Classifying Formulations and Tracking Accelerated Aging of Crosslinked Polyethylene Pipes by Applying Machine Learning Concepts to Infrared Spectra

by

Melanie Hiles

A Thesis

presented to

The University of Guelph

In partial fulfilment of requirements
for the degree of

Master of Science

in

Physics

Guelph, Ontario, Canada

ABSTRACT

CLASSIFYING FORMULATIONS AND TRACKING ACCELERATED AGING OF CROSSLINKED POLYETHYLENE PIPES BY APPLYING MACHINE LEARNING CONCEPTS TO INFRARED SPECTRA

Melanie Hiles **Advisor:**
University of Guelph, 2020 **John Dutcher**

Crosslinked polyethylene (PEX-a) pipes are emerging as promising replacements for traditional metal or concrete pipes used for water, gas, and sewage transport. However, PEX-a is susceptible to oxidative degradation during its manufacturing and end-use applications that can ultimately result in premature pipe failure. Therefore, understanding the relationship between pipe formulation and performance is critical to their proper design and implementation. We have developed a methodology using the machine learning techniques of principal component analysis (PCA), k-means clustering and support vector machines (SVM) to compare and classify different PEX-a pipe formulations based on characteristic infrared (IR) spectroscopy absorbance peaks. PEX-a pipes of one such formulation was subjected to accelerated aging under different conditions to examine the effect of external stresses on changes to the chemical composition of the pipe samples. PCA, Random Forest (RF) and Decision Tree (DT) based techniques were used to identify characteristic IR signatures related to aging of PEX-a pipes.

ACKNOWLEDGEMENTS

I would like to extend my sincerest gratitude to my supervisor, Dr. John Dutcher, for the guidance he provided throughout my degree. It has been a privilege to have completed this work under the supervision of someone who is not only an exceptional scientist, but also cares deeply about the success of his students. I feel very lucky to have been a part of his group and I am grateful for the opportunities I was given during my time here.

Thank you also to Dr. Robert Wickham and Dr. Michael Massa for their valuable insights and advice during my committee meetings and for their patience with my last-minute change of thesis topic.

I would like to acknowledge all of my amazing lab mates, who never once hesitated to help me with my research. I am grateful for the many times they made me laugh at work, and for always encouraging me to come out to rock climbing, boardgame nights and softball. They made this a great experience and I will miss being greeted by their friendly faces every day. In particular, would like to acknowledge Joseph and Fatemeh for carrying out the accelerated aging experiments and a large portion of the data collection for this project. Without their hard work this thesis would not have been possible. I would also like to express my sincere gratitude to Mike, who spent countless hours teaching me IR and helping me deliberate over findings. I owe a lot of my success in this project to his mentorship.

I would like to thank my parents, Lisa and Mason, my siblings, Steven and Lauren, and my husband, Tom. I have great respect for all of the times they patiently listened to presentations and put genuine effort into trying to understand what I was talking about. I will always be grateful for their love and support.

Finally, appreciation is due to our industrial partners for providing us with an abundance of samples to spark our curiosity, and for many great meetings that inspired new avenues of research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

1.1 High Density Polyethylene (HDPE)

1.1.1 Structure and Physical Properties

Polyolefins are a class of polymers used to manufacture what are known as thermoplastics. They are the most widely used synthetic polymers worldwide, with applications in toys, packaging, household items and disposables [1]. Semi-crystalline polyethylene (PE) is the most commonly used thermoplastic, consisting of high-molecular-weight hydrocarbon polymer chains that arrange to produce a flexible material with low weight, high tensile strength, and long-term stability. When PE is manufactured under low pressure, a polymer with a small degree of branching is produced, known as high-density PE (HDPE). The linear polymer chains of HDPE can pack into highly crystalline regions (Figure 1.1a) that produce desirable physical properties such as increased tensile strength. Although the thermoplastic nature of PE allows it to be repeatedly reprocessed by cycling temperature, its applications are often limited by degradation of its physical properties at elevated temperatures [2].

1.1.2 Crosslinking

To improve the overall performance of HDPE and extend its applications, HDPE is often crosslinked. Crosslinked HDPE, known as PEX, has many interchain C-C linkages, increasing

the melting temperature of HDPE, while reducing crystallinity by restricting chain packing as shown in Figure 1.1. Crosslinking has several advantages: it does not reduce the tensile strength of HDPE and it increases its resistance to environmental weathering [2]. However, the polymer can no longer be reprocessed by cycling temperature.

PEX is often classified according to its method of crosslinking, as either PEX-a, PEX-b or PEX-c. PEX-a is manufactured using the Engel process, whereby a peroxide is added into the raw material to activate a crosslinking reaction within the polymer upon exposure to infrared (IR) light. PEX-b uses silane, or moisture crosslinking; this is a process that involves incorporating a silane catalyst onto the PE backbone, which reacts to form crosslinks in the presence of moisture. Finally, PEX-c is manufactured by subjecting the PE to an electron radiation beam that promotes crosslinking reactions. PEX-c is the least prevalent of the three methods, since the penetration depth of the radiation is often smaller than the desired thickness of the material, which prevents the final product from being uniformly crosslinked. Despite this, all three materials are highly durable and meet the same ASTM (American Society for Testing and Materials) minimum performance requirements. The present study will focus on PEX-a.

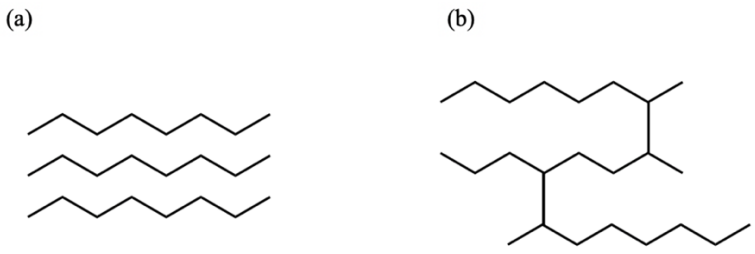

Figure 1.1: Schematic representation of the effect of crosslinking on HDPE. (a) Packing of uncrosslinked PE chains. (b) Network of PE chains after crosslinking illustrating restriction of chain packing.

1.1.3 Applications of Crosslinked HDPE – PEX-a Pipes

Peroxide crosslinked HDPE (PEX-a) is an attractive choice for plastic pipe manufacturers looking to replace traditional metal or concrete piping in applications such as water, gas, and sewage transport [3]. Its physical properties such as high melting temperature and tensile strength means PEX-a can be used in high-stress applications, while its resistance to environmental weathering means it can be reliably used in both indoor and outdoor applications for many years.

PEX-a pipes are manufactured by extrusion, a process by which the heated raw material is moulded to its desired shape. The pipe is then cooled and stretched to its desired size and thickness. This methodology must be carefully developed to prevent defects in the pipes that could lead to the formation of cracks. The pipes are also exposed to infrared (IR) light during or after the extrusion process in order to achieve crosslinking, which provides an added challenge to pipe manufacturers, as it can be difficult to achieve uniformly crosslinked pipes. To do so, the pipe is exposed to near infrared (NIR) light, avoiding the strong mid-infrared (mid-IR) absorption bands of polyethylene [1] while exploiting the weaker NIR absorption bands. This is a delicate balance between having enough absorption to activate the crosslinking reaction while having small enough absorption to allow penetration of the light to crosslink the entire wall thickness (several mm thick) of the pipe. The structure and stability of the PEX-a pipe is further complicated by the high degree of orientation introduced by the extrusion process, the semi-crystalline nature of PE, and the presence of other additives such as antioxidants and light stabilizers that improve stability of the pipe during manufacturing and in-service use. For these

reasons, it is important to carefully select manufacturing protocols that optimize the pipe's chemical and physical properties to achieve long-term stability, since oxidative degradation can result in chain scission that can ultimately result in premature pipe failure [4, 5].

1.2 Degradation Mechanisms of PEX

1.2.1 Overview

PEX-a pipe is susceptible to oxidative degradation during its manufacturing and end-use applications. During the manufacturing process, the PE is exposed to high temperature in the presence of oxygen, which can lead to thermooxidative degradation such as chain scission [6, 1, 7]. During installation and in service, PEX-a pipe can be exposed to factors such as oxygen, chlorine, elevated temperature, and UV light, which can induce thermooxidative and photooxidative degradation reactions [1, 4, 7, 8, 9]. These reactions proceed via established free-radical reaction pathways involving three steps: initiation, propagation and termination.

1.2.2 Initiation

In the initiation step, heat or light generates alkyl radicals via chain scission or hydrogen abstraction. Chain scission typically occurs during manufacturing, when PE is exposed to high temperatures and shear, such as during extrusion of PEX-a pipes [3, 7, 4]. These reactions, which occur along the PE backbone, result in the formation of terminal radicals such as the ones shown schematically in Figure 1.2a. Hydrogen abstraction, conversely, leaves the PE chain in tact and is

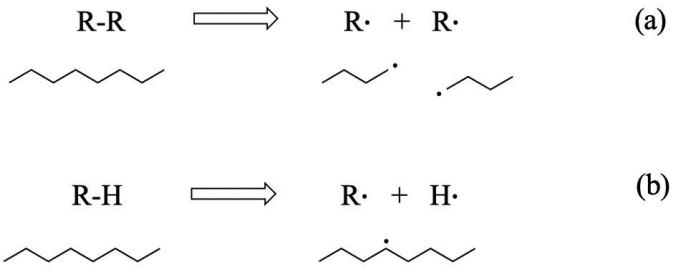

Figure 1.2: Schematic representation of possible initiation steps for degradation of PE modified from [7] including chain scission (a) and hydrogen abstraction (b).

therefore capable of producing mid-chain carbon radicals like those shown schematically in Figure 1.2b. This can occur as a result of chemical attack at any temperature but is expected to occur more readily at high temperatures [7, 8, 4]. The radicals produced by hydrogen abstraction are ideal candidates for subsequent crosslinking reactions, which help form the interconnected network of chains illustrated in Figure 1.1. Alternately, these radicals can continue in the degradation cycle by reacting with molecular oxygen in the propagation step [7, 1].

1.2.3 Propagation

Alkyl radicals, formed in the initiation step, can react with oxygen to form highly unstable peroxy radicals as shown schematically in Figure 1.3a. These peroxy radicals can subsequently remove hydrogen from the PE backbone, forming hydroperoxides, which are the key intermediate species in PEX-a oxidation (Figure 1.3b) [8, 7, 1, 4]. Hydroperoxides are unstable and can degrade to yield carbonyl compounds that can undergo further oxidative reactions (Figure 1.3c and Figure 1.3d) [8, 7, 1, 10, 4]. Additionally, vinyl unsaturations formed

during synthesis or produced during beta-scission of tertiary radicals can absorb UV light, leading to hydroperoxide formation [5]. Because oxygen cannot diffuse into crystalline regions, these degradation reactions are expected to occur in the amorphous regions of PEX-a or at crystal surfaces [8, 3].

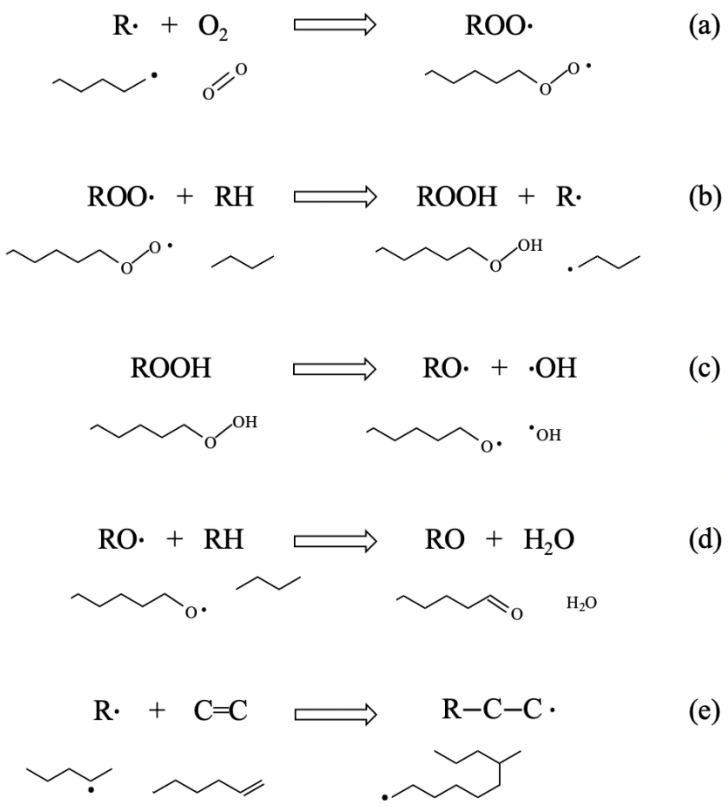

Figure 1.3: Schematic representation of some of the possible propagation steps for degradation of PE modified from [7]. Including the formation of (a) peroxy radicals, (b) hydroperoxides, (c) alkoxy- and hydroxyl radicals, (d) carbonyl compounds and (e) crosslinks via reactions with vinyl unsaturations.

1.2.4 Termination

Termination of the radical chain reactions may occur via a number of recombination and disproportionation reactions, a few of which are shown schematically in Figure 1.4 [7]. As mentioned in section 1.2.2, some of these may lead to the formation of crosslinks between PE chains [1, 4, 7]. Alternately, these reactions can result in the formation of carbonyl compounds such as ketones, aldehydes and carboxylic acids [1, 5, 4, 7].

The particular species formed in these steps are highly dependent on the external environment, which makes the study of degradation mechanisms extremely important in determining best-practices for manufacturing and in-service use of PEX-a pipes [8, 4]. For example, at high temperatures, such as those experienced during manufacturing, we expect a large number of free radicals to be present in the PE matrix [6]. In these cases, crosslinking reactions (Figure 1.3e) will dominate over hydroperoxide formation since diffusion of oxygen is limited [6]. Conversely, environments with high-oxygen and moderate-temperatures, such as those experienced during in-service use of the pipes, promote the formation of peroxyl radicals and hydroperoxides (Figure 1.3a and b) [6]. Since termination may occur directly after initiation or at any point during propagation, the distribution of radical species formed in these steps will determine how degradation of PEX-a is terminated and which final degradation products are present [8]. Furthermore, products formed via recombination of peroxy radicals (Figure 1.4d) are susceptible to further thermo-oxidative reactions after termination, which can expand the list of species present in the PE matrix [7].

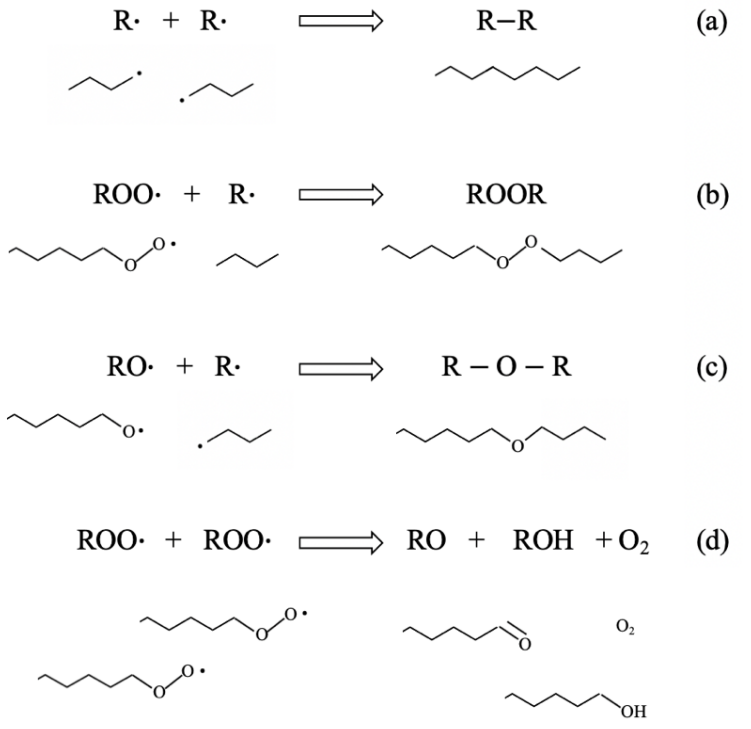

Figure 1.4: Schematic representation of some of the possible termination steps for degradation of PE modified from [7].

1.3 Stabilization of PEX-a Pipe

1.3.1 Overview

To achieve long-term stability, it is necessary to incorporate stabilizing agents into PEX-a pipes. These additives are classified according to their mechanism of action as primary antioxidants, secondary antioxidants, UV absorbers or hindered amine light stabilizers (HALS).

1.3.2 Antioxidants

Primary antioxidants are active at operating temperatures and are largely responsible for long-term pipe stability [1]. These stabilizing agents typically provide protection by acting either as H-donors that preferentially react with backbone peroxyl radicals, or as free-radical scavengers that form stable radical species, thereby terminating the degradation process. Irganox 1010 is an example of a commonly used, commercially available primary phenolic antioxidant. Its structure and possible mechanism of action is shown in Figure 1.5.

Secondary antioxidants are only active at high temperatures that occur during the manufacturing process. These species act in synergy with primary antioxidants by decomposing hydroperoxides that are produced either during the degradation process, or as a result of the action taken by primary antioxidants [7, 9, 1].

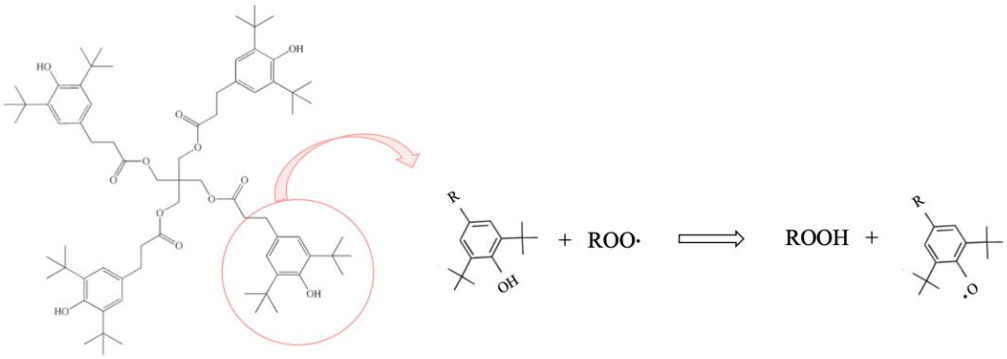

Figure 1.5: Schematic representation of Irganox 1010 and possible reaction mechanisms modified from [7].

1.3.3 Light stabilizers

Light stabilizers, such as HALS and UV absorbers, are also active at operating temperatures and are largely used to prevent degradation during installation or in outdoor applications. The reaction mechanisms of HALS are debated but their effectiveness in preventing both photo-oxidative and thermo-oxidative degradation has been attributed to their ability to react with hydroperoxides that, as discussed in section 1.2.3, may form as a result of the absorption of UV light by vinyl unsaturations [7, 11]. Chimassorb 944 is often selected for light stabilization due to its polymeric structure and high molecular weight, which prevents additive diffusion and volatilization from the pipe's surface [3, 7, 9, 12]. One possible reaction mechanism for Chimassorb 944 is shown in Figure 1.6 [7].

