Application of Principal Component Analysis for the Data Mining of High Resolution Mass Spectrometry Datasets

by

Yuan Chen

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Environmental Engineering

Department of Civil and Environmental Engineering
University of Alberta

Abstract

The release of oil sands process-affected water (OSPW) from tailings ponds is a major environmental issue that oil sands companies must consider over the next decade. Advanced oxidation processes (AOPs) and biological treatment processes have been shown to be able to degrade contaminant compounds and reduce toxicity of this OSPW. However, these processes are also associated with by-products generation which may be of environmental concern. This study successfully combined High Resolution Mass Spectrometry (HRMS) as an analytical tool to detect organic compounds (markers) in OSPW samples and Principal Component Analysis (PCA) as a statistical tool to manage the extensive HRMS datasets (over 1000 markers per sample). The HRMS and PCA were used to determine the markers most significantly changed during ozonation in different conditions and biological treatment processes and to determine their potential by-products. Based on *m/z* values, all the significant markers selected by PCA were designated into groups including naphthenic acids (NAs), oxidized NAs and unknown compounds. Of these markers, the main focus in this study was the unknown compounds given their trends in OSPW treatment processes have not been assessed previously. The significant unknown markers which decreased over treatment time were degraded during treatments; while those which increased over time were by-products of organic compounds found in raw OSPW treated by ozonation in different conditions and biological treatment processes. There were negligible or very low concentrations of compounds which were identified as by-products in ozonation in different conditions and biological treatment processes found in different raw OSPWs (Syncrude West in Pit, Suncor Pond 7 and CNRL OSPW). These compounds in raw OSPWs showed negative correlations to NAs concentrations and positive correlations to oxidized NAs concentrations, which indicate their close association with NAs degradation via oxidation. This study demonstrates an advanced approach to determining by-products that can be further used for any chemical or biological treatment process. Further research aimed at the identification of by-products structures and determination of potential degradation mechanisms will be useful in assessing treatment efficiency of OSPW compounds.

Acknowledgements

I would like to express my sincere appreciation to my supervisors: Dr. Mohamed Gamal El-Din and Dr. Leónidas Pérez-Estrada for their guidance, support and encouragement throughout my research project. I appreciate Dr. Leónidas Pérez-Estrada for sharing his experimental information with me for my thesis project

My sincere appreciation goes to our research group member MD Shahinoor Islam for sharing his experimental information with me for my thesis work, as well as the technicians from our research group Maria Demeter and especially Nian Sun who helped me for HRMS samples injections.

My sincere appreciation extends to the post-doctoral fellows of our research group, Dr. Kerry McPhedran for his critical and constructive manuscript review, as well as Dr. Pamela Chelme-Ayala for her support during my research. I greatly appreciate their valuable help.

I would also like to acknowledge the technique support from Waters Inc. for their field training and recommendations on the software application.

Finally, I am deeply thankful to my family and friends for their support, encouragement and sacrifice over the years.

Table of Contents

List of Tables

List of Figures

Abbreviations & Symbols

$\angle \alpha$	angle alpha
$^{\bullet}OH$	hydroxyl radicals
AOPs	advanced oxidation processes
BOD	biological oxygen demand
CO_3^{2-}	carbonate
COD	chemical oxygen demand
Da	daltons
DBE	double bonds equivalent
DO	dissolved oxygen
Fe (II) or Fe^{2+}	ferrous iron
Fe (III) or Fe^{3+}	ferric iron
Fe	zero-valent iron
Fe_3O_4	magnetite
FWHM	full width at half maximum
H_2O_2	hydrogen peroxide
HRMS	high resolution mass spectrometry
ISTD	internal standard
m/z	exact mass to charge ratio
MVA	multivariate analysis
NA+Ox	oxidized naphthenic acids
NAs	naphthenic acids
NIPALS	non-linear iterative partial least squares
O_3	ozone
OSPW	oil sands process-affected water
PAHs	polycyclic aromatic hydrocarbons
PC1, PC2	principal component 1 & 2
PCA	principal component analysis
PPM	part per million
R	resolution
SD	standard deviation

SNV	standard normal variate
SVD	single value decomposition
TBA	*tert*-butyl-alcohol
TDS	total dissolved solids
TKN	total Kjeldahl nitrogen
TNM	tetranitromethane
TOC	total organic carbon
ToF	time-of-flight
TSS	total suspended solids
UPLC-HRMS	ultra performance liquid chromatography - high resolution mass spectra
UV	ultraviolet
V1, V2	variable 1 & 2
α-Fe_2O_3	hematite
Δm	mass in difference

1.0 Introduction

This chapter will give a brief introduction to the background of oil sands process-affected water (OSPW) with related environmental issues and identified knowledge gaps. Secondly, the tools applied in this project will be introduced, including the High Resolution Mass Spectrometry (HRMS) technique that was applied to detect organic compounds in the samples, and Principal Component Analysis (PCA) which was applied to manage HRMS datasets. Finally, the overall research objectives of the project will be defined.

1.1 Oil Sands Process-affected Water Background

Athabasca Basin in northeastern Alberta, Canada, compromising of approximately 169 billion barrels of recoverable bitumen, is one of the largest oil sands deposits in the world (Energy Resources Conservation Board, 2010). Oil sands are a mixture of roughly 6% to 16% bitumen, 1% to 7% water and 80% to 87% mineral solids like sands, clay, and fine silts (Liu et al, 2005). Each year, more than 200 million barrels of crude oil are produced (Corinne, 2010). However, although oil sands production provides huge economic benefits, it is necessary to pay attention to the environmental impact associated with the oil sands mining operations. Water usage and remediation is one of the major environmental concerns. The Clarke caustic hot water extraction method is commonly used by oil sands industries to obtain high commercial value oil products. For each ton of oil sands, approximately 0.6 – 0.7 m^3 of hot water is required to mix with sodium hydroxide to release naphthenates which act as surfactants to help extract bitumen (Brient et al, 1995; Hadwin et al, 2006). Approximately 2 to 4.5 barrels of water are required for producing one barrel of synthetic crude oil. The waste water generated during the extraction process is called oil sands process-affected water (OSPW), which is a complex mixture of suspended solids, salts, inorganic compounds, dissolved organic compounds, and trace metals (Corinne, 2010). Inorganic compounds include calcium, magnesium, sodium, chloride, bicarbonate, sulphate and ammonia (Allen, 2008). Meanwhile the major organic compounds are benzene, toluene, polycyclic aromatic hydrocarbons (PAHs) and naphthenic acids (NAs).

The NAs are a complex mixture of predominantly alkyl-substituted alicyclic carboxylic acids，cycloaliphatic acids, and a small amount of aromatic acids (Cornie, 2010; Allen, 2008). The general formula for NAs is $C_nH_{2n+z}O_2$, where n is the carbon number, and Z represents the hydrogen deficiency because of the ring formation (Brient et al., 1995). OSPW are acutely and chronically toxic to aquatic lives, so a “zero discharge” policy for OSPW has been adopted, and all OSPW to date has been stored in tailings ponds (Brient et al., 1995; Allen, 2008). More than 90% of water demand for surface mining is recycled from settling basins, less than 10% of water demand is required from Athabasca River, and less than 10% of freshwater withdrawn from Athabasca River is returned to the nature. The National Energy Board (2006) reported that about 370 million m^3 of freshwater from the Athabasca River was used for oil sands activities in 2006, and Energy Resources Conservation Board (2010) estimated that approximately 720 million m^3 of OSPW were currently stored in the Athabasca oil sands region. Thus, in order to meet future reclaiming strategies, it is necessary to find efficient remediation methods for the decontamination and detoxification of OSPW for safe release to the environment.

Due to the persistence and toxicity of some chemical compounds, such as NAs, the natural degradation of OSPW by microbial activity is limited (Holowenko et al., 2002). However, previous studies have shown that treatment processes such as coagulation-flocculation, adsorption, advanced oxidation processes (AOPs) and biological treatment are able to remove or degrade OSPW contaminants and to reduce toxicity (Afzal et al., 2012; Perez-Estrada et al., 2011; Pourrezaei et al., 2011). Treatment processes like coagulation-flocculation and adsorption physically remove major organic contaminants (Pourrezaei et al., 2011). Biological treatment and advanced oxidation processes only partially oxidize organic contaminants into lower molecular species rather than completely degrading them into CO_2 and H_2O (Scott et al., 2008; Afzal et al., 2012;

Perez-Estrada et al., 2011). Thus, those lower molecular weight species or reaction intermediates formed during AOPs and biological treatment could be potential contaminants of environmental concern. For example, bromate was reported as a harmful by-product of ozonation (Sohn et al., 2004). Moreover, previous studies have shown that profiles of chemical compounds (e.g., NAs) change after treatment (Scott et al., 2008; Drzewicz et al., 2010; Han et al, 2008), however, until recently there has been limited study reporting on how the organic compounds, other than NAs and oxidized NAs, respond to treatment processes including AOPs and/or biological treatment.

