Fourier Transform Infrared Spectra Clustering for Biochar: A Principal Component Analysis Approach

by

© Vikentiy Pashuk

A thesis submitted to the Memorial University of Newfoundland in partial fulfillment of the requirements for the degree of Bachelor of Science Honors.

Department of Physics and Physical Oceanography

Memorial University

April 2024

St. John's, Newfoundland and Labrador, Canada

Abstract

Biochar, recognized for its porous structure and functional groups, holds promise as a tool for mitigating greenhouse gas transmissions, particularly CO_2. This study acts as a precursor for future exploration of the efficacy of Principal Component Analysis (PCA) on Fourier Transform Infrared spectra for sample categorization for CO_2 adsorption. Utilizing RStudio, spectra from feedstock and biochar auger wood and snow crab samples were subjected to PCA. Results indicate that, in smaller sample systems, overall spectral intensity outweighs chemical differences in peak structure, while larger systems exhibit increased significance of peak structure due to comparable intensities. Future research should investigate the influence of experimental conditions, such as temperature and exposure time, on spectral intensity for conclusive PCA clustering. Although PCA effectively distinguishes spectral features in diverse samples, its applicability to larger systems with colinear features requires further exploration.

Acknowledgements

A land acknowledgement is offered to recognise Indigenous peoples' enduring connection to their traditional territories; to recognise the history of the land that is currently shared by many peoples, and to recognise stewardship as a shared commitment of all those who reside therein. The practice of territorial acknowledgement is itself a replication of an Aboriginal practice, predating European contact.

We acknowledge that the lands on which Memorial University's campuses are situated are in the traditional territories of diverse Indigenous groups, and we acknowledge with respect the diverse histories and cultures of the Beothuk, Mi'kmaq, Innu, and Inuit of this province.

I employed artificial intelligence over the course of the project. I used ChatGPT to familiarize myself with the new coding language of RStudio. ChatGPT was also used to brainstorm the layout and help structure the thesis, and DeepL was used for editorial work to correct simple grammar errors, spelling, and run-on sentences.

Statement of contribution

I extend my sincere appreciation to Jake Breen for his invaluable contributions to the preliminary research on biochar and PCA analysis, including his initial work on developing the R code for PCA computation. I am also grateful to Dr. Kelly Hawboldt of the Department of Chemistry at Memorial University and her group, particularly Nadyana Incan and Zahra Ghanbarpour Mamaghani, for providing the essential samples of crab and auger wood in raw and biochar forms. Furthermore, I thank Dr. Chen of the Department of Physics and Physical Oceanography for his support in keeping me on schedule and guiding me through the initial thesis process. Lastly, I acknowledge Dr. Tarasov of the Department of Physics and Physical Oceanography for his late contributions to the project's coding component and insightful feedback on this paper, which have been invaluable in shaping the research direction. A special thanks is extended to Dr. Shivani Semwal for her constructive criticisms of the paper and assistance in obtaining the samples.

I would like to extend a separate gesture of gratitude to my supervisor, Dr. Kris Poduska of the Department of Physics and Physical Oceanography, whose unwavering support and keenness were instrumental in the success of this project, despite the multiple leaves and a demanding schedule. I am particularly grateful for her generosity in granting me access to the Fourier Transform Infra-red Spectroscopy apparatus, allowing me to gather data at my convenience for analysis. Moreover, I commend her for sharing a vision of rigorous and collaborative academic research with me, which has greatly enriched our scholarly journey. It is truly thanks to her exceptional supervision and mentorship that this paper is complete and well-written. I once again would like to collectively thank every friend, student, and faculty member who played a role in facilitating my success.

Table of contents

List of abbreviations

BC	Biochar Crab
BAW	Biochar Auger Wood
CAWB	Crushed Auger Wood Biochar
FTIR	Fourier Transform Infra-red
HTCB	Hydrothermal Carbonization Crab Biochar
HTCBT	Hydrothermal Carbonization Crab Biochar Temperature
IR	Infrared
PCA	Principal Component Analysis
PCB	(Fast) Pyrolysis Crab Biochar
RAW	Raw Auger Wood Biomass
RDC	Raw Dried Crab Biomass
UAWB	Uncrushed Auger Wood Biochar

Chapter 1

Introduction

Effective carbon sequestration methods are imperative in combating climate change, with biochar emerging as a promising solution [15]. Derived from organic matter, biochar converts waste products into robust, graphite-like carbon, which, coupled with its exceptional adsorption qualities, garners significant attention for carbon sequestration.

Understanding biochar's complex physical and chemical properties is crucial for maximizing its efficacy. Fourier transform Infrared (FTIR) spectroscopy serves as a key analytical tool, offering insights into biochar's chemical structure by analyzing its composition and functional groups.

Despite the utility of FTIR spectroscopy, the application of statistical methods to interpret biochar spectra remains underexplored in the academic literature. Principal Component Analysis (PCA) presents an opportunity to address this gap, offering a systematic approach to analyzing biochar spectra and enhancing our understanding. Implementing PCA and evaluating its effectiveness in spectral clustering can help improve global efforts at climate change mitigation. In this thesis, I aim to bridge this research gap by establishing the foundation for future studies or research on biochar spectral analysis using PCA.

1.1 Biochar

Biochar, a graphite-related material resembling charcoal, is produced through a unique process to minimize contamination and securely store carbon [18]. Derived from organic biomass like agricultural waste or forestry byproducts, biochar serves diverse purposes, from mitigating greenhouse gas emissions to improving soil fertility. Other advantageous properties of biochar include its inexpensiveness and ease of production. Amidst the pressing challenges of climate change, biochar emerges as a promising solution. By effectively capturing and storing carbon dioxide from various sources, it offers a sustainable approach to combating atmospheric pollution. Understanding biochar production intricacies and its wide-ranging applications is crucial for addressing environmental concerns and promoting sustainability.

The production is determined by the biomass types, combustion conditions, residence times, temperatures, heating rates, and reactor type [15]. As mentioned earlier, biochar is produced in controllable conditions from the near vacuum combustion of organic biomass, referred to as feedstock, at temperatures not higher than 700 °C.

Here, I highlight a few summaries of recent papers related to biochar, aiming to provide the reader with an overview of the current state of biochar research. For example, slow pyrolysis operates at a temperature range of 300−700 °C and the increase in temperature is gradual at 10 °C per minute [19]. Fast pyrolysis is similar but has a higher rate of temperature increase, ranging from 10 °C to 200 °C per second. Flash carbonization is a process in which a flash fire is ignited under the bed of biomass at a raised pressure of 1 MPa for less than half an hour. Lastly, hydrothermal carbonization is performed in subcritical water heated to around 200 °C at pressures nearing 20 MPa.

Though the temperature of biochar production has been observed to modify functionalization groups, other essential properties that add to the sorption of biochar are the modifications that take place before and/or after the carbonization process. The most studied methods of modification are chemical, physical, impregnatin with mineral sorbent, and magnetic. Chemical modification is either a one- or a two-step process. The single-step process allows for chemical modifications at the time of carbonization. The two-step method requires mixing the product with the chemical agent before and after the carbonization. The treatment is done by adding an acid, a base, or an oxidizing agent, which modify surface functional groups and sometimes improve porosity [19].

Physical modifications are less effective than chemical modifications and refer to porous structure changes. The leading method is steam activation, which increases the surface area and improves the carbonaceous structure as the trapped gas escapes and creates micropores while enlarging the diameters of existing ones [19].

Gai *et al.* [10] shows the effectiveness of nitrogen (N) and ammonium (NH_3) adsorption of chemically treated biochar. They had twelve biochars produced from wheat straw, corn straw, and peanut shell at temperatures ranging from $400 - 700$ °C. The findings suggested that at lower temperatures, the content of N was larger as compared to higher temperatures, which favoured carbon. Corn-straw biochar was the most effective at ammonium adsorption and decreased as the temperature of pyrolysis increased. All the biochar samples failed to reduce NO_3 and, in some cases, released it.

Liu *et al.* [16] took biomass waste in the form of ground coffee to capture CO_2 for stationary sources such as power plants and steel factories. Chemical modifications such as ammoxidation and potassium hydroxide activatons were used. Multiple analytical techniques, such as electron microscopy, Fourier-transform infrared spectroscopy, and X-ray photoelectron spectroscopy, were used to evaluate the physiochemical properties. It was found that biochar, which was nitrogen-doped by melamine, followed by potassium hydroxide (KOH), developed microporosity and had more active adsorption sites, increasing the CO_2 uptake.

Zhang *et al.* [24] studied high-temperature fast pyrolysis of bio-oil mass for two treatments: a CO_2-ammonia mixture and a conventional CO_2 and ammonia mixture, which were chemically and structurally compared. It was found that the mixture greatly increased the surface area and increased the number of N-containing groups. A temperature dependence was also found, suggesting that the adsorption at low temperatures (20 °C) was proportional to micropore volume and at higher temperatures (120 °C) it correlated to N content.

Zhang *et al.* [6] highlighted biochar adsorption values of CO_2 for bamboo charcoal biochar. Adsorption can reach up to 2.35 millimoles per gram, which is equivalent to 103.4 grams of CO_2 for one kilogram of biochar. Given biochar's adsorption capabilities, its chemical composition and structure are analyzed using Fourier Transform

Infrared (FTIR) Spectroscopy.

1.2 Fourier Transform Infrared Spectroscopy

Fourier Transform Infrared (FTIR) spectroscopy can help determine the structures of solid, liquid, and gas samples due to the fact that the vibrations within these materials are at the same energy as IR light [1]. When exposed to infrared radiation, solids absorb radiation at specific wavelengths, resulting in a change in the dipole moment. This transition occurs from an initial vibrational excited state, which is readily populated even at room temperature due to thermal energy, to another vibrational excited state. The wavenumber of the absorption peak corresponds to the energy difference between these vibrational states. However, it's important to note that not all the vibrational modes exhibit a dipole moment, and thus, not all vibrations contribute to observable peaks in the IR spectrum. Therefore, the number of peaks observed is not directly correlated with the number of vibrational degrees of freedom of the solid. The intensity in FTIR spectroscopy is related to the change in dipole moment during excitation. In the context of FTIR, excitation refers to the process by which molecules absorb infrared radiation and undergo transitions between different vibrational energy states. Most solids exhibit infrared activity due to their molecular vibrations, except for those with highly symmetric crystal structures or perfectly arranged atoms, such as certain crystalline forms of carbon [2]. In this thesis, we will investigate mid-range Fourier Transform Infra-red spectra, which are characterized in the range of $4000-400$ cm^{-1} as biochar is solid graphite-like.

Spectroscopy can reveal information about surface adsorbed species, such as CO_2, as the C-O bond is infrared visible [9].

Fig. 1.1 shows the mussel shell infrared spectrum, which has four peaks that define unique vibrational modes. The chemical bonds to which the peaks correspond can easily be determined from tables of vibrational mode values. The study on mussel shell FTIR points to the fact that the largest peak appears at 1467 cm^{-1} [7]. They also found that the second largest peak at 1080 cm^{-1} in the study. An alternative study [17] concluded that all the peaks are features of aragonite, a mineral form of calcium carbonate ($CaCO_3$), as the peak at 1460 cm^{-1} corresponding to anti-symmetry stretching vibration ($\nu3$), the peak at 1084 cm^{-1} is the symmetric carbonate

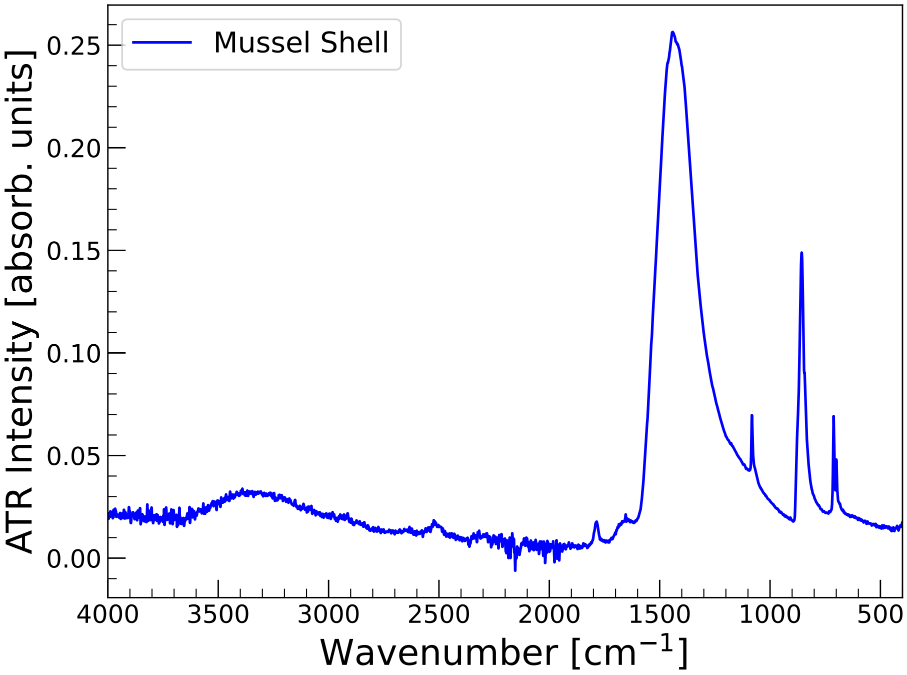

Figure 1.1: ATR intensity (absorbance units) vs wavenumber from 4000 – 400 cm^{-1} of Fourier Transform Infra-red Spectra obtained from a mussel shell on a Bruker Alpha II Spectrometer. The mussel shell spectrum is characterized by visible high intensity peaks which correspond to chemical bonds of different vibration modes.

stretching vibration ($\nu1$), and the peak at 854 cm^{-1} is the out-of-plane bending ($\nu2$) vibration. The pair of peaks at 700 cm^{-1} and 713 cm^{-1} are in-plane bending modes ($\nu4$). These peaks are all characteristic of aragonite.