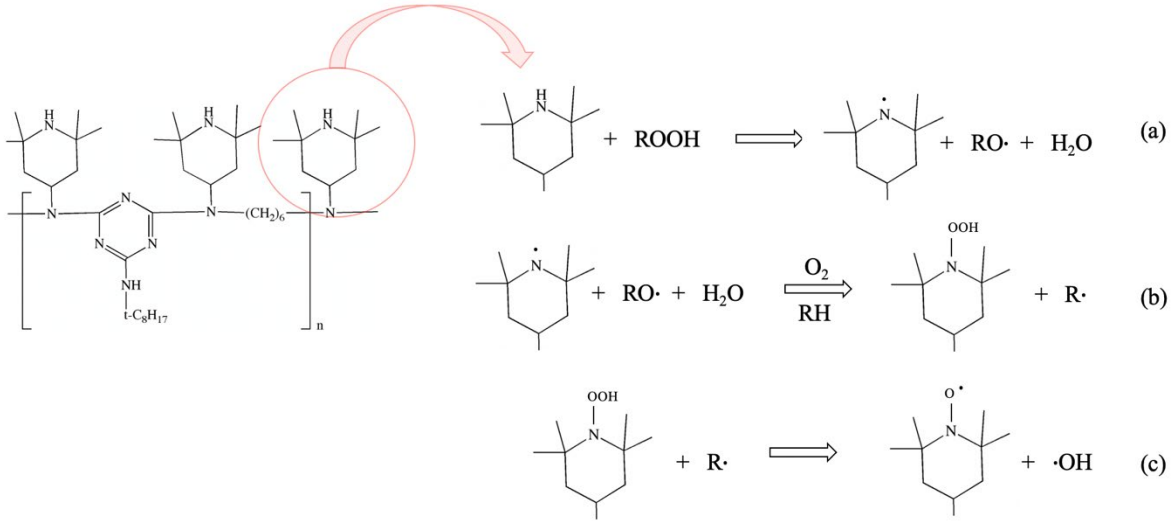

Figure 1.6: Schematic representation of Chimassorb 944 and possible reaction mechanisms with hydroperoxides modified from [7].

UV absorbers, on the other hand, act to protect the pipe from degradation by absorbing radiation before it has the chance to initiate the degradation. Often this reaction proceeds via absorption of UV light and molecular rearrangement of the additive molecule. One example of a commonly used, commercially available UV light stabilizer is Tinuvin 622, for which the chemical structure is shown schematically in Figure 1.7.

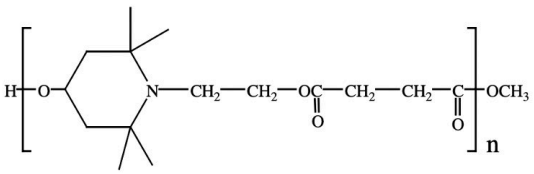

Figure 1.7: Schematic representation of Tinuvin 622 from [12].

1.3.4 Additive Hydrolysis and Leaching

Stabilizing agents are additives that can also undergo degradation reactions that reduce their efficacy. Many important additives have ester linkages (see, e.g., Figure 1.7) that are vulnerable to hydrolysis and photodegradation [7, 9, 12, 13, 14, 3]. These reactions reduce the volume pervaded by the additive molecule, which can facilitate additive loss via mechanisms such as leaching into the water flowing through the pipe [3, 15, 16]. This can also lead to concentration gradients across the radial profile of a pipe, hindering the performance of the additive. The diffusion of oxygen and stabilizing agents within the amorphous regions of the pipe play a fundamental role in these processes, which makes the degree of crystallinity an important factor in degradation kinetics [3, 9, 5].

1.4 Research Goals

IR absorption spectroscopy has been used extensively to characterize polymers [17]. In these experiments, the frequencies of characteristic molecular vibrations are measured as peaks in the absorption spectra. These frequencies are sensitive to the local environment of the molecules, and therefore can provide detailed information on the conformation, orientation, and phase state of the molecules within a sample [17]. Specifically, it is possible to track additive content by using IR absorptions from specific functional groups, such as phenols (Irganox), esters (Tinuvin), and triazine rings (Chimassorb), associated with additive species. Chemical species that play an important role in degradation, such carbonyls and vinyl unsaturations, as well as other properties such as the degree of crystallinity, can also be quantified using distinct IR bands.

In the present study, we use IR absorption spectroscopy to evaluate the chemical composition, degree of crystallinity and the presence of additives for three different formulations of PEX-a pipe. We use principal component analysis (PCA) and the machine learning techniques of k-means cluster analysis and support vector machines (SVM) to analyze and classify the data, with the goal of demonstrating that the IR absorption data can be used to unambiguously distinguish different pipe formulations. This methodology will aid development of a comprehensive database that can be used to optimize manufacturing protocols and carefully select processing conditions.

We also use IR microscopy, which combines the chemical specificity of IR spectroscopy with the high spatial resolution of light microscopy, to evaluate the uniformity and aging properties of PEX-a pipe. Specifically, we use this technique to characterize the spatial distribution of additives and defects within the pipes. These studies ultimately allow us to track the extent to which degradation is occurring that can be used to ensure the presence and proper function of additives throughout the wall thickness of the pipes. We use principal component analysis (PCA) and Decision Tree (DT) machine learning techniques to identify characteristic signatures of aging and to gain insight into the IR bands that are relevant to the degradation process.

2 Background

Parts of this chapter are based on work described in reference [23].

2.1 Fourier Transform Infrared (FTIR) Spectroscopy

2.1.1 Transmission FTIR

Infrared (IR) spectroscopy is an analytical technique that has been used extensively to characterize polymers [17]. In Fourier Transform IR (FTIR) spectroscopy, the frequencies of characteristic molecular vibrations are measured as peaks in the absorption spectrum, over a wavenumber range of 4000–600 cm^{-1}. These frequencies are sensitive to the local environment of the molecules, and therefore can provide detailed information on the conformation, orientation, and phase state of the molecules. It is therefore well suited to the characterization of PEX-a pipe properties and degradation, since both the physical and chemical properties of the pipe can be studied simultaneously from a single IR spectrum. For example, comparing IR absorptions that arise exclusively from the crystalline or amorphous regions allows the determination of the crystallinity of the sample, which is an important parameter in determining the mechanical strength of the pipe, the effectiveness of additives, and the long-term stability of the pipe. The amount of chemical species that play an important role in degradation, such as carbonyl and vinyl groups, can also be quantified. In many cases, it is possible to track additive content by using IR absorptions from specific functional groups, such as phenol groups, and triazine rings, that are associated with additive compounds in the pipes. In the present study,

transmission FTIR spectroscopy was used to evaluate the properties of bulk PEX-a pipe samples. A schematic of the FTIR experimental setup used in these experiments is shown in Figure 2.1.

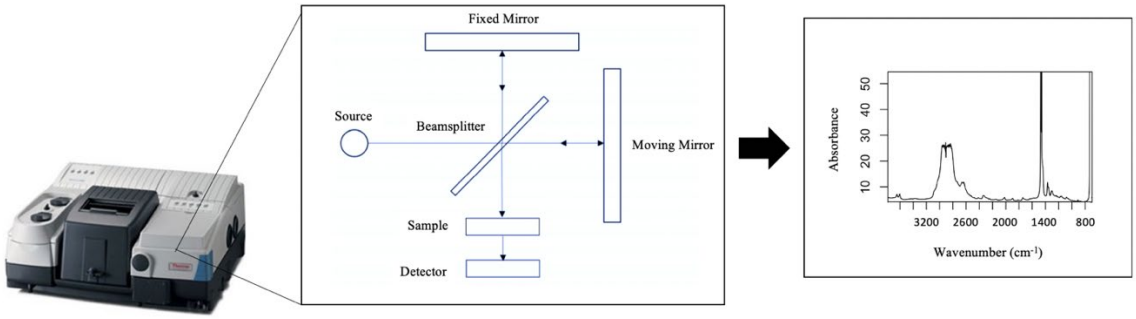

Figure 2.1: Photograph, schematic diagram and representative spectrum for the transmission FTIR spectroscopy experiment. Photograph adapted from [18]. Schematic adapted from [19].

2.1.2 FTIR Microscopy

FTIR microscopy combines the chemical specificity of FTIR spectroscopy with the high spatial resolution of light microscopy, allowing the mapping of the chemical composition of a sample with a spatial resolution of tens of micrometers. Typically, an FTIR microscope consists of an FTIR spectrometer, an optical microscope and a photoelectric IR detector, as shown in Figure 2.2. An aperture in the microscope can be tuned to adjust the sampling area of each spectrum. For studies of PEX-a pipe, FTIR microscopy provides a simple, powerful method of characterizing the spatial distribution of additive compounds and other chemical species within the pipes, and their correlation with the presence of microscopic defects on and within the pipe samples.

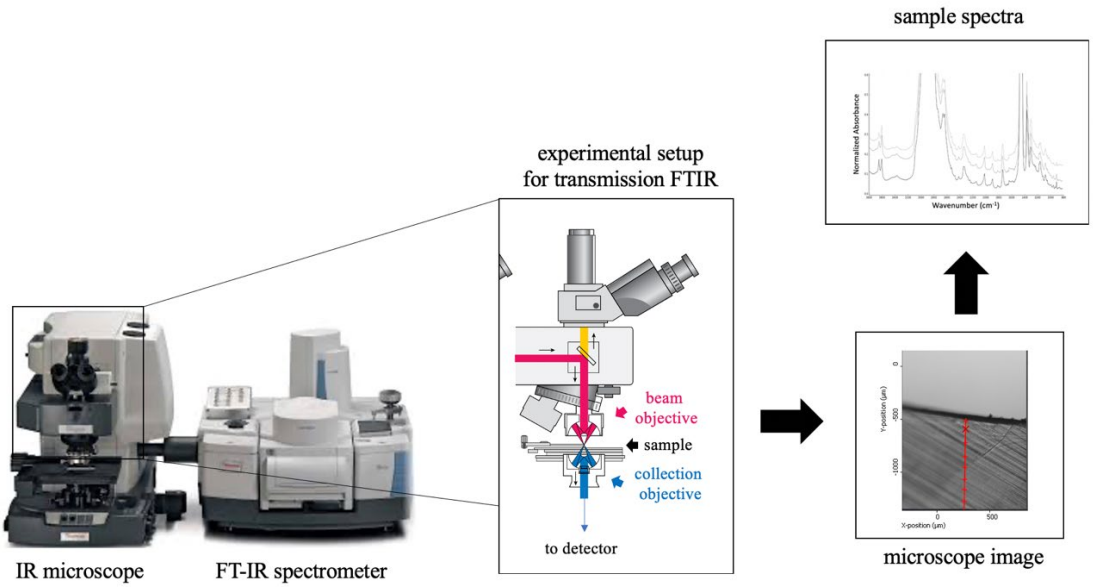

Figure 2.2: Photograph, schematic diagram, sample 1500×1500 μm microscope image and representative transmission FTIR spectrum collected through a 50×50 μm aperture for a Nicolet™ Continuμm™ Infrared Microscope. Photo and schematic adapted from [20].

2.2 Multivariate Machine Learning Techniques

2.2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an unsupervised, statistical multivariate data analysis technique often used for feature extraction and dimensionality reduction [21]. These are both processes by which complex, multivariate data sets are reduced from many correlated variables to fewer, uncorrelated variables, allowing trends in the data to be more easily identified [21]. PCA reduces the dimensionality of a data set by defining new variables that describe the

most important features of the data. The new variables, known as principal components (PCs), are defined orthogonally along directions of maximum variance in the original data and are linear combinations of the original variables. This can be seen visually in Figure 2.3, which shows a 3-dimensional data set being reduced to a 2-dimensional data set by defining the first two PCs along orthogonal directions of the greatest spread and second greatest spread in the data, respectively. In general, each subsequent PC is defined orthogonal to the previous PCs, along the direction of leftover variance, so to capture a unique and uncorrelated source of variance in the original data.

Mathematically, it is easiest to explain PCA in terms of a matrix representation of the data. For a given $n \times m$ data matrix X (with elements consisting of peak maximum absorbance values in the present study), composed of m variables (wavenumbers in the present study) and n observations (IR spectra in the present study), the covariance between each pair of variables x_i and x_j is represented by the corresponding element of the covariance matrix C, computed as

$$C_{ij} = \frac{\sum_{k=1}^{n}(x_{ik} - \bar{x_i})(x_{jk} - \bar{x_j})}{n-1} \qquad (2.1)$$

where x_{ik} is the i^{th} variable in the k^{th} observation, and $\bar{x_i}$ is the average of the variable x_i. PCA decomposes the covariance matrix into its eigenvectors and eigenvalues by singular value decomposition (SVD) or eigen-decomposition (EVD) according to $C = UVU$, where U is a $m \times m$ matrix in which the m rows represent the eigenvectors of the covariance matrix C, and

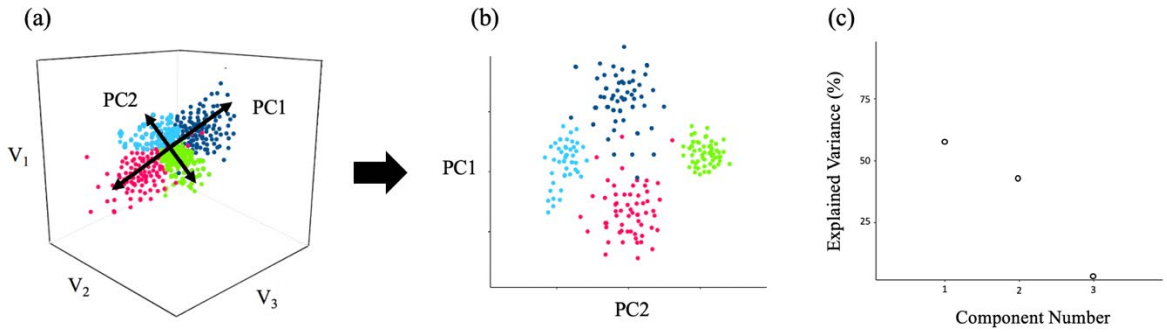

Figure 2.3: a) Illustration of the first two principal components (PCs) based on the spread of the data within a data set characterized by three variables V_1, V_2 and V_3. (b) Sample scatter plot of score values for PC1 and PC2. (c) Sample scree plot showing the percentage of explained variance for each PC.

V is a diagonal $m \times m$ matrix of the corresponding eigenvalues. The eigenvectors of the covariance matrix are the PCs that define the direction vectors of a new orthogonal m-dimensional space, pointing in the directions of maximum variance in the original data. The original data matrix X is then projected into PC-space by calculating $P = XU$, where each of the n rows in the PC-space data matrix P corresponds to a single observation in PC-space.

The eigenvalues of the covariance matrix represent the total amount of variance explained by each PC. The PC eigenvalues are often ordered in descending magnitude and visualized in what is known as a scree plot (an illustrative example of which can be seen in Figure 2.3c) [22]. This is done so that a small subset of $p \leq m$ components can be retained, in place of the m original variables for visualization or subsequent analysis without significant loss of information. A threshold PC value can be chosen such that a large amount of explained variance is described by a small number of PCs. Traditionally, the threshold is selected to retain

only components whose eigenvalues occur before an ‘elbow’ or change of slope in the scree plot [22]. However, given that this method has been criticized for its subjectivity, a variety of complementary techniques can and should be used to ensure minimal loss of information [22]. One such technique is known as a parallel analysis of the eigenvalues. The covariance matrix of an infinite random, uncorrelated data set is expected to have eigenvalues of 1 [22]. However, Horn pointed out that, for finite random data sets, the eigenvalues can exceed unity by an amount that depends on the size of the data set [22]. Therefore, Horn’s criteria for component retention involves dividing the observed eigenvalues by the corresponding parallel random eigenvalue. The adjusted eigenvalues are then retained according to Kaiser’s rule [22], which recommends retaining only components with an eigenvalue greater than 1. Using a combination of empirical and mathematically motivated methods offers the best approach for component selection and retention.

Furthermore, analysis of the linear combinations that define each PC can be used in exploratory analysis to accurately and concisely summarize relationships between the original variables [21]. Often, they can provide insight into the underlying structure of the data and, in some cases, can be used to assign physical significance to the components themselves [23].

PCA has proven to be an invaluable tool for many different types of problems and data structures, with applications ranging from exploratory analysis and outlier detection to image processing and compression [21, 24, 25, 13]. More recently it has found use in spectroscopic studies, where hundreds of variables corresponding to a variety of physical and chemical

properties can be measured with high resolution and are often correlated [14, 26, 27, 28, 29]. In these and many other cases, PCA offers a computationally inexpensive method for pattern recognition, noise reduction and feature extraction. For example, in the present study PCA has been used to compare the entire IR spectrum of many different samples at once and has ultimately allowed us to identify patterns and correlations in the data set that would be more challenging to describe using traditional methods. Furthermore, the mathematical simplicity of PCA and the ability to adapt PCA to many different derivative techniques for specialized use, makes it one of the most powerful data analysis tools available [13, 30, 31]. It can easily be used in conjunction with machine learning techniques to solve problems related to multicollinearity or overfitting and has actually been shown to produce predictive models that perform faster and with higher accuracy than those built using data in its original representation [32, 33, 34].

2.2.2 k-Means Clustering

Cluster analysis is an exploratory data analysis technique that is used to assess the similarity of items in a data set by grouping them in the absence of categorical information (i.e. it is a classification scheme) [35]. Often, clustering algorithms will use mathematical distance metrics to assess similarity between pairs of observations within an n-dimensional space. This method allows the algorithm to identify natural groupings within data sets by mathematically optimizing the chosen metric over all observations [35]. Its applications span three primary categories: data structure exploration, classification and data compression or reduction [35]. This has allowed researchers to use clustering to reach across many fields and provide a good

computational alternative to problems that have traditionally been dealt with by human intervention. However, like many other data mining tools, the broad range of uses and corresponding variety in data types has promoted many variations of clustering algorithms. This work will focus on one of the most popular clustering algorithms: *k*-means clustering [35].

k-means clustering is an unsupervised learning technique that is used to sort each observation within a data set into one of *k* unique groups. The technique allows the determination of the similarity between data points over all variables/dimensions. Initially, *k* data points are randomly selected and used as cluster centroids. Each of the remaining observations is assigned to its nearest centroid and this process is repeated iteratively until the sum of the squared errors (SSE) in n-dimensions, between the mean of a cluster μ_k and each of the m points in the cluster x_{ki}, over all *k* clusters,

$$SSE = \sum_{k=1}^{K} \left(\sum_{i=1}^{m} \left(\sum_{v=1}^{n} |x_{kiv} - \mu_{kv}|^2 \right) \right), \qquad (2.2)$$

is locally minimized [35].

The *k*-means algorithm has three parameters which may be tuned by the user: (1) the initial choice of cluster centroids, (2) the number of clusters *k*, and (3) the distance metric [35].

1. The algorithm is initialized either by (a) by selecting specific points to be used as centroids, (a procedure which may be appropriate when the data set is partially labelled, or (b) by selecting points at random so that the algorithm will recursively find the best choice of cluster centroids through an optimization process. In the present study, *k*-means clustering is used as an exploratory technique and is therefore always initiated randomly.

2. The choice of *k,* the number of clusters is often ambiguous, but it should be guided by the structure of the data and domain knowledge [35]. In exploratory analysis, it is not uncommon to run the algorithm using a variety of *k* values. The optimal number of clusters can then be selected based on predefined criteria, such as the minimum number of points per cluster [35].

3. The most commonly used similarity metric for *k*-means clustering is the Euclidean distance metric, which tends to produce clusters that are spherical in shape [35]. This may be ideal in cases where variation within a given group can be attributed to experimental error or spread about an average value. However, it should be noted that, depending on the choice of input variables and distribution of data points within the chosen subspace, alternate metrics may result in more useful or accurate clustering [35]. An example of spherical and non-spherical clusters can be seen in Figure 2.4.

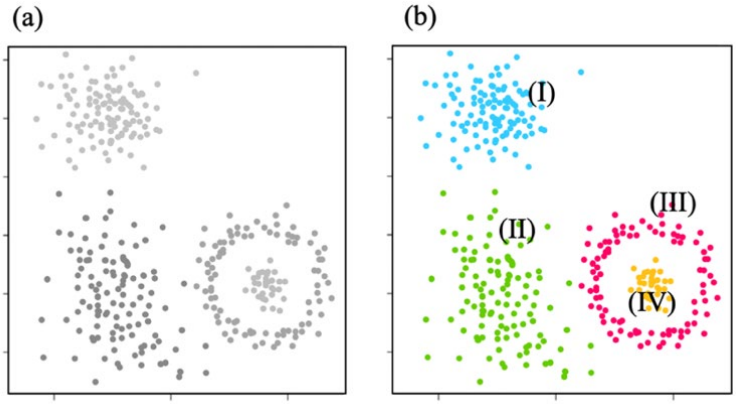

Figure 2.4: Diversity of cluster types illustrating the impact of the representation and the similarity metric on clustering. In (b), a visually intuitive cluster assignment is shown for the input data set in (a). Modified from [18].