Since many of the organic compounds remaining in OSPW after treatment are unknown, it is necessary to identify those compounds and their corresponding environmental concerns and health effects before subsequent release of OSPW effluents into the environment. Of most concern are those compounds which increased or formed during treatment processes which can potentially be more toxic than their parent compounds. In order to fill this knowledge gap, this project is the first step to determine those chemical compounds, other than NAs or oxidized NAs, which are increased or formed in OSPW during AOPs and/or biological treatment processes.

1.2 Characterization of Organic Compounds in OSPW

1.2.1 High Resolution Mass Spectrometry (HRMS)

Mass Spectrometry as an analytic tool has been widely used for both quantitative and qualitative applications (El-Aneed et al., 2009). In order to detect chemical compounds with low concentrations in OSPW, more sensitive instruments and techniques, such as High Resolution Mass Spectrometry (HRMS), have recently been applied to detect and quantify the organic compounds such as NAs in OSPW (Drzewicz et al., 2010; Matthew et al., 2012). In this project, Ultra Performance Liquid Chromatography-High Resolution Mass Spectra (UPLC-HRMS) from Waters Inc. was applied to analyze the organic contents in OSPW samples. The Mass Spectrometry technique electrically isolates molecules based on their exact mass-to-charge ratio (*m/z*) by converting molecules into either positively or negatively charged ions. In a typical mass spectrum, the x-axis is the *m/z* value, and the y-axis represents total ion counts (El-Aneed et al., 2009).

The electrospray ionization mode was used in this project. Literal to the name of "electrospray", injected samples are dissolved in solvent which enters the mass spectrometer in the form of spray. When the sample goes through a charged capillary, charged droplets are formed. With the stream of nitrogen, the charged droplets will further evaporate, and Coulomb explosions will break the charged droplets into smaller species. As a result, either ions will be desorbed from the surface based on the ion evaporation theory, or solvent will be completely evaporated as speculated by the charge residue theory (El-Aneed et al., 2009).

After ionization, the ions will be separated based on their *m/z* values in the mass analyzer (El-Aneed et al., 2009). The Time-of-Flight (ToF) analyzer was used in this project which separates the charged species based on their mobility through a strong electric field drift column. When the mobile phase (moving liquid or gas) mixture passes through the column, the stationary material phase (liquid or solid) influences the mobility of different compounds in the mixture (Skoog et al., 1998). Generally, all ions entering the column have the same kinetic energy, so ions with higher *m/z* ratios move slower through the column than those ions with lower *m/z* ratios. As a result, ions with different *m/z* ratios reach the detector at different times (El-Aneed et al., 2009) where the signals generated are translated into three dimensional mass spectra with intensity versus time versus *m/z* ratios. Each peak with the intensity greater than the noise level represents one ion with corresponding *m/z* ratio detected at specific time (Kenl, 2003; El-Aneed et al., 2009).

Resolution is defined as the capability of a mass spectrometer to separate *m/z* ratios, as expressed in Equation (1-1), where R is resolution, m is the normal mass of the first peak, and Δm is the mass difference between two adjacent peaks.

$$\boldsymbol{R = \frac{m}{\Delta \mathrm{m}}} \text{ ....} \textbf{Equation 1- 1}$$

(Adapted from Skoog et al., 1998)

If the height of the valley between two peaks is less than the critical height (usually 10% of the height), two peaks are considered as reliably separated. For example, to resolve two peaks with m/z ratios of 100.0 and 100.1, the spectrometer has to have a resolution of 1,000 (Skoog et al., 1998). For the instrument applied in this project, the observed resolution power was 400,000 FWHM (full width at half maximum) (Waters Inc., 2010).

The analysis was full scan in this project, so all fragments formed in the ion source were detected. The information gathered from the HRMS technique using full scan was very large, so appropriate multivariate analysis had to be applied to manage and extract the useful information from the datasets.

1.2.2 HRMS Data Mining by Multivariate Analysis

The concentrations of OSPW organic compounds vary between samples, treatments and from different sites. Analysis of these compounds creates extensive information datasets as detected by HRMS. Therefore, the study of the variations of OSPW organic compounds has to be done with a holistic approach. Multivariate Analysis (MVA) such as Principal Component Analysis (PCA) is a tool to identify which chemical compound or classes of compounds are most significant (e.g., change the most) under certain treatment conditions (CAMO, 2011).

Generally, a multivariate data matrix can be plotted into a multi-dimensional plot where one dimension represents one variable. The objective of a PCA is to transfer that sample into a new plan with fewer dimensions formed by principal components. For example, samples with three variables were plotted into three dimensions in Figure 1-1. The PCA finds the first principal component which explains the most variance in the original data, and second component will explain the second largest variance, and the remaining components will explain less variance. Typically, only the first two or three components are selected since the remaining components are not significant, so the multivariate data matrix can be reduced into lower dimensions. As shown in Figure 1-1, samples with three variables were transferred into a new plan formed by principal components (PC1 and PC2), so three dimensions (variables) were reduced into two dimensions (components). The new plot formed by principal components is called the PCA score plot. Score is defined as the distance from the sample on PC1 and PC2 axis to the score plot origin. For instance, PC1 score value is the horizontal distance from the origin to the sample on PC1 axis, and PC2 score value is the vertical distance on the PC2 axis from the sample to the origin as shown in Figure 1-1. In the score plot, the closer the distance between samples indicates greater similarities between the samples, and the greater the distance reflects larger differences between the samples. Thus, samples can be clustered into groups based on their relative distances in the score plot, and the similarities or differences between samples can be determined based on these distances.

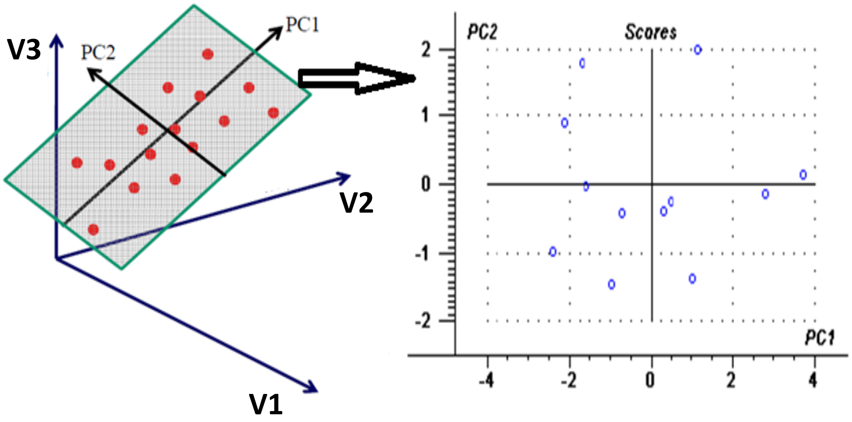


Figure 1- 1: Projection of Original Variables into PCA Score Plot, where V1, V2 and V3 = variable 1, 2 and 3; PC1 and PC2 = Principal Component 1 and 2 (revised from CAMO, 2011).

Given the score plot used to describe samples, the PCA also provides a loading plot to study sample variables. After graphing principal component in the original plot, there will be an angle (∠α) formed between the principal component and the variable as shown in Figure 1-2 which was generated after 90 degree clockwise rotation of Figure 1-1. Theoretically, the variable is more important to the corresponding principal component if the angle between the variable and the component is smaller. Figure 1-2 shows the angle α formed between component PC1 and variable V2. The loading of V2 on PC1 can be calculated as the cosine of angle α, with smaller angles giving larger cosine value, therefore the variable with larger loading value is more important to the corresponding principal component. Similarly, the loading value of each variable on each principal component can be calculated, and PCA loading plot as shown in Figure 1-3b can be generated based on the variable loading values on each principal component. Variables with higher PC1 or PC2 loading values are more significant, and are the major variables describing the differences between samples.

