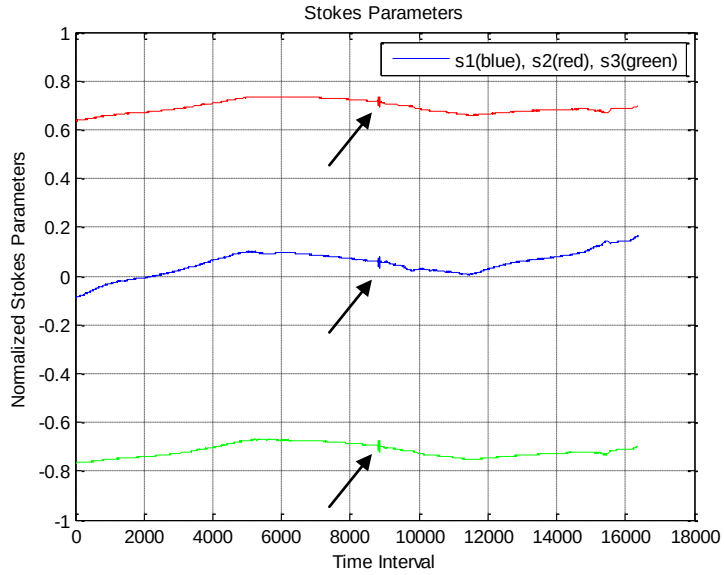


Figure 45 Moderate magnitude synthetic anomalies for dataset #2.

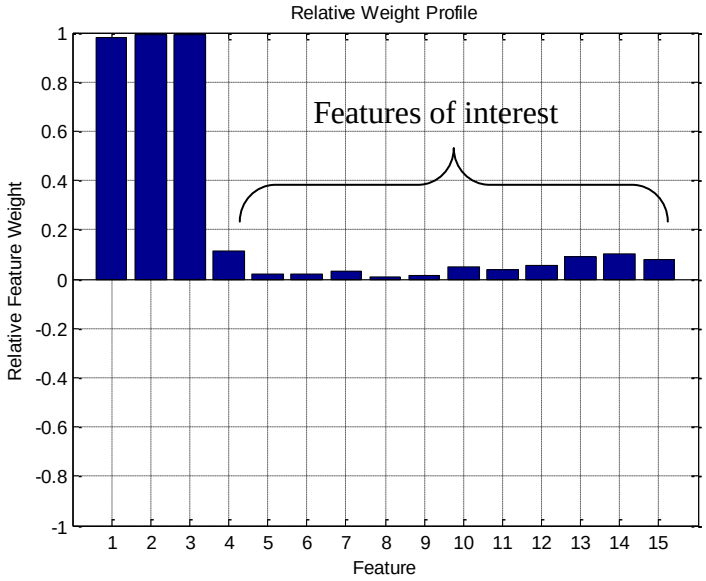


Figure 46 Relative feature weights for moderate synthetic anomalies.

It would appear the most relevant features in descending order are:

4 14 13 15 12 10 11 7 5 6 9 8

Running this analysis for a third time with still larger anomalies (figure 47) yields the relative feature weights graph shown in figure 48. It would appear the most relevant features in descending order are:

14 13 15 12 10 11 5 6 7 8 4 9

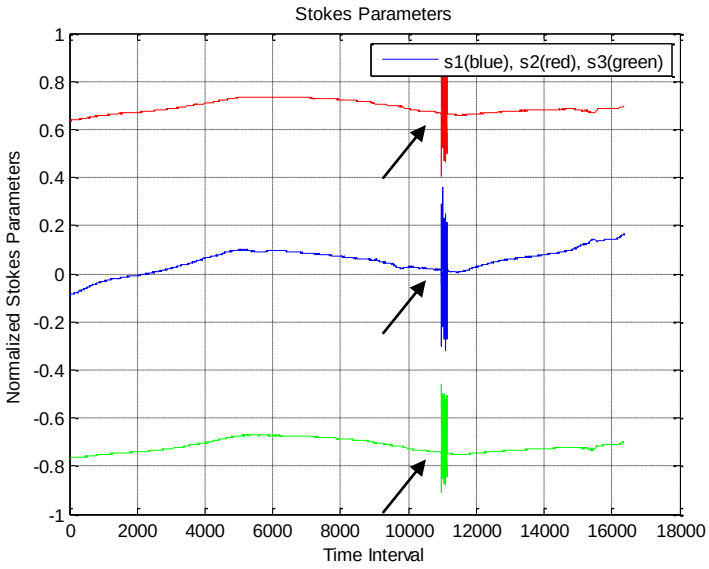


Figure 47 Large magnitude synthetic anomalies for dataset #2.

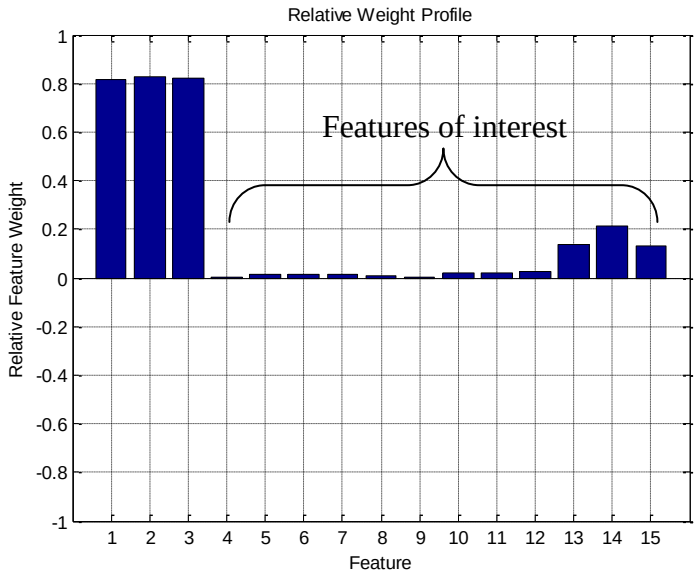


Figure 48 Relative feature weights large synthetic anomalies.

Generally speaking, the Hopkins statistic seems to be the feature with the highest relevance. However, considering relative feature weights range from -1 to 1, relative feature weights of 0.1 - 0.2 are not all that impressive. Based on feature response to anomalies given in section 5.3, higher relative feature weights are to be expected.