It is clear from Figure 2.4b that two of the clusters in the plot window (I and IV) are spherical in shape while the two others (II and III) are not. Clusters II and III can be more accurately classified as having ellipsoidal and ring-like structures, respectively. In these cases, applying the Euclidean distance metric in the original representation of the data may not achieve an optimal cluster assignment. For example, the Mahalanobis distance metric has been shown to outperform the Euclidean distance metric in the detection of ellipsoidal clusters, such as cluster II Figure 2.4b [35]. Furthermore, it is clear that a linear distance metric would not be sufficient for distinguishing cluster III from cluster IV, since points from opposite sides of the cluster III ring are further in Euclidean distance from each other than they are to any point in cluster IV. In these cases, the representation of the data must be modified prior to clustering, for example, by using kernel tricks and nonlinear dimensionality reduction techniques [35]. It should be noted that it is likely not possible to simultaneously achieve the above cluster assignment of all 4 groups shown in Figure 2.4b. However, this example shows the extent to which the accuracy of *k*-means clustering is highly dependent on the representation of the data, i.e., the choice of the input variables, and furthermore, illustrates the potential for misclassification if the distance metric is not based on a clear understanding of the existing data structures [35]. In the present study, we will explore the use of PCA as a feature extraction technique to optimize the representation of the data for *k*-means clustering of IR data for PEX-a pipes of different formulations.

2.2.3 Support Vector Machines (SVM)

Support vector machines (SVM) is a supervised machine learning technique often used to build predictive models. When SVM is used for predictive modelling, the data is randomly divided into training and testing data sets and the model is generated using only data from the training set. During training, the data is mapped onto a higher, n-dimensional feature space in which data of different (known) classes can be separated by n-dimensional linear decision boundaries, known as linear hyperplanes [36, 37]. The set of mathematical functions that is used to optimally transform the data is known as an SVM kernel. An example of such a transformation T is shown for data in two dimensions in Figure 2.5.

The decision boundary for the model is the linear hyperplane defined in the training stage by maximizing the orthogonal distance between each data point and the plane, or by maximizing the margin between the closest points in each of the defined classes [36, 37]. An example of the optimal linear hyperplane for an arbitrary data set is shown in two-dimensional feature space in Figure 2.6. During the prediction part of the analysis, the same transformation is performed on the testing data set and the optimal hyperplane is used as a decision boundary to predict the class of the testing data, or to predict the class of completely new observations. Data points that lie on the margins help to determine the location of the optimal hyperplane and are therefore known as support vectors.

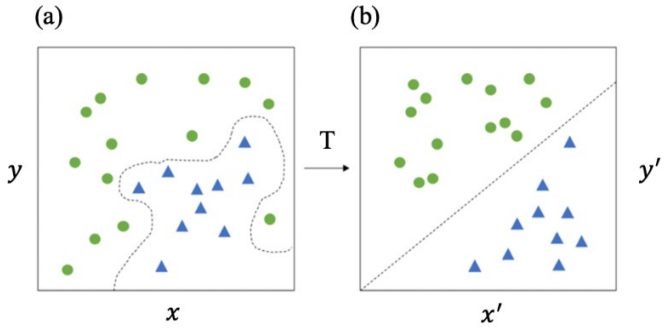

Figure 2.5: Schematic mapping of a data set (coordinates $\boldsymbol{x}$ and $\boldsymbol{y}$) to a new two-dimensional feature space (coordinates $\boldsymbol{x}'$ and $\boldsymbol{y}'$) via a transformation T. This mapping allows a linear boundary to be drawn between points of different classes (b). The boundary is often visualized in the original feature space to generate what is known as a contour plot (a).

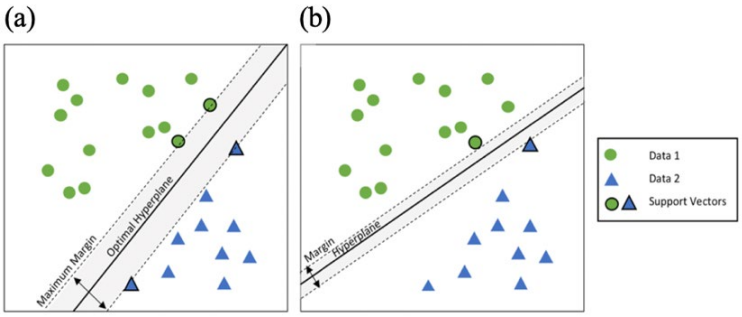

Figure 2.6: Contour plot showing the SVM decision boundary selection in two dimensions. Two different hyperplanes separating Data 1 (circles) and Data 2 (triangles) are shown in (a) and (b). The data points lying on the hyperplanes are indicated with a black outline. The margin between the hyperplane and the nearest data points is maximized in (a), corresponding to the optimal hyperplane.

2.2.4 Decision Tree (DT) Classification Analysis

Decision tree (DT) algorithms are supervised classification algorithms used to build predictive models. The algorithm identifies features in the existing data that can be used to derive a set of if-then classification rules. These rules can then be used to construct a model capable of

classifying new data. One benefit of this algorithm is that, unlike many black box machine learning algorithms, the classification criteria can be extracted and visualized in a tree-like flow chart. This serves as a way to visualize the specific data features extracted by the algorithm and has been extremely useful in allowing the user to better understand the underlying patterns and structure of the original data [38, 39]. The main objective of DT analysis is to partition the data objects into groups with a unique class label based on natural separations in the predictor variables. This is done by identifying the variable that best splits a given data set into two subgroups that can be labelled according to the majority class in that subgroup. In the present study, class labels have been assigned to each IR spectrum according to the aging method and exposure time used for that sample. In the case of continuous variables, such as the infrared absorbance intensities in the present study, split criteria will often be given as threshold values for the chosen variables. The split criteria are selected by maximizing the "purity" of the resulting subgroups, or nodes, for a given split using entropy methods [40]. This involves minimizing the impurity

$$I(A) = \sum_{i=1}^{C} - p_i(\mathrm{A}) log_2(p_i(A)) \tag{2.3}$$

of each node, A, by maximizing the information gain

$$\Delta I = p_i(A)\mathrm{I}(\mathrm{A}) - p_i(A_L)\mathrm{I}(A_L) - p_i(A_R)\mathrm{I}(A_R) \tag{2.4}$$

that occurs when node A is split into left and right daughter nodes, A_L and A_R respectively. Here $p_i(A)$ is the proportion of total data objects which belong to class i in node A [40][1]. The subgroups created by the initial split of the root node are then separated and the same algorithm is performed on each subgroup separately. The process of recursively splitting each subgroup into two groups is known as recursive partitioning and is repeated until the entire data set is split into pure groups (with each subgroup consisting of only one class), or until there is no statistically significant difference between classes in a given subgroup. This process will result in a cascading, tree-like decision map similar to that shown in Figure 2.7.

As stated above, decision trees have the competitive advantage of not being a 'black box' method, which makes both the predictive capability and classification process of the model

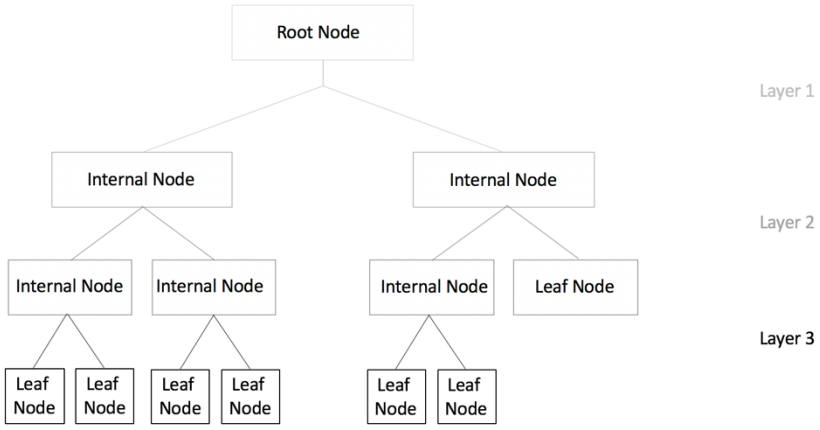

Figure 2.7: Illustration of the decision tree concept, in which the root node is split into two internal nodes, which are then split into two internal nodes and so on until only leaf (completely pure) nodes are obtained.

[1] The impurity, I has a maximum of $log_2(C)$, for a completely impure node, and a minimum of 0, for a completely pure node, also known as a leaf node.

particularly useful [38, 39]. It performs extremely well on high dimensional data sets, in part because of its ability to ignore irrelevant variables [41]. However, these advantages come at some expense. The algorithm proceeds in what is known as a 'one-step-optimal' fashion, with no ability to look ahead to assess the quality of future splits [40]. This makes the overall accuracy of the tree highly dependent on the initial (root node) splits. The DT algorithm is also sometimes criticized for being a what is known as a low-bias, high-variance technique. The bias-variance trade-off exists for many descriptive models and arises from the tendency for low-complexity modelling techniques to underfit the data and for high-complexity modelling techniques to overfit the data. Underfitting occurs when the model has a high amount of prediction bias and thus low prediction accuracy. However, these modelling techniques often have the advantage of generating consistent and reproducible models when a training data set is varied slightly and are therefore referred to as "high-bias, low-variance" techniques. On the other hand, overfitting can occur when high-complexity modelling techniques learn the training data set almost perfectly. This results in low bias models with a high prediction accuracy; however, these models also have a tendency to vary significantly when the training data set is modified slightly. The low bias and high variance tendencies of the DT algorithm often leads to significant variation between trees constructed with different variable and data subsets – a problem that impacts the reproducibility of the analysis. Furthermore, decision trees suffer from two intrinsic selection biases: (1) bias towards variables with a large number of values (possible splits), and (2) bias towards splits that are optimal for classes that have large number of observations in the data set [39, 40, 42]. In the present study the effect of the variable bias is minimized, since the precision of the measurement

allows for each observation to take a unique value for each variable, making the number of possible splits equal to the number of observations for every variable included in the analysis. The uneven data bias was minimized by ensuring that the data for each sample included approximately the same number of spectra.

Another technique that can be used to reduce overfitting of the DT model is cost-complexity pruning. In cost-complexity pruning, the fully-grown DT model is trimmed back to include only a subset of the total number of leaf nodes. The trimming begins by selecting and removing the leaf node with the lowest purity. Next, a complexity parameter (cp) is defined in order to restrict the DT algorithm from growing past this point. This is repeated on the pruned tree until the root node is reached and a full set of cp values is obtained for the DT model. At each pruning step a 10-fold cross validation is performed wherein the training data is split into 10 random subsets (folds) of equal size and a tree is grown subject to the cp restriction using 9/10 folds. The remaining fold is used to calculate the relative total impurity[2] of the tree and this process is repeated until each of the 10 folds have been used for testing. This allows for the calculation of an average relative total impurity for each cp value. The average relative total impurity can then be plotted as a function of cp value (or number of pruned nodes) in order to reveal how the extent of pruning affects the overall accuracy of the DT model. The cp value (and resulting pruned tree,) can then be selected such that the relative total impurity is minimized.

[2] Relative total impurity is the total number of misclassified samples normalized by the impurity of the root node.

2.2.5 Random Forest (RF) Classification Analysis

A recently developed solution to the problems associated with DT analysis is the use of ensemble methods such as the Random Forest (RF) classification analysis [24]. The RF algorithm is used to build a predictive model using a user-specified number of decision trees, each with its own set of unique if-then classification rules, capable of independently classifying new data. Each decision tree in the forest is generated using a randomly selected subset of variables, which allows for some trees to be built without the use of particular biasing variables. Further, each tree is trained using a randomly selected subset of observations (typically chosen to be 2/3 of the full data set), which solves the bias experienced by individual trees in the case of unequal data sets [41]. The remaining 1/3 of the data for each tree is used to test the individual tree's accuracy. The probability of misclassification for each tree is known as the 'out of bag' (OOB) error estimate, a quantity that is typically measured for each class separately. The OOB error rate for the RF model can be calculated for each class by averaging the OOB error over all trees in the forest. The total OOB error for the RF model can also be calculated by averaging the OOB error over all classes. When RF models are used for prediction, previously unseen data are fed into the model and forwarded to each tree. Each tree will classify the observation according to its own if-then classification rules and the overall predicted class is determined to be that which obtains the majority vote across all trees.

RF models are tuned using two variables: the number of trees, t and the number of randomly selected variables used to generate each tree, q. Each of these parameters can be optimized by assessing the change in total OOB error as the value of each variable is increased.

The optimal number of trees is selected to be that which produces a minimum total OOB error that ceases to fluctuate with an increase in t. The optimal number of variables is selected at the global minimum in the total OOB error rate; it is typically found to occur at $\sqrt{Q}$, where Q is the total number of variables in the data set [41].

As stated above, the RF algorithm solves many of the problems associated with the DT algorithm. However, since it is nearly impossible to visualize hundreds (if not thousands) of trees, the RF algorithm lacks the transparency that makes DT algorithms so appealing. There are a number of metrics that have been proposed to make up for this lack of transparency by assessing how each variable contributes to the prediction accuracy and quality of the model [41]. In the present study, we focus on four such metrics: (1) the mean minimal depth, (2) the Gini index, (3) the mean decrease in accuracy, and (4) the p-value.

1. The minimal depth is the minimum number of splits required to produce the first occurrence of a leaf node when the root node is split according to a given variable. The mean minimal depth is calculated by averaging the minimal depth of each variable over all trees. This metric provides a measure of the amount of information contained in each variable, since a split along a variable with a small mean minimal depth will produce trees which reach pure nodes quicker and are thus simpler. The mean minimal depth therefore does not assess variable importance according to prediction accuracy, but is instead based on the ability of the variable to minimize the size of the resulting tree (something that has been noted to improve the problem of overfitting) [40].

2. The Gini index is the average decrease in node impurity (across all nodes and trees) when a subset is split according to a given variable.
3. The mean decrease in accuracy measures the decrease in accuracy over all trees, or increase in OOB error rate, that occurs when a given variable is removed from the data set or replaced with random noise [41].
4. The p-value measures the evidence against the null hypothesis. It can be understood as the probability of obtaining a response or trend as extreme as those observed in the data set if the data is assumed to be completely random. Typically, a p-value < 0.05 is considered statistically significant. That is, a trend is considered statistically significant when the chances of observing it in a completely random phenomenon is 5% or less.

3 Materials, Methods and Analysis

Parts of this chapter are based on work described in reference [23].

3.1 Sample Preparation

3.1.1 Overview of Samples

Three different formulations of extruded, 2-mm wall thickness PEX-a pipe were used in these experiments, each manufactured using different polyethylene pellets, additive packages, and crosslinking peroxides (Table 3.1). This sample set was selected to allow for a comprehensive study of the effect of the additive package and crosslinking peroxide on the physical and chemical properties of PEX-a pipe as measured using FTIR.

Table 3.1: Pipe formulations specified by the additive package and the crosslinking peroxide. Additive packages I and II differ in the relative amounts of specific primary antioxidants, secondary antioxidants, or hindered amine light stabilizers (HALS). The specific additives and relative amounts have been withheld as the formulations used in the present study are proprietary. Modified from [23].

Pipe Formulation	A	B	C
Additive Package	I	II	I
Crosslinking Peroxide	a	a	b

3.1.2 Preparation of Radial Slices for Transmission IR Spectroscopy Experiments

Samples outlined in Table 3.1 were prepared for the transmission IR spectroscopy experiments by slicing along the direction of extrusion of the pipes using an American Optical

model 820 rotary microtome. This resulted in a series of 6–12 consecutive radial slices (~100 μm thick) across the thickness of each pipe as shown in Figure 3.1. Series of radial slices were prepared for three different, randomly chosen locations along the length of each of the three pipe formulations. The thickness of each slice was measured using a micrometer, and the thickness values were used to calculate the radial depth corresponding to each slice within the pipe. These samples were used to develop methodology to compare and classify pipe formulations based on characteristic IR spectroscopy absorbance peaks using PCA, SVM and *k*-means clustering.

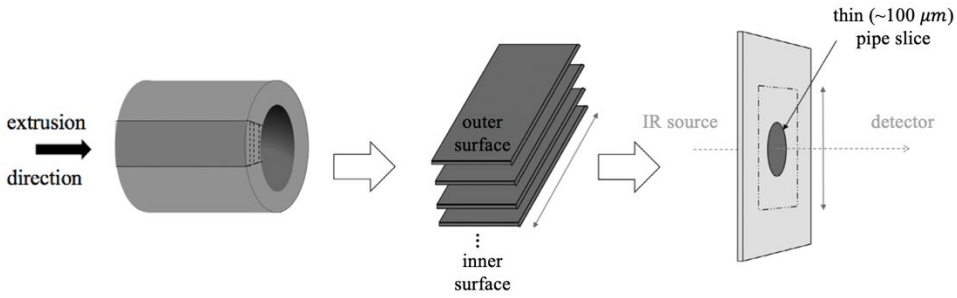

Figure 3.1: Sample preparation and experimental geometry used for the transmission FTIR spectroscopy experiments performed on radial slices of PEX-a pipe.

3.1.3 Accelerated Aging of PEX-a Pipe

The long-term performance of PEX-a pipes is, in large part, dependent on the structural properties of the polyethylene (PE) as well as the proper function of antioxidant additives that are incorporated into the PE matrix. Characterizing the time-correlated degradation of polymer and additive compounds with in-service use is therefore paramount to predicting pipe failure.

Accelerated aging is a technique that is used to determine the response of a material to the normal aging process, by exposing the material to more severe or frequent stresses over a shorter period of time. The primary assumption is that if accelerated aging is done in a controlled manner, the cumulative degradation to key chemical and physical properties of the pipe will proceed along the same trajectory as it would if the pipe were exposed to the conditions that produce natural weathering.

PEX-a pipes of formulation A were exposed to air and Milli-Q water at elevated temperatures for extended periods of time to observe and compare the effect of these external stresses on changes to the chemical composition of the pipe samples. The conditions chosen for accelerated aging in air were chosen to expose the pipe samples to similar but more extreme stresses, such as the presence of oxygen and the use of elevated temperatures, than would be experienced by the outer surface of the pipe wall in real world use. Similarly, the conditions chosen for accelerated aging in Milli-Q water were chosen to expose the pipe samples to similar but more extreme stresses experienced by the inner surface of the pipe wall. The goal of the accelerated aging studies was to assess whether it is possible to induce the same type of changes in pipe properties experienced in real world installations at relatively low hot water temperatures for extended periods of time by using higher temperatures for shorter periods of time.

Segments of extruded, 3 mm wall thickness PEX-a pipes, of approximately 6 cm in length were prepared for accelerated aging experiments. The segments were placed in glass beakers and either exposed to hot air at 85 $^{\circ}$C or exposed to hot Milli-Q water at 85 $^{\circ}$C. Samples were aged at elevated temperatures for a total of 21 days. The progression of aging was

monitored at 3, 6, 8, 10, 14 and 21 days. At each time point, the samples were removed from the oven and allowed to cool for a minimum of 20 minutes under room temperature conditions before being prepared for FTIR microscopy measurements. This set of samples was used to characterize the effects of accelerated aging of PEX-a pipe using PCA and decision tree-based methods.

3.1.4 Preparation of Axial Slices for Transmission FTIR Microscopy

Samples were prepared by slicing PEX-a pipe samples perpendicular to the extrusion direction using an American Optical model 820 rotary microtome, resulting in ~100 μm thick axial slices, as shown schematically in Figure 3.2. To ensure that the measured radial changes to pipe properties were due to exposure of the outer and inner pipe walls, and not the ends of the pipe samples, to the accelerated aging conditions, ~500 μm of the pipe samples were removed from the end of the pipe sample before slicing the axial slice used in the FTIR microscopy experiments.