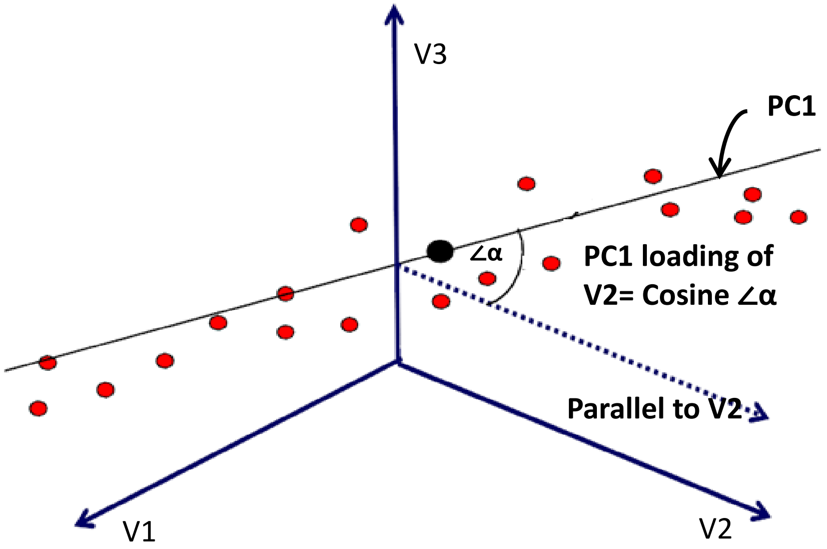


Figure 1- 2: Interpretation of PCA Loadings, where PC1=principal component 1; V1, V2 and V3=variable 1, 2 and 3 (revised from CAMO, 2011).

The key to a PCA study is to combine the score and loading plots, allowing correlations between samples and variables to be interpreted. Theoretically, the variables with more positive PC1 loadings are more positively correlated to samples with more positive PC1 scores, and anti-correlated to samples with negative PC1 scores. Similarly, variables with higher PC2 loadings are more correlated to samples with higher PC2 scores, but anti-correlated to samples with negative PC2 scores. However, PC1 and PC2 are perpendicular, so there is no correlation between those two components (CAMO, 2011). As a result, the variables with higher PC1 loadings may reflect their unique properties in the samples with higher PC1 scores. In contrast, variables with small loading values near the origin are common variables found in the majority of samples and cannot be used to describe differences. Therefore, by combining PCA score and loading plots, a researcher is able to determine which variables are specific to certain samples, and which can be used to describe the differences between other samples (CAMO, 2011; Jolliffe, 2002; Johnson and Wichern, 2007; Sanguansat, 2012). More detail of the PCA theories will be discussed in in Chapter 2.

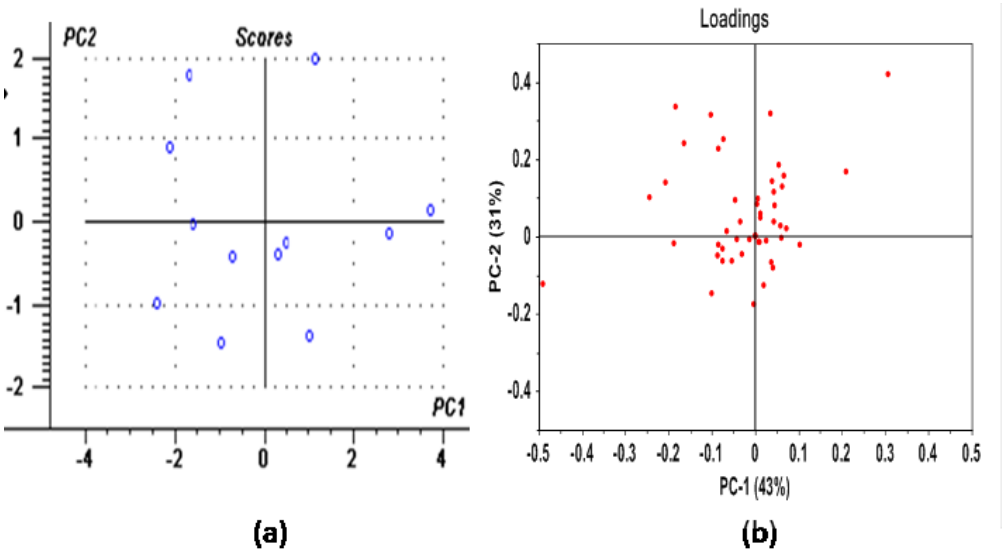


Figure 1- 3: Typical (a) PCA Score and (b) Loading Plots (revised from CAMO, 2011). Score plot (a) showed samples (blue circles); loading plot (b) showed variables (red points) in samples.

1.3 Research Objectives

There is a great need for the development of a high-throughput data analysis method to study the variation of the OSPW organic compounds before and after different treatment processes. The organic compounds of OSPW received from different regions and mining operations sites are variable (Allen, 2008; Pourrezaei et al., 2011), so treatment processes may have altered efficiencies on OSPW from these sites. The general objective of this research is to use Multivariate Analysis (MVA) software to identify compounds decreasing, increasing or forming in OSPWs during various treatment processes and originating from different sites. Specific objectives are:

- Mine the information from high-throughput HRMS datasets coming from OSPW treated by ozonation in different conditions, biological treatment processes, and raw OSPW from different sites samples by using the statistical tool Principal Component Analysis (PCA).

- Classify the datasets information into different organic groups (i.e., NAs, oxidized NAs and unknown compounds) to compare the OSPW samples from different processes and sites.

- Identify which compounds are increasing, decreasing or being formed in OSPW during treatment processes.

2.0 OSPW Treatment Processes and Analyses Review

In this chapter a literature review is presented regarding the different advanced oxidation processes (AOPs) and biological treatment used in the remediation of the OSPW. Additionally, High Resolution Mass Spectrometry (HRMS) as an analysis method for the characterization of the OSPW organic compounds is described. The application of PCA as an analysis method of HRMS datasets is also reviewed.

2.1 Advanced Oxidation Processes (AOPs)

Ozonation involves direct and indirect metabolic pathways for degradation of compounds. The direct pathway, as shown in Equation (2-1), is where ozone reacts directly with organic compounds, but this reaction can be slow and selective. For example, the compounds containing aldehydes and carboxylic acid are not reactive with ozone. Also, direct reactions typically lead to the incomplete oxidation of organic compounds. The indirect pathway, as shown in Equation (2-2) and (2-3), is where the ozone decomposes and forms very reactive hydroxyl radicals, in which R represents the remaining elements in the chemical compound and $\cdot OH$ is the hydroxyl radical. The reaction of hydroxyl radicals with organic compounds is fast, relatively non-selective and typically leads to the complete oxidation of organic compounds to CO_2 (Catalkaya and Kargi, 2009). In other words, the indirect pathway supplements the direct pathway (Nawrocki and Kasprzyk-Hordern, 2010).

$$\boldsymbol{O_3 + R - C \rightarrow R - C = O = C - R + O_2} \text{ ……. } \textbf{Equation 2- 1}$$

$$\boldsymbol{O_3 + OH^- \rightarrow O_2 + \mathrm{H}O_2^-} \text{ ……………………….. } \textbf{Equation 2- 2}$$

$$\boldsymbol{O_3 + HO_2^- \rightarrow O_2 + O_2^{\cdot -} + \cdot OH} \text{ ……………………….. } \textbf{Equation 2- 3}$$

Where R represents the remaining elements in the chemical compound and $\cdot OH$ is the hydroxyl radical (revised from Catalkaya and Kargi, 2009).

The use of ozonation for OSPW treatment has been the subject of many recent studies. Gamal El-Din et al. (2011) reported that during the first five minutes of ozonation of OSPW, 0.1 mg/L/s total acid-extractable organics were degraded, but after that initial time period, the degradation rate was only 0.04 mg/L/s. Thus the change of the degradation rate indicated that fast ozone-reacting compounds would react with ozone directly at the beginning of the reaction followed by the formed hydroxyl radical which was very active and reacted with the remaining compounds.

Perez-Estrada et al (2011) recommended that ozone preferentially degraded the NAs with higher carbon number and more rings. It was explained that more carbons in the molecules indicated that more H atoms were available for $^{\bullet}$OH radicals abstraction. Also, as the number of rings increased, the number of tertiary carbon (three carbon neighbors) also increased in the NAs molecules. The H atoms on the tertiary carbon were found to be more reactive compared with H atoms on the primary carbon (one carbon neighbours) and secondary carbon (two carbon neighbours). However, the mechanisms of degradation of OSPW NAs were difficult to study due to the presence of matrix complexity including dissolved organics and saline matrix in OSPW (Perez-Estrada et al, 2011).