In the examples above, all of the simulated disturbances were created to last longer than the window size of 50 observations. The expected higher feature weights were not observed due to the fact most of the features tend to react to long-term disturbances as if they had become the "new normal." In other words, the ranges of the various statistics tend to revert back to the original ranges once a "new normal" is established. This behavior dilutes the amount of information associated with the feature value. A "new normal" is established when the duration of the disturbances is greater than the size of the observation window.

There is another troublesome issue introduced by the window approach. Consider the case where a single anomalous point enters the window, followed by 49 normal points. EVT will label the 49 points as normal but since the features apply to the window, feature values continue to react in a manner that signals the presence of anomalous data. This will continue as long as the one anomalous point remains in the window. This causes feature values for the window (which continue to signal anomaly) to be applied to the points labeled as normal. This behavior is a second reason for the dilution of information associated with feature values.

This suggests a more accurate model may be possible if the EVT rule for anomaly identification is modified such that the first anomaly encountered in the window triggers the anomaly label for this point and the next $n - 1$ points, where n is the window size. The influence of this point will persist thru the entire window. This approach can be justified based on the importance of detecting the very beginning of fiber-optic tapping activity. The resulting interpretation of an anomalous point is that it is present in a window in which anomalous behavior has been observed, but may not be anomalous itself.

Consider the following situation where synthetic anomalies lasted for only 4 observations (intervals 7171 through 1724). Again, the disturbances are relatively small (figure 49). From here on out features 1 – 3 will be skewness, 4 – 6 will be kurtosis, 7 – 9 will be the $\boldsymbol{L}$ function values, and 10 – 12 will be the Hopkins statistic.

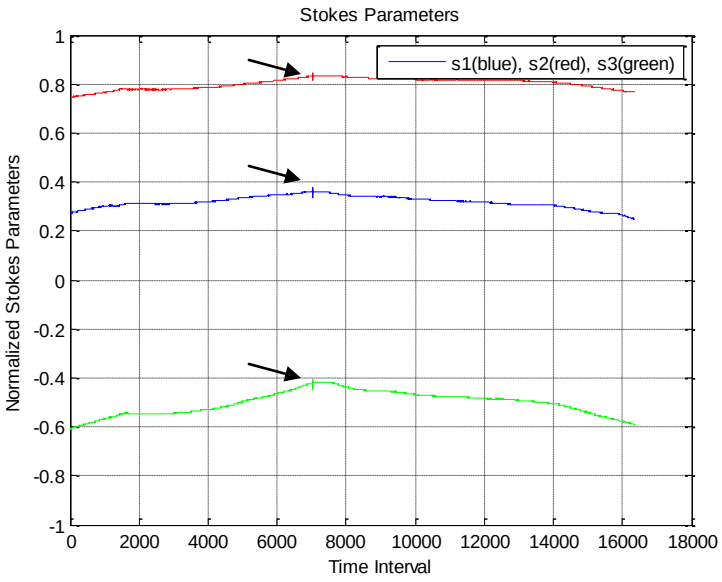


Figure 49 Small magnitude synthetic anomalies (short term).

ReliefF assessed the relative importance of the features as shown in figure 50.

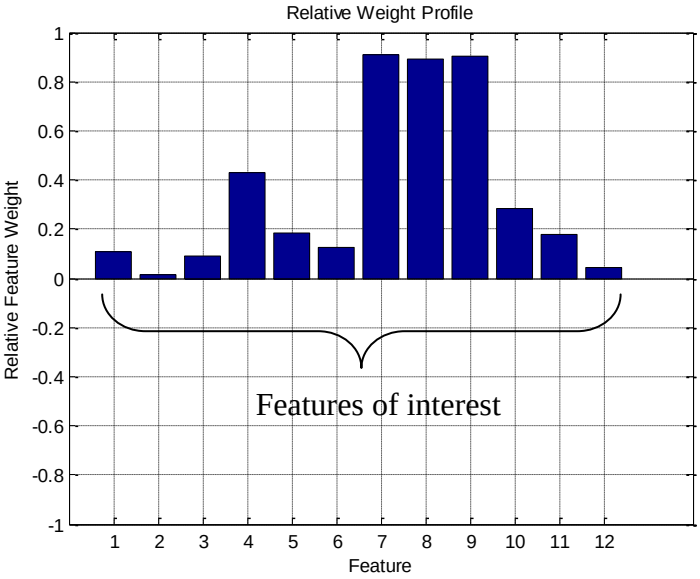


Figure 50 Relative feature weights for small synthetic anomalies.

Consider the moderate anomalies from interval 7037 to 7039 (figure 51)

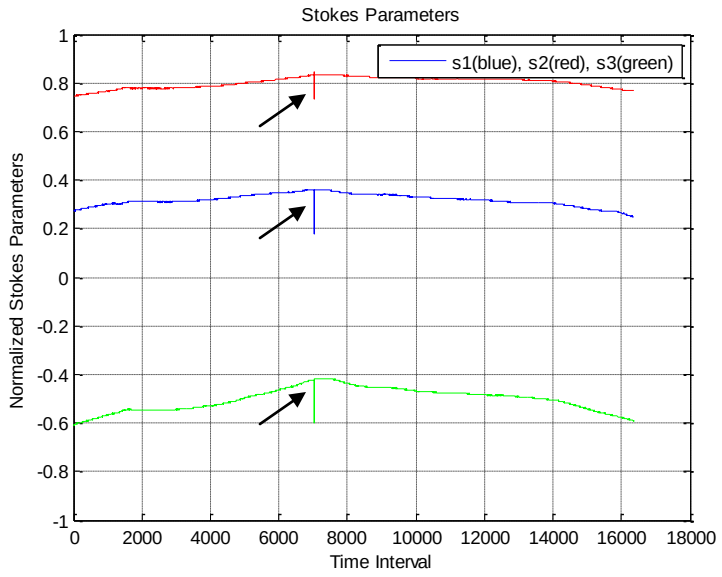


Figure 51 Relative feature weights for moderate synthetic anomalies.