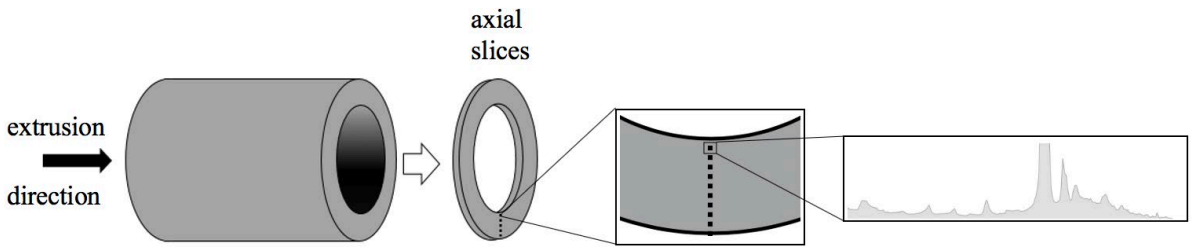

Figure 3.2: Sample preparation and experimental geometry used for the FTIR-microscopy experiments performed on axial slices of PEX-a pipe.

3.2 Transmission IR Spectroscopy Experiments

3.2.1 Transmission IR Measurements of Radial Slices

For radial slices, transmission IR absorption spectra were collected using a Thermo/Nicolet Nexus model 470 FT-IR ESP spectrometer equipped with a DTGS detector and a ZnSe wire-grid polarizer. The geometry of the measurement is shown schematically in Figure 3.1. The spectra were acquired using 32 co-added scans in the wavenumber range 4000–600 cm^{-1} at a resolution of 4 cm^{-1} using light polarized parallel and perpendicular to the direction of the extruded pipe as shown in Figure 3.1. IR absorption is expressed in absorbance units

$$A = -\log (I/ I_0) \tag{3.1}$$

where I and I_0 are the single beam intensities transmitted through the pipe slice and air respectively. Parallel ($A_{\parallel}$) and perpendicular ($A_{\perp}$) polarized spectra were collected for each slice and used to calculate the corresponding orientation-independent spectrum according to [43, 44]:

$$A_0 = \frac{A_{\parallel}+2A_{\perp}}{3} \quad . \tag{3.2}$$

3.2.2 Transmission IR Spectroscopy Measurements of Axial Slices

Transmission IR absorption spectra were collected for axial slices using a Thermo/Nicolet Continuμm Infrared Microscope equipped with an MCT detector. Each spectrum is the result of 32 co-added scans in the wavenumber range 4000–650 cm^{-1} at a resolution of 4 cm^{-1}. Approximately 30 spectra were collected, in 100 μm increments, across the wall thickness of

each axial slice to obtain a radial profile of each sample as shown schematically in Figure 3.2. Each spectrum represents the average IR absorption over a 50 × 50 μm area (at a given radial depth), expressed in absorbance units as outlined in section 3.2.1.

3.3 Data Analysis Techniques

3.3.1 Data Pre-Processing for IR Data Collected from Radial and Axial Pipe Slices

A baseline subtraction was performed separately for each absorbance spectrum using the CRAN 'baseline' package in R. For spectra collected from radial slices as described in section 3.2.1, the "modpolyfit" correction method was used to fit an n^{th} order polynomial to segments within the wavenumber range 3700–850 cm^{-1}. For spectra collected from axial slices as described in section 3.2.2, the asymmetric least squares ("als") correction method was used to individually baseline the 9 wavenumber intervals shown in Figure 3.4. In both cases, saturated regions of the spectrum were removed prior to baselining. The IR absorption band at 2019 cm^{-1}, which can be attributed solely to polyethylene and arises from both amorphous and crystalline regions, was used as an internal reference to account for sample-to-sample intensity variations arising from differences in sample thickness [2]. The absorbance values at each wavenumber in a spectrum were divided by the maximum intensity of the 2019 cm^{-1} band. The resulting data from each experiment were then used to construct an $n \times m$ data matrix X in which each of the n independent observations represent a single spectrum of absorbance values for m selected

wavenumbers. In this work, spectral regions of interest were selected to be centred on vibrational frequencies of chemical species in the polyethylene and the antioxidant additives in PEX-a as shown in Figure 3.5. The individual wavenumbers used in each experiment can be seen in Table 4.1 and Table 4.2.

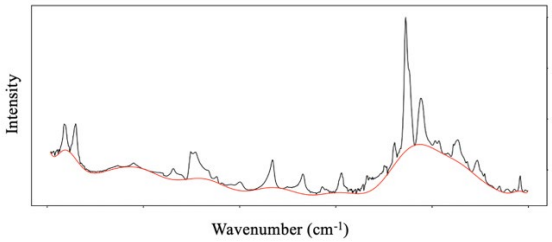

Figure 3.3: Example of polynomial baseline fit (red) for spectra (black) collected from radial slices, using the “modpolyfit” correction method for the CRAN ‘baseline’ package in R.

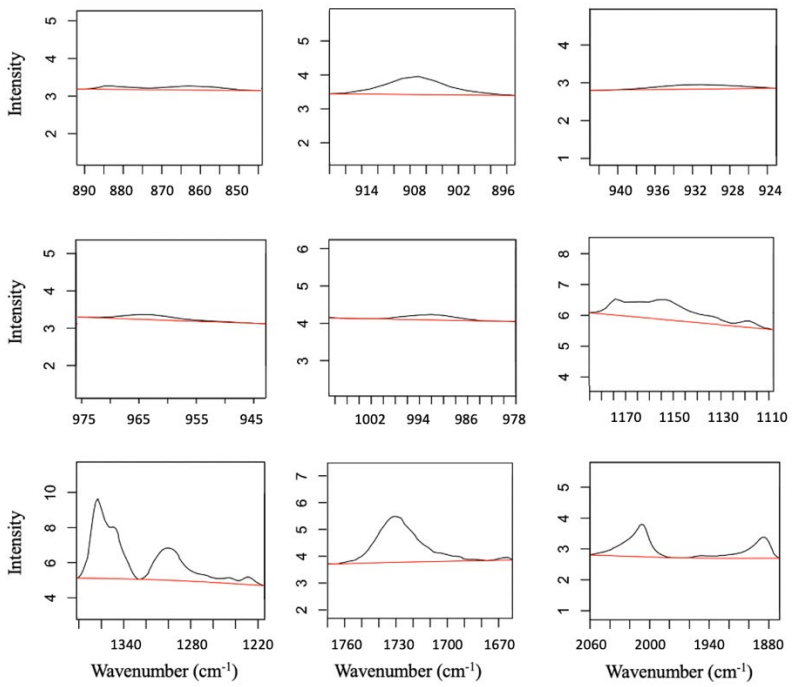

Figure 3.4: Example of the nine-section baseline fit (red) for spectra (black) collected from axial slices using the asymmetric least squares method in R. Each of the nine spectral regions shown are from distinct, randomly selected spectra within the dataset.

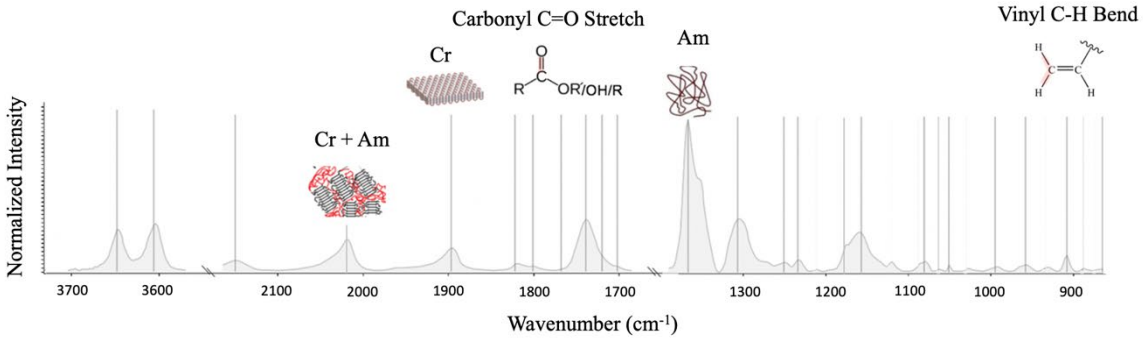

Figure 3.5: Representative IR spectrum of PEX-a pipe sample after applying the baselining procedure. The vertical lines correspond to the wavenumber values used in the PCA of IR data collected from radial and axial pipe slices.

3.3.2 PCA on IR Data Collected from Radial and Axial Pipe Slices

In each experiment, the m columns of X were centered by subtracting the column mean from each of the n observations in the column, which ensured that the maximum variance directions were defined with respect to the column mean. The n values for each of the m selected wavenumbers were then scaled by dividing each value by the standard deviation of its column, which ensured an equal weighting of all wavenumbers and ensured that the PCA was not dominated by vibrational frequencies with larger intrinsic absorption intensities.

PCA was used to convert the m original variables from each data set into a different set of m uncorrelated variables or PCs. In each case, the original data matrix X was geometrically projected onto the matrix composed of eigenvectors U, to obtain a matrix of transformed data points P, whose elements are known as PC scores. The eigenvalues (or proportion of explained variance) for each PC were converted to percentages and visualized in descending order as a scree plot, similar to the one shown in Figure 2.3c [22]. The scree plot was used to observe the distribution of variance across the PCs and to select a subset of $p \leq m$, high-variance PCs to be retained for further analysis. In the present study, we complemented the results of the PCA scree plot analysis with a parallel analysis based on Horn's method using the CRAN "paran" package in R [22]. The score values of the retained PCs were visualized in scatter plots, similar to the one shown in Figure 2.3b. In the present study, a physical significance was assigned to the retained PCs by observing the relative weights of each of the wavenumbers in the linear combinations that form each PC.

In both experiments, involving both axial and radial slices, a parallel $n \times m$ data matrix, X' was constructed in which each of the n independent observations represent a single baseline fit for all m baselined wavenumbers. A PCA was performed on the scaled and centred baseline matrix to ensure that artifacts and biases were not introduced via the baselining procedure. The score values for PC1 and PC2 were visualized in a 2D scatter plot of PC1 versus PC2. We observed a random distribution of PC score values for PC1, which captured > 98% of the total variance in the data. This result indicates that no significant trends or correlations existed between the individual baselines.

3.3.3 *k*-means Clustering on IR Data Collected from Axial Pipe Slices

IR spectra collected from radial slices of the three pipe formulations listed in Table 3.1 showed significant differences in the intensities of several IR absorbance bands. Because of this, we performed an unsupervised, *k*-means clustering analysis to evaluate if the IR data in the original representation (IR absorbance intensities) or in the transformed representation (PC space) could be accurately classified into three distinct clusters for the three different pipe formulations. In the present study, *k*-means clustering was performed using the CRAN "kmeans" package in R. The *k*-means analysis was performed on the original data matrix X, by assigning each observation to one of three groups or clusters based on its value in m-dimensional space. The *k*-means analysis was also performed separately on the transformed data matrix P, by assigning each PC score to one of three groups or clusters based on its value in p-dimensional

space. The assigned clusters in both clustering experiments were visualized in PC space by labelling points according to their cluster assignment, so that the results of each clustering analysis could be compared directly.

3.3.4 SVM Classification of IR Data Collected from Axial Pipe Slices

The results of the PCA and *k*-means clustering analysis on IR data collected from radial pipe slices were complemented by using a supervised machine learning classification technique known as an SVM model. To generate the SVM model, the data set of IR intensities for the m selected bands (Table 4.1) were randomly divided into a training data set, consisting of two-thirds of the data points, and a testing data set, comprising the remaining one-third of the data points. The model was trained to predict the response variable (pipe formulation type) class based on the values of predictor variables (IR intensities) using data from the training data set. The results of the predictive model were visualized in a contour plot, similar to the one shown in Figure 2.5a, in which the decision boundaries for the SVM model were superimposed on the original data points to evaluate the fidelity of the classification scheme, that is, to ensure that errors in the classification scheme were not biased to a specific set of outlier points. The model was used to predict the class of the data in the testing data set and the results were compared to the actual class label to assess the accuracy of the SVM model.

3.3.5 Tracking Spectral Regions of Interest (ROI) in IR Spectra Collected from Axial Pipe Slices using Indices

Functional groups from both the PE and the additives can contribute to a given absorption band, making it challenging to determine the accurate cause for differences in the overall band shape or intensity. This can limit the amount of aging information that can be extracted from changes to important IR bands since, as the pipes age, degradation can proceed via a number of complex mechanisms, each effecting the pipe in slightly different ways. This is particularly problematic for bands in the carbonyl region (1775–1765 cm^{-1}), since oxidative degradation can lead to an increase in the intensity of overlapping bands at low wavenumbers (1730–1675 cm^{-1}), while leaching of the additives that contain ester species can cause the intensity of overlapping bands at high wavenumbers (1775–1720 cm^{-1}) to decrease as shown in Figure 3.6. Furthermore, it is important to carefully monitor the behavior of both amorphous and crystalline components of the PEX-a pipes during the aging process. While crystalline regions act as barriers to additive diffusion, the amorphous regions are most susceptible to oxidative degradation, allowing diffusion of additives and oxygen within the pipe wall.

In order to accurately characterize key responses of PEX-a pipes to accelerated aging, we defined three indices based on peaks in IR spectra of PEX-a pipe that we used to track the stability and presence of additives, as well as the uniformity of the cross-linking and signs of degradation across the wall thickness of the pipes. By comparing IR absorptions that arise exclusively from the crystalline or amorphous regions of the PE, that is, bands at arise

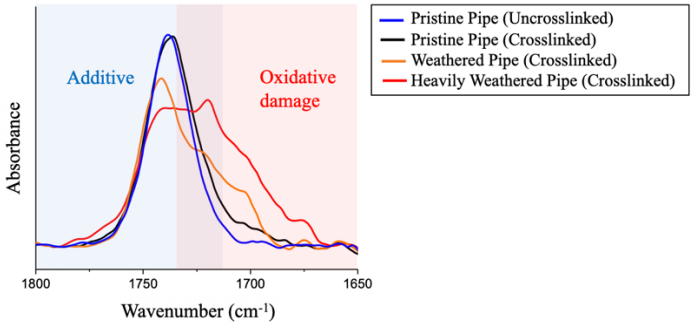

Figure 3.6: Sample IR spectra from aged and unaged (crosslinked and uncrosslinked) PEX-a pipe to illustrate the types of changes that can arise in the carbonyl region (1675-1775 cm^{-1}) with intense weathering of PEX-a pipe. Figure by M. Grossutti.

exclusively from the crystalline or amorphous regions of the PE, that is, bands at 1305 cm^{-1} and 1895 cm^{-1} in Table 4.1, we obtained a measure for the degree of crystallinity, which is an important factor in determining the mechanical strength of the pipe, the effectiveness of the stabilizing additives, and the long-term stability of the pipe. In the present study, the crystallinity index was defined according to

$$Crystallinity\ Index = \left(\frac{\frac{A_{1895}}{A_{1305}}}{\frac{A_{1895}}{A_{1305}} + \frac{C1}{C2}} \right) \times 100, \tag{3.3}$$

where A_{1895} and A_{1305} are the integrated areas of the IR bands arising from the crystalline and amorphous regions of PE, respectively. The coefficients, $C1 = 6.1$ and $C2 = 17$ are the absorption coefficients for the 1895 cm^{-1} and 1305 cm^{-1} bands, respectively. As stated above, the specific carbonyl content within the pipe can be monitored by carefully tracking the different IR

absorbance intensities within the range of 1675-1730 cm^{-1}. In order to obtain independent measures of additive leaching and oxidative degradation we defined two indices: the carbonyl index and center of gravity (COG), which measure the overall amount of carbonyl species and the shape of the carbonyl band, respectively. The carbonyl index was defined as the integrated area of the baselined and normalized spectral region from 1775-1675 cm^{-1}. The carbonyl COG was defined as the weighted average of all wavenumbers in range of 1775-1675 cm^{-1}, where each wavenumber is weighted by their respective baselined and normalized IR intensity. When used together, these indices allow us to assign changes to the carbonyl band to the presence of either additives or oxidative species within the pipes.

3.3.6 Random Forest Classification of IR Data Collected from Axial Pipe Slices

The results of the PCA analysis on the IR data collected from axial pipe slices was supplemented by using a supervised machine learning classification technique known as a Random Forest (RF) [45]. In the present study, the m selected variables (IR intensities) that were used in the PCA analysis of axial pipe data were used as predictor variables to generate a RF model using the CRAN "randomForest" package in R. The number of decision trees t, used in the RF model and the number of variables q, used in constructing each decision tree was selected by determining the respective values that gave a minimum total OOB error[3], when averaged over

[3] The 'out of bag' (OOB) error estimate for a tree in the RF algorithm is the number of observations that are misclassified by that particular tree, calculated using the remaining 1/3 of the data that was not used in the tree's construction.

100 RF models[4]. The number of decision trees t, that were sampled in this analysis ranged from 100-2500 and the number of variables in each tree q, that were sampled in this analysis ranged from 1-7. The RF model was used as an exploratory tool to determine which IR bands contained the most relevant information for predicting the pipe class, which was assigned based on the aging time and method. The most relevant IR bands were determined using the mean minimal depth, the Gini index, the mean decrease in accuracy, and the p-values, which were plotted in two multi-way plots.

3.3.7 Decision Tree Classification of IR Data Collected from Axial Pipe Slices

The results of the PCA and RF analyses on the IR data collected from axial pipe slices was complemented by using a supervised machine learning classification technique known as a Decision Tree (DT). In the present study, the 7 variables that were found to be the most relevant in the RF analysis were used as predictor variables to generate a DT model with the CRAN "rpart" package in R. The DT model was pruned using a cost-complexity pruning with a 10-fold cross validation as outlined in section 2.2.4. The resulting data pruned DT model was visualized in a flowchart.

[4] The total OOB error is the OOB error for all classes averaged over all trees. This quantity is averaged over all 100 RF models generated at each sampled value of n, to obtain the average total OOB error.

4 Results and Discussion

Parts of this chapter are based on work described in reference [23].

4.1 Classifying Formulations of PEX-a Pipes Using PCA and *k*-means Clustering on IR Data Collected from Radial Slices

A representative IR absorbance spectrum for pipe A (Table 3.1) is shown in Figure 3.5. Many of the absorption peaks can be attributed to molecular vibrations characteristic of the PE and/or the additives within the pipe. In the present study, we have identified 22 distinct IR bands that play an important role in determining the mechanical strength, resistance to degradation and long-term stability of the pipes. The selected bands include those that arise exclusively from the crystalline or amorphous regions of the PE as well as chemical species that play an important role in degradation, such as carbonyl and vinyl groups. In Table 4.1, we list the wavenumber values corresponding to the 22 distinct bands that we have used, together with the molecular groups that can be attributed to the bands.

As mentioned in section 3.3.5, it is not uncommon for bands of different molecular origin to overlap, making it challenging to determine accurately the cause for differences in the overall band shape or intensity. For example, it is difficult to determine whether differences in the band near 1160 cm^{-1} is due to the differences in additives or polyethylene. Unbiased approaches to pattern recognition are thus extremely useful in these situations. In the present study, PCA, *k*-means clustering and SVM analyses were performed on a data set consisting of 85 IR spectra

collected from roughly equal numbers of radial slices from PEX-a pipes of formulations A, B and C (Table 3.1).

Table 4.1: Peak wavelengths of the absorbance bands used in the study of radial slices of PEX-a pipe by transmission FTIR-spectroscopy. Modified from [23].