Scott et al. (2008) used ozone to treat the NAs in sediment-free OSPW from the Recycle Water Pond at Syncrude Canada Ltd., Fort McMurray, Alberta, Canada. By using a Seair Diffusion System (Seair Diffusion Inc.), the ozone concentration in the reactor was kept at 35 mg/L. Approximately 70% of NAs were degraded after 50 minutes, and the effluent was considered as non-toxic by Microtox bioassay. While after 130 minutes of ozonation, the hydrocarbon concentration decreased from 19 to 0.6 mg/L, and the residual NAs concentration was only 2 mg/L which was less than 5% of the initial concentration. However, Scott et al. (2008) pointed out that there was no significant change in total organic carbon (TOC decreased from 60 mg/L to 46 mg/L), which indicated that ozone only oxidized NAs into other compounds rather than completely degrading into CO_2. Moreover, 50% of chemical oxygen demand (COD) was decreased and biological oxygen demand (BOD) increased from 2 mg/L to 15 mg/L. As a result, the BOD/COD ratio increased from 0.01 to 0.15 by ozonation, which indicated that the biodegradability of OSPW was improved. Martin et al. (2010) suggested that ozonation was able to accelerate the NAs biodegradation and toxicity removal by indigenous microbes. Basically, ozone selectively oxidized bio-persistent NAs into their oxidized form like hydroxyl- or keto-NAs, and later the microorganisms could easily degrade the oxidized NAs and decrease the OSPW toxicity.

Advanced oxidation processes (AOPs) methods utilize strong oxidants (e.g., hydroxyl radicals) to more efficiently oxidize contaminants by combining a strong oxidizing agent such as hydrogen peroxide (H_2O_2) and ozone (O_3) with catalysts and ultraviolet (UV) irradiation (Tarr, 2003; Parsons, 2004; Robertson, 2010). The chemical reactions of AOPs are essentially the same as ozonation, but the reaction rates are much faster than ozonation alone (Catalkaya and Kargi, 2009).

During photolysis, parent compounds are broken down into small molecular compounds by absorbing light. When photolysis is combined with catalysts such as hydrogen peroxide (H_2O_2), the process is called a photocatalytic AOP. Drzewicz et al. (2010) used vacuum UV irradiation (VUV at 172 nm) and UV irradiation (UV at 254 nm) combined with H_2O_2 to degrade alicyclic carboxylic acid (cyclohexanoic acid (CHA), a model NA compound). After 80 minutes of irradiation with a UV dose of 450 mJ/cm^2 and H_2O_2 concentration of 60 mg/L, 10 mg/L (86% of initial concentration) of CHA were degraded, and undefined by-products were found.

Liang et al. (2011) compared four AOPs (UV/TiO_2, UV/IO_4^-, $UV/S_2O_8^{2-}$ and UV/H_2O_2) to degrade NAs in a model OSPW containing high total dissolved solids (TDS) and total suspended solids (TSS) concentrations. To achieve target residuals of 5 mg/L NAs and 3.4 mg/L TOC, UV/H_2O_2 (50 mM) at pH 8 were found as the optimal conditions, and the $UV/S_2O_8^{2-}$ process was found to leave significant residual sulfate in the water, therefore not recommended.

Catalytic ozonation that utilizes the catalysts such as Fe (II) and Fe (III) to increase the efficiency of ozone decomposition and hydroxyl radical formation at low pH is an AOP, which leads to the faster degradation and more effective mineralization of organic contaminants (Nawrocki and Kasprzylk-Hordern, 2010). For example, Kishimoto and Ueno (2012) investigated the catalytic degradation of 1, 4-dioxane in a synthetic wastewater with zero-valent iron (Fe), ferrous ion (Fe^{2+}), ferric ion (Fe^{3+}), hematite (α-Fe_2O_3) and magnetite (Fe_3O_4). It was observed that Fe accelerated the degradation rate in pH from 3 to 9, but Fe^{2+} and Fe^{3+} only enhanced the degradation at pH of 3. The effects of α-Fe_2O_3 and Fe_3O_4 on the degradation efficiencies were minor. Thus, Kishimoto and Ueno (2012) suggested the reaction of ozone with Fe^{2+}/Fe^{3+} enhanced the hydroxyl radical production to accelerate the degradation rate of contaminants. In addition, the catalytic ozonation was suggested to have the additional advantage of fewer by-products generation (Nawrocki and Kasprzylk-Hordern, 2010).

The presence of scavengers decreases the ozonation efficiencies on the organic contaminants degradation (Nawrocki and Kasprzylk-Hordern, 2010). For example, Qi et al. (2009) achieved 80% degradation of *2,4,6*-trichloroanisole (TCE) by 10 minutes of ozonation with addition of 0.2 g/L alumina at pH 5.8. However, the addition of 10^{-3} M *tert*-butyl alcohol (TBA) as a radical scavenger reduced overall degradation to less than 30% removal. Also, Hiner et al. (2001) suggested that tetranitromethane (TNM) as a strong scavenger of superoxide radical reacted with $O_2^{\cdot -}$ radicals in solution, as 200 μM and 2mM TNM were observed to rise oxygen generation of 50% and 500% respectively. Grebel et al. (2010) evaluated the effects of scavengers such as Cl^-, Br^- and carbonates (H_2CO_3 + HCO_3^- + CO_3^{2-}) on UV/H_2O_2 treatment of phenol in saline water. It was observed that the presence of > 0.2 mM Br^- reduced phenol removal rate by 75%; the maximum scavenging effects reduced phenol removal rate by 35% with the presence of 400 mM Cl^-; and the maximum scavenging effects of carbonates reduced phenol degradation rate by 28% when the carbonate concentration reached > 100 mM.

Therefore, the recent studies suggest that advanced oxidation processes (AOPs) are able to partially degrade the organic contaminants such as NAs in OSPW into oxidized forms and smaller compounds which are more easily biodegradable, and by-products are potentially formed during the AOPs treatment. However, there have been no previous studies focusing on the detection and identification of unknown organic by-products.

2.2 Biological Treatment Processes

Compared with other remediation methods, bioremediation has the advantage of being both low cost and environmental friendly. The NAs of OSPW have been considered as the most toxic component of OSPW, therefore most OSPW studies have focused on NAs biodegradation. Given the toxicity of OSPW and difficulty in separating single compounds from the mixture of OSPW NAs, commercial NAs manufactured as surrogates for NAs in tailings water have been used extensively in the investigation of bioremediation (Whitby, 2010).

Clemente et al. (2004) used microorganisms from Mildred Lake Settling Basin of Syncrude Canada Ltd. to degrade commercial NAs ("Kodak salts") and refined naphthenic acids ("Merichem acids"). Approximately 90% of Kodak NAs in viable cultures were degraded in the first 10 days. The biodegradation of Merichem acids showed similar results, with NAs concentrations in viable cultures decreased from 109 mg/L to 8 mg/L in the first 10 days. Lo et al. (2006) observed that NAs with more rings in the structure were more resistant to microbial degradation, and microorganisms tended to degrade lower molecular weight NAs more readily than higher molecular weight NAs (Clement et al., 2004).

Han et al. (2008) studied the mechanisms of NAs biodegradation, and suggested that the cyclic and alkyl branching prevent biodegradation of NAs, because when the alkyl branching attached to the β carbon, the tertiary or quaternary carbon appeared at α or β position prevent the β-oxidation which is the most common or first mechanism preferred by most microorganisms to degrade aliphatic and alicyclic carboxylic acids. The β oxidation pathway involves the formation of new carboxylic acids with two carbons fewer than parent compounds (Whitby, 2010). Other mechanisms include α-oxidation followed by β-oxidation as a second mechanism, known as combined α- and β-oxidation that further improves the biodegradation of NAs. The third mechanism is the cyclic NAs degradation by aromatization which forms an aromatic intermediate (Han et al., 2008).

Biryukova et al (2007) and Lai et al. (1996) tested factors such as type of microorganisms, temperature, dissolved oxygen (DO), phosphate concentration and type of microorganisms that affected the biodegradation of commercial NAs. It was found that higher temperature, higher DO concentration and higher phosphate concentration provided microorganisms better environmental conditions to grow and degrade more NAs (Lai et al, 1996). Microorganisms from a non-contaminated region had no ability to degrade commercial NAs, but microorganisms from tailings water were able to partially degrade commercial NAs with low molecular weight (Biryukova et al, 2007).