FTIR spectroscopy is a significant tool in biochar analysis because it highlights the bonding features of the samples. Gurwinder Singh [20] carried out a chemical treatment on biochar and analyzed it using FTIR spectrography. The feedstock used was Arundo donax, a type of giant reed, activated with potassium hydroxide (KOH). Upon FTIR spectroscopy, aromatic C-C stretching modes, O-H and N-H stretching vibrations of carboxyl and amine groups were found.

However challenges arise from the heterogeneous nature of charcoal compared to biochar [22], lacking a well-defined chemical formula and often exhibiting poor crystallization. Consequently, the IR spectra of biochar tend to feature broad peaks that offer limited detail regarding its composition. To overcome this limitation and gain a clearer understanding of biochar properties, complementary analytical techniques are essential.

Techniques like scanning electron microscopy, X-ray diffraction, and photoelectron spectroscopy have been employed to understand the structure and elemental composition of biochar [23]. Scanning electron microscopy is employed to investigate the porosity and surface morphologies of both the feedstock and biochar, and x-ray diffraction is used to determine the crystalline structure. While photoelectron spectroscopy is used for distinguishing and measuring functional groups as well as for the elemental composition and O/C molar ratio, which corresponds to biochar stability, it allows for surface analysis [23]. These techniques collectively aid in identifying characteristic chemical bonds, porosity, crystalline structure, functional groups, and elemental composition, thereby allowing for differentiation based on various structural features of biochar.

Despite these methods serving as a means of compound structure analysis of biochar [21] they will not be investigated in this paper. Instead, Fourier-transform Infrared (FTIR) spectroscopy is the primary tool looked at for spectral chemical analysis and principal component analysis.

1.3 Principal Component Analysis

Principal component analysis (PCA) represents a versatile statistical technique employed in reducing the complexity of multivariate datasets. It achieves this by amalgamating interrelated variables into composite variables termed principal components. These principal components constitute linear combinations of the initial variables, designed to capture the maximum variance across all variables within the dataset [11]. This method can help make data more interpretable by creating new uncorrelated variables that arise from eigenvalue and eigenvector analysis [14]. However, PCA is ineffective when handling highly non-linear relationships. In the context of analyzing biochar spectra, PCA holds significant promise, as it can provide greater analytical depth by distinguishing subtle variations in spectral features, enabling the differentiation of biochar samples based on their chemical composition and structural characteristics. If PCA is successful, it could be used for prediction and observation of CO_2 adsorption qualities, allowing researchers to infer the potential carbon sequestration capacity and environmental impact of biochar samples.

Though not all studies discussed below are directly related to PCA of spectra, it's important to emphasize that they serve as examples demonstrating the versatility of the PCA approach in sorting and deriving physical insights from data.

Similarly, PCA was used in a study done by Amy Iezzoni and Marvin Pritts [13] in horticultural research. Due to the large number of traits, classical statistical analysis is inefficient, but due to the correlation of the many variables, the PCA method to reduce the dimension is practical. The resulting components, which end up being a combination of complex traits, suggest biological significance enhancing grouping and interpretability. One example explored in the paper involved extracting new variables from the PC scores, which identified lateral vigor and reproductive vigor from raspberry growth and reproductive attributes.

Barnston [5] used rotated principal component analysis (RPCA) of the Northern Hemisphere for different months of the year to describe the seasonality and persistence of major modes of interannual variability. The RPCA was able to identify and describe the seasonality and persistence of major modes as the loadings of the RPCA revealed major circulation patterns such as the North Atlantic Oscillation and the Pacific/North American Pattern. The results agreed with previous studies and concluded that RPCA provides physical interpretability and is statistically stable for climate investigations.

Finally, the only paper to combine PCA and FTIR was conducted by J. M. Andrade [4], where the principal component analysis of mid-IR indexes was performed in three different ways of controlling oil spills to evaluate chemical weathering. Multiple crude oils were evaluated and compared using the PC analysis. Initially, all the samples were similar, which is why mid-IR was used. The 3-way datasets allowed for easy comparison of six different hydrocarbon complex mixtures and their evolution over time. They found that the aromaticity index was almost exclusive in characterizing the weathering process. Other IR indexes were related to the evolution of products such as carbonyl and sulfoxide and were not relevant to extracting common weathering patterns.

Biochar serves as an effective means of eliminating various toxic compounds, including greenhouse gases such as CO_2. While Principal Component Analysis has also found its use in large-variable systems such as atmospheric science and some clinical studies. However, previous studies have not explored the use of PCA to investigate the physical significance behind the observed sample groupings or whether such groupings provide meaningful insights. Since a reduction of FTIR spectra is possible using the PCA showcase by J. M. Andrade, it can serve to identify key characteristics and better classify complex biochar samples. This paper aims to close that gap and investigate whether PCA could be effective in identifying chemical differences and grouping biochar together based on spectrographic data.

Chapter 2

Theory

2.1 Illustrative Example

Principal Component Analysis (PCA) is a widely used statistical technique for exploratory data analysis and dimensionality reduction. Its primary goal is to condense high-dimensional datasets into a smaller set of principal components while retaining as much original information as possible. This reduction enables easier visualization and interpretation of complex data in a lower-dimensional space.

In (Fig. 2.1), PCA is applied to the well-known Iris dataset, which comprises measurements of sepal and petal dimensions for three species of iris flowers: Setosa, Versicolor, and Virginica [8]. Each data point in the PCA plot represents an individual iris flower, positioned according to its values along the first two principal components (PC1 and PC2). Distinct clusters emerge along PC1, with Setosa exhibiting negative values and Versicolor and Virginica displaying positive values. This indicates clear differences among the species, albeit with some overlap between Versicolor and Virginica along PC1.

This example underscores PCA's role in revealing patterns and clusters within datasets like the Iris dataset. In FTIR spectroscopy, PCA may offer insights by simplifying spectra data, potentially aiding in analysis and interpretation. While not definitive, exploring clustering in FTIR data through PCA could provide valuable insights.

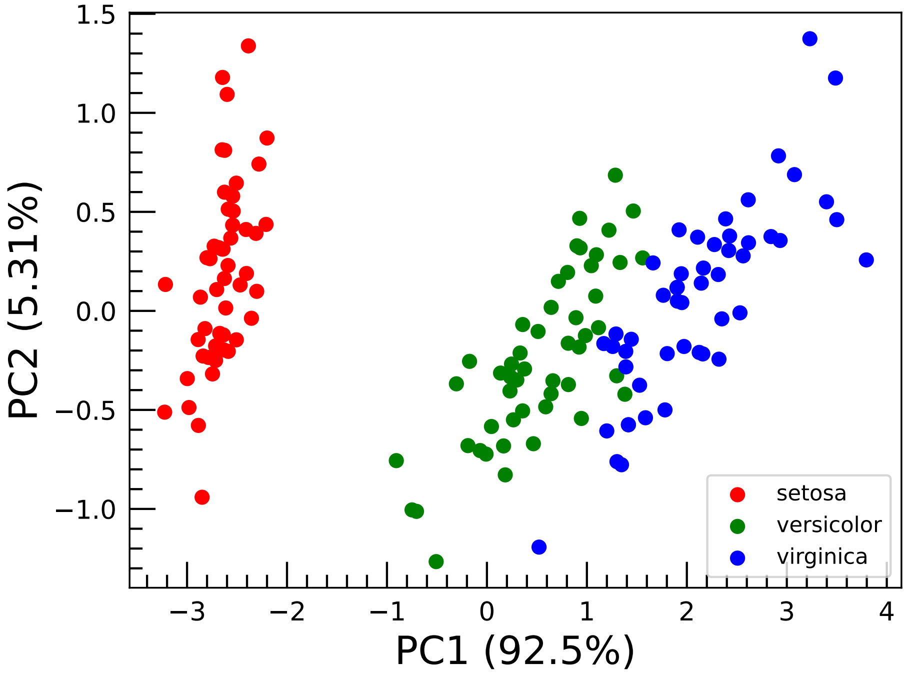

Figure 2.1: PCA plot illustrating the distribution of Iris species—Setosa (red), Versicolor (green), and Virginica (blue)—based on morphological features. PC1 and PC2 represent the primary and secondary axes of variation, respectively. Each data point corresponds to an individual Iris specimen revealing clusters of the individual species along PC1.

2.2 Mathematical Interpretation

Principal component analysis is a practical statistical technique that reduces dimension of variables of a system. For standard PCA a dataset with p number of variables for each n sample or individual. This creates dimensional vectors $\mathbf{x}_1, ..., \mathbf{x}_p$ otherwise expressed as a n by p matrix $\mathbf{X}$, whose jth column is the vector $\mathbf{x}_j$ of the jth variable. PCA seeks a linear combination of columns of $\mathbf{X}$ with maximum variance which is a measure of the spread of a set of data points. The linear combination with the maximum variance are expressed as $\sum_{j=1}^{p} a_j \mathbf{x}_j = \mathbf{Xa}$. Here $\mathbf{a}$ is a vector of constants $a_1, a_2, ...a_p$. The variance, represented by var($\mathbf{Xa}$)=$\mathbf{a}'\mathbf{Sa}$, where the var() function follows the conventional statistical definition, characterizes the spread or variability of the data projected onto the direction defined by the eigenvector $\mathbf{a}$. Here, $\mathbf{S}$ denotes the covariance matrix associated with the dataset, capturing the pairwise covariances between variables, and the transpose operation ($'$) indicates the transposition of the vector $\mathbf{a}$.

Finding the linear combination with maximum variance is the same as obtaining p-dimensional vector $\mathbf{a}$ which maximizes $\mathbf{a}'\mathbf{Sa}$. One requirement of the above statement is normalization such that $\mathbf{a}'\mathbf{a}$=1. Using the Lagrange multiplier λ, the equation becomes $\mathbf{a}'\mathbf{Sa}$-λ($\mathbf{a}'\mathbf{a}$-1), where $\mathbf{S}$ is the covariance matrix and $\mathbf{a}$ is the eigenvector. The Lagrange multiplier is introduced to enforce the constraint that the eigenvectors must be unit vectors ($\mathbf{a}'\mathbf{a} = 1$), ensuring that they form an orthonormal basis. Differentiating with respect to the vector $\mathbf{a}$, and equating to zero produces an eigenvector equation. $\mathbf{Sa}$-$\lambda\mathbf{a}$=0 or alternatively $\mathbf{Sa}$=$\lambda\mathbf{a}$. Therefore, $\mathbf{a}$ must be the orthonormal eigenvector and λ the associated eigenvalue of the covariance matrix $\mathbf{S}$. The largest eigenvalues signify the linear combinations that capture the most significant variance within the data. It also corresponds to the biggest principal component that captures the largest variance. The transform data will have principal component number of new dimensions. Only relative magnitude and sign patterns are meaningful as multiplying by -1 doesn't affect the loadings [14].

Any p by p real symmetric matrix, such as a covariance matrix $\mathbf{S}$, will have p real eigenvalues λ_k. This results in the formation of an orthonormal basis such that for each eigenvalue λ_k, there exists a corresponding eigenvector $\mathbf{a}_k$. The eigenvectors form an orthonormal basis, meaning that they are mutually orthogonal and each has a unit norm. Mathematically, this implies that $\mathbf{a}'_{k'}\mathbf{a}_k$ equals 0 if $k' \neq k$ and equals 1 if $k' = k$.

This approach shows that the full set of eigenvectors of $\mathbf{S}$ are solutions to the linear combinations $\mathbf{X}\mathbf{a}_k = \sum_{j=1}^{p} a_{jk}\mathbf{x}_j$ which maximize variance. The result is subject to uncorrelatedness with previous linear combinations. The uncorrelatedness results from the fact that the linear combination variance $\mathbf{X}\mathbf{a}_k$ and $\mathbf{X}\mathbf{a}_{k'}$ is given by $\mathbf{a}'_{k'}\mathbf{S}\mathbf{a}_k$=$\lambda_k\mathbf{a}'_{k'}\mathbf{a}_k$=0 if $k' \neq k$.

It follows that, the principal components (PCs), $\mathbf{a}_k$, are a set of orthonormal eigenvector, components of which are referred to as loadings. While the projection of the loadings onto PC, $\mathbf{X}\mathbf{a}_k$, represents the vector score of the kth PC [14].

Chapter 3

Samples

3.1 Documentation

Samples of biochar, raw softwood, and snow crab (Chionoecetes Opilio) biomass were provided by Dr. Kelly Hawboldt's group (Memorial University Process Engineering). They prepared the samples using different processing methods, as described below:

Softwood auger biomass (RAW) was subjected to the auger method at 500 °C (BAW), where the continuous feeding of biomass at a rate of approximately 1 kg/hr occurred within an auger reactor under slight vacuum conditions, thereby inducing fast pyrolysis.

Biochar from softwood biomass was subjected to fast pyrolysis within the Atefeh Furnace (UAWB). This involved batch processing of softwood biomass within a furnace maintained at 500 °C while continuously purged with nitrogen gas to avoid oxygen exposure.

To investigate the effect of particle size, biochar produced from the Atefeh Furnace was subjected to post-production grinding (CAWB) to achieve reduced particle dimensions, allowing assessment of the effect on biochar properties.

Crab biochar (PCB/BC) was produced from snow crab meat residue, and the shell (RDC) underwent pyrolysis at 500 °C after a drying period of approximately 3 hours.

Hydrothermal crab biochar (HTCB) was derived from dried snow crab that underwent hydrothermal treatment. The pretreatment consisted of oven drying followed by grinding, while the post-treatment consisted of filtration and drying, with no chemical intervention throughout the process.

Another crab hydrothermal carbonization biochar (HTCBT) sample was produced for different temperatures and exposure times with no chemical modifications and similar drying procedures described for HTCB.