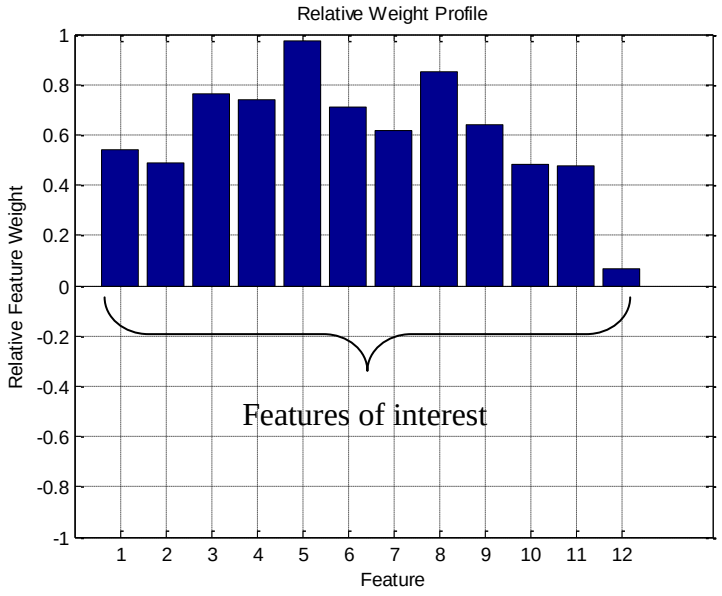


Figure 52 Relative feature weights for moderate synthetic anomalies.

As a final example, consider the following anomalies from interval 12162 to 12175 (figure 53).

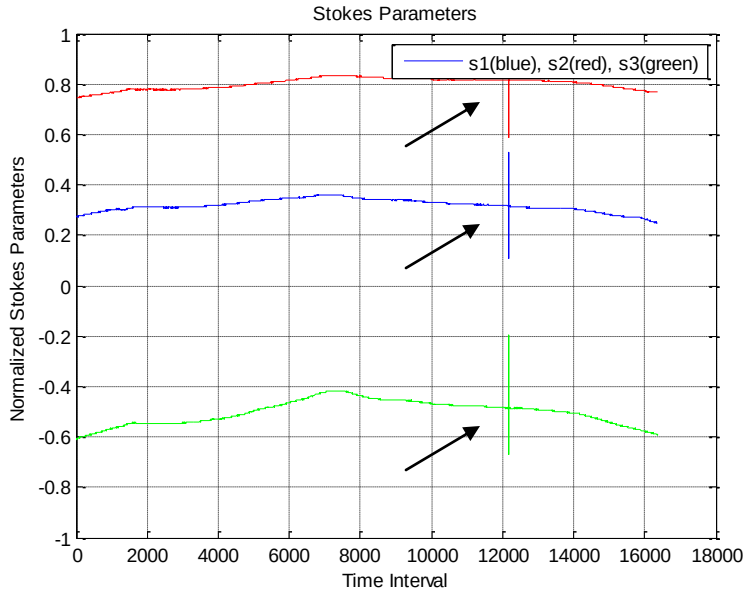


Figure 53 Relative feature weights for large synthetic anomalies.

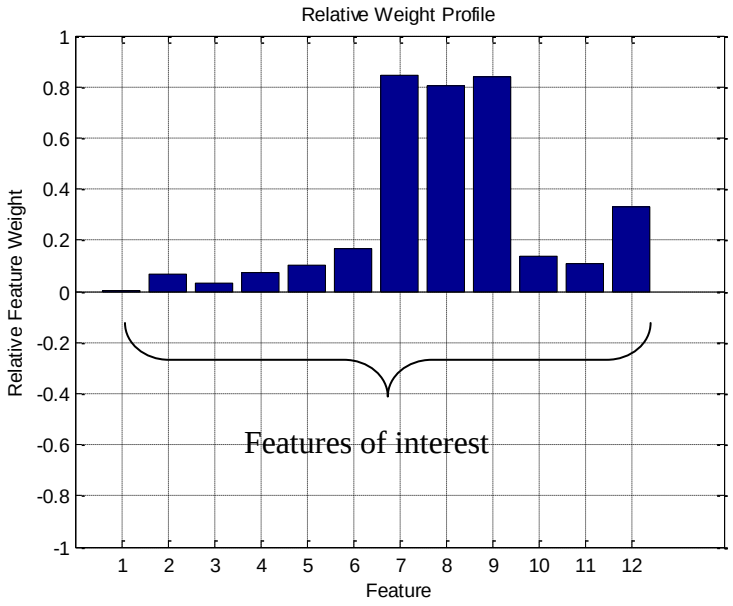


Figure 54 Relative feature weights for large synthetic anomalies.

It is clear from the previous examples that feature weight profiles are not all alike. Most features appear to have some relevancy from time to time but obvious consistency is lacking. If any general observation can be made, it would seem that the ***L*** function feature appears to be a leading candidate.

8.3 The Learning Algorithm

Before continuing on, it would be good to make the following observation. From the extended data acquisitions discussed earlier, and the studies characterizing polarization fluctuations in buried fiber as being slow, the expectation is that an anomaly in the absence of a physical disturbance is extremely rare. This means a simple model that arbitrarily predicts all observations as being normal, is highly accurate. Consider the case where only 1 anomaly occurs in 1000 blocks. With 50 observations per block this is 1 anomaly in 50,000 observations. A model that arbitrarily predicts 50,000 normal events and 0 abnormal events is correct 99.998% of the time. This sounds impressive until it is realized the model missed the only important event in the collection. The accuracy of a model for this application domain should be assessed based on the resulting confusion matrix (true positives, false positives, true negative, false negatives).

The variation in feature weight profiles shown in section 8.2 suggest that most all of the features have something to say from time to time. The relative importance of a feature tends to manifest itself differently depending on characteristics of the data collection. The total number of features is 12. This is not considered to be high dimensional and so using all of the features should not be problematic.

The approach will be as follows: A set of 5 decision trees will be constructed, with each tree using a random subset of 3 features. Each tree will provide a classification for each observation. The final classification will be based on majority vote[17]. It is well known that a majority decision by committee of weak learners[18] tends to improve the overall accuracy of the learning algorithm, provided there is independency amongst the committee members. The majority vote "correctness" profile for a collection of weak learners is shown in figure 55. The blue line is the "correctness" profile for weak learners having a 51% success rate; the red line is the "correctness" profile for weak learners having a 56% success rate. The use of random features for each tree tends to promote independence between the trees.