Unlike commercial NAs which contain a substantial fraction of rapidly biodegradable NAs with low molecular weight, the dominant compounds in OSPW are predominantly recalcitrant with multiple branches and rings (Lo et al., 2006; Scott et al., 2005). Thus, the biodegradation of OSPW NAs was found to be much slower and more persistent due to the presence of high molecular weight NAs (Corinne, 2010). In addition, Han et al. (2009) observed that more biodegradation of NAs occurred under aerobic conditions than in anaerobic conditions. Unfortunately, most OSPW tailing ponds are anaerobic, especially in the subsurface, so there would be limited biodegradation processes occurring in the tailing ponds (Holowenko et al., 2001). However, the methane production in Syncrude Canada Ltd's tailing ponds was found to increase in past years. Theoretically, acetate and H_2 for methanogens could be produced by β-oxidation of long-chain carboxylic acids, as proved by Jeris and McCarty (1965) with anaerobic sewage digesters. Holowenko et al. (2001) studied if NAs were substrate to support methanogenesis in the OSPW tailings, but there was no evidence to prove that NAs from OSPW were the direct source of methane production in tailing ponds of Syncrude Canada Ltd. Also, Corrine (2010) pointed out that it was not understood yet if the NAs in the tailing ponds were the substrate in methane biogenesis *in situ*.

In summary, recent studies show that indigenous microorganisms from OSPW contaminated regions are able to easily degrade commercial NAs with relatively less carbons and rings under aerobic conditions. However, the biodegradation of OSPW NAs are slower due to the presence of higher molecular weight NAs and anaerobic conditions in tailing ponds. As reviewed in AOPs literatures, AOPs are able to degrade larger molecular compounds into lower molecular weight compounds which are more readily biodegradable. Thus, AOPs are suggested to be used as a complementary treatment to further biological treatment processes (Martin et al., 2010).

2.3 Review of OSPW Characterization by HRMS

Mass spectrometry as a quantitative and qualitative tool has been widely used in the field of environmental engineering to detect chemical compounds in various waters. In recent years, a few studies have applied HRMS to detect the NAs and oxidized NAs, and to estimate or quantify their concentrations in OSPW. By comparing compound concentrations before and after treatment processes, the overall degradation and formation of compounds during treatments can be determined.

Martin et al. (2010), Gamal El-Din et al. (2011) and Perez-Estrada et al. (2011) successfully quantified and assessed NAs concentrations before and after ozonation of OSPW by using the HRMS technique. Their analysis was done using a Waters ACQUITY UPLC system (Water, MA, USA) and the detection was performed in-line with a high-resolution (7,000-10,000) mass spectrometer equipped with a TurboIon Spray source operating in negative ion mode. NAs with carbon number from 7 to 22 and Z number from 0 to -12 were identified by matching their exact masses to the high-resolution mass measurements, and were further quantified by using relative responses to the internal standard. The OSPW NAs profiles before and after ozonation were plotted into three dimensions including carbon number, Z number and relative response on each axis as shown in Figure 2-1 for comparison. The relative response was considered as being directly proportional to the concentrations. Clearly, the magnitudes of NAs in relative response in fresh OSPW in Figure 2-1(a) were much higher than those in ozonated OSPW in Figure 2-1(b), so it was concluded that NAs were almost completely degraded after ozonation of 80 mg/L of utilized ozone dose (Gamal El-Din et al., 2011). With similar approach, Gamal El-Din et al. (2011) and Perez-Estrada et al. (2011) reported that the percentages of oxidized NAs were increased after ozonation.

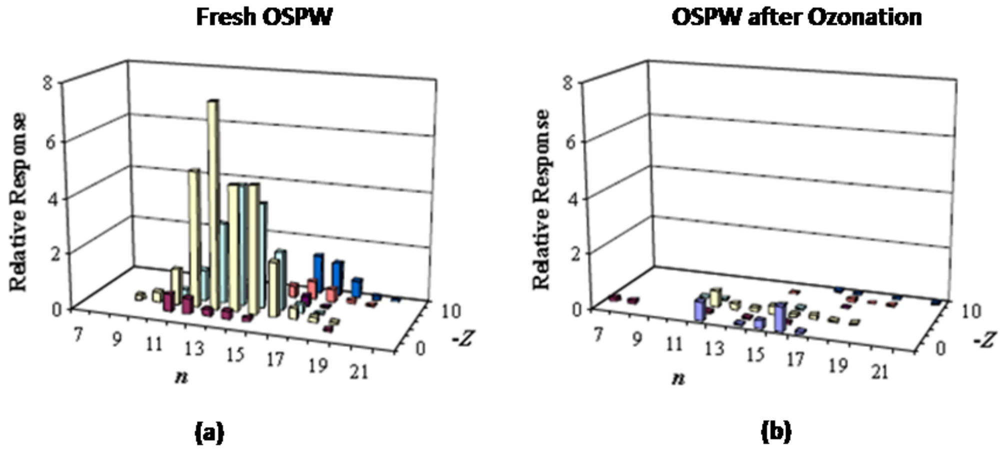


Figure 2- 1: NAs Profiles in (a) Fresh OSPW, and (b) OSPW after Ozonation, where n= carbon number (7 to 22), Z= hydrogen deficiency due to the ring formation (0 to -12) (revised from Gamal El-Din et al., 2011).

Matthew et al. (2012) quantified NAs concentrations in samples collected from the Athabasca River region using a Shimadzu LC 20XR LC system with a time-of-flight high-resolution (about 30,000 at *m/z* 250) mass spectrometer with an electrospray source, operating in negative ionization mode. NAs with carbon number from 7 to 22, and Z number from 0 to -20 were identified by matching their theoretical masses to the high-resolution mass measurements. Matthew et al. (2012) integrated NAs peaks with signal-to-noise ratio greater than or equal to 3:1, and concentrations were calculated based on the calibration curve. With Principal Component Analysis on the NAs quantification results to study the differences between samples collected along Athabasca River region, Matthew et al. (2012) concluded that regional samples with higher NAs concentrations indicated that the regions were potentially contaminated by oil sands mining activities.

2.4 Principal Component Analysis

Multivariate Analysis such as Principal Component Analysis (PCA) is a statistical tool to help manage and extract information from large datasets affected by multiple variables (Sanguansat, 2012). As reviewed previously, Matthew et al. (2012) successfully applied PCA to manage a complex HRMS data matrix to locate NAs contaminated regions. Similarly, PCA was the major strategy applied in this project to manage HRMS datasets, to classify the samples into groups and to extract significant variables.

Table 2- 1: Typical PCA Datasets Structure (revised from CAMO, 2011).

	Variable 1	**Variable 2**	**…**	**Variable n**
Sample 1				
Sample 2				
…				
Sample n				

Since there are many variables that are not significant, it is a challenge to accurately select useful variables from datasets. The purpose of using PCA is to extract useful variables only and remove noise data so that the sample dimensionality will be reduced (Jolliffe, 2002; Johnson and Wichern, 2007; Sanguansat, 2012). Usually, the multivariate datasets for PCA will be organized into a datasets structure or matrix similar to Table 2-1. Each column is one variable, and each row is one observation or sample (CAMO, 2011).

The PCA will transfer the original datasets into new latent components as previously shown in Figure 1-1. Basically, the new latent component or principle component is the linear combination of the original variables. Each component can be expressed in Equation 2-1:

$$\mathbf{PC = b_1\,(X_1) + b_2\,(X_2) + \ldots\ b_p\,(X_p)\ \ldots\ldots\ldots\ldots\ (Equation\ 2\text{-}1)}$$

(Revised from CAMO, 2011)

In Equation 2-1, PC is the principal component; b_p is the regression coefficient for observed variable X_p (CAMO, 2011). Generally, a larger regression coefficient b_p indicates that the corresponding variable more significantly contributes to the component. Technically, the variable loads significantly on the component (Jackson, 1991; SAS, 2012). It is recommended that each component represents at least three variables for a reliable PCA without losing too much of the original information (Jolliffe, 2002, SAS, 2012). As introduced in Chapter 1, the first component PC1 will cover the most variance in original datasets, and second component PC2 will be perpendicular to PC1 and explain the second largest variance. The remaining components will explain declining variances, so they are less significant to the overall PCA model. Typically, the first two to three components are sufficient enough to generate a good PCA model (CAMO, 2011; Jolliffe, 2002).

The basic theory of PCA such as score and loading plots for PCA results interpretations was introduced in Chapter 1 and further the details are discussed in Chapter 4. This section mainly reviews theories not covered in Chapter 1, including general data pre-treatment methods, PCA algorithms, validation methods, and applications in environmental engineering areas will be discussed.