Chapter 4

Methods

4.1 Sample Measurements

I collected the data for each sample by completing three measurements on a plate, with five repetitions per sample. An exception was observed for samples of HTCBT, which utilized a distinct geometry and produced transmission spectra. Furthermore, due to a discrepancy in data points compared to the rest of the HTCBT dataset, one measurement had to be excluded.

4.2 Experimental Measurements

For internal reflection FTIR, for each sample, I collected thirty-six scans averaged within the Bruker ALPHA II FTIR Spectrometer. Before measurements a background spectrum was done and served in measuring the signal contribution of both the instrument and the environment to the overall spectrum. This process involves taking the ratio of the sample beam spectrum to the background spectrum, ensuring accurate and reliable data analysis by removing these irregularities. I selected the resolution such that it would give 5022 spectral points in the range of 4000 to 400 cm^{-1}. Data was stored in both OPUS and .csv formats, allowing for convenient access. OPUS, the Bruker IR software, facilitated original data examination, while .csv files were used in PCA processing in RStudio. The .csv files organize data with the first column representing wavenumbers ranging from 4000 to 400 cm^{-1} and the second column denoting relative intensity or absorption/transmission spectra.

4.3 Implementation of PCA

Prior to R-based PCA, in the preliminary analysis, I used a reference PCA plot created in QuasarOrange [3] by Jake Breen to verify my results with R. The largest eigenvalues of the PCA of a dataset are referred to as 'principal components' (PCs). Upon adequate data loading and processing in R, the principal component score plots were compared with those from QuasarOrange to ensure consistency. 'Loadings' are used instead of eigenvectors of the principal component. Individual principal component loadings indicate component dependencies on wavenumbers, depending on whether they are negative or positive. The projection of the principal component eigenvector onto the original dataset is referred to as a 'score.' A positive loading (eigenvector) indicates a positive principal component score (projection), and vice versa. Score plots visualize the distribution of samples in the principal component space, providing insights into the relationships and clustering patterns within the dataset.

For PCA processing, the initial column served as the basis of the merged dataset, which housed the wavenumber, given that spectra comparison was conducted using identical instruments and resolution. Subsequently, the data was merged into a matrix, the column of wavenumber removed, transposed, and dimensionally reduced using PCA. This results in three variables as compared to the 5022 intensity variables of the original matrix.

I did exploratory checks, including evaluating outcomes when employing correlation versus covariance matrices, with identical results prompting the selection of the covariance matrix for its variance maximization, central for dimensional reduction.

I normalized and removed NaN values from the dataset as part of the preprocessing, aiding in mitigating discrepancies and errors in PCA and enhancing interpretability.

I utilized scree plots to assess component variance, often revealing that the first two components collectively accounted for over 99% of the variance. This necessitated the utilization of two components unless the third component had a sufficient percentage contribution.

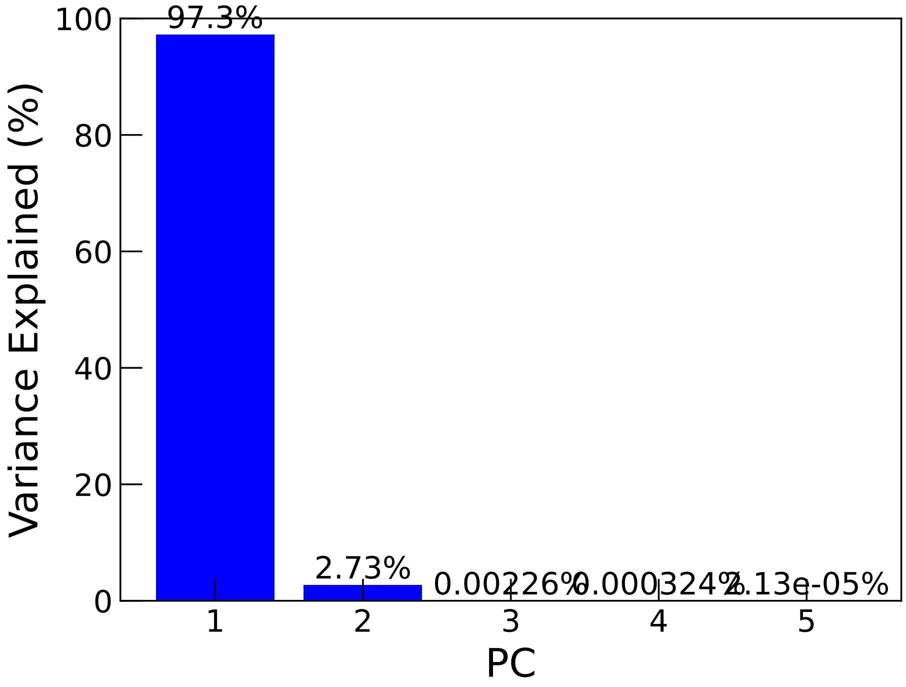

Figure 4.1: This bar graph illustrates the percentage variance attributed to each principal component identified through principal component analysis (PCA) for the combined Raw Auger Wood (RAW) and Biochar Auger Wood (BAW) datasets. The graph highlights the principal component that contributes the most variance to the system. It is noteworthy that the first component predominantly explains the variance.

As can be seen from Fig. 4.1, the initial scree plot from R reveals the variance of the principal component for varying datasets. The dataset pictured is the raw and biochar auger dataset, with most of its variance explained by the first component with 97.3% of the variance. The only other component is the second component with a variance of 2.73%, which is insignificant for analysis and is ignored.

While BiPlots provide insights into component-variable relationships, their utility is limited by the extensive dataset, rendering interpretations challenging due to the clustering of over five thousand variables on one plot.

For grouping interpretations, I leveraged sample projections onto principal component space through dot product computations. Scores were normalized, and both loadings and projections were saved in .csv format for subsequent analysis and visualization in Python. Python facilitated one-, two-, or three-dimensional visualization of the results.

Chapter 5

Results

5.1 Raw and Biochar Analysis

5.1.1 Raw and Biochar Auger Wood Comparison

The first combined set of data that I analysed was the raw and biochar samples of the same material with fifteen measurements each, with intensity as the variable. Then, the principal component routine was performed on the combined dataset, which was comprised of raw auger wood (RAW) and biochar auger wood (BAW) produced at 500 °C by fast pyrolysis. The raw sample was similar to a fine wood filling, while the biochar resembled charred, black powder.

The original spectra shown in Fig. 5.1 a) for the RAW sample had two characteristic peaks at 3400 cm^{-1} wavenumber, which is typically where the N-H or O-H stretching bonds are observed in the wood, whereas the BAW does not have these peaks as the water was evaporated in the process of biochar formation. In addition, the RAW has a sharp peak at about 1000 cm^{-1} wavenumber, which is the energy range in which the O-C stretch bond occurs as many other vibrational modes share this energy, which the biochar does not exhibit due to the fact that the combustion process requires oxygen and it was mostly consumed in the pyrolysis. The BAW has a peak at 1712 cm^{-1} corresponding to C=O stretching and a peak at 1581 cm^{-1} which likely suggests the presence of the carbon-carbon double bond (C=C) typically observed in coal-like material and hence biochar [22].

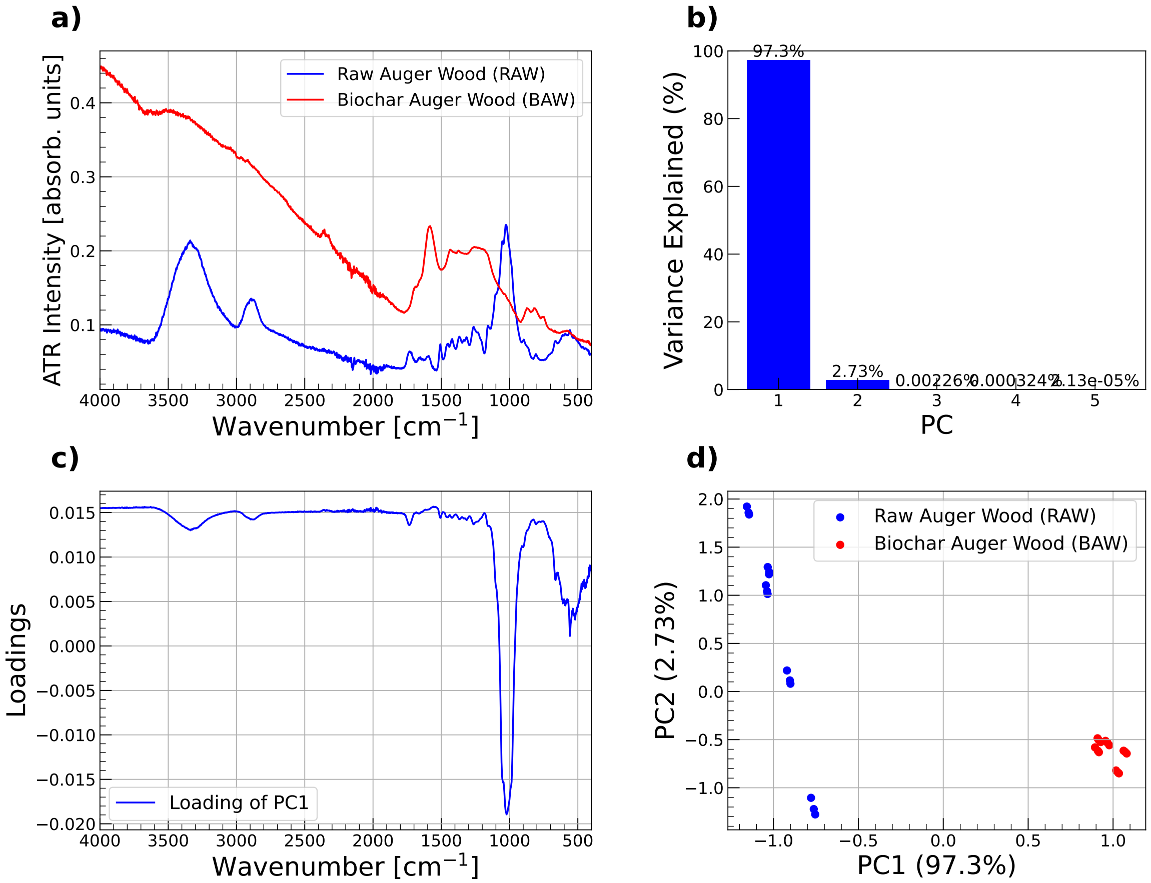

Figure 5.1: **a)** illustrates ATR intensity with absorbance units plotted against wavenumbers between 4000 and 400 cm^{-}1 of spectra of raw and biochar samples of auger wood obtained from FTIR spectroscopy. The blue line represents the raw spectra (RAW), while the red line represents the biochar spectra (BAW). It is noteworthy that biochar exhibits a distinctly different spectrum as compared to raw auger wood due to the combustion of organic material during pyrolysis. **b)** shows the percentage variance captured by each principal component (PC). The first explains 97.3% variance, with the second one explaining only 2.73%. **c)** depicts the loading plotted against the variable of wavenumber for the first principal component. The loading of the first component (blue) demonstrates predominantly positive values across most spectra, with a notable sharp negative loading observed at approximately 1000 cm^{-1}. **d)** illustrates the distribution of two discrete sample sets in one-dimensional principal component space. The data points representing BAW samples (red triangles) cluster tightly around a PC1 score of 1, whereas the RAW samples (blue circles) are grouped around a PC1 score of −1. This observation suggests an inverse loading of RAW samples with respect to the first component.

This spectral feature aligns with the typical infrared spectra of modern wood charcoal. Specifically, the absorption bands at 3400 cm^{-1}, 1725-1705 cm^{-1}, and 1581 cm^{-1} correspond to the medium N-H stretching of aliphatic primary amine, the strong C=O stretching of aliphatic ketone or cyclohexanone/cyclopentenone, and the presence of a carbon-carbon double bond (C=C), respectively [22].

Fig. 5.1 b) shows the variance explained by each principal component. The first component, with 97.3% variance, dominates the analysis, reducing the problem to one dimension.

Fig. 5.1 c) shows the loadings, confirming that the first component is explained by the majority of the wavenumbers as it has high positive loadings in almost all the wavenumbers from 4000 to 1100 cm^{-1}. However, the loading has highly negative values at about 1000 cm^{-1}, which suggests that the first component is not influenced by this region and the wavenumbers in this region have an inverse relationship to the first component. Overall, the first component loading seems to capture most of the variance, confirming the scree plot, despite moderate variance for low wavenumber. Specifically, the peak at 1000 cm^{-1} exhibits a negative loading, prominently affecting RAW samples, thereby attributing negative scores to them. Conversely, BAW samples exhibit positive scores, reflecting their alignment with predominant positive loadings. The absence of the peak at 1000 cm^{-1} in the BAW dataset contributes to its positive PC1 scores, clearly marking a separation from RAW samples. However, the main qualifying factor is the intensity around this peak, as it is the only time the RAW sample has a higher intensity than BAW.

Given the dominance of variance explained by the first principal component (PC1), a one-dimensional representation adequately captures the data's variability, as illustrated in Fig. 5.1 d). The scores along PC1 reveal distinct patterns: RAW samples are centered around -1, while the BAW dataset exhibits a PC1 score of 1.

The PCA managed to capture the difference between RAW and BAW samples associated with the fact that the spectra of the samples exhibited mostly different intensities that are also coupled with the chemical features of pre- and post-combustion.

5.1.2 Raw and Biochar Dried Crab Comparison

Another combined data set of raw dried crab (RDC) and biochar crab (BC) samples was analysed using PCA and confirmed previously established findings. The variance analysis in R showed in Fig. 5.2 b) that for the dataset containing RDC and BC, 66.7% of the variance was explained in the first component, 32.0% in the second component, and 1.30% in the third component. The third component was ignored for this data due to its insignificance. I anticipated, based on previous findings for RAW and BAW, that the two spectra of RDC and BC would be successfully grouped.