Band Frequency	Source
908 cm^{-1}	PE vinyl unsaturation $\mathbf{CH = CH_2\ H}$ out of plane bending
958 cm^{-1}	PE trans unsaturation $CH = CH$ H out of plane bending
964 cm^{-1}	PE trans unsaturation $CH = CH$ H out of plane bending
1051 cm^{-1}	Crystalline PE
1079 cm^{-1}	Amorphous PE $C - C$ stretching
1160 cm^{-1}	Ester $C - O - C$ from additive
1176 cm^{-1}	Crystalline PE
1234 cm^{-1}	Phenolic $C - OH$ from additive
1249 cm^{-1}	Unassigned
1305 cm^{-1}	Amorphous PE CH_2 wagging
1367 cm^{-1}	Amorphous PE CH_2 wagging
1531 cm^{-1}	Triazine ring from additive
1569 cm^{-1}	Triazine ring from additive
1639 cm^{-1}	PE chain unsaturation $C = C$
1739 cm^{-1}	Carbonyl stretching from additive
1799 cm^{-1}	Unassigned
1818 cm^{-1}	Unassigned
1895 cm^{-1}	Crystalline PE
3371 cm^{-1}	Unassigned
3436 cm^{-1}	Unassigned
3606 cm^{-1}	Phenol $O - H$ stretching from additive
3648 cm^{-1}	Phenol $O - H$ stretching from additive

4.1.1 Principal Component Analysis (PCA)

PCA was performed using the IR absorption peak maximum values of 22 selected absorbance bands (Table 4.1) as input variables. The total amount of variance explained by each

PC is shown in the scree plot in Figure 4.1a. The results indicate that PC1, PC2, and PC3 account for a very significant amount (89%) of the total variance in the data, with the remaining variance (11%) approximately equally distributed between the other 19 PCs. Parallel analysis performed on the eigenvalues of the PCA confirmed that it was appropriate to retain the first three PCs for further analysis (Figure 4.1b), since PC1-PC3 are the only adjusted eigenvalues that are greater than one. The score values for the first three PCs are shown in a 3D scatter plot in Figure 4.2a, and distinct clusters are observed corresponding to the three different pipe formulations. The relative weights for each of the wavenumbers in the linear combinations for the first three PCs are shown in Figure 4.3, in which the IR band weights are labeled by color according to their physical origin: crystalline PE, amorphous PE, PE chain unsaturations, and additives.

By examining the most heavily weighted IR bands (and their physical origins, as described in Table 4.1) for PC1-PC3 (Figure 4.3), together with two-dimensional plots of PC scores for PC1-PC3 (Figure 4.2), we can gain considerable insight into the physical significance of the PCs. In Figure 4.2b, in which PC1 is plotted as a function of PC2, it can be seen that the PC1 values for pipe C are offset relative to those for pipes A and B. Since pipe C differs from pipes A and B in the peroxide that is used to initiate crosslinking, the separation of the PC1 values in Figure 4.2b suggests that PC1 captures the variance in the data that results from the use

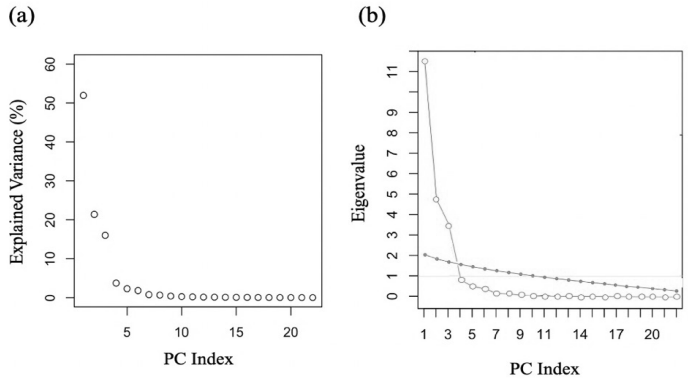

Figure 4.1: Eigenvalue analysis for the PCA on IR data collected from radial pipe slices modified from [23]. (a) Scree plot for PC1-PC22. (b) Adjusted PCA eigenvalues (open circles) calculated by scaling by the corresponding parallel eigenvalues (solid circles) plotted as a function of PC number. The adjusted PCA eigenvalues (open circles) that are greater than one correspond to the PCs that were retained for further analysis.

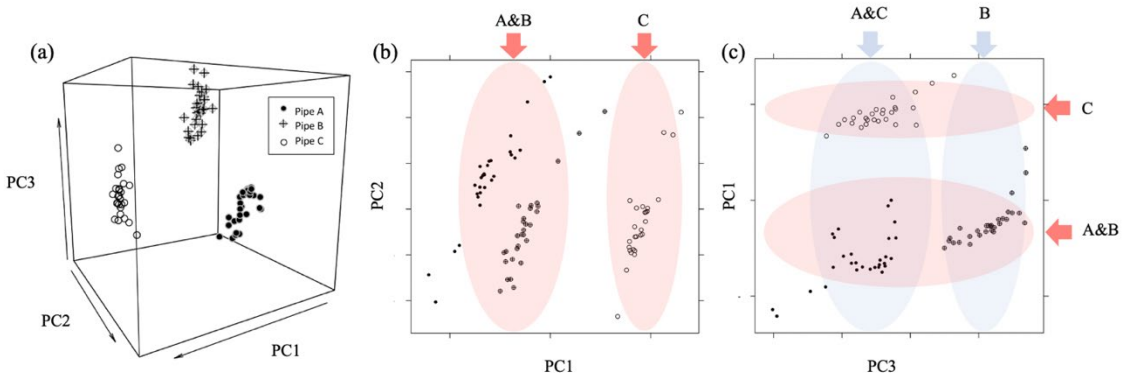

Figure 4.2: (a) 3D scatter plot of PC score values for PC1-PC3. Two-dimensional scatter plots of PC scores are shown for (b) PC2 versus PC1, projected onto the PC3 = 0 plane, and (c) PC1 versus PC3, projected onto the PC2 = 0 plane. The pink and blue ellipses highlight the groupings of the data points for the different pipe formulations A, B, and C on each plot. The points are labelled according to formulation type. Figure has been modified from [23].

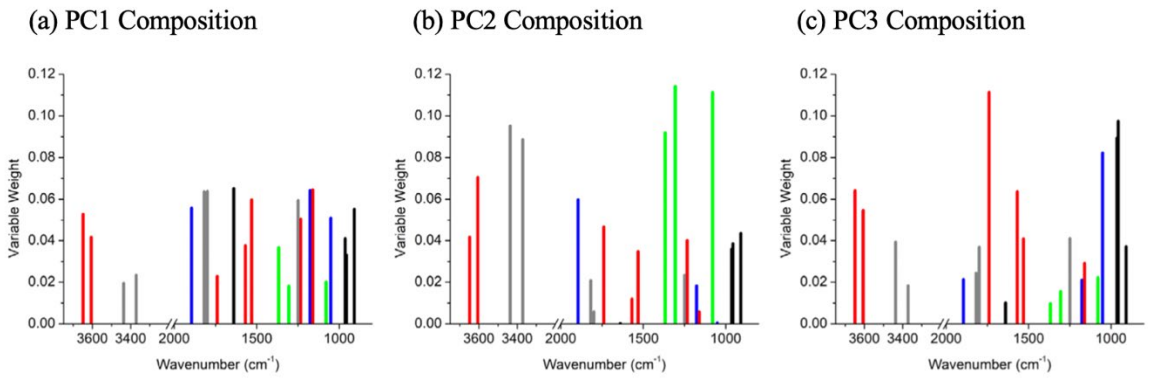

Figure 4.3: Relative weights of the IR bands in the linear combinations that define PC1-PC3. IR band weights are labeled by color according to their physical origin: crystalline PE (blue), amorphous PE (green), PE chain unsaturations (black), additives (red), and unassigned (gray). Figure has been modified from [23].

of different peroxides. Although there are no IR bands that are due to the peroxide, since it is consumed in the crosslinking process, the choice of peroxide clearly has a significant effect on a variety of pipe properties. The underlying PC1 IR band weights (color coded according to their physical origins) are shown in Figure 4.3a , and it can be seen that no single IR band dominates, with the band weightings from crystalline PE, PE unsaturations, and additives contributing roughly equally to PC1. Clearly, the choice of crosslinking peroxide impacts a broad range of pipe properties and explains the greatest variance between pipe formulations. This is not surprising, given that the crosslinking process is central to the PEX-a pipe manufacturing process, with crosslinking having a large impact on a broad range of pipe properties including the degree of crosslinking, crystallinity, and unsaturations. Interestingly, the amorphous PE IR bands have a very low weighting in PC1, as does the carbonyl band that derives its intensity

primarily from the additives in virgin (unused) pipe but can also contain contributions from oxidative damage.

In Figure 4.2c, in which PC1 is plotted versus PC3, we can also see that the values of PC3 distinguish pipe B (additive II) from pipes A and C (additive I). Therefore, in addition to the ability to distinguish the different pipe formulations according to the choice of crosslinking peroxide based on the PC1 values, PC3 makes further distinctions between the pipes according to the choice of additive package. Furthermore, IR bands that are primarily determined by the additives are the most heavily weighted variables in PC3, accounting collectively for over 30% of PC3 (Figure 4.3c). This result supports our interpretation of the source of variance captured by PC3 and its physical significance.

Figure 4.2b shows a large spread in PC2 score values for all three formulations so that this component does not clearly distinguish between formulations. Instead, the PC2 values seem to be dominated by systematic variations that arises due to differences in the spectra collected from different radial depths of each pipe. This experiment could be improved by using only data collected from the same radial depth of each pipe, which would eliminate this source of variance from the data set. This type of refined analysis could also result in a reordering of PC2 and PC3.

In the above analysis, we used the maximum value of each of the 22 selected IR absorbance peaks as the input variables for the PCA. We also performed PCA on a data set that included ranges of absorbance values spanning each IR peak (IR bands). In comparing the results of the analysis using the IR bands to that obtained using only the IR absorbance peak maximum

values, we found that visually similar groupings were obtained. However, PCA applied to the IR bands produced lower accuracy results, with seven PCs required to describe the same total variance (89%), which was likely due to the overlap of bands and noise due to the presence of water vapor. This comparison confirmed that our choice of using the IR absorbance peak maxima as input to the PCA provided sufficient information to distinguish between different pipe formulations, while resulting in PCs that were not dominated by noise.

4.1.2 *k*-means Clustering

To obtain an independent assessment of the classification of the PEX-a pipes, we used *k*-means clustering. Data were clustered into three groups using the results of the PCA, that is, score values of PC1, PC2, and PC3, as input data. This analysis produced clusters, which classified the data points with other data points corresponding to the same pipe formulation with 98.8% accuracy. The assigned clusters are shown in PC space in Figure 4.4a with points labelled according to cluster assignment. We repeated this analysis iteratively, each time applying the clustering algorithm to input data sets composed of PC1-PCn where n ranged from 3 to 22. We observed that the accuracy of clustering reached a maximum for three variables (PC1-PC3), with no improvement in accuracy when the clustering algorithm was applied to more than these three components (Figure 4.5). This observation further supports the results of the scree plot and parallel analysis (Figure 4.1), confirming that PC1-PC3 are sufficient to adequately describe the entire data set.

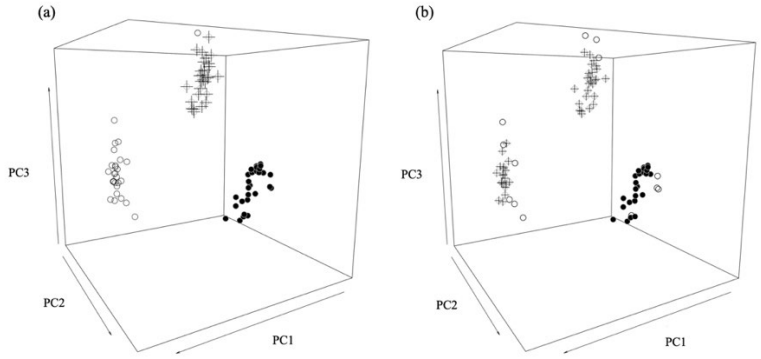

Figure 4.4: 3D plots of the results of *k*-means clustering for k = 3 using (a) PCA score values as input data, and (b) 22 selected (normalized) absorbance values as input data. The results of both clustering analyses for pipe formulations A, B, and C are represented by solid circles, crosses, and open circles, respectively. Figure has been modified from [23].

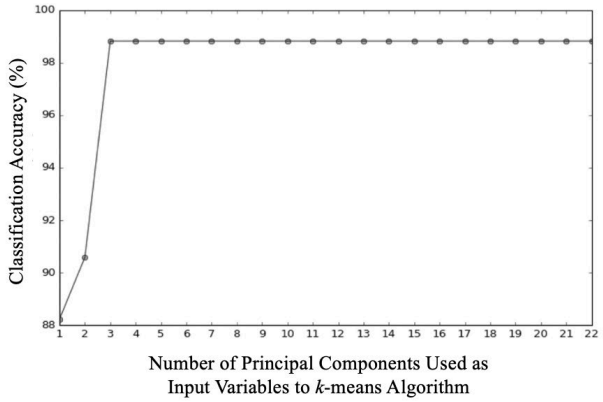

Figure 4.5: Accuracy of clustering versus number of principal components used in the *k*-means clustering algorithm for the data in the present study. Figure has been modified from [23].

As mentioned in section 2.2.2, the accuracy of classifying data using *k*-means clustering is highly dependent on the representation of the data, that is, the choice and structure of the input variables [35]. For example, it is possible that clustering the PC score data would produce

different results than clustering the raw IR absorbance values. Data were clustered into three groups, using the 22 selected IR absorbance bands (Table 4.1) as input data. As shown in Figure 4.4b, we find that *k*-means clustering of the raw IR absorbance values results in three distinct clusters, as for *k*-means clustering of the PC score data shown in Figure 4.4a. Although one of the clusters resulting from analyzing the raw IR absorbance values corresponds primarily to data for pipe A, the other two clusters are mixtures of data for pipes B and C. This result is problematic since we know from PCA (Figure 4.2a) that there is more variance between data for different pipes than within the data for a single pipe. In contrast, *k*-means clustering of the PCA score data reliably groups the data for the pipes according to their formulation. This result demonstrates the advantage of using PCA as a data pre-processing technique for the *k*-means cluster analysis to achieve proper classification of pipe data.

4.1.3 Support Vector Machine (SVM) Classification

Given our success in using PCA and *k*-means clustering to accurately classify different pipe formulations, it is reasonable to expect that it is possible to generate predictive models, which can accurately classify new samples based on the presence of distinct features and patterns in the relative IR absorbance intensities. To evaluate this approach, we used the 22-peak maximum absorbance values corresponding to the IR bands as indicated in Table 4.1, and their known pipe formulation classifications, to build an SVM model. In the present study, a radial basis function (RBF) was used for the SVM kernel. The data items (IR spectra) were randomly divided into two subsets: one subset contained two-thirds of the data, and a second subset

contained the remaining one-third of the data. A predictive model was generated using the larger subset as training data, which allowed the definition of decision boundaries. The model was evaluated using the remaining data by categorizing each new data point relative to the decision boundaries. We found that our SVM model could be used to classify the formulation of additional pipe data based on IR absorbance data with 100% accuracy.

As an example of how the SVM model could be used to classify new data sets, we used the SVM model to classify 10 IR spectra collected from pipe with the same formulation as pipe A but with a larger, 4-mm wall thickness. In all cases, the SVM model correctly predicted the new samples to be pipe formulation A.

Since we found that it was possible to use an SVM model based on IR absorption data to classify new samples into the correct pipe formulation with very high accuracy, it should be possible to use SVM models for quality control during pipe production and to evaluate new PEX-a pipe formulations. However, we note that the accuracy of SVM models can be compromised when the amount of data used for training the model is limited. For example, if there is little variation in the data set between classes but a large amount of spread within each class, then it may be necessary to increase the amount of training data to ensure the model is trained on representative data. In this case, increasing the amount of data would allow the model to use a large number of outlier points in its determination of decision boundaries and will thus be more likely to select a decision boundary that can accurately classify the outliers for each class. Additionally, if the randomly chosen training data subset was partitioned unevenly, e.g.

consisting primarily of data from pipes A and B, then the SVM model would likely yield poor classifications of pipe C. As the size of the training data set is increased, the biasing of the model by randomly selecting training data becomes less likely, and the accuracy of the model will be increased. If an SVM model is to be used for quality control purposes than we advise using a larger data set than the one used in the present study to train the model. This data set should include data from pipes manufactured on different dates and should span a larger variety of manufacturing conditions to ensure that to model is trained on a representative data set.

4.2 Tracking Accelerated Aging of PEX-a Pipes by Applying PCA and Decision Tree Based Classification Techniques to IR Data Collected from Axial Slices

In the present study, PEX-a pipes were subjected to accelerated aging according to procedures outlined in section 3.1.3 to observe and compare the effect of external stresses on changes to the chemical composition of the pipe samples. PCA, Random Forest and Decision Tree analyses were performed on a data set consisting of 369 IR spectra collected from roughly equal numbers of axial slices collected from: virgin (unused) PEX-a pipe, PEX-a pipe that was exposed to air and Milli-Q water at 85°C for 3, 6, 8, 10, 14 and 21 days, and in-service (naturally aged) PEX-a pipe. All of the samples used in the present study are PEX-a pipes of formulation A (Table 3.1).

4.2.1 Tracking Spectral Regions of Interest (ROI) using Indices

The three index values outlined in section 3.3.5 (Crystallinity Index, Carbonyl Index and Carbonyl COG) were calculated from IR data collected on axial pipe slices, and they are plotted in Figure 4.6 as a function of radial depth. As discussed in section 3.3.5, the IR intensity on the high-wavenumber side of the carbonyl region can be attributed primarily to ester bonds from the additive species within the pipe. Therefore, we expect hydrolysis of this bond from exposure to Milli-Q water at elevated temperature to cause a decrease in the high-wavenumber IR intensity. If there is little or no increase in the low-wavenumber IR intensity, corresponding to bands that can be attributed to degradation products, we can expect that: 1) the Carbonyl COG will shift to lower values, and 2) the total area of the carbonyl band will decrease with exposure time. Figure 4.6 shows clear and distinct trends for the Carbonyl Index (b and e) and Carbonyl COG (c and f) values for both air- and Milli-Q-aged pipe samples. As expected, when pipes were exposed to Milli-Q water at high temperature, the total integrated area of the carbonyl band decreased dramatically with exposure time (Figure 4.6b). Furthermore, we observed a decrease in the Carbonyl COG with increasing exposure time to Milli-Q water at elevated temperature (Figure 4.6c). These results, when taken together, serve as evidence for additive leaching with exposure to Milli-Q water at elevated temperature. The decrease in the Carbonyl COG allows us to attribute the decrease in Carbonyl Index to the loss of ester carbonyl species that occur in the wavenumber range from 1775–1720 cm^{-1}, rather than an increase in the bands that occur in the wavenumber range from 1720–1675 cm^{-1} which arise primarily from products of oxidative degradation. This behaviour is mirrored in the data collected on the in-service pipe sample,

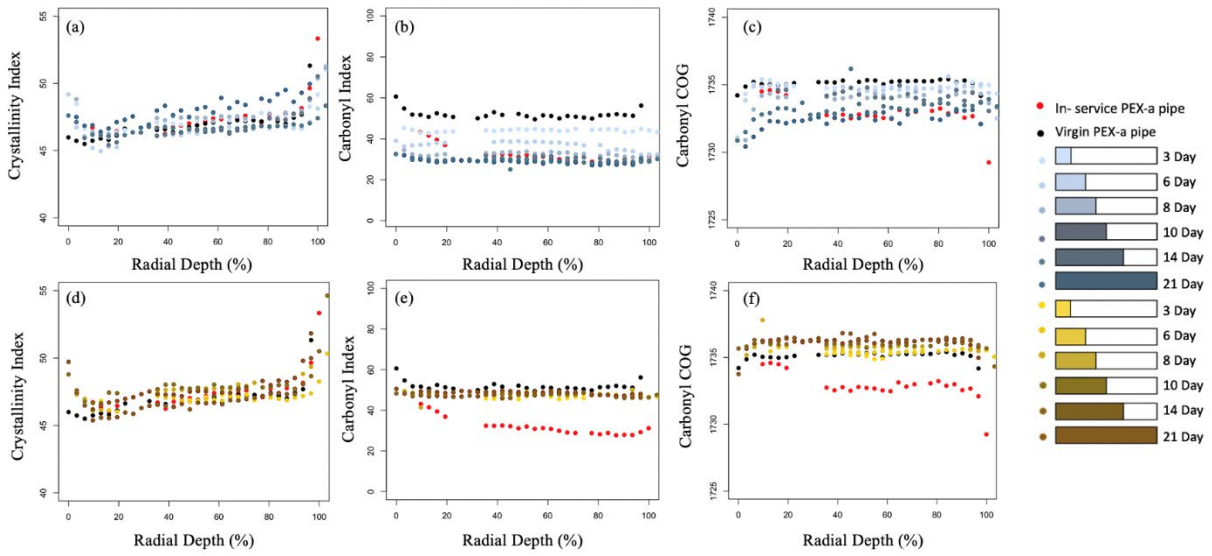

Figure 4.6: Calculated values of the Crystallinity Index, the Carbonyl Index and the Carbonyl COG for virgin PEX-a pipe (black), in-service PEX-a pipe (red), and pipe that was exposed to air (gold) and Milli-Q water (blue) at 85°C for 3, 6, 8, 10, 14 and 21 days. Plots a, b and c show the effect of aging in Milli-Q water whereas plots d, e and f show the effect of aging in air. Data for virgin and in-service pipe are also shown in each plot.

which further supports this reasoning. Furthermore, our results illustrate the importance of using both the overall intensity and shape of the carbonyl band to study the effects of aging on PEX-a pipe.