2.4.1 Data Pre-treatment

For a proper PCA, the original datasets generally need to be pretreated by transferring into other scales. Typical pre-treatments include Standard Normal Variate, pareto scaling, and log transformation (Van den Berg et al, 2006). For more accurate analysis of spectra data it is recommended to do a Standard Normal Variate (SNV) transformation. The original datasets will be subtracted from the mean in each dimension and divided by the standard deviation of each dimension. The new dataset will have a mean of zero and standard deviation of one. Such transformation is able to remove multiplicative interferences of scatter and particle size effects from spectral data (CAMO, 2011). Pareto scaling has a similar concept as SNV, but uses the square root of the standard deviation rather than the standard deviation. It also has the advantage of keeping partial relative importance of each variable in the original datasets as compared to SNV. Log transformation is one of the most common pre-treatment methods, because such transformation can set the data into a normal probability if the log scales relationship exists in the original datasets (Van den Berg et al, 2006). Webster (2001) suggested checking the skewness coefficient as the critical value to decide if the transformation of original data was required to get the normal distribution, because the skewed distributed data might fail to properly estimate the model. If the skewness coefficient is greater than 1, log transformation is recommended. For skewness coefficient in the range from 0.5 to 1, square root transformation is suggested. If skewness coefficient is less than 0.5, the original datasets is considered as the normally distributed, so no transformation is required (Webster, 2001).

Overall, the purpose of any transformation is to modify datasets in order to get better PCA results, but each pre-treatment method has its own strength and weaknesses (Van den Berg et al, 2006). For instance, SNV transformation simplifies the data structure with mean value of zero and standard deviation of one, but it assigns equal weights to all variables so that their relative importance in the original data structure is lost. Though pareto scaling has the advantage of keeping partial relative importance of variables, it is very sensitive to large fold (magnitude) change. Log transformation is unable to deal with datasets with relatively large standard deviations and zeroes (Van den Berg, 2006). The application of data pre-treatment should be done carefully, because PCA is sensitive to the data scales and sometimes transformations may make the PCA results worse (Gao et al, 1999; Praveena et al, 2012; Van den Berg et al, 2006).

2.4.2 PCA Algorithm

Non-linear Iterative Partial Least Squares (NIPALS) and Single Value Decomposition (SVD) are two commonly used algorithms for PCA. NIPALS is used when the datasets have missing data and is only accurate to compute the first few components of large datasets because it will accumulate more and more errors with higher components. SVD is appropriate for smaller datasets without missing data. Unlike NIPALS that computes only the first few components, SVD will calculate all components, so such algorithm is time consuming and not appropriate for datasets with either large samples or variables (CAMO, 2011; Jackson, 1991).

2.4.3 PCA Validation

The validation step estimates the uncertainty of the model prediction on new datasets and the model is considered to be valid if the uncertainty is within an acceptable range (Jackson, 1991; Johnson and Wichern, 2007). Validation procedures for PCA involve cross validation and test set validation (CAMO, 2011). Cross validation is the most common method used when there are not enough samples or the selected samples do not have large variations. This validation procedure uses all samples for both calibration and validation steps where some samples are first randomly selected for model calibration and the remaining samples are left out for validation. This process is repeated several times until every sample is left out once, and the uncertainty test results for iterations will be combined together to generate the final cross validation results (CAMO, 2011; Johnson and Wichern, 2007). Alternatively, the test set validation procedure is recommended for datasets with large sample sizes because samples used in the calibration step are not applied in validation step. Thus, the test set validation method will provide a more representative assessment of the model (CAMO, 2011).

2.4.5 Application of PCA in Environmental Engineering

PCA has been widely applied as a statistical tool in environmental engineering research to manage both relatively simple datasets containing sample populations much greater than number of variables, and more complex datasets containing number of variables much greater than the sample populations. The use of HRMS with extensive datasets creates the need for PCA for data mining as evaluation of these datasets without multivariate analysis would be impractical.

With the help of PCA, several studies were performed for water quality monitoring and identification of membrane fouling sources (Helena et al., 2000; Peldszus et al., 2011). For instance, Helena et al. (2000) used PCA to evaluate the water composition of an alluvial aquifer of Pisuerga River in Spain using a 64×16 matrix (64 samples combined with 16 physico-chemical variables), and Helena et al. (2000) found hydro-chemical variables were highly correlated, and PC1 explained 33% of the variance highly correlated to the mineralization which continuously occurred during the survey period. Similarly, Peldszus et al. (2011) successfully applied three principle components to estimate the surface water constituents on reversible and irreversible membrane fouling. It was found that both protein-like substances and particulate/colloidal maters in feed water highly correlated to reversible fouling. However, for the irreversible fouling, protein-like substances were the only significant factor.

PCA has demonstrated a strong ability to extract information from a simple data matrix with sample populations much greater than number of variables. For instance, Parinet et al. (2004) applied PCA to extract four significant variables including pH, conductivity, UV absorbance at 254 nm and permanganate index for raw water from the 2,310 samples with 18 analytical variables, and the first two PCs that explained 62% of original variance were sufficient to accurately describe the trophic state of eutrophic lake systems. Similarly, by applying PCA on datasets of 573 samples with 26 water quality variables, Olsen et al. (2012) used the first two PCs to explain 60% of original data variances, and the significant variables (i.e. chloride, sodium, sulfate and etc.) with loading values greater than 0.75 on both PC1 and PC2 axis were extracted. These results lead to the conclusion that the runoff from fields containing land applied poultry waste and wastewater treatment plant effluent were the potential sources for surface water pollution (Olsen et al., 2012).

PCA has not only been applied to simple datasets, it has been shown to successfully manage complex mass spectrometry data matrix with a number of variables much greater than sample populations to differentiate samples based on their chemical compounds variations. For example, Sleighter et al. (2010) applied

PCA to explore large datasets encountered from Fourier transform ion cyclotron resonance mass spectra of dissolved organic matter (DOM). 38 water samples were collected along a terrestrial to marine transect of lower Chesapeake Bay. Overall, 500 dominant peaks in each sample were selected to represent the characteristics of each sample with each peak assigned a chemical formula based on *m/z* values. After removing duplicates, the matrix was created with 2,143 peaks named by formulas in 38 samples, and their corresponding magnitudes were used for multivariate analysis. Sleighter et al. (2010) applied both hierarchal cluster analysis (HCA) and Principal Component Analysis (PCA) to get similar clustering results, but PCA had the advantage of studying the variables responsible for groupings by assessing the colocation of variables and samples in loading and score plots respectively. Due to the large number of variables in the datasets, PC1 (28%) and PC2 (19%) only covered a total variance of 47% in original datasets. However, Sleighter et al. (2010) suggested that the amount of variance was sufficient to indicate the linear relationship between variables, and the inclusion of PC3 (13%) variance could not significantly provide additional relevant information while further complicating plot interpretations. In order to further understand samples and organic compounds, the variables significantly contributing to the clustering were selected to study their molecular structures based on *m/z* values, suggested formulas and calculated double bond equivalent (DBE) values.

Appropriate data pre-treatment typically improves PCA results on HRMS datasets. For example, Fraser et al. (2013) combined HRMS and PCA techniques to profile tea samples for determining the potential compounds associated with fermented and unfermented tea. A total of 88 samples with 57 black tea (fully fermented) samples, 11 oolong tea (10-80% semi-fermented) samples, and 20 green tea (unfermented) samples were analyzed by Ultra-Performance Liquid

Chromatography-Mass Spectrometry (UPLC-MS) to generate 690 and 359 peaks detected with positive and negative ionization mode respectively. Overall the auto-scaling, pareto scaling and log transformation all improved clustering compared with PCA alone, and Fraser et al. (2010) selected log transformation for date pretreatment due to its convenience in application. Similar to Sleighter et al. (2010), low variances were found with PC1 (26%) and PC2 (11%) for a total of 37% of original datasets variances were explained. By collocated samples and variables in score and loading plots, Fraser et al. (2010) extracted and identified significant markers (variables) associated with fermentation to describe differences between tea samples by comparing their *m/z* values, relative retention time and source induced fragmentations. Fraser et al. (2010) indicated that such analytical approaches had substantial power to distinguish differences in metabolite profiles of tea samples.

Matthew et al (2012) applied PCA to study NAs profiles, source determinations, and correlations to other water quality variables in surface and ground water in oil sands regions. 58 samples from lower Athabasca Region were collected as surface water samples, 6 Athabasca River sediment pore water samples were collected as regional groundwater samples, and 2 samples from active tailing ponds were collected to represent the OSPW samples. A data matrix of NAs homologue peaks identified by HRMS was pre-treated by logarithm transformation to satisfy the assumption of normal distribution. The first three PCs (PC1=33%, PC2=16%, PC3=9%) covered 58% of total variance. Further, by assessing correlations between score and loading plots, Matthew et al. (2012) concluded that natural fatty acids with even-number of carbons and bitumen-derived acids were identified as two categories of NAs. Bitumen-derived acids highly contributed to PC1 and natural fatty acids highly contributed to PC2. Surface water samples had lower bitumen-derived acids but higher natural fatty acids because of the contribution of microorganisms. In contrast, sediment pore water samples had lower natural fatty acids but higher bitumen-derived acids. Variations between samples were mainly due to sampling locations (upstream, downstream and depth). By studying other water quality parameters, Matthew et al. (2012) found PC1 was positively correlated to total dissolved solids (TDS), hardness, total alkalinity, bicarbonate, calcium, barium, magnesium, magnesium, manganese, chloride, and ammonia. PC2 was negatively correlated to total Kjeldahl nitrogen (TKN).