Fig. 5.2 a) displays the original spectra of the RDC in blue and BC in red. The RDC exhibits broad peaks between 3700 cm^{-1} and 2500 cm^{-1}, which are likely due to the water OH- stretch bond. In contrast, the BC spectra has a smooth 'bump' and still displays the peak at 2900 cm^{-1}, which could be due to the -C-H stretch or -C-H aldehydic bond stretching. The water peak in RDC is sharp, as expected for a raw sample. However, the intensity of the BC in that region is irregularly high and difficult to quantify, as one would expect water to be removed in pyrolysis. The three small peaks between $1600 - 1400$ cm^{-1} could be the C=O amide bonds in the RDC, which were removed in the combustion, leaving only the sharp aragonite peak at 1460 cm^{-1} after combustion. The aragonite peak is expected as it is part of the chemical make-up of the crab shell. The peak at approximately 1050 cm^{-1} is likely due to another vibrational mode of aragonite as well as the peak at 856 cm^{-1}, a common characteristic of crustrations like crab.

Hao *et al.* [12] collected FTIR of chitosan of swimming crab (Portunus trituberculatus) shell pretreated in subcritical water. FTIR spectra had bands at 3443–3416 cm^{-1} corresponding to OH- stretching. Furthermore, the amide-I band in the α-chitin split at 1660 and 1625 cm^{-1} attributed to intermolecular hydrogen bonds. Amide II chitosan peaks are observed at approximately 1560 cm^{-1}. There are some similarities between the spectra I obtained and the findings from the study by Hao.

The loadings plot in Fig. 5.2 c) shows the variance explained by the wavenumbers for the first two components. PC1 loading in blue is highly positive for the wavenumber of 1600 cm^{-1}, 1000 cm^{-1}, and 900 cm^{-1}, neutral to the region between 2500 cm^{-1} and 1700 cm^{-1} and negative for all other wavenumbers. PC2 loading in red is positive for all wavenumbers with neutral values corresponding to 1600 cm^{-1}, 1000 cm^{-1}, and

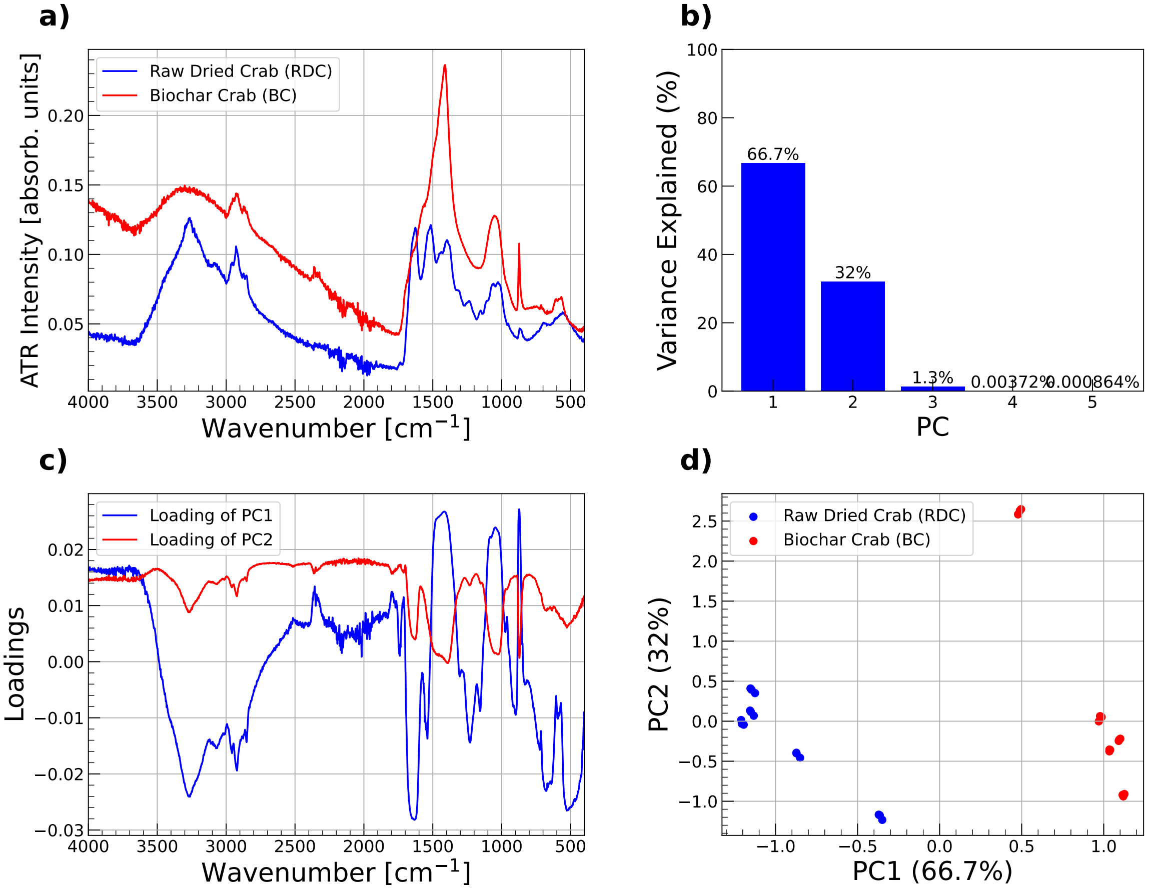

Figure 5.2: **a)** illustrates ATR intesity (absorbance units) vs wavenumber between 4000 and 400 cm^{-}1 of the original spectra of raw dried crab (RDC/blue) and biochar crab (BC/red) obtained from FTIR spectroscopy. It is noteworthy that while the biochar exhibits a spectrum similar to that of the raw material due to the lower flammability of crab biomass, the intensity of the biochar spectrum is generally higher across most wavenumbers. **b)** shows percentage variance captured by each principal component (PC). The first captures 66.7% variance, the second one explains only 32.0%, and third explains the remaining 1.30%. **c)** presents the loading plotted against the variable of wavenumber for the first and second principal components. The loading of the first component (blue) showcases positive values for peaks around 1500 cm^{-1}, 1000 cm^{-1}, and 900 cm^{-1}, while exhibiting negative values for wavenumbers ranging from 3500 cm^{-1} to 2800 cm^{-1}, as well as around 1600 cm^{-1} and in the lower range from 800 cm^{-1} to 400 cm^{-1}. **d)** displays the distribution of two discrete sample sets in one-dimensional principal component space. Note that while the BC data (red triangles) may contain outliers, they are predominantly grouped, as are the RDC data along PC1 score with BC centered 1, and RDC mostly centered around −1.

900 cm^{-1} wavenumbers. The two components seem to be inverses of each other for the peaks of 1600 cm^{-1}, 1000 cm^{-1}, and 900 cm^{-1} wavenumbers.

Fig. 5.2 d) illustrates the PC space plots, revealing distinct groupings despite the presence of outliers resulting from measurement anomalies. BC tends to cluster around a score of 1 on PC1, reflecting their predominantly positive loadings across most wavenumbers. Conversely, RDC samples are predominantly clustered around a score of -1 on PC1, indicating an inverse relationship with PC1 characterised by predominantly negative loadings. This disparity can be attributed to the higher intensity of BC spectra compared to RDC. The high intensity can be attributed to the fact that biochar spectra have increased porosity and surface area due to pyrolysis processes, leading to greater interactions with IR. The score plot suggests that PC1 predominantly captures variations in spectral intensity.

Regarding PC2, the majority of data points for both BC and RDC exhibit negative scores. However, three outlier BC measurements display positive scores on PC2, corresponding to their positive loadings on this component. These outlier spectra exhibit distinct features around wavenumbers 1600 cm^{-1}, 1000 cm^{-1}, and 900 cm^{-1}, setting them apart from the rest of the dataset. This observation implies that the outlier spectra are primarily represented by PC2, suggesting unique spectral characteristics associated with these outliers.

To ensure the validity of the evaluation, I removed the outlier measurements from the combined dataset before conducting the PC analysis in R. The outliers in the dataset were identified as the three measurements of BC with positive scores of 2.5 for PC2, as they significantly influence the variance explained by PC2 and deviate from the typical trends in the dataset. Upon removal of these measurements, the results showed that the first component dominated, explaining 90.8% of the variance, while the second component explained 8.7% and the third component had minimal variance at 0.5%.

My analysis confirms that outliers for the combined crab dataset are associated with the second component. Once the outliers are removed, the first component becomes the only significant one, producing distinct groupings.

The key difference between raw and biochar samples is the increase in intensity and reduction of peak structure due to combustion processes, resulting in the variance being explained by one component unless radical outliers are recorded in the data.

5.1.3 Combining the Crab and Wood Biochar Samples

As observed in subsection 5.1, the raw and biochar samples of wood and crab are well-grouped due to intensity differences. The groupings will be clear when comparing the two biochars, and the first component should explain most of the variance. As previously noted, BAW is mostly dependent on carbon bonds formed after combustion, while the crab shell is heavily explained by the calcium carbonate spectral peaks. The initial analysis in R confirmed the suspicion, as Fig. 5.3 b) shows that 99.8% of the variance was explained solely by the first component, leaving only 0.16% variance for the second component. This indicates that the problem is one-dimensional.

Fig. 5.3 a) shows that the spectra features differ only in the region between 1750 cm^{-1} and 400 cm^{-1} wavenumbers. Moreover, The BAW (blue) has a higher overall intensity for most of the data and has a broad peak from 1750 to 1200 cm^{-1}, as well as a sharp peak at 1600 cm^{-1} corresponding to the C=C stretch bond, as explained in the previous subsection 5.1.1. The three peaks around 1460 cm^{-1}, 1100 cm^{-1}, and 856 cm^{-1} for crab biochar are significant, with the first indicating the presence of $CaCO_3$ and the last qualifying the calcium carbonate to be of the aragonite mineral form.

In Fig. 5.3 b), the loading plot indicates predominantly positive values across all wavenumbers, except for slightly less positive values at 1000 cm^{-1} and 900 cm^{-1}, along with a negative dip at 1450 cm^{-1}. Spectral analysis of the loadings reveals a negative loading coinciding with regions where BC spectra exhibit higher intensity compared to BAW.

Furthermore, Fig. 5.3 c) visualises the scores, clearly delineating BC samples on the negative side and BAW samples on the positive side of the axis due to their differing intensity. Once more, the score plots indicate that BAW samples are centred around 1, while BC samples are centred around -1. This observation underscores that the intensity of the respective biochars serves as the primary determinant for sample classification and not the chemical bond peak structure.

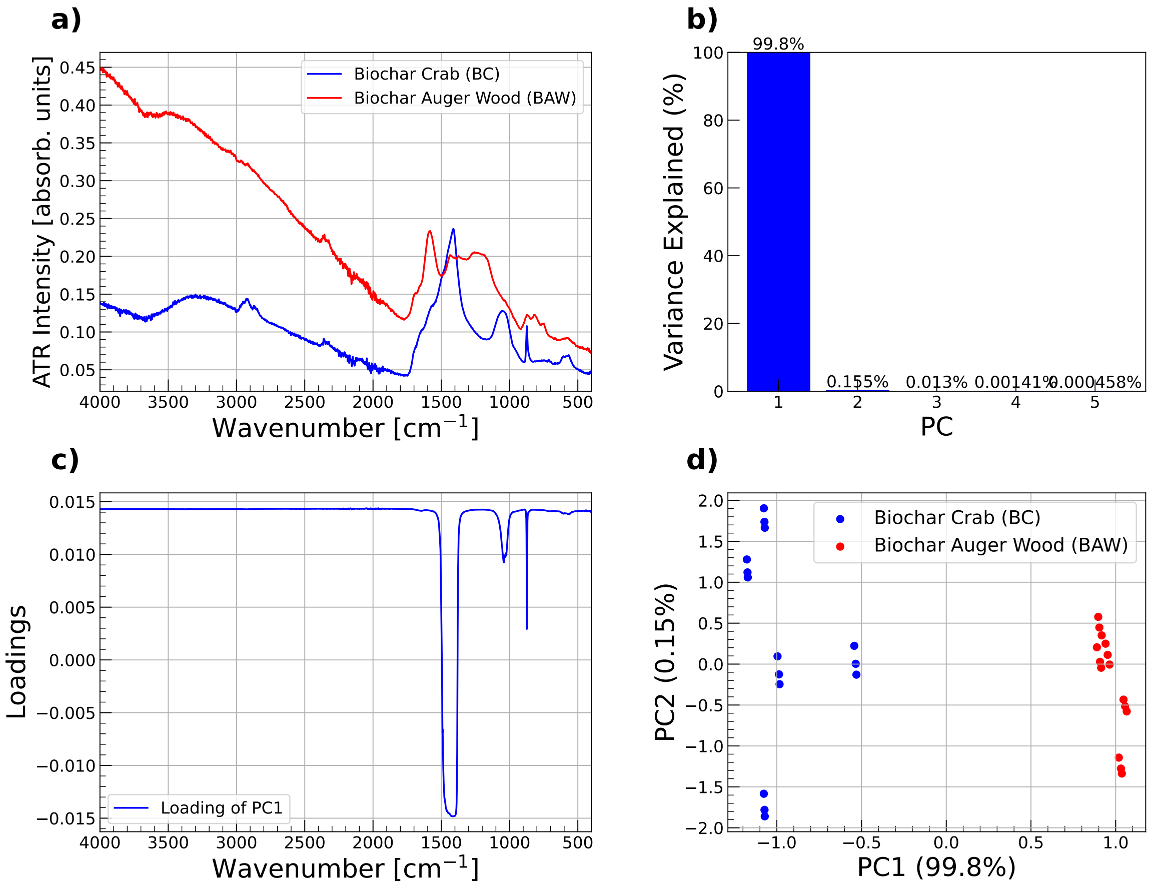

Figure 5.3: **a)** displays the FTIR spectra of biochar auger wood (BAW) in blue and biochar crab (BC) in red plotted for ATR intensity (absorbance units) vs wavenumber between 4000 cm^{-1} and 400 cm^{-}1. While the spectra exhibit similarity in the range from 4000 cm^{-1} to 2000 cm^{-1}, differences in intensity are observed. The peak structures differ significantly; BAW displays a single sharp peak with higher intensity compared to BC, which exhibits three visible peaks in its spectra. **b)** shows percentage variance captured by each principal component (PC). The first captures all the variance at 99.8%. **c)** illustrating the loadings of the first principal component (blue) plotted against the variable of wavenumber. The loading profile remains predominantly positive until 1500 cm^{-1}, with minor fluctuations. Note, the three small depressions at 900 cm^{-1}, 1000 cm^{-1}, and 1460 cm^{-1} with the latter one having the only negative loading. **d)** depicts the distribution of samples in one-dimensional principal component space. Samples are positioned on opposite sides of the PC1 axis at -1 and 1, attributed to the negative loading of one peak in BC that surpasses BAW in intensity.