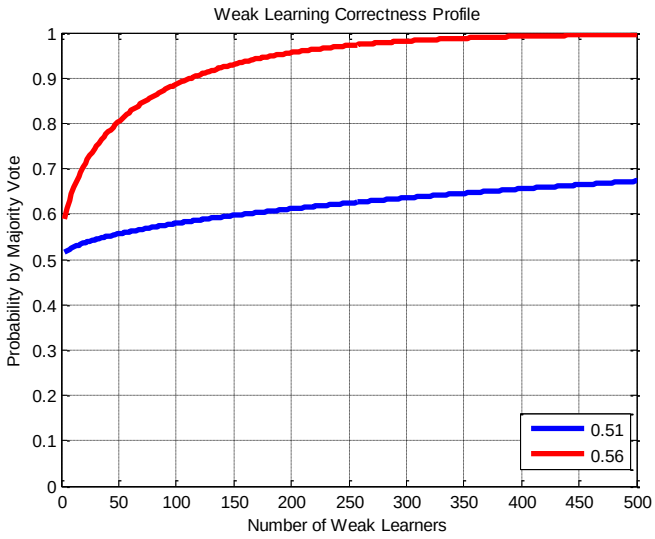
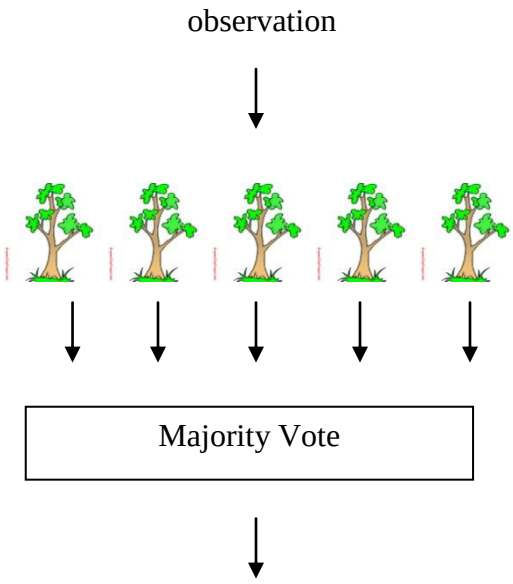


Figure 55 Majority vote "correctness" profile for *p*=0.51 and *p*=0.56.

This figure shows the probability of a correct decision by majority vote of a collection of weak independent learners. Each learner is an independent trial to the binomial distribution.

[17] Majority vote is one form of "ensemble learning" or learning by a "committee of experts."
[18] A weak learner is one having accuracy > 0.50.

The use of a random subset of features for trees in a forest is known as Random Forests™ [76]. The approach used here is a simpler version of Random Forests™ in that the selection of random features is set for each tree, and not randomized within the tree.



Each tree is aware of a random subset of 3 features

The dataset used to develop the forest of trees consists of 32,761 observations. This is actually two extended data collections merged together with injection of synthetic anomalies of various magnitudes. The merged dataset is shown in figure 56.

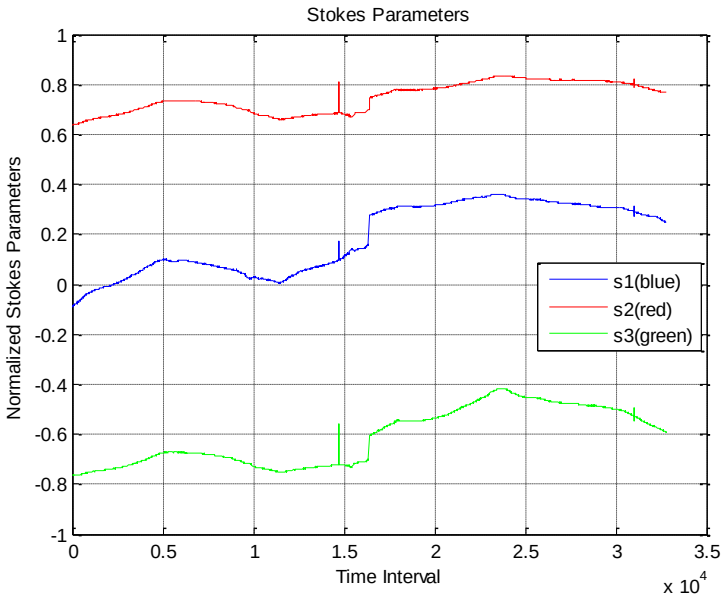


Figure 56 Training dataset for forest development.

Tree 1 is aware of features 6, 15, and 12 but uses only feature 12. Tree 2 is aware of features 9, 3, and 15 but uses only 9 and 13. Tree 3 is aware of features 6, 10, and 14 but uses only 10. Tree 4 is aware of features 5, 6, and 8 and uses all of them. Tree 5 is aware of features 9, 11, and 15 but uses only 11. Tree 1 has only 1 level, tree 2 has 2 levels, tree 3 has only 1 level, tree 4 has 2 levels and tree 5 has 1 level. The results for each tree are given in table 1. A majority vote by the forest corrects one 1 false positive in tree 1.

Table 1 Tree classification results (training dataset).

Tree 1	Levels - 1	Predicted Class	Predicted Class
Features	6,**12**,15	Anomaly	No Anomaly
Actual Class	Anomaly	160	0
Actual Class	No anomaly	1	31100

Tree 2	Levels - 2	Predicted Class	Predicted Class
Features	**9**, **13**,15	Anomaly	No Anomaly
Actual Class	Anomaly	160	0
Actual Class	No anomaly	0	31101

Tree 3	Levels - 1	Predicted Class	Predicted Class
Features	6,**10**,14	Anomaly	No Anomaly
Actual Class	Anomaly	160	0
Actual Class	No anomaly	0	31101

Tree 4	Levels - 2	Predicted Class	Predicted Class
Features	**5**,**6**,**8**	Anomaly	No Anomaly
Actual Class	Anomaly	160	0
Actual Class	No anomaly	0	31101

Tree 5	Levels - 1	Predicted Class	Predicted Class
Features	9,**11**,15	Anomaly	No Anomaly
Actual Class	Anomaly	160	0
Actual Class	No anomaly	0	31101

As expected, this works well on the training dataset. The results of the 5 trees and the resulting forest will now be assessed using data not seen by the forest during training.