3.0 Materials and Methodologies

The entire scope of data analysis was demonstrated in the flow chart shown in Figure 3-1. Starting with the center vertical stream, three raw OSPW (Suncor, CNRL and Syncrude OSPW) samples were first analyzed by HRMS to detect markers in the samples. PCA was then applied to simplify the data matrix of detected markers and determine the significant markers describing the major variations between different raw OSPWs. The stream from Syncrude OSPW (thin orange line) demonstrates this sample was treated by ozonation in different conditions and/or biological treatment processes. The samples collected during the various treatment processes were analyzed by HRMS with PCA applied to determine markers that were significantly changed during treatment processes. The behaviours of markers indicated if they were degraded or formed as by-products during treatment processes. Finally, these significant markers were further tracked in their presence/absence and correlations to NAs and oxidized NAs in the raw OSPW from the three different sites.

The ozonation in different conditions and biological treatment processes experiment samples were provided by various researchers in Civil and Environmental Engineering at the University of Alberta. A brief overview of the experiment information is introduced in following section. Samples including raw OSPW and OSPW treated by ozonation in different conditions and biological treatment processes were analyzed using an Ultra Performance Liquid Chromatograms Mass Spectrometry (UPLC-HRMS) instrument from Waters Inc. with raw data recorded onto the online computer. The injections were performed by a professional laboratory technician, thus the sample injection procedures are only briefly reviewed here. The main focus of this chapter is on the methods setup of Masslynx Software which was applied for HRM dataset analysis.

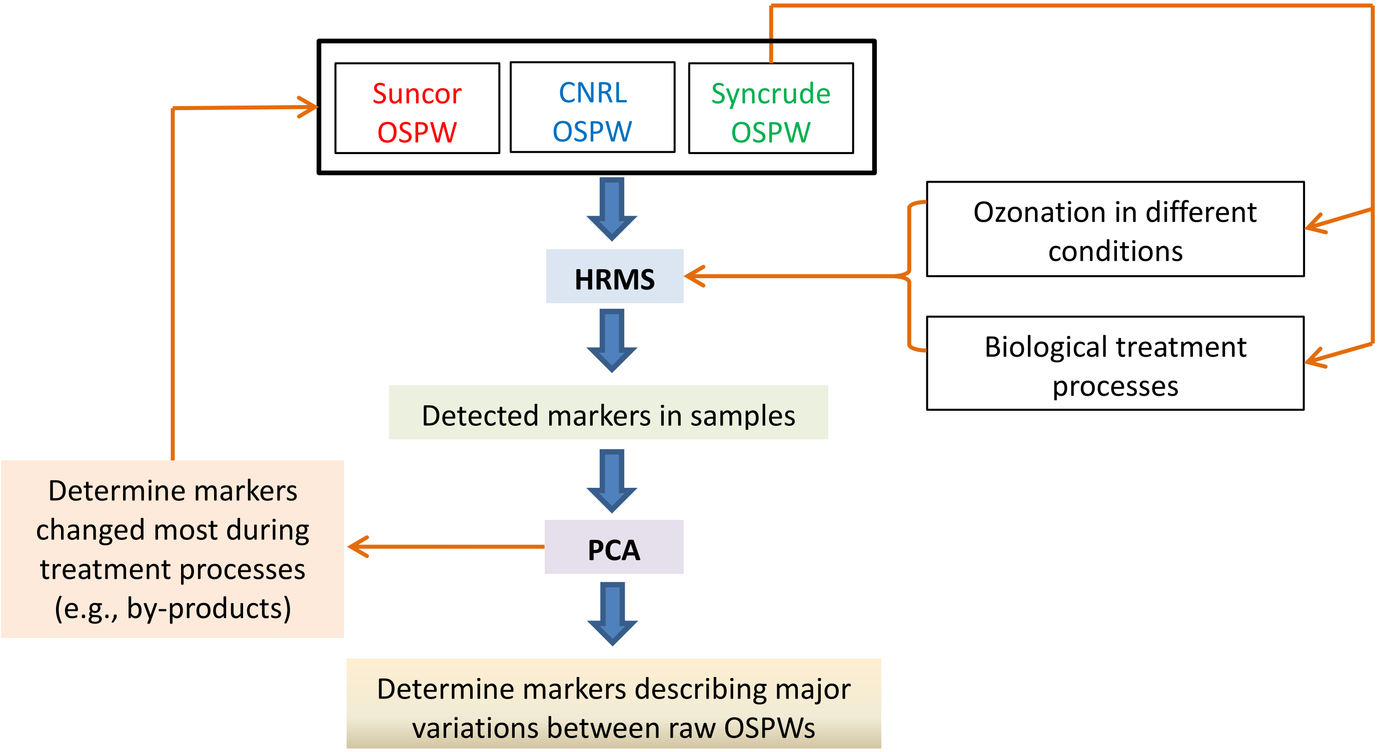

Suncor OSPW
CNRL OSPW
Syncrude OSPW
HRMS
Ozonation in different conditions
Biological treatment processes
Detected markers in samples
Determine markers changed most during treatment processes (e.g., by-products)
PCA
Determine markers describing major variations between raw OSPWs

Figure 3- 1: The Flow Chart of Entire Scope of Data Analysis Procedures, where HRMS = high resolution mass spectrometry; PCA = principal component analysis.

3.1 Sample Information

3.1.1 Samples from Ozonation in Different Conditions Experiments

The ozonation in different conditions experiment information was provided by the research group member who completed the experiments. Basically, six various ozonation condition experiments were performed which included the addition of different scavengers and catalysts. Experiments including proposed impacts included:

(1) Raw OSPW + O_3 (general ozonation);

(2) Raw OSPW + O_3 + CO_3^{2-} (carbonate from $NaHCO_3$; impedes ozone decomposition);

(3) Raw OSPW + O_3 + TBA (*tert*-butyl alcohol; hydroxyl radical quencher);

(4) Raw OSPW + O_3 + TNM (tetranitromethane; free radical quencher);

(5) Raw OSPW + O_3 + TBA + CO_3^{2-};

(6) Raw OSPW + O_3 + Fe (II) (Ferrous iron from Fe_2SO_4; ozonation catalyst).

The ozonation was performed in a batch reactor which was a 1L Pyrex glass bottle with a gas diffuser and a custom made external loop to mix the volume of liquid. A peristaltic pump was utilized to move the liquid though the external loop which had a sampling valve and a bypass valve installed. A GSO-40 Effizon ozone generator (WEDECO AG Water Technology, Herford, Germany) fed with extra dry high purity oxygen was used to produce ozone gas. The ozone concentration (inlet and outlet) was monitored using two HC500 ozone monitors (WEDECO, USA).

Raw OSPW was from Syncrude West in Pit in 2010 that was stored at 4 °C until used. In each ozonation experiment, 1L of raw OSPW was allowed to reach room temperature by pouring into the reactor without further preparation. Meanwhile each scavenger and catalyst was added into OSPW at concentrations of 20 mg/L and 1 mM, respectively. Previous to ozonation, the inlet gas flow meter was set to 1 L/min and the ozone monitors at both inlet and outlet were reset to zero flushing oxygen, and a time zero sample (t_0) was taken. The ozone generator was activated and samples were collected using 20 mL glass vials with 1 mL of 1 M $NaNO_2$ used to quench the ozone and prevent further reactions. The ozone treatment was applied for 5 minutes, in which samples were taken every 10 seconds during the first minute, and afterwards at 2, 3 and 5 minutes. All the samples were kept at 4 ºC until HRMS analysis. However, due to the economic cost and time constraints, only OSPW control, 0s, 20s, 40s, 60s, 180s, and 300s samples were used for HRMS analysis.

For statistical analysis, each sample was injected 5 times so that instrumental replicates were obtained in each sample group. The 5 repetitions were used because the sample responses to the instrument signal might be varied due to the errors during samples preparation and injection, and instrument sensitivities to the surroundings. As a result, some injections might be detected as outliers by PCA, which need to be removed during PCA. After removing outliers, it was necessary to have at least 3 replicates for a reliable statistical analysis.