5.1.4 Combining the Raw and Biochar Sample of Wood and Crab

As a further extension, I combined the datasets of both raw and biochar samples of wood and crab. The initial analysis in RStudio revealed that the variance was spread over the first four components. Fig. 5.4 b) reveals that the first two components dominated and explained 97.9% variance, with the first one having 83.1% and the second one having 14.8%. The last two components are ignored for the analysis as their shared variance is combined to 2.1% which is too insignificant to capture information about the dataset.

Figure 5.4 a) shows the spectra of the individual samples. Each spectrum has unique characteristics and peaks, which were described in detail in sections 5.1.2 and 5.1.1. It is important to note that the BAW spectrum is the only one that does not exhibit the 'bump' around 3400 cm^{-1} and has a higher intensity than all other samples, partly due to the fact that it was the only finely grounded sample. The dependence of the samples on certain bonds and peaks remains true, such as one of the C-O stretch bonds (around 1000 cm^{-1}) for RAW, one of the vibrational modes of aragonite (1460 cm^{-1}) for BC, and C=C (1581 cm^{-1}) for BAW.

In Fig. 5.4 c), the loadings reveal the significant components and their dependence on each wavenumber. The first component predominantly features positive values across most wavenumbers, with neutral dips observed around 1000 cm^{-1} and a negative dip at 1460 cm^{-1}. Conversely, the second component's loading exhibits highly positive values at 1460 cm^{-1} and mostly positive values elsewhere, except for a notable negative dip at 3400 cm^{-1}, a highly negative value at 1000 cm^{-1}, and negative values between 600 cm^{-1} and 400 cm^{-1}. The second component also features a very high positive loading at around 856 cm^{-1} and positive loadings for the peak observed at around 1600 cm^{-1}.

Analyzing the two-dimensional scores plot from Fig. 5.4 d), the samples exhibit distinct groupings along the scores of each component, with a few outliers observed within the BC group. As indicated by the loadings analysis, only one sample, BAW, possesses positive loadings along the first component, consistent with its generally higher intensity across most data points, except for a couple of peaks. This discrepancy arises from the higher intensity of BAW data compared to all other samples,

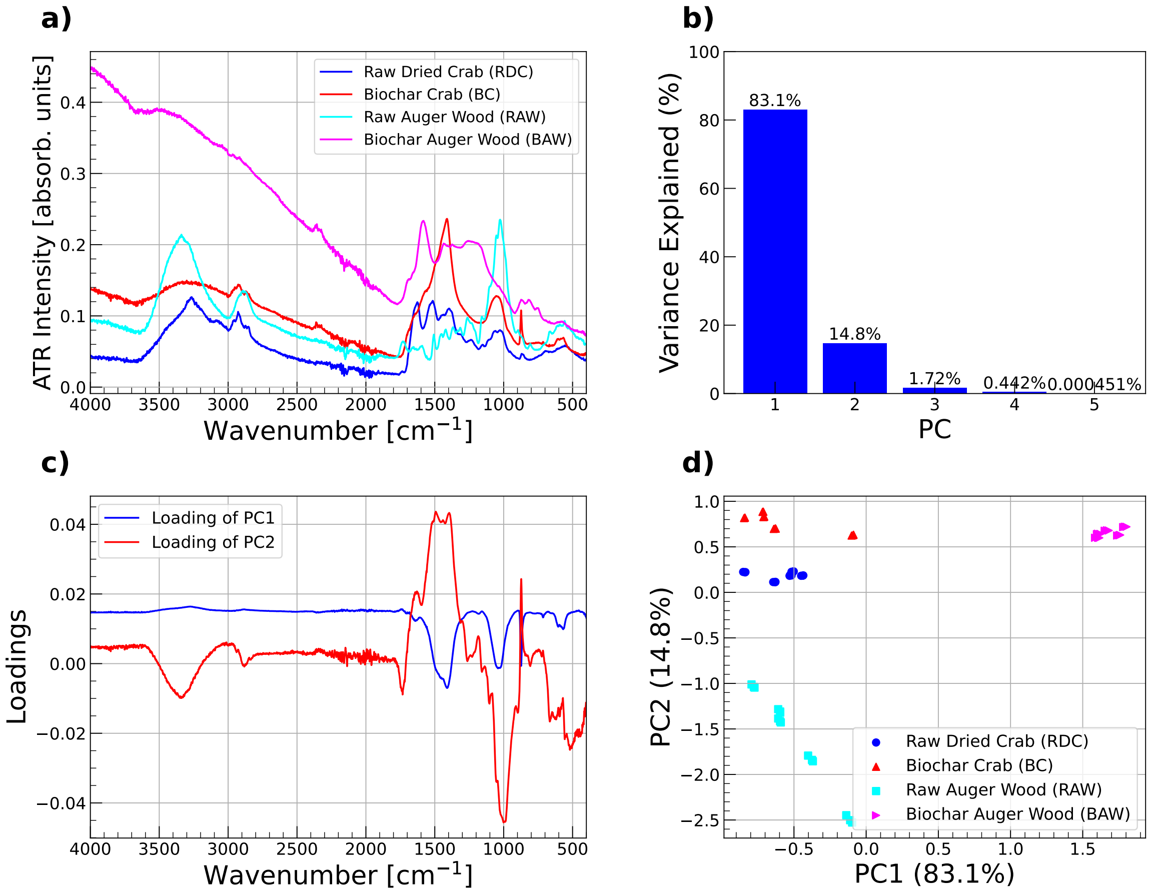

Figure 5.4: **a)** the FTIR spectra of BAW (magenta), RAW (cyan), BC (red), and RDC (blue) samples are plotted for ATR intensity (absorbance units) vs wavenumber between 4000 cm^{-1} and 400 cm^{-1}. BAW exhibits distinct properties compared to the other samples within the range of 4000 to 1700 wavenumbers. The peaks are dispersed, with a notable concentration around 1500 cm^{-1} and 1000 cm^{-1}, displaying varying levels of intensity. **b)** shows percentage variance captured by each principal component (PC). The first captures 83.1% variance, the second one explains 14.8%, with the rest captured by the remaining components. **c)** presents the loadings of the two main components plotted against the wavenumber. The first loading (blue) closely resembles the loading observed in section 5.1.3 of the one-dimensional analysis of Wood and Crab biochar, exhibiting predominantly positive loadings except for two depressions where BC and RAW surpass it in intensity. The second loading (red) is slight positive with fluctuations at 1500 cm^{-1} and 900 cm^{-1}, along with negative loadings at 3300 cm^{-1}, 1000 cm^{-1}, and within the low wavenumber range of 700 – 400 cm^{-1}. **d)** illustrates the measurements of the samples plotted in principal component space, revealing their distribution across the two-dimensional space and clear visual groupings. It is noteworthy that outliers are observed within the BC and RAW groups. Additionally, BAW exhibits a positive PC1 score, while the rest of the samples are distributed along the negative PC1 score and the entire range of the second component score.

barring two peaks in BC at 1460 cm^{-1} and RAW at 1000 cm^{-1}.

Refer to Supplementary Material A for an analysis concerning K-mean clustering algorithms and outlier removal.

The remaining samples manifest a negative PC1 score of -0.5 and are predominantly distributed along the second component. The second component appears to capture more nuanced chemical differences between samples, as they exhibit comparable intensities. RDC and BC display positive values of PC2 scores, indicating their positive loadings at 1460 cm^{-1} and 900 cm^{-1}, both featuring higher intensity than RAW, but are inversely rated due to the lower intensity 'bump' at 3400 cm^{-1}. Subsequently, the crab samples are further distinguished by their intensities, yielding either high or low positive scores on the second component. RAW predominantly manifests negative values on the second component owing to its peak at 1000 cm^{-1}, characterised by a negative loading.

I showed earlier that if the samples are greatly different in intensity, like raw and biochar, the high-intensity samples will dominate the loadings. In systems with numerous spectra of similar intensity, such as those discussed in the later section 5.3, structural chemical differences, such as the characteristic peaks of bond vibration modes, become more significant.

An extension could be carried out where the number of variables is not all the wavenumber but the select few that represent the peaks, which will help if chemical bonds are the focus of analysis for groupings.

5.2 Different Carbonization Processes

To test production methods I combined two samples that shared the crab shell biomass but were produced using different procedures into a dataset. Hydrothermal carbonisation biochar (HTCB) required the sample to be heated in subcritical water for long periods of time, while pyrolysis crab biochar (PCB) was produced at high temperatures over a short period of time.

Upon completion of the PCA, the Fig. 5.5 b) revealed that the first component explained 70.0% of the variance and the second component had 27.2% variance while all other components were too small for consideration.

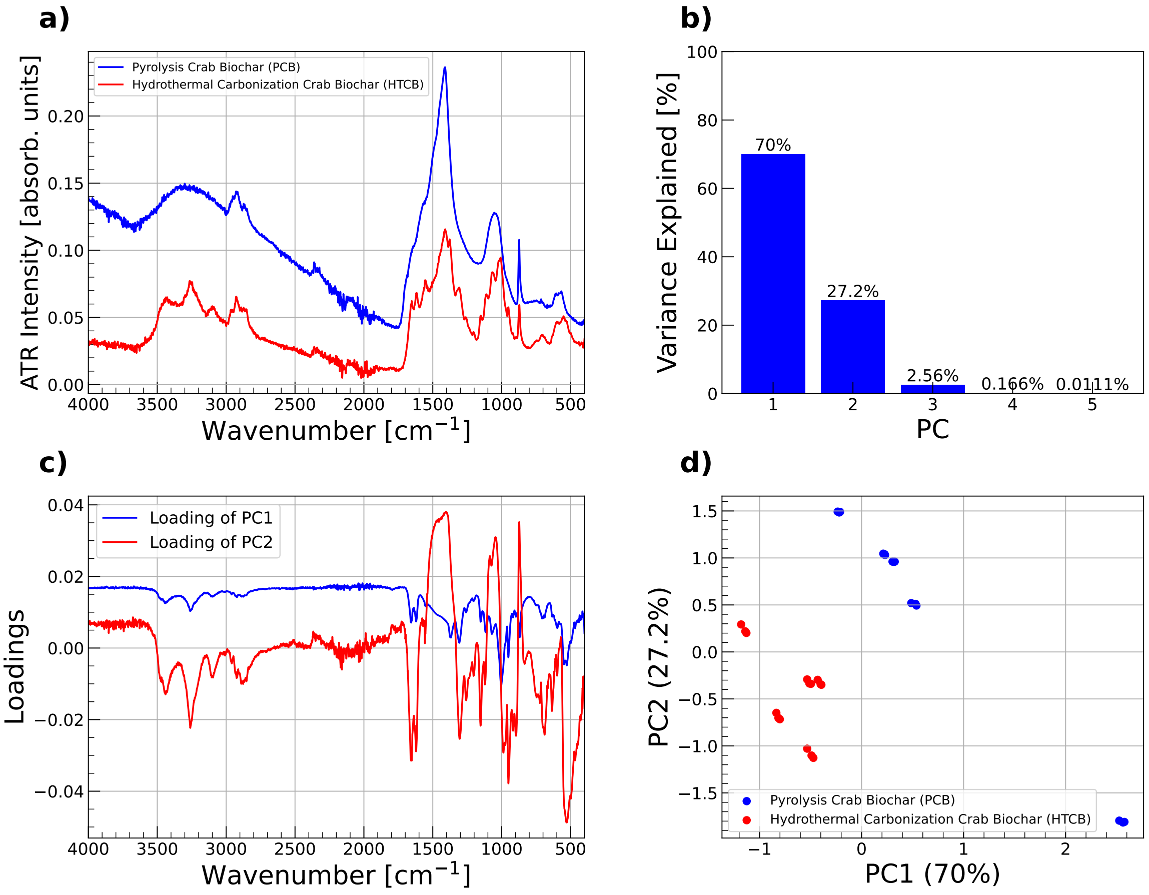

Figure 5.5: **a)** displays ATR intensity with absorbance units against the wavenumber ranging from 4000 to 400 cm^{-1} of the FTIR spectra of crab biochar subjected to different treatments, with the blue line represents pyrolysis crab biochar (PCB), and the red line representing the hydrothermal carbonization biochar (HTCB). Despite the outlier status of the depicted PCB spectra, the PCB samples consistently exhibits higher intensity compared to HTCB across all other samples. **b)** shows percentage variance captured by each principal component (PC). The first captures 70.0% variance, the second one explains 27.2%, the remainder is captured by the third, forth, and firth component at 2.7%. **c)** illustrating the loadings of the first (blue) and second (red) components plotted against the wavenumber. The first component loading predominantly displays positive values across most data points, with smaller positive values observed in the region corresponding to peaks for HTCB. Meanwhile, the second component loading demonstrates positive values aligning with the peaks of HTCB at 1500 cm^{-1}, 1000 cm^{-1}, and 900 cm^{-1}, along with negative loadings for the triad of peaks around 3300 cm^{-1}, as well as between the aforementioned HTCB peaks. **d)** depicts the scores plot in two-dimensional principal component space for HTCB and PCB. Apart from outliers of PCB, distinct groupings of the two treatment groups emerge, with PCB samples clustered around positive values of both PC1 and PC2 scores, while HTCB samples predominantly reside on the opposite side with negative PC1 and PC2 scores.