The dataset used below have been held out from all previous work. Dataset #1 contains of 7200 observations. It is known to contain *no* anomalies. The second dataset is dataset #1 with some small magnitude anomalies.

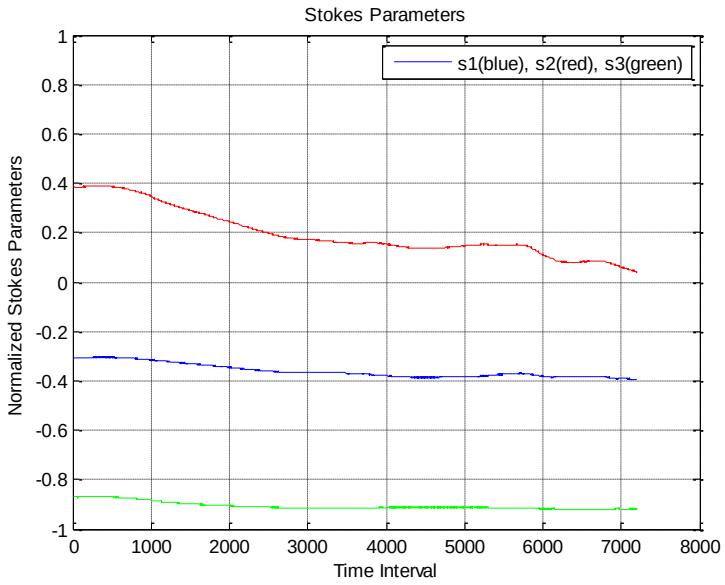


Figure 57 Example test dataset #1 (no anomalies).

Classification results for the 5 trees are shown in table 2. The forest labeled all cases correctly (table 3).

Table 2 Tree classification results (dataset #1 - no anomalies).

Tree 1		Predicted Class	Predicted Class
		Anomaly	No Anomaly
Actual Class	Anomaly	0	0
Actual Class	No anomaly	0	7199

Tree 2		Predicted Class	Predicted Class
		Anomaly	No Anomaly
Actual Class	Anomaly	0	0
Actual Class	No anomaly	0	7199

Tree 3		Predicted Class	Predicted Class
		Anomaly	No Anomaly
Actual Class	Anomaly	0	0
Actual Class	No anomaly	0	7199

Tree 4		Predicted Class	Predicted Class
		Anomaly	No Anomaly
Actual Class	Anomaly	0	0
Actual Class	No anomaly	17[19]	7182

Tree 5		Predicted Class	Predicted Class
		Anomaly	No Anomaly
Actual Class	Anomaly	0	0
Actual Class	No anomaly	0	7199

[19] The false positives were caused by an elevation in the kurtosis feature. The kurtosis values were slightly above the tree decision rule.

Table 3 Forest classification results (dataset #1 - no anomalies).

Forest	Majority Vote	Predicted Class	Predicted Class
		Anomaly	No Anomaly
Actual Class	Anomaly	0	0
Actual Class	No anomaly	0	7199

Synthetic anomalies of small magnitude were introduced to the dataset #1 beginning at interval 4651 and ending at interval 4568 (referred to as dataset #2).

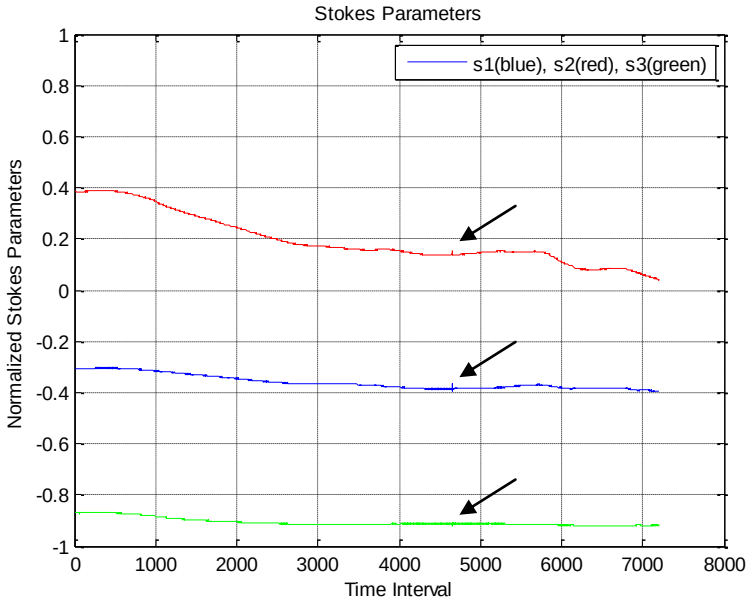


Figure 58 Example test dataset #2 (dataset #1 with small magnitude anomalies).

Classification results for the 5 trees are shown in table 4. Results for the forest are shown in table 5. The forest labeled all cases correctly.

Table 4 Tree classification results (dataset #2 with anomalies).

Tree 1		Predicted Class	Predicted Class
		Anomaly	No Anomaly
Actual Class	Anomaly	57	1
Actual Class	No anomaly	1	7140

Tree 2		Predicted Class	Predicted Class
		Anomaly	No Anomaly
Actual Class	Anomaly	12	1
Actual Class	No anomaly	46	7140

Tree 3		Predicted Class	Predicted Class
		Anomaly	No Anomaly
Actual Class	Anomaly	57	1
Actual Class	No anomaly	1	7140

Tree 4		Predicted Class	Predicted Class
		Anomaly	No Anomaly
Actual Class	Anomaly	57	18
Actual Class	No anomaly	1	7123

Tree 5		Predicted Class	Predicted Class
		Anomaly	No Anomaly
Actual Class	Anomaly	57	1
Actual Class	No anomaly	1	7140

Table 5 Forest classification results (dataset #2 with anomalies).