Conversely, when the pipe is exposed to air at elevated temperatures, we observed a slight decrease in the total carbonyl area after 3 days, followed by a small, gradual increase with exposure time from 3-21 days (Figure 4.6e). We also observed a small increase in the Carbonyl COG with exposure time from 3-21 days (Figure 4.6f). These results seem to suggest that aging of PEX-a pipe in air at elevated temperatures results in an initial decrease in all carbonyl species. However, the changes observed in the spectra in response to air aging are much smaller than

those observed in response to aging in Milli-Q water at the same temperature. Since we do not expect oxygen to diffuse significantly into the wall of the pipe, this result is not surprising, and it suggests that significant oxidation does not occur throughout the bulk of the pipes.

As can be seen in Figure 4.6a and d, we observed a large amount of spread within the Crystallinity Index values calculated for all axial pipe samples, with values ranging from 45% to 55% crystallinity. These results showed no clear trend with aging time; however, we did observe an increase in the values at the outer (0%) and inner (100%) surfaces for all pipe samples. Likewise, for all samples at 0% and 100% radial depth, we observed corresponding increases and decreases in the calculated Carbonyl Index and Carbonyl COG values, respectively. This indicates that there is likely different or more extreme degradation occurring in all samples at the inner and outer surfaces of the pipe wall. For this reason, we have chosen to focus our accelerated aging studies on data collected from the middle 50% of the pipe, to examine the effects of aging on the bulk properties of the pipes. A parallel study should be completed in which high resolution measurements at the outer and inner surfaces of the pipes are performed to assess the extent of oxidative degradation in these regions where oxygen can react.

This methodology ultimately allows us to evaluate the efficacy of our aging protocol, by enabling us to accurately measure important changes to the degree of crystallinity and specific chemical signatures of leaching and oxidative degradation throughout the pipe wall. The results shown in Figure 4.6 demonstrate the benefit of using specialized indices to measure the effect of aging on specific chemical species. For this reason, we have used a combination of normalized

peak maximum absorbance values and index values in our machine learning studies of axial slices of PEX-a pipe. Although it is also possible to use normalized intensities from multiple wavenumbers within an IR band to obtain the same information, we expect that band overlap would result in unwanted correlations between low and high wavenumber intensities that could affect the results of the PCA. Furthermore, we expect that band overlap would increase the noise associated with each of the selected wavenumbers within the band, limiting their usefulness in providing valuable classification information in the RF and DT analyses. We used the three indices and 7 peak maximum absorbance values listed in Table 4.2, collected from 25%-75% radial depths, in the machine learning studies to examine the effects of aging on PEX-a pipes of formulation A.

Table 4.2: Peak wavelengths of the absorbance bands used in the study of axial slices of PEX-a pipe using transmission FTIR-microscopy. The indices are defined in section 3.3.5.

Band Frequency	Source
865 cm^{-1}	Unassigned
908 cm^{-1}	PE vinyl unsaturation $CH = CH_2$ H out of plane bending
958 cm^{-1}	PE trans unsaturation $CH = CH$ H out of plane bending
991 cm^{-1}	PE terminal unsaturation
1160 cm^{-1}	Ester $C - O - C$ from additive
1234 cm^{-1}	Phenolic $C - OH$ from additive
1249 cm^{-1}	Unassigned
1305 cm^{-1}	Crystallinity Index as defined in section 3.3.5
1720 cm^{-1}	Carbonyl COG Index as defined in section 3.3.5
1740 cm^{-1}	Carbonyl Index as defined in section 3.3.5

4.2.2 Principal Component Analysis (PCA)

PCA was performed on a data set consisting of virgin pipes and pipes subjected to accelerated aging, using the IR absorption peak maximum values of 7 selected absorbance bands and 3 specialized indices (Table 4.2) as input variables. The total amount of variance explained by each PC is shown in the scree plot in Figure 4.7a. The results indicate that PC1 and PC2 account for a significant amount (79%) of the total variance in the data. We performed Horn's parallel analysis on the eigenvalues of the PCA, and this confirmed that it was appropriate to retain the first two PCs for further analysis (Figure 4.7b), since PC1 and PC2 are the only adjusted eigenvalues that are greater than one. The score values for PC1 and PC2 are shown in a 2D scatter plot in Figure 4.8, which shows distinct aging trajectories in PC-space for both air- and Milli-Q-aged pipe. To examine the time dependence of pipe degradation in PC-space, we have plotted the average PC1 and PC2 score values separately, for each pipe sample, as a function of aging time in Figure 4.9. The relative weights for each of the wavenumbers in the linear combinations for the first two PCs are shown in Figure 4.10, in which the IR band weights are labeled by color according to their (positive or negative) covariance within the PC.

By examining the most heavily weighted IR bands (and their physical origins, as described in Table 4.2) shown in Figure 4.10, together with two-dimensional plots of PC scores (Figure 4.8 and Figure 4.9), we have assigned a physical significance to both PC1 and PC2. In Figure 4.8 and Figure 4.9a, it can be seen that the PC1 score values increased with exposure time

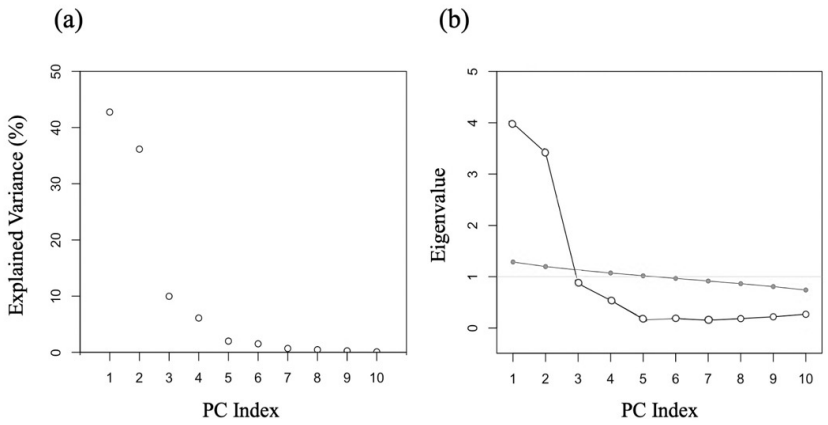

Figure 4.7: Eigenvalue analysis for the PCA on IR data collected from axial pipe slices. (a) Scree plot for PC1-PC10. (b) Adjusted PCA eigenvalues (open circles) calculated by scaling by the corresponding parallel eigenvalues (solid circles) plotted as a function of PC number. The adjusted PCA eigenvalues (open circles) that are greater than one correspond to the PCs that were retained for further analysis.

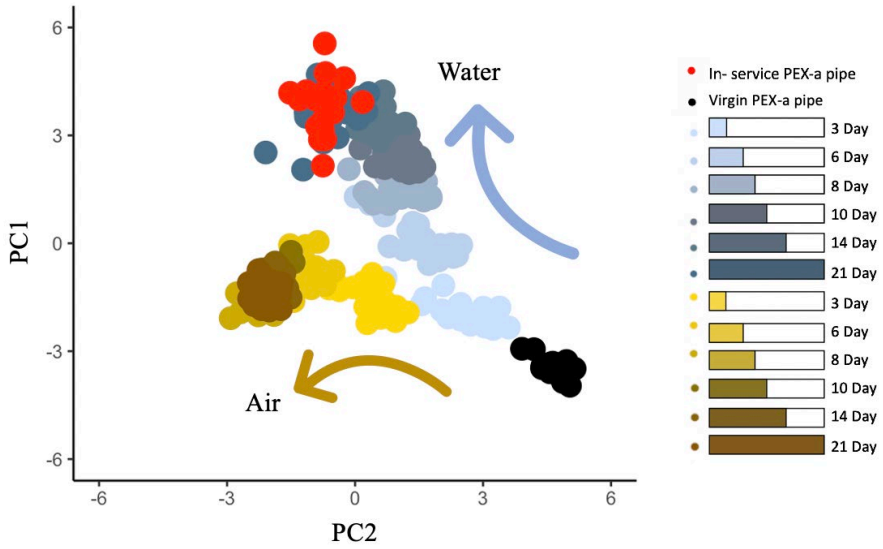

Figure 4.8: 2D plot of PC score values for PC1 and PC2 for aging in Milli-Q water (blues) and air (yellows and browns) at 85°C for different aging times. Data for virgin pipe (black) and in-service pipe (red) are also shown.

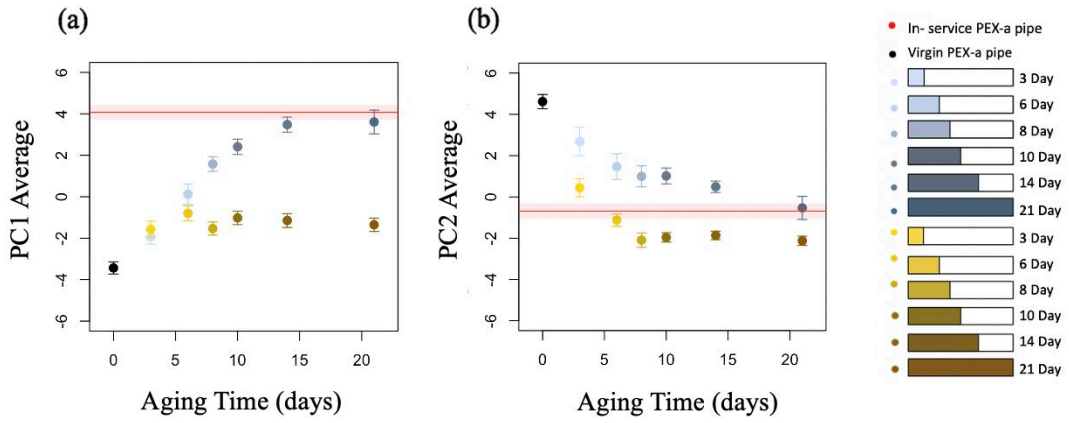

Figure 4.9: Average PC scores for aging in Milli-Q water (blues) and air (yellows and browns) at 85°C as a function of days aged for (a) PC1 and (b) PC2. Data for virgin pipe (black) and in-service pipe (red) are also shown.

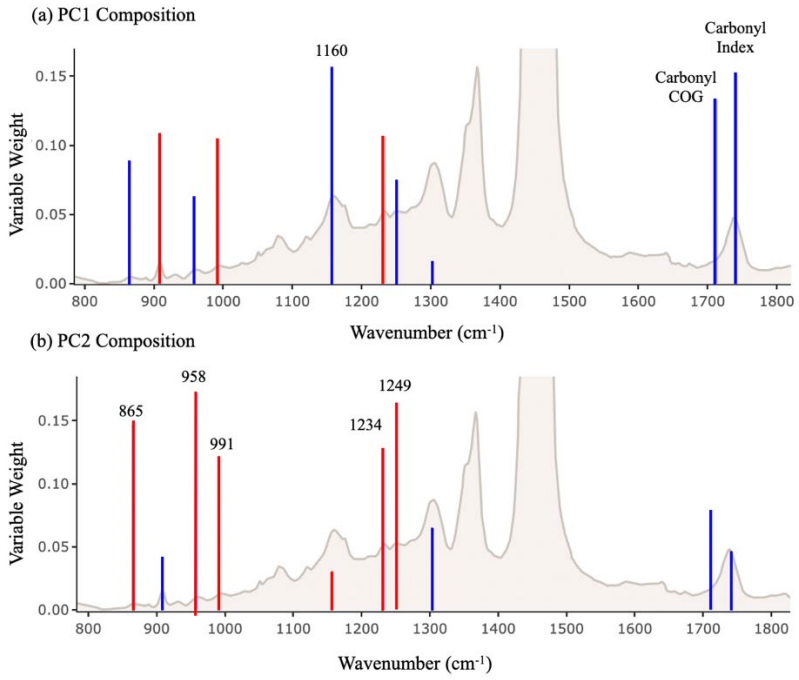

Figure 4.10: Relative weights of the IR bands in the linear combinations that define the PC1 and PC2 values for the PEX-a data set. High IR band weights are labeled according to their physical origin and colored based on negative (blue) and positive (red) correlations in the original data. Weights for the COG, crystallinity and carbonyl area index values are labelled and superimposed onto the spectra at 1720 cm^{-1}, 1740 cm^{-1} and 1305 cm^{-1} respectively.

for Milli-Q-aged samples, approaching the value measured for the in-service pipe at the longest aging times. In contrast, for the samples aged in air, a much smaller increase in the PC1 score values was observed. By examining the underlying PC1 relative band weights in Figure 4.10a, it is clear that the Carbonyl COG Index, the Carbonyl Index and the ester band at 1160 cm^{-1} dominate in their contributions to PC1. These bands can all, at least in part, be attributed to additive species within the pipe. Furthermore, all three variables have a negative covariance contribution to PC1, indicating that a decrease in these bands causes an increase in the PC1 score values. As outlined in the analysis of indices described in section 4.2.1, a decrease in these bands with exposure to Milli-Q water is indicative of additive leaching. We have therefore attributed the source of variance captured by PC1 as being the result of hydrolysis and leaching of additives within the pipe wall. We also note that PC1 (and the physical phenomenon that it describes,) accounts for ~ 44% of the total variance (Figure 4.7a), making it the most significant feature of the data set.

Given our interpretation of the physical origin of PC1, it is important for us to address the small increase observed in PC1 for air-aged pipe from 0 to 3 days, followed by a constant value for longer aging times. Although, we do not expect that additive leaching can occur during exposure to air at elevated temperatures, the PC1 results are consistent with our observation of a decrease in the ester carbonyl species in for air-aged samples between 0 and 3 days, as described in section 4.2.1. Since we observe small increases in the Carbonyl Index and Carbonyl COG Index from 3-21 days that are not mirrored by a decrease in the PC1 score values for these data, as discussed in section 4.2.1, we suggest that the plateau observed in PC1 in Figure 4.9a is the

result of a balance between negative contributions to PC1 from the highly-weighted bands and positive contributions to PC1 from the bands with low weights.

In Figure 4.8 and Figure 4.9b, we can also see that PC2 score values decrease with exposure time for both air- and Milli-Q-aged samples, reaching plateaus for both types of aging after ~ 8 days. The plateau in PC2 score values for Milli-Q aging is consistent with the value measured for the in-service pipe at the longest aging times, whereas the plateau for air aging occurs at a more negative value. This result will be studied further in section 4.2.3. Examining the underlying PC2 relative band weights in Figure 4.10b, it is clear that a mixture of bands arising from vinyl unsaturations and phenolic antioxidants dominate in their contributions to PC2 (Table 4.2). Additionally, we observed that all of these bands have positive correlations in PC2, which implies that the observed decrease in PC2 score values with aging time occurs because of decreases in the intensity of these bands. This result is not surprising, since we expect that aging of PEX-a pipes will result in a decrease in the intensity of the band at 1234 cm^{-1}, as additives perform their intended anti-oxidative function. Furthermore, vinyl unsaturations are highly reactive sites within the PE, and are also expected to decrease with aging as described in section 1.2.3. However, as discussed in the section 4.2.1, we do not expect oxygen to diffuse into the bulk of the pipe during the aging process, nor do we observe the signature of oxidative degradation within the pipes. It is, however, possible that these reactions are occurring due to a small amount of trapped oxygen or radical species that were incorporated in the PE matrix during the manufacturing process. For these reasons, we believe that PC2 represents the variance

that arises from small changes, induced by elevated temperature, to a number of relatively reactive species within the PE and additives.

As mentioned above, the 21-day Milli-Q-aged PC1 and PC2 score values coincide quite closely with those corresponding to the in-service pipe data. This suggests that the bulk changes to PEX-a that are observed in real-world installations are likely due to exposure to water at elevated temperatures. We note that, since the PCs are measures of the variance that exists within the data set, there is a limit to the types of conclusions we can draw from the results of the PCA. Since PC1 and PC2 are defined specifically by the variance that exists between air-aged, Milli-Q-aged and virgin pipe data, we cannot expect a geometric projection of in-service pipe to reveal any additional variation that has not already been captured by the PCs. Rather, we should interpret these results as evidence that the in-service and 21-day Milli-Q-aged samples differ in the same way from the virgin and air-aged pipe samples, and they should not be taken to mean that the two samples do not vary in any way with respect to each other. In other words, although we might expect some variance to exist between the in-service and 21-day Milli-Q-aged samples, the PCs used in the present study are not defined in such a way that allows this variance to be observed. The analysis performed in the present study only allows us to conclude that exposure to high temperature water is necessary for accelerated aging of PEX-a pipes but may not be sufficient. This will be explored further in the RF and DT analyses outlined in sections 4.2.3 and 4.2.4.

4.2.3 Analysis of Random Forest (RF) Classification

The application of PCA to IR data has allowed us to observe direct and unique aging trajectories in PC-space for PEX-a pipes aged in different ways (Figure 4.8). Furthermore, an analysis of the variables that contribute most significantly to each PC revealed the IR bands that are most relevant to the pipe degradation process. However, this analysis was limited since PCA only allows us to identify groups of IR bands which covary; PCA is not capable of providing detailed information about the exact cause of the separation of data points in PC-space. In section 4.2.4 we expand on this analysis, using Decision Tree (DT)-based methods to identify the specific IR bands that are responsible for the movement along the trajectories during accelerated aging. As outlined in section 2.2.4, the DT algorithm is a low-bias, high-variance technique and thus suffers from high variability with small changes to the training data set. To offset some of this bias, we have performed a Random Forest (RF) classification analysis. This preliminary classification allows us to identify, using ensemble methods, variables that frequently produce high accuracy trees. These variables can then be used to build an optimized DT model.

RF analysis was performed on a data set consisting of virgin, lab-aged and in-service pipe samples, using the IR absorption peak maximum values of 7 selected absorbance bands and 3 specialized indices (Table 4.2) as input variables. Methods outlined in section 2.2.5 were used to select values for the two user-specified parameters: (a) the number of variables q used in each tree, and (b) the number of trees t used to build the RF model. As shown in Figure 4.11a and b, the optimized values of q and t were found to be 3 and 2200, respectively. Using these values,

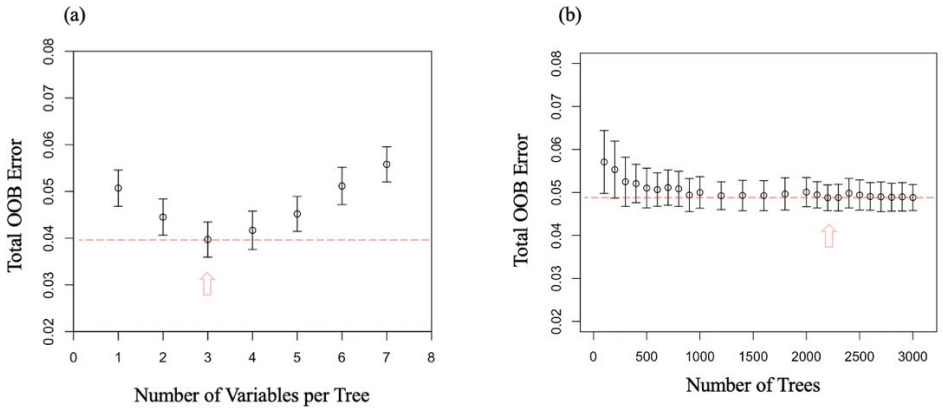

Figure 4.11: Average total OOB error rate over for 100 iterations of the RF algorithm as a function of (a) the number of variables per tree and (b) the number of trees. The horizontal dotted lines indicate the minimum value of average total OOB error, and the red arrows indicate the parameter values they occur at.

the RF model has a relatively small average total OOB error of 4.9%. In Figure 4.12, we show two multi-way plots comparing, (a) the mean decrease in accuracy and the mean minimum depth, and (b) the Gini index and the mean minimum depth, with the points sorted according to their p-value.