Fig. 5.5 a) shows that the difference between the two methods is the magnitude of the intensity of the peaks. As the two spectra have the characteristic aragonite peaks for $CaCO_3$ near 1460 cm^{-1} and856 cm^{-1}. The other difference between the spectra is the behaviour between the 3500 cm^{-1} and 3000 cm^{-1} wavenumber. The figure also shows that one spectra has much lower intensity than the other. This could be due to the fact that pyrolysis is more effective at increasing the surface area of the biochar, allowing for greater IR interactions as compared to hydrothermal carbonisation processes. However, it was evaluated that the PCB spectra pictured is an outlier sample with abnormally high intensity and is responsible for the first component, as seen from the fact that in Fig. 5.5 c) has a positive value of PC1 twice that of any other sample. As seen before in Section 5.1.2 outliers, by themselves, contribute a lot of variance to PCs. Refer to Section A Fig. A.1 for the confirmation of the assertion that the pictured PCB has an uncharacteristically high intensity.

In Fig. 5.5 c), the loadings plot reveals the primary components for assessing two different preparation methods of crab shell biochar. The first component exhibits predominantly strong positive values across most wavenumbers, with slight negative loadings observed at 1000 cm^{-1} and 500 cm^{-1}. This deviation stems from the outlier sample, which exhibits lower intensity than the HTCB, specifically at 1000 cm^{-1} and 500 cm^{-1}. Conversely, the second loading displays values hovering just above zero within the range of wavenumbers from 2700 to 1700 cm^{-1}, with highly positive values observed at 1500 cm^{-1}, near 1000 cm^{-1}, and 900 cm^{-1}. Negative loadings are evident for the three peaks from 3500 cm^{-1} to 2800 cm^{-1} in the original spectra, as well as around 1700 cm^{-1}, 1300 cm^{-1}, 1000 cm^{-1}, and nearly anything below 800 cm^{-1}. This pattern suggests that the first component encodes information pertaining to the outliers, while the second component captures information regarding the remaining spectra. Indeed, upon removal of the PCB dataset outliers, the scree plot in R reflected 89.5% and 9.5% of variance in the first and second components, respectively.

Examining the score plot in Fig. 5.5 d) for hydrothermal carbonisation and pyrolysis crab biochar in two dimensions, the samples exhibit distinct groupings, barring the outliers within the PCB group. Specifically, the PCB group clusters around positive PC1 and PC2 scores, while the HTCB group clusters around negative PC1 and PC2 scores. The loadings facilitate interpretation by indicating that the overall intensity of PCB samples is higher, as evidenced by their positive score on PC1, whereas HTCB samples exhibit a negative score on PC1. Furthermore, the concentration of PCB samples around positive values for PC2 scores, as captured by the second loadings, corresponds to the positive values for the three sharp peaks characteristic of PCB samples, while HTCB samples are inversely rated, with lower intensities observed between the peaks, contributing to their negative scores.

This is the second case of outliers having a large impact on the system, suggesting that a base line correction implementation is necessary. This also poses an interesting question about why the intensity jumped in such a great fashion for the same sample arranged slightly differently on the FTIR plate. A more rigorous approach to measuring data must be streamlined to avoid intensity differences by a factor of 10 and to eliminate errors in measurement methods.

The outliers significantly impact the component space because the principal components capture the variance of these samples themselves, as seen in the analysis conducted in Section 5.1.2. Therefore, it's essential to consider averaging the measurements to assess their impact on subsequent data clustering.

5.3 Experimental Conditions

5.3.1 Crushed vs Uncrushed

To investigate similar spectra, I created a dataset with two samples of biochar of the same auger wood biomass, one that was finely crushed (CAWB) and one that wasn't (UAWB), and subjected them to PCA. Preliminary analysis from Fig. 5.6 b) attributes 88.3% and 9.90% variance to the first and second components, respectively, reducing the dimensionality of the problem to two.

The CAWB and UAWB samples share nearly identical spectra, as captured by Fig. 5.6 a). There are slight differences in intensity in some intervals of the wavenumber near 4000 cm^{-1}, but the spectra analysis does not reveal any clear intensity trends that can be captured by PCA. As the samples have the same chemical make-up, their spectra share peaks and dominant features. The most distinguishable feature of the different samples is the C=C peak at 1600 cm^{-1} to 1200 cm^{-1}. It is important to recognize that only one measurement out of fifteen for each sample is illustrated, and not all samples may display spectra following the identified pattern of switching intensity after the C=C peak.

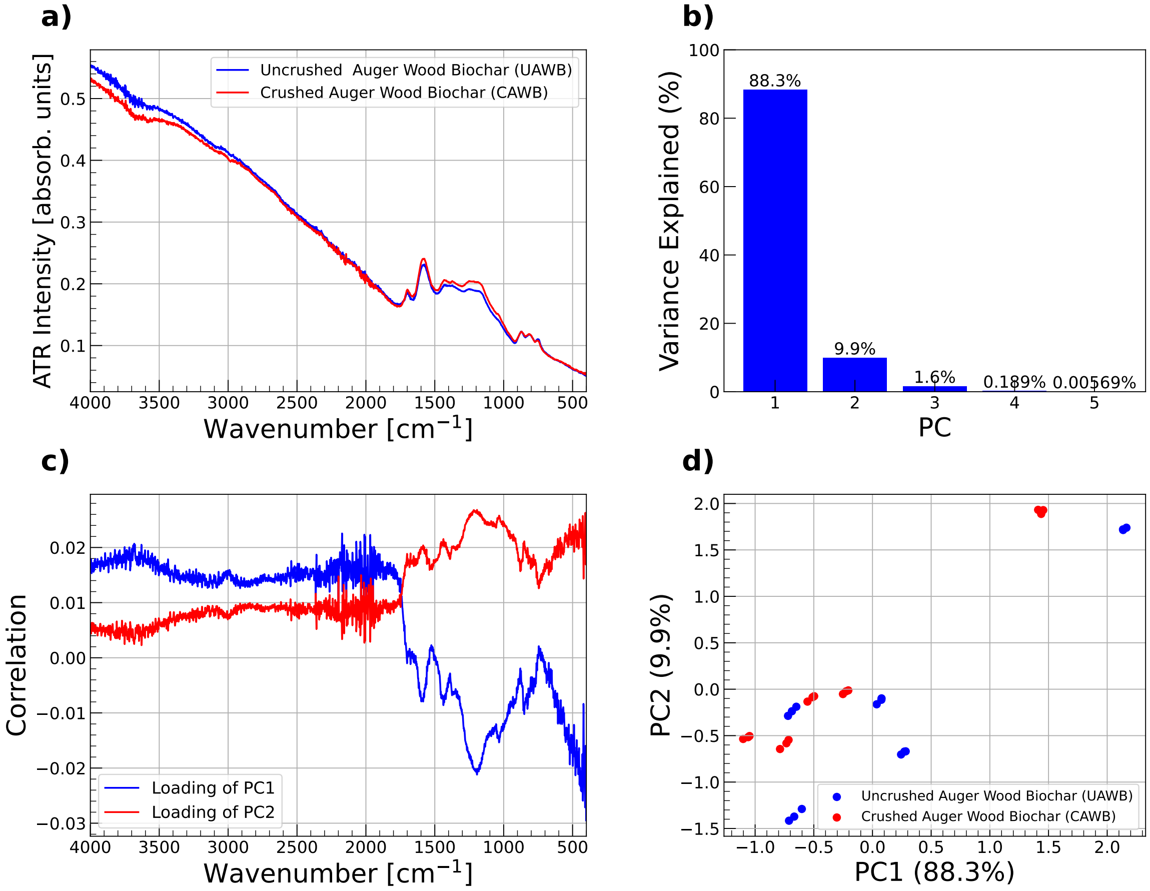

Figure 5.6: **a)** reveals that the FTIR spectra of Auger Wood Biochar samples plotted for ATR intensity against wavenumber from 4000 cm^{-1} to 400 cm^{-1}, whether crushed (CAWB/Red) or not (UAWB/Blue), are nearly identical, with minor differences in intensity observed hear 4000 cm^{-1} and between 1500 cm^{-1} and 1200 cm^{-1}, specifically for the two measurement instances pictured. However, this pattern does not hold for other CAWB and UAWB samples due to variable measurements. **b)** shows percentage variance captured by each principal component (PC). The first two components explain 98.21% of the variance. **c)** presents the loading of the first (blue) and second (red) components plotted against the wavenumber shows both components being positive until around 1750 cm^{-1}. Subsequently, the first component becomes negative while the second component becomes positive, creating a near mirror image. **d)** displays most measurements of the samples confined to negative scores of PC1 and PC2, with outliers residing in positive PC1 and PC2 scores. No clear grouping is evident.

In Fig.5.6 c), the loadings plots appear as mirror images of each other with considerable fluctuations. This observation suggests the possibility that the two loadings could pertain to entirely different samples, which initially seems promising. However, as depicted in Fig.5.6 d), this assumption does not hold true as the samples are mixed in PC space and no clear clusters emerge.

Both loadings exhibit positive values from 4000 to 1600 cm^{-1}, with the first loading displaying a notably stronger positivity. Subsequently, beyond 1600 cm^{-1}, the relationship reverses, with the second loading becoming positive and the first loading becoming negative. This pattern is corroborated by the intensity progression, where the blue line representing CAWB demonstrates higher intensity at higher wavenumbers, followed by an increase in intensity for the UAWB sample.

Due to variations in intensities across measurements, loadings can oscillate between negative and positive values, resulting in Fig. 5.6 c) lacking clear groupings. Some samples of CAWB exhibit negative scores on the first component, while others display positive scores. The first component loadings indicate that samples with negative scores have lower intensities after the C=C bond, whereas those with positive scores exhibit higher intensities at higher wavenumbers, inversely rated for their intensity after the C=C bond, thereby yielding higher scores.

A similar assessment applies to the second component, albeit with all loadings being positive, indicating that the majority of the data is inversely rated on this component. Two outliers exhibit positive scores on both PC1 and PC2. While the implications for the first component remain unclear, for the second component, the positive loadings correspond to positive scores, suggesting significantly higher intensity levels compared to other measurements. It could be hypothesized that the intensity magnitude is substantially higher for the first 2300 cm^{-1} wavenumbers, becoming less negative for subsequent wavenumbers down to 400 cm^{-1}. The scores of the outliers are 2 in both PC1 and PC2, surpassing the highest positive score of approximately 0.25 observed in other measurements on either component, indicating a measurement anomaly attributed to the experimental set-up.

The physical state of the sample fails to show up in intensity as the spectra in Fig. 5.6 a) only differ in the region between 1500 cm^{-1} to 1200 cm^{-1} wavenumber, which is not enough for good PCA results.

Despite the fact that the third component was insignificant, a three-dimensional analysis was carried out, which, as expected, did not yield any meaningful results due to the insignificance of the variance captured by the third component.

Additionally, clustering was attempted and did not reveal any clear evidence of grouping already established from visual analysis. Due to low intensity differences and intensity variability between the two samples, the principal component failed to cluster.

This is an expected result as FTIR captures the chemical qualities of the samples, which do not show in the intensity if the samples are crushed or not. This has to be tested further in better-controlled environments to confirm the assertion. A further analysis of experimental conditions for temperature and exposure time was conducted and can be found in Supplementary Material A.3.

Chapter 6

Discussion

In the process of analysing the data for this thesis, several considerations emerged regarding data treatment and interpretation. Despite efforts, the removal of samples deemed outliers was not achieved, impacting the robustness of the analysis. Additionally, a methodological improvement could involve averaging multiple measurements to enhance data reliability. For instance, adopting an approach such as averaging one-third of the samples in sets of five could provide more robust insights. By increasing the dataset to accommodate more data points, Principal Component Analysis (PCA) could benefit from a richer pool of information, thereby improving the accuracy of categorization and facilitating regression algorithms for predictive modelling.

While our investigation did not directly assess CO_2 adsorption, the PCA analysis unveiled distinct groupings within the experimental data. Although variations in FTIR intensity could potentially correlate with CO_2 adsorption, our study provides the foundational groundwork necessary before attempting such analyses. Future research endeavours could focus on correlating FTIR intensity changes with CO_2 adsorption across different samples, paving the way for classifying materials based on their adsorption capabilities.

The chemical composition analysis of biochar received limited attention in this study. However, an alternative approach could involve structuring the dataset with peak features as variables rather than considering the entire range of wavenumbers. This adjustment would enable grouping based on the chemical bonds present, facilitating the identification of functionalization groups within samples. Such an approach could aid in categorising biochar samples based on their ability to remove specific compounds, thereby enhancing the applicability of the research in environmental remediation efforts.

For spectral PCA, preprocessing techniques such as normalizing spectral intensities to certain (common) peaks could mitigate variability in intensity values across measurements. As demonstrated in Section 5.3, inconsistent physical conditions pose challenges in data grouping. To address this, employing robust preprocessing techniques is essential to enhancing the reliability of results.

Furthermore, expanding the sample size is imperative for conclusive evaluations. Section A.3 underscores the limitations of large sample sizes but low measurement counts in drawing definitive conclusions. Investigating how dimensions change and exploring loadings across a larger sample pool would provide more comprehensive insights into the dataset.