Forest		Predicted Class	Predicted Class
		Anomaly	No Anomaly
Actual Class	Anomaly	58	0
Actual Class	No anomaly	0	7141

Discovering a simple algorithm that makes no mistakes is unusual. But consider the following: no tree exceeded a depth of 2 and most trees used only 1 of the available features. These two observations suggest good discrimination by multiple features occurring at a high level in the tree. This is consistent with the theory behind a recently proposed unsupervised anomaly identification method known as *iforest* [52]. The isolation forest method employs random features for multiple trees and makes use of the fact anomalies are separated out at high levels in the tree. The *iforest* anomaly score is partly based on the tree depth at which separation occurs.

The sensitivity of the proposed features to anomalous polarization fluctuations is such that a highly accurate classifier is achieved using a simple forest of just a few trees.

Chapter 9: Future Work and Summary

Summary

This dissertation has investigated different methods for detecting fiber optic tapping activity. Empirical data suggest that polarization is especially sensitive to fiber geometry and therefore an effective method of monitoring fiber-optic links for tapping activity. Different representations of polarization suggested different methods of detection; these included measures from ordinary statistics (skewness and kurtosis), measures from point process statistics (***L*** function) and spatial statistics (Hopkins statistic ***h***). The use of the last two for anomaly detection appears to be new. Framed in a spatial context, the emergence of a cluster was construed to indicate the presence of a spatial anomaly, presumably caused by disturbances to fiber geometry. Both the ***L*** function and Hopkins statistic ***h*** have been shown to be capable of identifying spatial anomalies because of their cluster detection attributes. In the case of the Hopkins statistic, detection is made possible by redefining the use and interpretation of the metric. In the case of the ***L*** function, anomalies produce different function profiles which are indicative of changes in inter-point distances.

Most features demonstrated a high ability to distinguish between normal and abnormal behavior. This ability to distinguish is further enhanced by a small set of

independent decision trees using a majority voting scheme. No observations were misclassified.

The contributions of this dissertation are the following:

1) The use of Extreme Value Theory to characterize the naturally occurring polarization fluctuations in unperturbed fiber.
2) The use of Extreme Value Theory as an automated domain expert for supervised learning problems.
3) The use of the ***L*** function and Hopkins statistic as spatial anomaly identifiers.
4) The use of polarization as an ultra-sensitive fiber-optic tapping activity sensor.

This dissertation has raised a few questions which may be worthy further study.

Is EVT an anomaly detector?

While all features demonstrated some ability to identify spatial anomalies, the question must be asked: "Why not just use EVT as the anomaly detector"?

The series of 3-day long data acquisitions showed no anomalous periods (no false positives). This is encouraging because it suggests EVT performs well when it comes to characterizing the natural state of fiber, at least in this particular setting[20]. This setting is likely to mimic those where the cable is buried or otherwise installed in

[20] Continuous 68 hour surveillance was not performed. However, it was fairly certain that no fiber perturbations took place.

relatively stable environments. Any false alarms generated by EVT could probably be handled by adjusting the return period. Longer return periods are associated with higher thresholds. Higher thresholds are associated with less false positives but may produce more false negatives (missed anomalies).

Since EVT can be used to estimate probabilities of maxima *and* minima, it may be possible to periodically "re-characterize" the natural behavior of fiber by examining the minimum expected fluctuation. Falling below the minimum threshold may indicate a period of increased stability which would then trigger a new assessment and update of the maximum threshold. This approach would provide adaptation to changing conditions.

What about less stable environments?

Short of actually investigating aerial and undersea cables, it may be possible to simulate the behavior of polarization fluctuations in these types of environments simply by taking every n^{th} measurement from the current collections. Longer time intervals would likely result in larger fluctuations. It is not clear however, that this truly mimics a noisier setting. This may only represent stable environments with larger naturally occurring fluctuations. This is probably a worthwhile pursuit in any case. Consider the fact that naturally occurring fluctuations, as observed in this dissertation, were rather small. As such, they almost certainly included instrument error. Similar investigations with larger naturally occurring fluctuations would tend to mitigate instrument error.

How does the length of the fiber modify the method?

Many characteristics of fiber depend on the length of the fiber. Relationships between the Stokes parameters, the degree of polarization, and the power of the received signal in optical fiber of greater lengths may require different approaches. Greater lengths of fiber should produce some degradation in the degree of polarization of the received signal. Degradation in the degree of polarization may require some relaxation in the relationship between the Stokes parameters. This in turn, may require modifications to the EVT rules used for labeling anomalous cases.

What added information does the dual fiber approach yield?

The earlier mentioned dual-fiber study noted that polarization fluctuations are similar for fibers sharing a common path. This sort of configuration would certainly include two stands of fiber in the same breakout cable. Consider the process of placing an optical tap in this environment. The first step would be to remove a portion of the breakout cable in order to gain access to the two fibers. This step is likely to produce similar changes to the geometry of both fibers, and similar magnitude polarization fluctuations in both fibers. At this point the outer jacket must be removed from at least one of the two fibers. Removal of the outer jacket should produce different polarization signatures between the two fibers.

Similar magnitude changes without subsequent differing polarization signatures would not be as threatening as the one just described. The first scenario suggests the breakout cable is being physically disturbed but not necessarily compromised. This

may be reflective of innocent cable maintenance activity. The second suggests the breakout cable has been penetrated, increasing the probability of suspicious behavior.

References

1. Shannon C., "Communication Theory of Secrecy Systems." *Bell System Technical Journal*, Vol. 28, Issue 4, pp. 656-715, 1949. Available at: http://netlab.cs.ucla.edu/wiki/files/shannon1949.pdf
2. Shor, P., "Polynomial Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer." *SIAM J. Computing*, Vol. 26, No. 5, pp. 1484-1509, October, 1997.
3. http://www2.fz-juelich.de/jsc/qip/research/QC/idealQC (Julich Supercomputer Centre).
4. Prucnal, P. et al., "Physical Layer Security in Fiber-Optic Networks Using Optical Signal Processing." *Communications and Photonics Conference and Exhibition, 2009*, ACP 2009, Asia Vol. 2009-Supplement, pp. 1-10.
5. http://www.youtube.com/watch?v=2fP-j4XCuFs
6. http://www.youtube.com/watch?v=GSj-8UO3SDQ
7. http://www.youtube.com/watch?v=2DvaubDDbss
8. http://www.blackhat.com/presentations/bh-federal-03/bh-fed-03-gross-up.pdf
9. Fowles, G., *Introduction to Modern Optics.* Dover Publications, Inc., 2nd Ed., 1989.
10. Gerrard, A. and Burch, J. *Introduction to Matrix Methods in Optics.* Dover Publications, Inc., 1994.
11. Green, P. *Fiber Optic Networks.* Princeton Hall PTR, 1993.