Each of the four parameters shown in Figure 4.12 provides a unique measure for the importance of, and amount of information contained within, a given variable in the RF model. In Figure 4.12, it is clear that the Carbonyl Index, the Carbonyl COG Index, the 958 cm^{-1} peak maximum, the 1160 cm^{-1} peak maximum and the Crystallinity Index are all statistically significant, with p-values less than 0.05. In addition, the Carbonyl Index, the Carbonyl COG Index, the 958 cm^{-1} peak maximum and the 1160 cm^{-1} peak maximum values have the highest

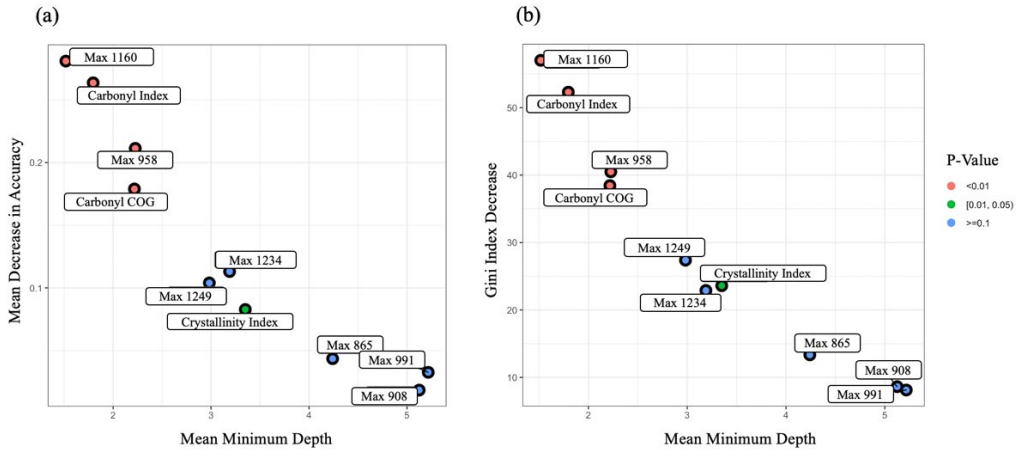

Figure 4.12: Multi-way plots showing four independent measures of the importance of each of the variables used in the RF model. The plots compare (a) the mean decrease in accuracy versus mean minimum depth with the points sorted according to their $\boldsymbol{p}$-value, and (b) Gini index versus mean minimum depth with the points sorted according to their $\boldsymbol{p}$-value.

mean decrease in accuracy, largest Gini index decrease, and smallest mean minimum depth. The high mean decrease in accuracy suggests that splits made according to these four variables, on average, produced trees with high accuracy, since removing (or 'noising') them increased the number of misclassified OOB observations. The high Gini index decrease for these four variables indicates that splits performed frequently result in pure (leaf) nodes. Finally, the low mean minimum depth indicates that splitting of the root node according to any of the four variables resulted in daughter branches that reach leaf nodes with few additional splits. From this it is apparent that the Gini index and mean minimum depth both measure the extent to which specific variables contribute to the production of short trees. Shortness is seen as a desirable quality in a decision tree (DT) model, since these trees tend to sort previously unseen data better [46]. Although the peak maximum values from the 1234 cm^{-1} and 1249 cm^{-1} bands have p-

values > 0.05, indicating low statistical significance, we observed that based on the other 3 statistics (mean decrease in accuracy, Gini index and mean minimum depth), these variables were important. For this reason, we have decided to remove only the three, least important variables for our DT classification analysis: the peak maximum values of the bands at 991 cm^{-1}, 908 cm^{-1} and 865 cm^{-1}. In addition, we examined the effect of removing these three bands on the PCA.

Based on the results of our RF classification analysis, we performed a second PCA on data collected from virgin and aged pipe samples, using the top 7 of the 10 most important variables[5] from the RF analysis as input variables. As in our previous PCA (section 4.2.2), data from the in-service pipe was geometrically projected onto the PC-space. To compare the results of this PCA to those obtained in section 4.2.2, the average PC1 and PC2 score values for all samples as a function of aging time are shown in Figure 4.13 for both the 7 variable PCA (plots a and b) and the 10 variable PCA presented in section 4.2.2 (plots c and d). The total amount of variance explained by each PC and the relative weights for each of the wavenumbers in the linear combinations for the first two PCs are provided in the Appendix in Figure A1 and Figure A2, respectively. We note that the removal of the three bands in the 7-variable PCA did not significantly change the distribution of explained variance or relative variable weights/correlations so that the physical interpretations given in section 4.2.2 are still valid.

[5] Importance was determined by using a holistic analysis of the 4 information measures shown in Figure 4.12.

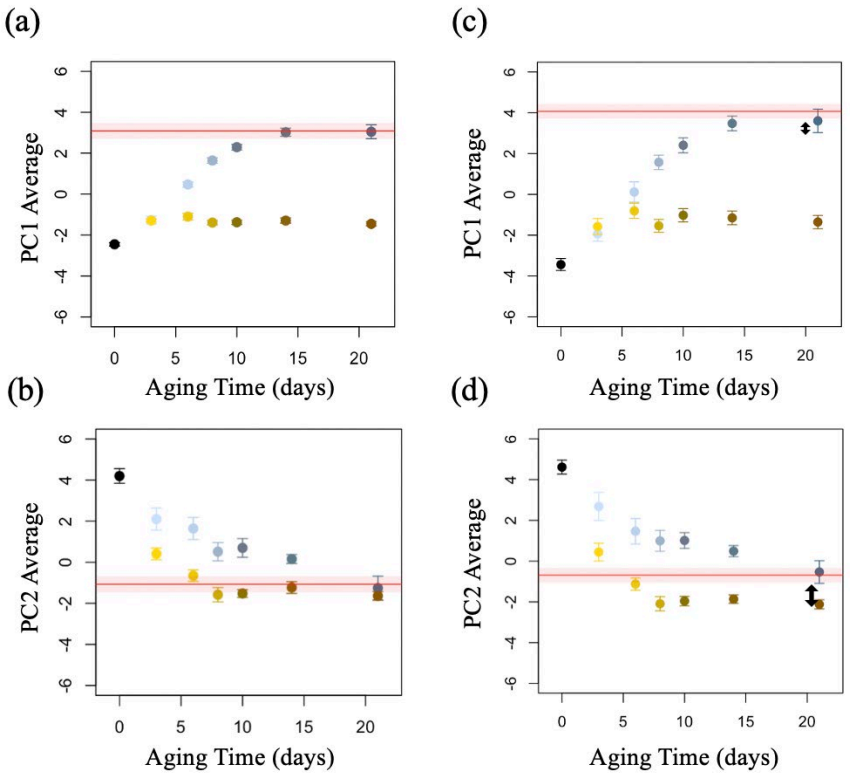

Figure 4.13: PCA comparison using (a, b) only the top 7 most important variables (determined by random forest classification analysis) and (c, d) all 10 variables (selected for use in section 4.2.2) as input variables. The thick vertical arrows indicate the primary change to the original PCs that occurs when the 3 least important variables are removed from the PCA analysis.

In comparing the average PC1 score values for the two PCAs, we can see that removing the three bands in the 7-variable PCA resulted in very little change to the shapes of the PC1 aging trajectories (PC1 average score values). Interestingly, we do see a reduction in the standard deviation for each sample; the three bands removed in the 7-variable PCA appear to account only for the intra-group spread observed in the original (10-variable PCA) PC1 values. Therefore, it seems that removing these bands allows us to access a clearer and more direct PC1 aging pathway. In contrast to this, the removal of the three bands caused a significant change to the

aging trajectories in PC2. We observed that removing the three bands caused the plateau of PC2 score values for both air- and Milli-Q-aged samples to converge (Figure 4.13b), resulting in both 21-day values to overlap with that of the in-service pipe. Based on these results, we are confident that the 865 cm^{-1}, 908 cm^{-1} and 991 cm^{-1} bands are primarily responsible for the gap between the observed PC2 score values for the air- and Milli-Q-aged data[6] and that this feature of PC2 has little statistical significance. The convergence suggests that the reactions that are described by PC2 occur to the same extent for pipes are aged in air or water, and that they occur relatively quickly with exposure to high temperature. Overall, these results suggest that elevated temperature is the primary driver for the spectral changes described by PC2 and that the choice of aging medium (Milli-Q or air) used does not play an important role in these reactions. This is clearly not the case for PC1, which describes sources of variance that are almost entirely driven by the aging medium.

To test the prediction accuracy of the RF technique, another RF model was generated by dividing the data into a training set, consisting of virgin and aged pipe data, and a testing set, consisting of data from in-service pipes. This model had a total average OOB error of 4.7%, which is comparable to the RF model generated using the full dataset. When the new RF model was used to predict the class of the data in the testing set, we observed that 100% of the test observations were classified as 21-day Milli-Q-aged pipe, which further supports the results of the PCA outlined in section 4.2.2.

[6] Indicated by the thick vertical arrow in Figure 4.13d.

4.2.4 Analysis of Decision Tree (DT) Classification

Based on the results from our RF classification analysis described in section 4.2.3, we performed a decision tree (DT) classification of data collected from virgin and aged pipe samples, using the top 7 of 10 most important variables from the RF analysis as input variables. The DT model was trained using the entire data set consisting of virgin, lab-aged and in-service pipe samples. Cost-complexity pruning was performed to determine an optimal control parameter (cp) value for pruning the tree. The set of cp values used in this analysis and their resulting relative are shown in Figure 4.14. CRAN documentation recommends selecting the maximum cp value for which the average relative total impurity falls within one standard deviation of the minimum average relative total impurity, indicated by the horizontal dashed line in Figure 4.14 [47]. In Figure 4.15 we have shown a simplified flowchart which includes only the root node and first internal node splits for the DT model. A cp value of 0.028 was selected and a flowchart of the resulting fully grown and pruned DT model is shown in Figure 4.16.

Examining the flowchart shown in Figure 4.16, we can see that although two of the variables included in the analysis (1234 cm^{-1} and 1249 cm^{-1}) that had low statistical significance (high p-values) were not used to split nodes within the tree. Additionally, many of the splits selected for the Milli-Q-aged data are consistent with results obtained in the PCA. For example, the DT algorithm selected the ester band at 1160 cm^{-1} for splits of Milli-Q-aging data between 6, 8, 10 and 14 and 21 days. It is clear in Figure 4.16 that the threshold for the ester band decreases as the algorithm separates progressively more aged pipes. Interestingly, this result indicates that

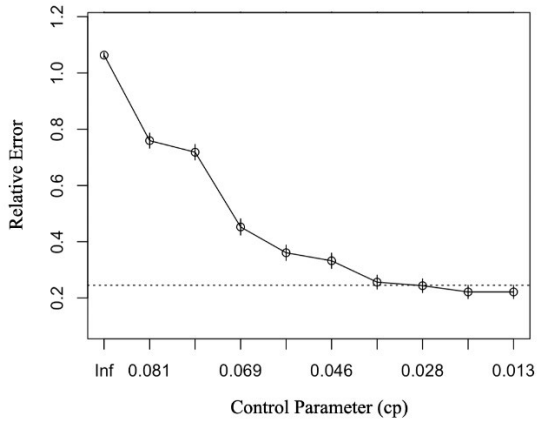

Figure 4.14: 10-fold cross validation results for the cost-complexity pruning of our DT model. The optimal complexity parameter (cp) was selected to be 0.028. CRAN documentation recommends selecting the maximum cp value for which the average relative total impurity falls within one standard deviation of the minimum average relative total impurity, indicated by the horizontal dashed line [47].

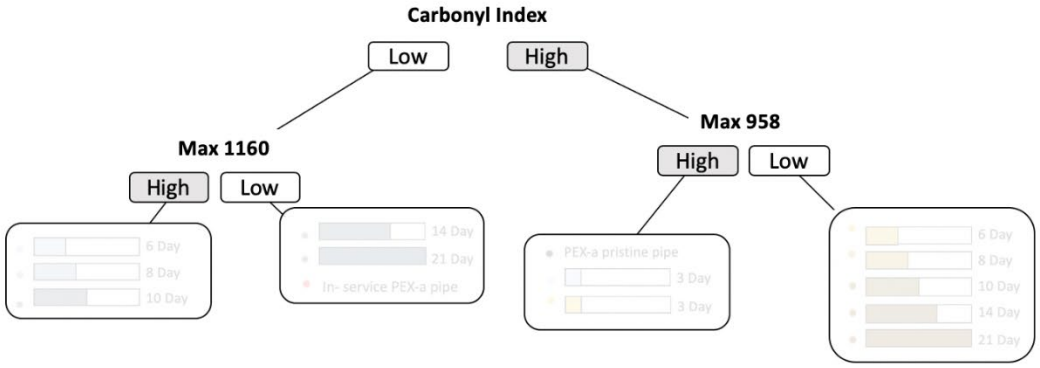

Figure 4.15: Root node split and first right and left internal node splits for the decision tree model. The exact threshold values have been replaced with labels indicating the IR band selected for the split, as well as ‘high’ and ‘low’ labels indicating the side of the threshold for each daughter node is on.

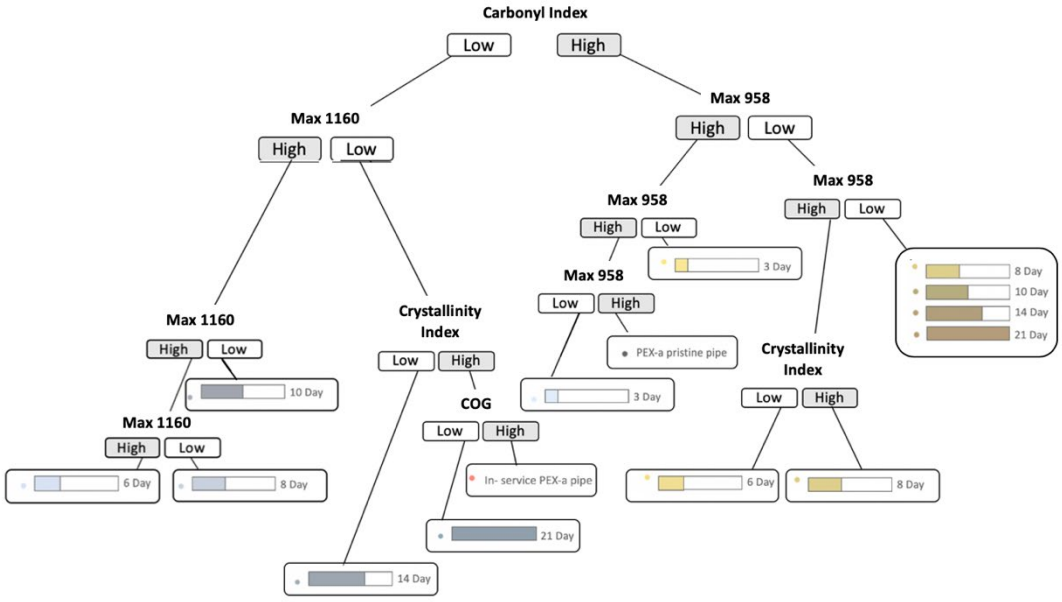

Figure 4.16: Decision tree model pruned using a complexity parameter (cp) of 0.028. The exact threshold values have been replaced with labels indicating the IR band selected for the split, as well as 'high' and 'low' labels indicating the side of the threshold for each daughter node is on.

the DT algorithm, like the PCA, has identified hydrolysis and leaching of the additives as being the most prominent feature in the Milli-Q-aged data set. Additionally, it seems that, after 14 days, the hydrolysis and leaching of the additives level off and the DT algorithm is forced to use less informative bands such as the Crystallinity Index and Carbonyl COG Index (Figure 4.12) to split the data. This result is consistent with the plateau observed in PC1 and PC2 for Milli-Q-aged data. Finally, we found that the DT algorithm grouped the data from the in-service pipe closest to the 21-day Milli-Q-aged data, as expected based on the results of the PCA and RF analysis.

Examining some of the other branches in the DT, we observed that the 3-day Milli-Q- and air-aged pipes were first grouped with the virgin pipes, as shown in Figure 4.15. Since the 3-day Milli-Q- and air-aged pipes have comparable score values for both PC1 and PC2 (Figure 4.13), this result is reasonable. It suggests that the effect of additive leaching becomes significant only after 3 days of accelerated aging. We also observed that the band at 958 cm^{-1} was often selected to differentiate between samples. This result that is not surprising since this band has high weighting in PC2 (Figure 4.10b), and it suggests that, before additive leaching begins, this band provides the largest source of variance within the IR spectra and is a result of exposure to high temperatures. We can therefore use our DT analysis to reinforce the conclusions drawn in section 4.2.2: reactions involving trans vinyl unsaturations within the PE occur relatively quickly with exposure to high temperatures and that the particular aging medium (Milli-Q versus air) used does not play an important role in these reactions.

The prediction accuracy and generalizability of the DT model were tested using a data set of 169, previously unseen Milli-Q- and air-aged observations. The model was capable of correctly classifying 155 of the 169 observations, corresponding to an overall prediction accuracy of 91.7%. Additionally, we observed that 11 of the 14 misclassified observations in the testing data were from 21-day Milli-Q-aged samples, and that they were all misclassified as in-service PEX-a pipe. By removing these samples from the data set, the DT model achieved an accuracy of 98.0% (148 of 151) for the remaining (non 21-day Milli-Q-aged) data. This 61% error rate for Milli-Q-aged and in-service pipe samples is not unreasonable, given the results of the PCA and RF classification analysis which also showed that the two samples are almost

indistinguishable. This result is actually quite informative; the inability of the DT model to distinguish between in-service and the long-time Milli-Q-aged samples suggests that there are no significant differences between these spectra. This implies (1) that high temperature Milli-Q water is sufficient for accelerated aging of the bulk of PEX-a pipes to simulate in-service use of the pipes, and (2) that the presence of chlorine during in-service use does not have a significant impact on the bulk of PEX-a pipes.

5 Summary and Future Work

5.1 Summary of Findings

In the present study, we have developed a methodology for preparing and measuring crosslinked polyethylene (PEX-a) pipes using transmission FTIR spectroscopy, and we have analyzed the results for three different pipe formulations using PCA and machine learning techniques. The application of PCA to the data set revealed that a large percentage (89%) of the total variance in the data set could be explained by the first three PCs (PC1-PC3). By examining the contribution of the individual IR bands to the PCs, we determined that PC1, which described the largest variance in the data, represented differences due to the choice of the peroxide crosslinker, whereas PC3 could be attributed to differences in the choice of additives. PCA enabled the visualization of differences between the different pipe formulations, with values of the three PCs resulting in distinct clustering of the data for each formulation. We confirmed the classification of the three pipe formulations obtained using PCA by applying two machine learning methods, k-means clustering and SVMs. Using the PCA results as input, the predictive models generated via k-means clustering and SVM classified the three different pipe formulations to within 98.8% and 100% accuracy, respectively. Our approach highlights the advantages of using machine learning techniques to characterize PEX-a pipes manufactured using new formulations and different processing conditions. This will allow pipe manufacturers

to achieve a detailed understanding of the pipe formulation and manufacturing process, and ultimately optimize pipe performance and durability.