An avenue for future research involves exploring non-standard PCA methods to assess their impact on clustering analysis. By evaluating the efficacy of different PCA techniques, researchers can gain a deeper understanding of the underlying data structures and refine analytical approaches accordingly.

Moreover, investigating the influence of various production conditions, such as temperature and exposure time, on spectral characteristics is vital. Given the demonstrated effectiveness of PCA in clustering chemically modified samples, extending this analysis to include diverse production parameters would enhance understanding.

In summary, this discussion highlights several avenues for refinement and expansion in future research endeavours aimed at enhancing the robustness and applicability of PCA in clustering biochar spectral data.

Chapter 7

Conclusion

In conclusion, Principal Component Analysis (PCA) emerges as a powerful tool for clustering spectra characterised by distinct intensity variations and prominent peaks. Particularly in scenarios where spectra exhibit significant intensity discrepancies, PCA effectively clusters data based on overall intensity differences, proving especially useful in smaller systems comprising a limited number of samples. In such cases, small systems often collapse into one dimension, which can subsequently be expanded to higher dimensions, facilitating effective grouping.

The application of PCA in this study confirmed itself in differentiating spectral variations resulting from different biochar production methods. By reducing the dimensionality of the data, PCA successfully distinguished between spectra associated with various production techniques, thereby enabling an insightful comparison.

However, it is important to acknowledge the limitations of PCA, particularly in cases where spectra exhibit similarities in intensity across chemical structures, as observed in the comparison between crushed and non-crushed samples. In such instances, PCA's effectiveness decreases, highlighting the need for alternative analytical approaches.

Moreover, the inconclusive results obtained highlight the necessity for further exploration of PCA's applicability in accommodating experimental conditions beyond chemical modifications. Factors such as temperature, exposure, and physical variables need to be investigated to discover whether PCA can effectively cluster spectra for these modifications.

In essence, while PCA excels at grouping spectra based on chemical and spectral disparities, its limitations become apparent when attempting to cluster samples produced with minor chemical modifications. As such, future research endeavors should aim to explain the boundaries of PCA's effectiveness and explore supplementary methodologies to address its inherent limitations, thereby advancing our ability to cluster spectral data for a diverse number of samples.

Bibliography

[1] Chem LibreTexts: How an FTIR Spectrometer Operates. https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Spectroscopy/Vibrational_Spectroscopy/Infrared_Spectroscopy/How_an_FTIR_Spectrometer_Operates. Accessed: April 1, 2024.

[2] Chem LibreTexts: Infrared: Interpretation. https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_%28Physical_and_Theoretical_Chemistry%29/Spectroscopy/Vibrational_Spectroscopy/Infrared_Spectroscopy/Infrared:_Interpretation. Accessed: April 5, 2024.

[3] PCA. https://orange3.readthedocs.io/projects/orange-visual-programming/en/master/widgets/unsupervised/PCA.html. Accessed: April 7, 2024.

[4] J. M. Andrade, P. Fresco, S. Muniategui, and D. Prada. Comparison of oil spillages using mid-IR indexes and 3-way procrustes rotation, matrix-augmented principal components analysis and parallel factor analysis. *Talanta*, 77(2):863–869, 2008.

[5] A.G. Barnston and R.E. Livezey. Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Monthly Weather Review*, 115(6):1083 – 1126, 1987. Cited by: 2956; All Open Access, Bronze Open Access.

[6] Zhang C., Ji Y., Li C., Sun S., Xu Y., Jiang L., and Wu C. The application of biochar for CO2 capture: Influence of biochar preparation and CO2 capture reactors. *Ind Eng Chem Res*, 2023.

[7] Anupam Chakraborty, Saida Parveen, Dipak Kr. Chanda, and Gautam Aditya. An insight into the structure, composition and hardness of a biological material: the shell of freshwater mussels. *RSC Adv.*, 10:29543–29554, 2020.

[8] scikit-learn Contributors. The Iris Dataset. https://scikit-learn.org/stable/auto$_{e}xamples/datasets/plot_{i}ris_{d}ataset.html. \quad Accessed : April7, 2024$.

[9] Aastha Dutta. Chapter 4 - Fourier Transform Infrared Spectroscopy. In Sabu Thomas, Raju Thomas, Ajesh K. Zachariah, and Raghvendra Kumar Mishra, editors, *Spectroscopic Methods for Nanomaterials Characterization*, Micro and Nano Technologies, pages 73–93. Elsevier, 2017.

[10] Xiapu Gai, Hongyuan Wang, Jian Liu, Limei Zhai, Shen Liu, Tianzhi Ren, and Hongbin Liu. Effects of feedstock and pyrolysis temperature on biochar adsorption of ammonium and nitrate. *PLoS ONE*, 9(12), 2014. Cited by: 446; All Open Access, Gold Open Access, Green Open Access.

[11] Michael Greenacre, Patrick J. F. Groenen, Trevor Hastie, Alfonso Iodice D'Enza, Angelos Markos, and Elena Tuzhilina. Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100, 12 2022.

[12] Gengxin Hao, Yanyu Hu, Linfan Shi, Jun Chen, Aixiu Cui, Wuyin Weng, and Kazufumi Osako. Physicochemical characteristics of chitosan from swimming crab (Portunus trituberculatus) shells prepared by subcritical water pretreatment. *Scientific Reports*, 11, 01 2021.

[13] Amy F. Iezzoni and Marvin P. Pritts. Applications of principal component analysis to horticultural research. *Hortscience*, 26:334–338, 1991.

[14] Ian Joliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *The Royal Society Publishing*, 2016.

[15] A. Krutof, H. Bamdad, K.A. Hawboldt, and S. MacQuarrie. Co-pyrolysis of softwood with waste mussel shells: Biochar analysis. *Fuel*, 282:118792, 2020.

[16] Shou-Heng Liu and Yi-Yang Huang. Valorization of coffee grounds to biochar-derived adsorbents for CO2 adsorption. *Journal of Cleaner Production*, 175:354 – 360, 2018. Cited by: 91.

[17] Barbara Myszka, Martina Schüssler, Katrin Hurle, Benedikt Demmert, Rainer Detsch, Aldo R. Boccaccini, and Stephan E. Wolf. Phase-specific bioactivity and altered Ostwald ripening pathways of calcium carbonate polymorphs in simulated body. *RSC Advances*, 2019. Electronic Supplementary Information (ESI).

[18] Muqing Qiu, Lijie Liu, Qian Ling, Yawen Cai, Shujun Yu, Shuqin Wang, Dong Fu, Baowei Hu, and Xiangke Wang. Biochar for the removal of contaminants from soil and water: a review. *Biochar*, 4:1–25, 2022.

[19] Anushka Upamali Rajapaksha, Season S. Chen, Daniel C.W. Tsang, Ming Zhang, Meththika Vithanage, Sanchita Mandal, Bin Gao, Nanthi S. Bolan, and Yong Sik Ok. Engineered/designer biochar for contaminant removal/immobilization from soil and water: Potential and implication of biochar modification. *Chemosphere*, 148:276 – 291, 2016. Cited by: 945.

[20] Gurwinder Singh, In Young Kim, Kripal S. Lakhi, Prashant Srivastava, Ravi Naidu, and Ajayan Vinu. Single step synthesis of activated bio-carbons with a high surface area and their excellent CO2 adsorption capacity. *Carbon*, 116:448 – 455, 2017. Cited by: 245.

[21] Fred Steive and Carrie Donley. Introduction to x-ray photoelectron spectroscopy. *Journal of Vacuum Science Technology A*, 2020.

[22] Stephen Weiner. *Infrared Spectroscopy in Archaeology*, page 275–316. Cambridge University Press, 2010.

[23] P.R. Yaashikaa, P. Senthil Kumar, Sunita Varjani, and A. Saravanan. A critical review on the biochar production techniques, characterization, stability and applications for circular bioeconomy. *Biotechnology Reports*, 28:e00570, 2020.

[24] Xiong Zhang, Shihong Zhang, Haiping Yang, Ye Feng, Yingquan Chen, Xianhua Wang, and Hanping Chen. Nitrogen enriched biochar modified by high temperature CO2-ammonia treatment: Characterization and adsorption of CO2. *Chemical Engineering Journal*, 257:20 – 27, 2014. Cited by: 183.

Appendix A

Supplementary Material

A.1 Intensity

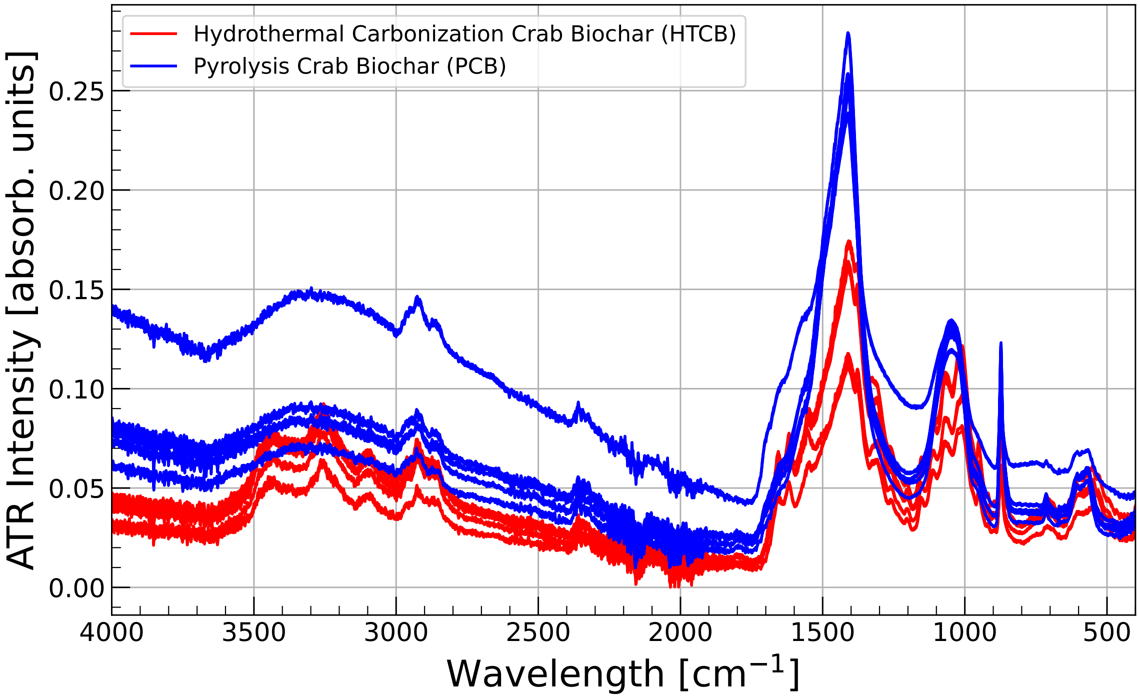

Figure A.1: The figure displays spectra measurements of HTCB (red) and PCB (blue) samples plotted against intensity and wavenumber. It is evident that despite rouge outliers the average spectra intensity of PCB is higher compared to HTCB, and PCB exhibits sharper and smoother peaks.

To aid in understanding the intensity variations between measurements of the same sample, all the measurements were plotted for samples of hydrothermal carbonisation crab biochar and paralysis biochar. The intensity analysis in Fig. A.1 confirms the previously stated assertion that the spectra considered in Section 5.2 is in fact an outlier with a greater intensity, but the plot also serves to show that intensity is generally higher for PCB measurements than HTCB.

A.2 Outliers

To implement K-means clustering, an extension to the analysis discussed in Section 5.1.3, outlier removal methods were explored to enhance the reliability of the results. As observed in the preceding section, the variability in intensity across measurements can introduce errors in the principal component analysis (PCA), necessitating the removal of certain measurements using established outlier detection techniques.

K-means clustering is a popular unsupervised machine learning technique used for partitioning a dataset into a predetermined number of clusters, where each data point belongs to the cluster with the nearest mean. In the context of this study, K-means clustering was employed to group similar samples together based on their principal component scores, thereby facilitating the identification of distinct clusters within the dataset.

Among the outlier removal methods, Mahalanobis distances emerged as the most effective approach, primarily due to their ability to account for the covariance structure of the data.

In the context of outlier detection, Mahalanobis distance measures the deviation of an observation from the mean distribution in terms of the number of standard deviations. Mathematically, it is defined as:

$$D(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}$$

Here, $\mathbf{x}$ represents the position of the sample in the principal component (PC) space, $\boldsymbol{\mu}$ denotes the mean vector of the sample distribution, $\boldsymbol{\Sigma}$ is the covariance matrix of the sample distribution, and $(\mathbf{x} - \boldsymbol{\mu})^T$ represents the transpose of the difference between $\mathbf{x}$ and $\boldsymbol{\mu}$.

In the approach utilized, Mahalanobis distances were computed for each observation in the dataset based on their principal component scores (denoted as `pc1_scores` and `pc2_scores`), which serve as the coordinates in the PC space. Subsequently, the covariance matrix and mean vector of the data were calculated, and the aforementioned formula was applied to each observation.

The thresholds for the Mahalanobis distances were then calibrated between 0.50 and 0.95 to remove trivial outliers, with the calibration process involving the determination of effective quantiles.

The Mahalanobis distances were computed for individual sample groups, enabling outlier removal to be performed within the distribution of each unique sample. This approach helps avoid the removal of entire sample groups during collective analysis, thereby yielding more accurate results.