12. Galarossa, A., Palmieri, L., Pizzinat, A., Schiano, M., and Tambosso, T., "Measurement of Local Beat Length and Differential Group Delay in Installed Single-Mode Fibers." *Journal of Lightwave Technology,* Vol. 18, No. 10, pp. 1389-1394, October 2000.
13. Kaminow, I. and Koch, T., (Eds). *Optical Fiber Telecommunications,* Vol. 3, Part 1, Chapter 6. Academic Press, 1997.
14. Kaminow, I. and Matsumoto, T., "Polarization in Optical Fibers." *Journal of Quantum Electronics,* Vol. 17, No. 1, January, 1981.
15. Cruz J., et al., "Faraday Effect in Standard Optical Fibers: Dispersion of the Effective Verdet Constant." *Applied Optics,* Vol. 35, No. 6, February 20, 1996.
16. Smith A., "Polarization and Magnetooptic Properties of Single-Mode Optical Fibers." *Journal of Lightwave Technology*, Vol. 17, No. 1, January 1, 1978.
17. Rashleigh, S., "Origins and Control of Polarization Effects in Single-Mode Fibers." *Journal of Lightwave Technology*, Vol. LT-1, No. 2, pp. 312-333, June 1983.
18. Sakai, J. and Kimura, T., "Polarization Behavior in Multiply Perturbed Single-Mode Fibers." *Journal of Quantum Electronics*, Vol. QE-18, No. 1, pp. 59-65, January 1982.
19. Tian, F., Wu. Y. and Ye, P., "Analysis of Polarization Fluctuation in Single-Mode Optical Fibers with Continuous Random Coupling." *Journal of Lightwave Technology,* Vol. LT-5, No. 9, pp. 1165-1168, September 1987.
20. Feng, T., "Random Coupling Theory of Single-Mode Fibers." *Journal of Lightwave Technology*, Vol. 8, No. 8, pp. 1235-1242, August, 1990.

21. Van Deventer, M., "Probability Density Functions of Optical Polarization States: Theory and Applications." *Journal of Lightwave Technology,* Vol. 12, No. 12, pp. 2147-2152, December, 1994.
22. Huang, W. and Yevick, D., "Improved Models of Optical Fiber Birefringence – Part 1." *Optical Quantum Electronics,* Vol. 39, pp. 91-103, 2007.
23. Huang W. and Yevick, D., "Improved Models of Optical Fiber Birefringence – Part 2." *Optical Quantum Electronics,* Vol. 39, pp. 105-117, 2007.
24. Galtarossa, A., "Statistical Characterization of Fiber Random Birefringence." *Optics Letters,* Vol. 25, No. 18, pp. 1322-1324, September 15, 2000.
25. Imai, T. and Matsumoto, T., "Polarization Fluctuations in a Single-Mode Optical Fiber." *Journal of Lightwave Technology,* Vol. 6, No. 9, pp. 1366-1375, September 1988.
26. Cameron J. et al., "Time Evolution of Polarization Mode Dispersion in Optical Fibers." *Photonics Technology Letters,* Vol. 39, No. 9, pp. 1265-1267, September, 1998.
27. De Angelis, C., et al., "Time Evolution of Polarization Mode Dispersion in Long Terrestrial Links." *Journal of Lightwave Technology,* Vol. 10, No. 5, pp. 552-555, May, 1992.
28. Nicholson, G. and Temple, D., "Polarization Fluctuation Measurements on Installed Single-Mode Optical Fiber Cables." *Journal of Lightwave Technology,* Vol. 7, No. 8, pp.1197-1200, August, 1989.
29. Harmon, R., "Polarization Stability in Long Lengths of Monomode Fiber." *Electronic Letters,* Vol. 18, No. 24, pp. 1058-1060, November 1982.

30. Poole, C., et al., "Polarization Dispersion and Principal States in a 147-km Undersea Lightwave Cable." *Journal of Lightwave Technology,* Vol. 6, No. 7, pp. 1185-1190, July 1988.
31. Nelson, L., et al., "Field Measurements of Polarization Transients on a Long-Haul Terrestrial Link." *Photonics Conference (PHO),* 2011, pp. 833-834.
32. Karlsson M., et al., "Long-Term Measurement of PMD and Polarization Drift in Installed Fibers." *Journal of Lightwave Technology,* Vol. 18, No. 7, pp. 941-951, July, 2000.
33. Shaneman, K. and Gray,S., "Optical Network Security: Technical Analysis of Fiber Tapping: Mechanisms and Methods for Detecting and Prevention." *2004 IEEE Military Communications Conference,* Vol. 2, pp. 711-716.
34. Box, G., Jenkins, G. and Reinsel, G. *Time Series Analysis: Forecasting and Control.* Prentice-Hall, Inc., 1994.
35. Heymann S. et al., "Outskewer: Using Skewness to Spot Outliers in Samples and Time Series.", *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONMAN 2012),* to appear.
36. Mumford, D. and Desolneux A., *Pattern Theory,* A. K. Peters, Ltd., 2010.
37. Babu, G. and Feigelson, E., "Spatial Point Processes in Astronomy." *Journal of Statistical Planning and Inference,* Vol. 50, Issue 3, pp. 331-326, March, 1996.
38. Law, R., et al., "Ecological Information from Spatial Patterns of Plants: Insights from Point Process Theory." *Journal of Ecology,* Vol. 97, pp. 616-628, 2009.

39. Gatrell, A., et al., "Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology." *Transactions of the Institute of British Geographers,* Vol. 21, No. 1, pp. 256-274, 1996.

40. Comas, C. and Mateu, J., "Modelling Forest Dynamics: A Perspective from Point Process Methods." *Biometric Journal,* Vol. 49, Issue 2, pp. 176-196, 2007.

41. Moran, P., "Notes on Continuous Stochastic Phenomena." *Biometrika,* Vol. 37, No. 1/2, pp. 17-23, June, 1950.

42. Geary R., "The Contiguity Ratio and Statistical Mapping." *The Incorporated Statistician,* Vol. 5, pp. 115-127 + 129-146, 1954.