To observe and compare the effect of external stresses on changes to the chemical composition of the pipe samples, we have developed a methodology to evaluate the uniformity and aging properties of PEX-a pipe using IR microscopy, and we have analyzed the results for pipes exposed to both air and water at elevated temperatures in using PCA, Random Forest (RF) and Decision Tree (DT) based learning techniques. The results of the PCA revealed that the aging behaviour of the pipes could be represented using only the first two PCs (PC1 and PC2), which account for a significant percentage (79%) of the total variance in the data. Visualization of these results revealed distinct aging trajectories in PC-space for both air- and Milli-Q-aged pipe. By examining the contribution of the individual IR bands to the PCs, we determined that PC1, which is responsible for moving the Milli-Q-aged pipes along their trajectory, represents differences that arise primarily due to hydrolysis and leaching of additives within the pipe wall. Furthermore, we have attributed the source of variance captured by PC2, with smaller changes induced by elevated temperature, to a number of relatively reactive species within the PE and additives. Furthermore, results of the PCA, RF and DT analysis revealed that pipes aged for 21 days in Milli-Q water were nearly identical to pipe samples that were used in real-world installations. These results suggest that additive leaching and hydrolysis plays a significant role in the aging of bulk PEX-a pipes, and that high temperature water is both necessary and sufficient to replicate the in-service conditions of bulk PEX-a pipes. Furthermore, signs of oxidative degradation were not observed within the bulk of the pipe. We therefore have

attributed any other changes that were observed in non-additive IR bands to reactions that occur relatively quickly and are likely driven by high temperatures rather than by diffusion of oxygen within the pipe walls. Overall, these studies have allowed us to identify and track characteristic signatures of aging and have provided insight into the IR bands that are relevant to the degradation process.

5.2 Future Work

The results outlined in section 4.2.1 indicate that there is likely different or more extreme degradation occurring in all samples at the inner and outer surfaces of the pipe wall. Future work should be done in which high-resolution measurements at the outer and inner surfaces of the pipes are performed to examine the effects of aging and the extent of oxidative degradation in those regions where oxygen can react. A parallel study using PCA and DT based techniques on this data would allow us to determine whether, at the surfaces of the pipes, oxidation is occurring more readily with exposure to air, where oxygen is readily available, or in water, where additives can be leached and depleted. This study also has the potential to provide insight into the IR bands that distinguish the different degradation mechanisms that are occurring in the two mediums.

Future accelerated aging experiments performed on these pipes will also make use of a novel, high-pressure recirculating system, which will allow water-based accelerated aging experiments to be performed quicker and with greater precision. This development will facilitate

experiments wherein temperature, pressure and chlorine concentration are each, systematically varied to help us to understand the effects of each of these stresses on the outer surfaces of the pipes. This work will work towards our goal of developing a time-temperature superposition that can be used in the future to assess the relative age of pipes used in real world installations. These experiments will be done on all three pipe formulations discussed in section 4.1, to assess the impact of the peroxide crosslinker and additive package on the pipe's resistance to degradation.

Finally, accelerated aging will be performed on virgin pipes as well as pipes with manually induced scratches/cracks. These experiments could help us to understand (a) mechanisms for crack growth and (b) whether the condition of the bulk of the pipe has any impact on the mechanical strength of the pipes, i.e. whether it contributes to the likelihood of bursting of the pipe. We expect that some of the changes highlighted in the present study do not have a detrimental impact on the mechanical properties of the pipes. Aging pipes to failure in our recirculating system will help us work towards making a more direct connection between the observed spectroscopic aging behavior and changes to the physical properties of PEX pipe.

REFERENCES

[1] T. Ojeda, A. Freitas, K. Birck, E. Dalmolin, R. Jacques, F. Bento and F. Camargo, "Degradability of linear polyolefins under natural weathering," *Polymer Degradation and Stability,* vol. 96, no. 4, pp. 703-707, 2011.

[2] H. A. Khonakdar, J. Morshedian, U. Wagenknechr and S. H. Jafari, "An investigation of chemical crosslinking effect on propertiesof high-density polyethylene," *Polymer,* vol. 44, no. 15, p. 4301, 2003.

[3] K. Thörnblom, M. Palmlöf and T. Hjertberg, "The extractability of phenolic antioxidants into water and organic solvents from polyethylene pipe materials - Part I," *Polymer Degradation and Stability,* vol. 96, no. 10, pp. 1751-1760, 2011.

[4] S. Al-Malaika, C. Goodwin, S. Issenhuth and D. Burdick, "Perspectives in Stabilization of Polyolefins," *Polymer Degradation and Stability,* vol. 64, no. 1, p. 145, 1999.

[5] L. C. Mendes, E. S. Rufino, F. O. de Paula and A. C. Torres, "Mechanical, thermal and microstructure evaluation of HDPE after weathering in Rio de Janeiro City," *Polymer Degradation and Stability,* vol. 79, pp. 371-383, 2003.

[6] M. Hakkarainen and A.-C. Albertsson, "Environmental Degradation of Polyethylene," *Advanced Polymer Science,* vol. 69, pp. 177-199, 2004.

[7] R. Maria, *Monitoring the degradation of PE pipes by IR-microscopy,* Aus Lissabon, Portugal: Technical University of Darmstadt, 2014.

[8] M. Hakkarainen and A.-C. Albertsson, "Long-Term Properties of Polyolefins," Berlin, Springer-Verlag, 2004, pp. 177-200.

[9] R. Maria, K. Rode, T. Schuster, G. Geertz, F. Malz, A. Sanoria , H. Oehler, R. Brüll, M. Wenzel, K. Engelsing, M. Bastian and E. , "Ageing study of different types of long-term pressure tested PE pipes by IR-microscopy," *Polym,* vol. 61, pp. 131-139, 2015.

[10] F. Gugumus, "Physico-chemical aspects of polyethlene processing in open mixers: Review of publihed work," *Polymer Degradation and Stability,* vol. 66, no. 2, pp. 161-172, 1999.

[11] R. Maria, K. Rode, R. Brüll, F. Dorbath, B. Baudrit, M. Bastian and E. Brendlé, "Monitoring the influence of different weathering conditions on polyethylene pipes by IR-microscopy," *Polym. Deg. Stabil.,* vol. 96, p. 1901, 2011.

[12] M. Scoponi, S. Cimmino and M. Kaci, "Photo-stabilisation mechanism under natural weathering and accelerated photo-oxidative conditions of LDPE films for agricultural applications," *Polymer,* vol. 41, pp. 7969-7980, 2000.

[13] I. T. Joliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Phil. Trans. R. Soc. A,* p. 374, 2015.

[14] A. G. Ryder, "Classification of narcotics in solid mixtures using principal component analysis and raman spectroscopy," *Journal of Forensic Science,* vol. 47, no. 2, 2002.

[15] N. S. Allen, L. M. Moore, G. P. Marchall, C. Vasiliou, J. Kotecha and B. Valange, "Diffusion and extractability characteristics of antioxidants in blue polyethylene water pipe: A DSC and radiolabelling study," *Polymer Degradation and Stability,* vol. 27, no. 2, p. 145, 1990.

[16] R. Spatafore and L. T. Pearson, "Migration and Blooming of Stabilizing Antioxidants in Polypropylene," *Polymer Engineering and Science,* vol. 31, no. 22, p. 1610, 1991.

[17] H. W. Siesler and K. Holland-Moritz, Infrared and raman spectroscopy of polymers, New York: M. Dekker, 1980.

[18] A. P. Review, "Nicolet 6700 FT-IR Spectrometer from Thermo Scientific," Compare Networks Inc, 2020. [Online]. Available: https://www.americanpharmaceuticalreview.com/25304-Pharmaceutical-Infrared-Spectroscopy-Equipment-Infrared-Spectrometers/5822268-Nicolet-6700-FT-IR-Spectrometer/. [Accessed July 2020].

[19] D. A. Skoog, F. J. Holler and T. A. Nieman, Principles of Instrumental Analysis 5th Edition, Orlando, Florida: Harcourt Brace & Company, 1998.

[20] T. Scientific, "Thermo Scientific Nicolet Continuμm Infrared Microscope," 2014. [Online]. Available: https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FCAD%2Fbrochures%2FBR51076-E-1214M-Continuum-L-1.pdf&title=Tmljb2xldCBDb250aW51Jm1pY3JvO20gSW5mcmFyZWQgTWljcm9zY29wZQ==. [Accessed 2020].

[21] S. Wold, K. Esbensen and P. Geladi, "Principal Conmponent Analysis," *Chemometrics and Intelligent Labratory Systems,* vol. 2, pp. 37-52, 1987.

[22] G. Y. Kanyongo, "Determining The Correct Number Of Components To Extract From A Principal Components Analysis: A Monte Carlo Study Of The Accuracy Of The Scree Plot," *Journal of Modern Applied Statistical Methods,* vol. 4, no. 1, p. 120, 2005.

[23] M. Hiles, M. Grosutti and J. R. Dutcher, "Classifying Formulations of Crosslinked Polyethylene Pipe by Applying Machine-Learning Concepts to Infrared Spectra," *Journal of Polymer Science, Part B: Polymer Physics,* vol. 57, pp. 1255-1262, 2019.

[24] T. Bouwmans, S. Javed, H. Zhang, Z. Lin and R. Otazo, "On the applications of robust PCA in image and video processing," *Proceedings of the IEEE,* vol. 206, no. 8, p. 1427, 2018.

[25] C. Clausen and H. Wechsler, "Color image compression using PCA and backpropagation learning," *Pattern Recognition,* vol. 33, p. 1555, 2000.

[26] H. H. Nieuwoudt, B. A. Prior, I. S. Pretorius, M. Manley and F. F. Bauer, "Principal component analysis applied to fourier transform infrared spectroscopy for the design of calibration sets for glycerol prediction models in wine and for the detection and classification of outlier samples," *Journal of Agricultural and Food Chemsitry,* vol. 52, p. 3726, 2004.

[27] E. Kaznowska, J. Depciuch, K. Szmuc and J. Cebulski, "Use of FTIR spectroscopy and PCA-LDC analysis to identify cancerous lesions within the human colon," *Journal of Pharmaceutical and Biomedical Analysis,* vol. 134, p. 259, 2017.

[28] K. Janne, J. Pettersen, N. O. Lindberg and T. Lundstedt, "Hierarchal principal component analysis and projection to latent structure technique on spectroscopic data as a data pretreatment for calibration," *Journal of Chemimetrics,* vol. 15, p. 203, 2001.

[29] J. Guicheteau, L. Argue, D. Emge, A. Hyre, M. Jacobson and S. Christesen, "Bacillus spore classification via surfaced-enhanced raman spectroscopy and principal component analysis," *Applied Spectroscopy,* vol. 62, no. 3, p. 267, 2008.

[30] U. Kruger, J. Zhang and L. Xie, Developments and applications of nonlinear principal component analysis - a review, Berlin: Springer, 2008.

[31] S. Ma and Y. Dai, "Principal component analysis based methods in bioinformatics studies," *Briefings in Bioinformatics,* vol. 12, no. 6, p. 714, 2011.

[32] S. T. Ikram and A. K. Cherukuri, "Improving accuracy of intrusion detection model using PCA and optimized SVM," *Journal of Computing and Information Technology,* vol. 24, no. 2, p. 133, 2016.

[33] X. Xu and X. Wang, "An adaptive network intrusion detection method mased on PCA and support vector machines," *Advanced Data Mining and Applications,* p. 696, 2005.

[34] M. Z. Nasution, O. S. Sitompul and M. Ramli, "PCA based feature reduction to improve the accuracy of decision tree c4.5 classification," *Journal of Physics: Conference Series,* vol. 978, 2018.

[35] A. K. Jain, "50 years beyond k-means clustering," *Pattern Recognition Letters,* vol. 31, no. 8, p. 651, 2010.

[36] Y. Liang, Q. S. Xu, H. D. Li and D. S. Cao, Support vector machines and their application in chemistry and biotechnology, Boca Raton: CRC PRess, 2011.

[37] I. O, "Applications of support vector machines in chemistry," *Reviews in Computational Chemsitry,* vol. 23, p. 291, 2007.

[38] T. S. Khoon, "Near-infrared raman spectroscopy with recursive partitioning techniques for precancer and cancer detection," National University of Singapore, 2009.

[39] T. Hothorn, K. Hornik and A. Zeileis, "Unbiased Recursive Partitioning: A Conditional Inference Framework," *Journal of Computational and Graphical Statistics,* vol. 15, no. 2, pp. 651-674, 2006.

[40] J. Mingers, "An empiracle comparison of selection measures for deciion-tree induction," *Machine Larning,* vol. 3, pp. 319-342, 1989.

[41] V. Svetnik, A. Liaw, C. Tong, J. Culberson, R. Sheridan and B. Feuston, "Random Forest: A Classification and Regresion Tool for Compound Classification and QSAP Modelling.," *Journal of Chem Inf Comput Sci,* vol. 43, no. 6, pp. 1947-1958, 2003.

[42] A. P. White and W. Z. Liu, "Bias in Information-Based Measures in Decision Tree Induction," *Machine Learning,* vol. 15, pp. 321-329, 1994.

[43] D. M. Bigg, M. M. Epstein, R. J. Fiorentino and E. G. Smith, "Coninuous extrusion of high-modulus semicrystalline polymers," *Polymer Engineering and Science,* vol. 21, no. 2, p. 76, 1981.

[44] K. S, "Infrared Spectra of High Polymers," *Fortschr. Hochpolym. -Forsch,* vol. 2, p. 51, 1960.

[45] R. E. Schapire, "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[46] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning,* vol. 1, pp. 81-106, 1986.

[47] T. Therneau, B. Atkinson and B. Ripley, "Package 'rpart'," 12 April 2019. [Online]. Available: https://cran.r-project.org/web/packages/rpart/rpart.pdf. [Accessed 8 August 2020].

[48] J. Peng, S. Peng, A. Jiang, J. Wei, C. Li and J. Tan, "Asymmetric Least Squares for Multiple Spectra Baseline Correction," *Analytica Chimica Acta,* pp. 63-68, 2010.

[49] K. H. Liland , B.-H. Mevi and R. Canteri , "Package 'baseline'," 11 05 2020. [Online]. Available: https://cran.r-project.org/web/packages/baseline/baseline.pdf.

[50] D. G. Lin and E. V. Vorob'eva, "Decrease in the Performance of a Phenolic Antioxidant in Preparation of Inhibited Polyethylene Films by Hot Pressing," *Russian Journal of Applied Chemistry,* vol. 90, no. 5, pp. 780-787, 2017.

[51] D. J. Hand and R. J. Till, "A simple generalization of the area under the ROC curve for multiple class classification problems," *Machine Learning,* vol. 45, pp. 171-186, 2001.

[52] L. Breiman and C. Chao, "Using Random Forest to Learn Imbalanced Data," University of California, Berkley, 2004.

[53] B. Bell, D. E. Beyer, N. L. Maecker, R. R. Papenfus and D. B. Priddy, "Permanence of polymer stabilizers in hostile environments," *Journal of Applied Polymer Science,* vol. 54, no. 11, p. 1605, 1994.

[54] K. Kerlsson, G. D. Smith and U. W. Gedde, "Molecular Structure, Morphology, and Antioxidant Consumption in Medium Density Polyethylene Pipes in Hot-Water Applications," *Polymer Engineering and Science,* vol. 32, no. 10, p. 659, 1992.

[55] A. Daffertshofer, C. J. Lamoth, O. G. Meijer and P. J. Beek, "PCA in Studying Coordination and Variability: A Tutorial," *Clinical Biomechanics,* vol. 19, no. 4, p. 415, 2004.

[56] P. Jiang-Qing and C. Song, "A study of the photolysis of a commercial hindered amine light stabilizer," *Polymer Degradation and Stability ,* vol. 40, pp. 375-378, 1993.

APPENDIX

The carbonyl index was calculated according to:

$$\text{Carbonyl Index} = \int_{1675}^{1775} \frac{\mathrm{I}_x}{\mathrm{I}_{2019}}\, dx\,, \tag{A1}$$

where I_x is the baselined absorbance intensity at wavenumber x.

The carbonyl COG was calculated according to:

$$\text{Carbonyl COG} = \frac{\sum_{i=1675}^{1775} i\left(\frac{\mathrm{I}_i}{\mathrm{I}_{2019}}\right)}{\mathrm{N}}, \tag{A2}$$

where I_i is the baselined absorbance intensity at wavenumber i, and N is the total number of values in the weighted average.

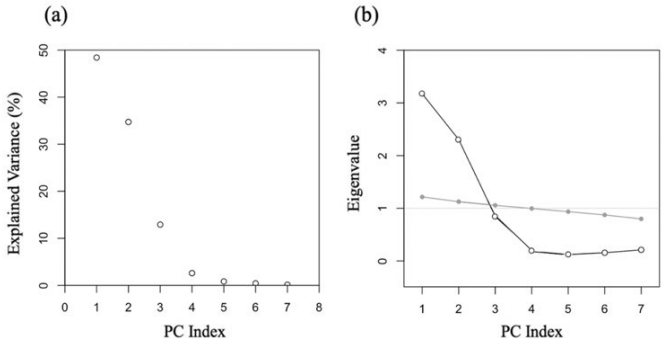

Figure A1: Eigenvalue analysis for the 7-variable PCA performed on IR data collected from axial pipe slices. (a) Scree plot for PC1-PC7. (b) Adjusted PCA eigenvalues (open circles) calculated by scaling by the corresponding parallel eigenvalues (solid circles) plotted as a function of PC number. The adjusted PCA eigenvalues (open circles) that are greater than one correspond to the PCs that should be retained for further analysis. It is clear that removal of the three bands in the 7-variable PCA did not significantly change the distribution of explained variance between the components.

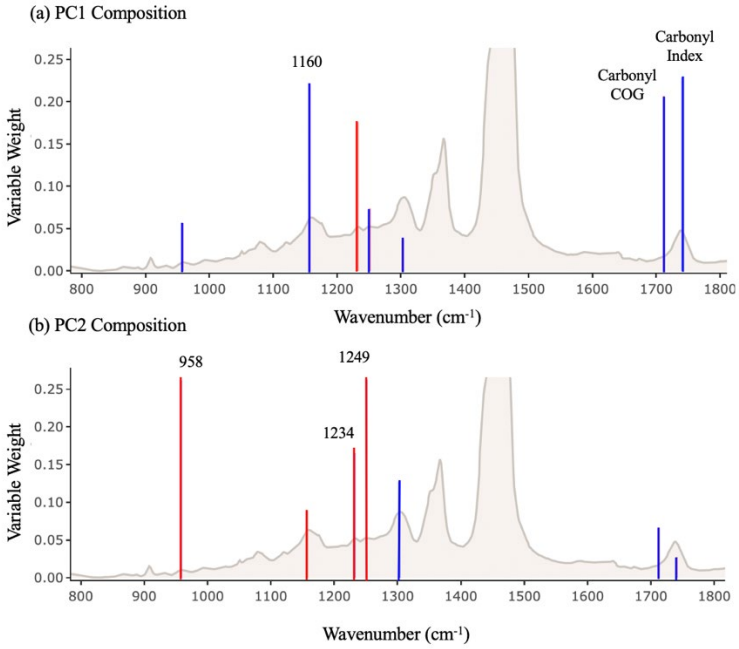

Figure A2: Relative weights of the IR bands in the linear combinations that define PC1 and PC2 in the 7-variable PCA. High IR band weights are labeled according to their physical origin and colored based on positive (blue) and negative (red) correlations in the original data. Weights for COG, crystallinity and carbonyl area index values are labelled and superimposed onto the spectra at 1720 cm^{-1}, 1740 cm^{-1} and 1305 cm^{-1} respectively. It is clear that removal of the three bands in this analysis did not significantly change the relative weights of, or correlations between the remaining bands.