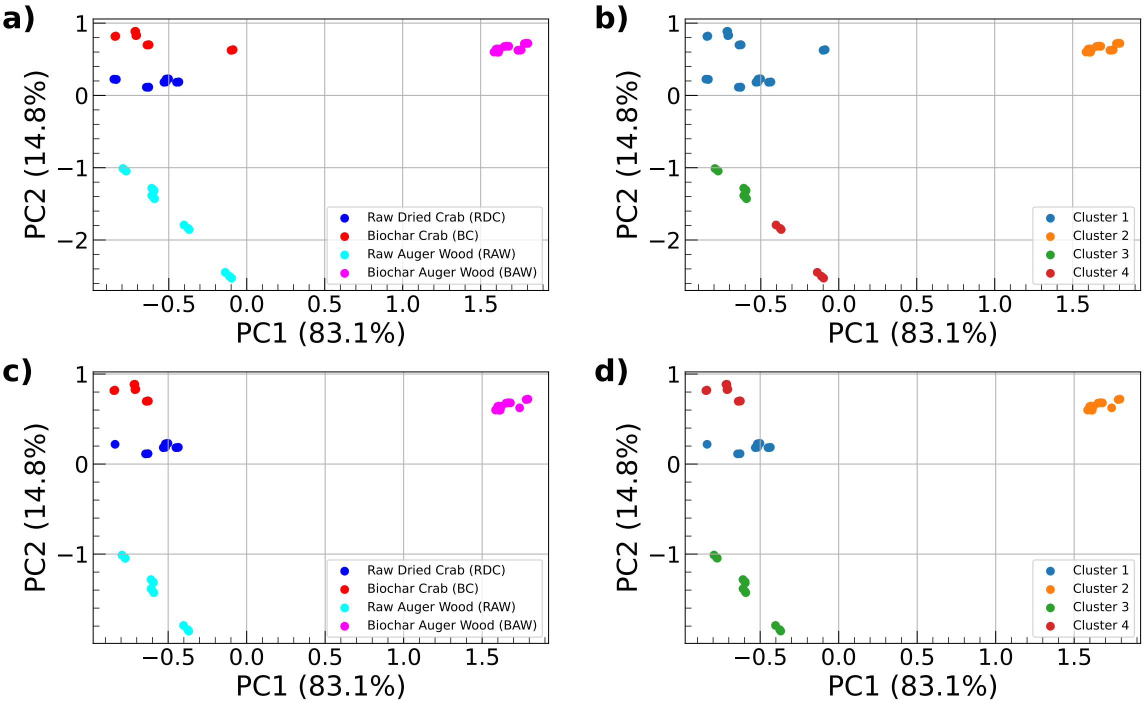

Figure A.2: **a)** Subplot shows the original distribution of Raw Dried Crab (blue), Biochar Crab (red), Raw Auger (cyan), and Biochar Auger (magenta) which are grouped well besides the outliers present in Biochar Crab samples. **b)** Subplot shows the K-Mean clustering applied to the original distribution of the samples which fails to identify the four groups. **c)** Subplot shows the distribution with at least three outliers removed from each group for the samples. The samples are now grouped better with the Raw Auger sample spread out more. **d)** Subplot shows the K-Mean clustering performed for the new distribution without the outliers. The algorithm successfully manages to cluster the four samples.

This analysis greatly improves the number of measurements done per sample as the distribution becomes more defined. This is to say that the grouping analysis done visually is also valid numerically and can be applied to further studies.

A.3 Transmittance Geometry FTIR temperature and exposure time analysis

Although the FTIR spectra considered in this section have a different geometry and record transmission rather than absorption, it is still possible to conduct a PC analysis. The data was collected for hydrothermal crab biochar (HTCBT) made from crab shells that underwent hydrothermal carbonisation at different temperatures and exposures at varying times. The dataset includes clear records of the temperatures and exposures at which the samples were created. This provides an opportunity to determine if PC analysis is effective in identifying small differences in spectroscopy data and relating them to the experimental condition of the samples.

An initial analysis from Fig. A.3 shows that three components accounted for 98.48% of the overall variance. The first component explained 74.39%, the second component was 17.01%, the third component was 7.08% and the rest of the components explained the remaining 1.52%. A two-dimensional or three-dimensional analysis is required to analyse this problem. A three-dimensional analysis was not included due to time limitations.

Fig. A.4 a) shows the FTIR for transmissions. Since the sample is made up of crab shell, we expect to see the aragonite featured in Section 5.1.2 depression characteristic for calcium carbonate. Moreover, the peak that corresponds to water at 3500 cm^{-1} is associated with the hydrothermal carbonisation preparation method of the biochar. Some samples have much higher transmittance dips than other samples, but the spectra follow the general trend with transmittance variations except for the low wavenumbers of $1000 - 400$ cm^{-1} where variations are most visible.

In Fig. A.4 b), the loadings exhibit simultaneous positive and negative values with significant swings across a range of wavenumbers. The first component loadings display positive values for the peak at 1500 cm^{-1} and the peak at 1000 cm^{-1} in the FTIR spectra. However, negative values are observed for wavenumbers between 2600

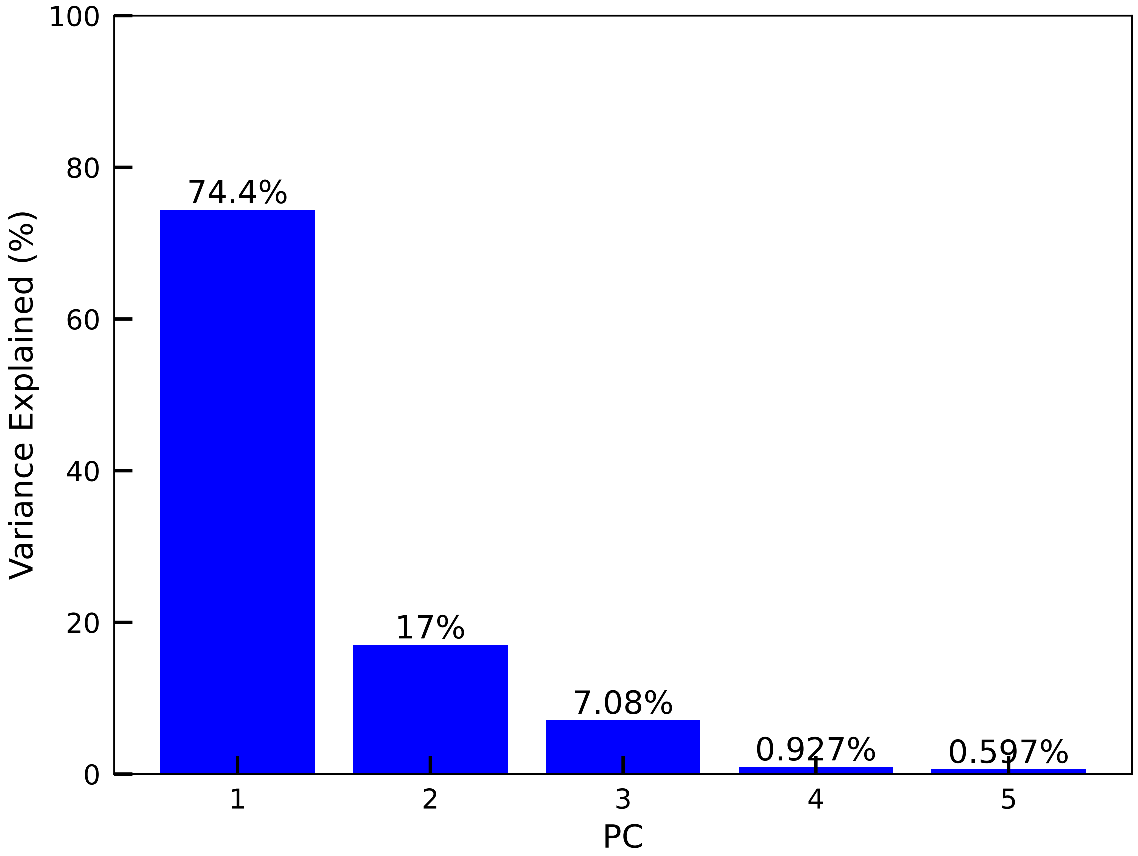

Figure A.3: The plot shows percentage variance explained for the Hydrothermal Crab Biochar dataset by each principal components (PC). The total variance is explained by the first three components with 74.39%, 17.01%, and 7.08% respectively.

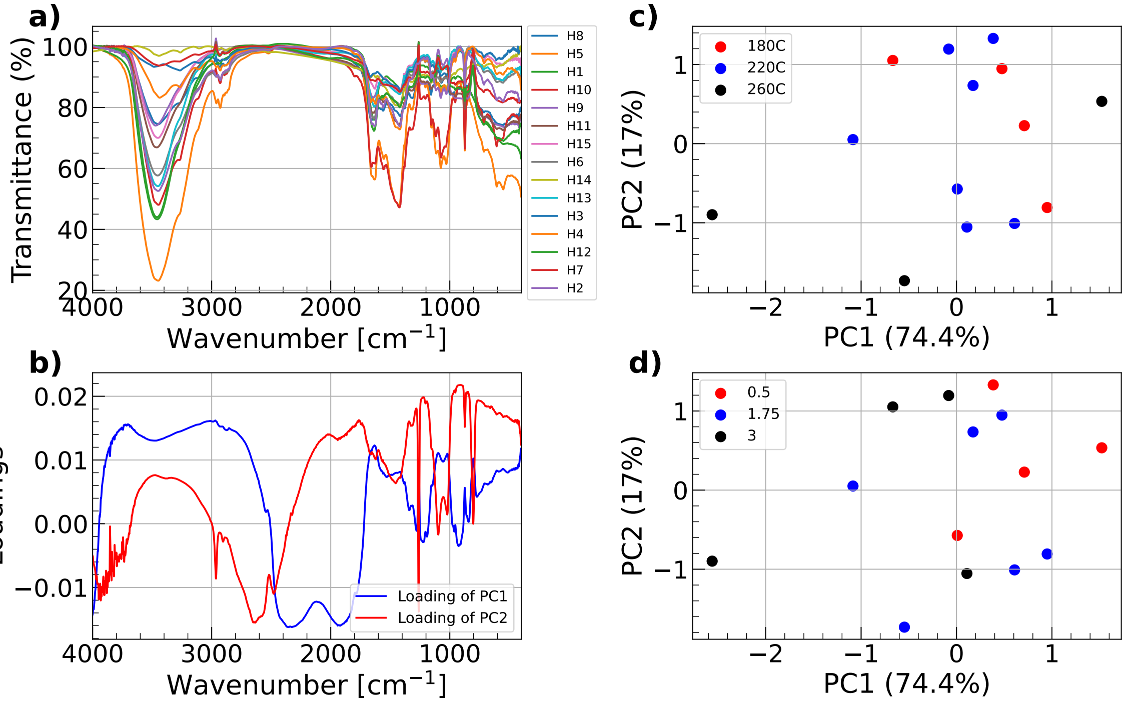

Figure A.4: **a)** displays the original transmission FTIR data obtained for fifteen samples of HTCBT plotted against wavenumber, ranging from 4000 cm^{-1} to 400 cm^{-1}. While the spectra are largely similar, noticeable differences in intensities are observed, particularly in the dips around 3500 cm^{-1}, 1500 cm^{-1}, 1000 cm^{-1}, and closer to 400 cm^{-1}. **b)** illustrating the loadings of the first (blue) and second (red) components, which exhibit extreme oscillations. The loading of the first component displays positive values for identified dips in the spectra and negative loadings in regions between them. Meanwhile, the second component demonstrates slightly less positive loadings for the first dip of the spectra, followed by positive values for the near 100% intensity between 2900 cm^{-1} and 2000 cm^{-1}. It then oscillates between large and small, or even slightly negative, values in between the dominant dips at 1500 cm^{-1} and 1000 cm^{-1}. **c)** illustrates the scores for samples at three different temperatures. Most samples exhibit positive scores with respect to PC1 and PC2, except for a few samples at 260 °C. No trivial clusters can be identified. **d)** displaying scores plots for the three different exposure time with the same distribution as the temperature due to just variable relabeling associated with the HTCBT production conditions. Exposure time didn't enhance the ability to identify clear groups either.

and 1900 cm^{-1}.

Similarly, the second component also displays positive values, albeit slightly smaller for the water peak, immediately followed by negative loadings for the spectral feature at 3000 cm^{-1}. Subsequently, the loadings remain negative for peaks that exhibit positive values for the first loading, sharply transitioning to positive values after 1000 cm^{-1}.

The scores plot for different temperatures in Fig. A.4 c) depicts scattered samples across the space of the first two components. A positive score on the first component indicates low transmittance for the water peak as well as subsequent peaks at 1500 cm^{-1} and 1000 cm^{-1}. Similarly, a positive score on the second component suggests low transmittance for the water peak, albeit with less pronounced features near 1300 cm^{-1}, 1100 cm^{-1}, and the end of the spectrum near 400 cm^{-1}. Some samples produced at 260 °C exhibit negative scores on both components, suggesting a nuanced behaviour that is challenging to quantify. A subset of samples produced at 220 °C displays a positive score on the first component and a negative score on the second component, indicating low overall transmittance without the behavior near 3000 cm^{-1}.

Similarly, Fig. A.4 d) demonstrates that exposure times independently do not yield distinct clustering. The distribution is a copy of the previous plot because multiple variables are associated with the same sample. It remains unclear which variable defines the distribution in PC space, indicating that a combination of production conditions is necessary for better groupings.

Since the samples have temperature and exposure time recorded together, I did not investigate PCA for a combination of these parameters. Moreover, the ratio of water to biomass was skipped entirely in the analysis but is a further parameter that must be included.

The above finding suggests that the PCA is not effective in grouping temperature and exposure time independently for biochar production. Due to the low number of samples, it isn't clear what effect temperature and exposure time play in altering the chemical properties of crab shell biochar and hence the transmittance.

PCA is effective at determining whether transmittance produces results with clear clusters; however, the impact of temperature and exposure time on transmittance and other pre- and post-production treatments need to be subjected to further analysis to determine PCA reliability for these conditions.