43. Ripley, B., "The Second-Order Analysis of Stationary Point Processes." *Journal of Applied Probability,* Vol. 13, No. 2, pp. 255-266, June, 1976.

44. Illian, J., et al., *Statistical Analysis and Modelling of Spatial Point Patterns,* John Wiley & Sons, Ltd, 2008.

45. Pfeiffer, D., et al., *Spatial Analysis in Epidemiology,* Oxford University Press, 2008.

46. O'Brien D., et al., "Spatial and Temporal Comparison of Selected Cancers in Dogs and Humans, Michigan, USA, 1964-1994." *Preventive Veterinary Medicine,* Vol. 37, Issue 3, pp. 187-204, November, 2000.

47. Abernethy, D., et al., "Evaluating Airborne Spread in a Newcastle Disease Epidemic in Northern Ireland." *Proceedings of the 9th Symposium of the International Society for Veterinary Epidemiology and Economics*, 2000.

48. Hopkins, B. and Skellam, J., "A New Method for Determining the Type of Distribution of Plant Individuals." *Annals of Botany*, Vol. 18, pp. 213-227, April, 1954.
49. Banerjee, A. and Dave, R., "Validating Clusters using the Hopkins Statistic." *Proceedings of the IEEE International Conference on Fuzzy Systems*, pp. 149-153, 2004.
50. Hawkins, D., *Identification of Outliers.* Chapman and Hall, 1980.
51. Hawkins, S., et al., "Outlier Detection using Replicator Neural Networks." *Proceedings of the Fifth International Conference and Data Warehousing and Knowledge Discovery,* 2002.
52. Liu, F., et al., "Isolation-Based Anomaly Detection." *ACM Transactions on Knowledge Discovery from Data,* Vol. 6, No. 1, pp. 3:2-3:39, March, 2012.
53. Chandola, V. et al., "Anomaly Detection: A Survey." *ACM Computing Surveys*, Vol. 41, No. 3, Article 15, July, 2009.
54. Markou, M. and Singh, S., "Novelty Detection: A Review – Part 1: Statistical Approaches." *Signal Processing,* Vol. 83, Issue 12, pp. 2481-2497, 2003.
55. Markou M. and Singh, S., "Novelty Detection: A Review – Part 2: Neural Network Based Approaches." *Signal Processing,* Vol. 83, Issue 23, pp. 2499-2521, 2003.
56. Hodge, V. and Austin, J., "A Survey of Outlier Detection Methodologies." *Artificial Intelligence Review,* Vol. 22, No. 2, pp. 85-126, October, 2004.

57. Banerjee, A. et al., "A Support Vector Method for Anomaly Detection in Hyperspectral Imagery." *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 44, No. 8, pp. 2282-2291, August, 2006.

58. Ma, J., et al., "Anomalous Payload Detection System using Analysis of Frequent Sequential Pattern." *IEEE Fifth International Conference on Information Assurance and Security,* pp. 75-78, 2009.

59. Ma, J., and Xu, Z., "Network Anomaly Detection using Dissimilarity-Based One-Class SVM Classifier." *IEEE International Conference on Parallel Processing Workshops,* pp. 409-414, 2009.

60. Zhong, J., et al., "An Unsupervised Network Intrusion Detection Based on Anomaly Analysis." *IEEE Second International Conference on Intelligent Computation Technology and Automation,* pp. 367-370, 2009.

61. Mukkamala, S., et al., "Intrusion Detection using Neural Networks and Support Vector Machines." *Proceedings of the International Joint Conference on Neural Networks,* Vol. 2, pp. 1702-1707, 2002.

62. Piciarelli, C., et al., "Trajectory-Based Anomalous Event Detection." *IEEE Transactions on Circuits and Systems for Video Technology,* Vol. 18, No. 11, pp. 1544-1554, November, 2008.

63. Cochin, V., et al., "Anomaly Detection in VHF Radar Measurements." *Proceedings of the Geoscience and Remote Sensing Symposium,* Vol. 2, pp. 1171-1174, 2004.

64. Hanczr, B., et al., "Improving Classification of Micro Array Data Using Prototype Based Feature Selection." *ACMSIGKDD Explorations Newsletter,* Vol. 5, Issue 2, pp. 23-30, December, 2003.

65. Goh, L., et al., "A Novel Feature Selection Method to Improve Classification of Gene Expression Data." *APBC'04: Proceedings of the Second Conference on Asia Pacific Bioinformatics,* Vol. 29, January, 2004.

66. Cover, T., Thomas, J., *Elements of Information Theory,* Wiley, 2006.

67. Theodoridis, S. and Koutroumbas, K., *Pattern Recognition,* Elsevier, 2009.

68. Lee, W., et al., "Information-Theoretic Measures for Anomaly Detection." *Proceedings of the 2001 IEEE Symposium on Security and Privacy,* pp. 130-143, 2001.

69. Guyon, I., Elisseeff, A., "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research,* Vol. 3, pp. 1157-1182, March, 2003.

70. Kira, K. and Rendell, L., "A Practical Approach to Feature Selection." *Proceedings of the 9th International Workshop on Machine Learning,* pp. 249-256, 1992.

71. Kononenko, I. "Estimating Attributes: Analysis and extensions of relief." *Proceedings of the European Conference on Machine Learning, 199: ECML'94.* pp. 171-183, Springer Verlag.

72. Robnick-Sikonja, M. and Konenenko, I., "Theoretical and Empirical Analysis of ReliefF and RelieFF." *Machine Learning Journal,* Vol. 53, pp. 23-69, 2003.

73. Coles, S. *An Introduction to Statistical Modeling of Extreme Values.* Springer-Verlag, 2001.

74. Katz, R., et al., "Statistics of Extremes in Hydrology." *Advances in Water Resources*, Vol. 25, pp. 1287-1304, 2002.

75. AghaKouchak, A. and Nasrollahi N., "Semi-Parametric and Parametric Inference of Extreme Value Models for Rainfall Data." *Water Resources Management*, Vol. 24, Issue 6, pp. 1229-1249, April, 2010.

76. Brieman, L, "Random Forests." *Machine Learning*, Vol. 45, No. 1, pp. 5-32, October, 2001.