3.1.2 Samples from Biological Treatment Processes Experiments

OSPW biodegradation samples were taken from a research group member who completed the experiments. Raw OSPW was from Syncrude West in Pit in 2010, and was stored at 4 °C. The biodegradation study of OSPW was carried out in a 1 L amber bottle at room temperature at 150 rpm on horizontal shaker (Innova^{TM} 2100, New Brunswick Scientific, USA). Both raw and ozonated OSPW without any pre-treatment, and without addition of any nutrient was used. The experiment was carried out for 28 days with samples taken at 0 day (OSPW control), 5 days, 14 days and 28days, and stored at 4 ºC until HRMS analysis.

As previously stated, due to the concerns of economic costs and time constraints, only triplicate injections (repeated injection from same sample) were performed. However, using less replicates weakened the potential for statistical analysis if any injection was identified as an outlier, which was an issue for the statistical analysis limitation shown in the OSPW biodegradation results in Chapter 4.

3.1.3 Samples of OSPW from Different Sites

Raw OSPW is known to be variable from different sites and different regions within the same tailing pond (Allen, 2008; Pourrezaei et al., 2011). The potential by-products increased or formed during ozonation in different conditions or biological treatment processes were tracked by their presence, absence or varying behaviour in comparison to the other raw OSPWs. Three different raw OSPWs were provided for a fingerprinting project to study the variations of raw OSPWs including Suncor Pond 7, Syncrude West in Pit (WIP) and CNRL OSPW. However, the sampling information such as procedures and locations was not provided. Samples were directly taken from the collection barrels of OSPW stored at 4 ºC for HRMS analysis. Only duplicate injections (repeated injection from same sample) were prepared for raw OSPWs due to the economic and time constraints. The lack of appropriate replicates limited the statistical analysis, and the results could only be shown as average values without error bars (standard deviations) in the raw OSPW variations section in Chapter 4.

3.2 High Resolution Mass Spectrometry

A Waters Acquity UPLC System (Milford, MA) was used for OSPW analysis since it had been shown to efficiently separate organic compounds such as NAs and their oxidized products (Gamal El-Din et al., 2011; Perez-Estrada et al., 2011). HRMS samples were prepared in 500 µl OSPW sample with an additional 100 µL ISTD and 400 µL methanol, for a total volume of 1 ml as injected through a Waters UPLC Phenyl BEH column (1.7 µm, 150 mm × 1 mm) for chromatographic separation. A 10 mM ammonium acetate solution prepared in Optima-grade water was used as mobile phase A; and 10 mM ammonium acetate in 50% methanol 50% acetonitrile, both Optima-grade, were used as mobile phase B. Gradient elution was performed as follows: 1% eluent B for first 2 minutes, ramped to 60% effluent B by 3 minutes, and to 70% eluent B by 7 minutes, next to 95% eluent B by 13 minutes, followed by a hold until 14 minutes, and finally returned to 1% eluent B again, finished by a further 5.8 minutes re-equilibration time. The flow during the whole process was controlled at 100 µL/min and column temperature was kept stable at 50 ^{0}C (Afzal et al., 2012).

The detection system was equipped with a high resolution Synapt G2 HDMS mass spectrometer with an electrospray ionization source operating in negative ion mode. A technician from Waters Corporation (Milford, MA) helped tuning and calibration by using standard solutions of lucine enkaphenlin and sodium formate. Masslynx ver. 4.1 was installed to control the Water Acquity UPLC System and for data analysis.

3.3 Software Method Setup

In Masslynx software, there are several sub-software programs including Markerlynx (XS V4.1 SCN803, Waters Inc.), EZ info (V2.0, Umetrics AB), Targetlynx (V4.1 SCN 803, Waters Inc.) and Elemental Composition calculator

(V4.0, Water Inc.) that can be used for different analyses.

The following five steps were included in the current data analysis process. Each step used different sub-software from Water Inc. or Microsoft Excel:

Step (1): Markerlynx software detected and collected the potential markers from the raw datasets of injected samples. Markers were presented as *m/z* values, and peaks areas were automatically integrated.

Step (2): EZ Info software applied PCA based on the markers area integration results from Markerlynx to determine the markers that exhibited the major differences across the samples.

Step (3): Markers were defined as significant and non-significant to describe the differences between samples. Only significant markers were further identified as NAs, oxidized NAs or unknown compounds based on *m/z* values. Microsoft Excel was used to plot trends in areas of significant unknown markers across samples to review the changes of markers during treatment processes.

Step (4): Targetlynx software could be used to manually integrate the peak areas of significant markers extracted by PCA to compare with results from Markerlynx.

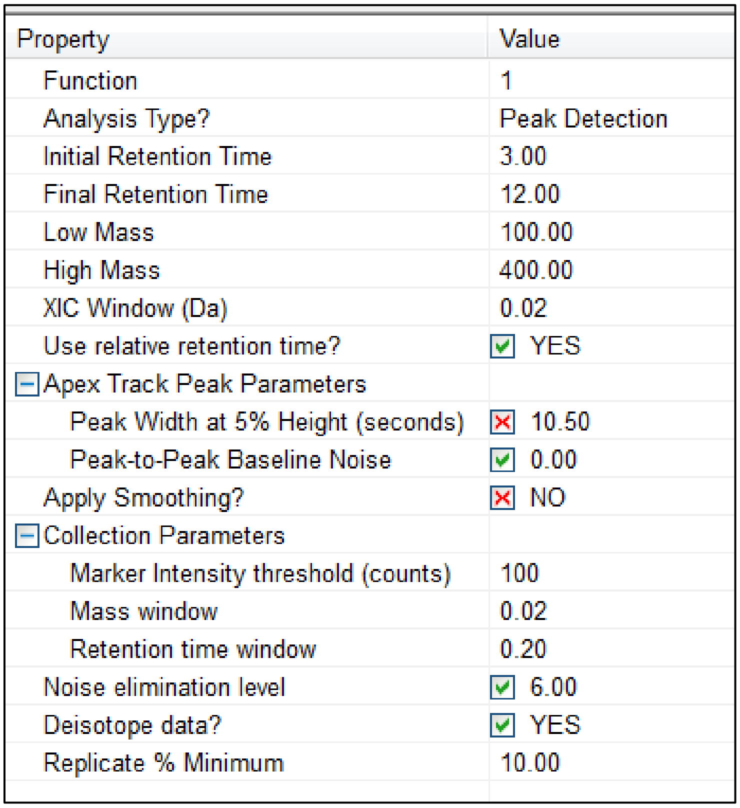
Step (5): Elemental Composition calculator gave possible elemental compositions of significant unknown markers based on their exact masses.

Each Masslynx sub-software method had to be individually setup in order to appropriately operate and reliably process the datasets. The specific method setup procedures are described in detail below.

3.3.1 Step 1: Markerlynx Method Setup

Samples to be analyzed were selected to create a sample list. The next step was to setup the method in Markerlynx to process the analysis of the sample list. The peak alignment and detection parameters are shown in Table 3-1.

Table 3- 1: Parameters in Markerlynx Method Setup Window (adapted from Waters Inc., 2010).



Property	Value
Function	1
Analysis Type?	Peak Detection
Initial Retention Time	3.00
Final Retention Time	12.00
Low Mass	100.00
High Mass	400.00
XIC Window (Da)	0.02
Use relative retention time?	☑ YES
⊟ Apex Track Peak Parameters	
Peak Width at 5% Height (seconds)	☒ 10.50
Peak-to-Peak Baseline Noise	☑ 0.00
Apply Smoothing?	☒ NO
⊟ Collection Parameters	
Marker Intensity threshold (counts)	100
Mass window	0.02
Retention time window	0.20
Noise elimination level	☑ 6.00
Deisotope data?	☑ YES
Replicate % Minimum	10.00

Function: Function 1 was set as the default function by Waters Inc. to process mass spectra data files.

Analysis Type: Peak detection was selected as the analysis type in order to detect markers in the samples.

Initial and Final Retention Time: The entire chromatogram of the marker was shown in a range of 0 to 15 minutes. However, only 3-12 minutes were set as the initial and final retention time to specify the range of retention time window in which the data to be processed, because little peaks and mostly noises were shown in the chromatogram out of this range.

Low and High Mass: Although the instrument is able to detect up to a mass of 100,000 *m/z* unit, only 100 – 400 *m/z* units (equivalent to daltons; Da) was selected as the mass range over which the software detected makers, because all NAs and oxidized NAs analyzed (n = 7 to 22; Z= 0 to -12) are within this range.

XIC Window (Da): XIC window limited the mass accuracy (Da) of acquired data, and 0.02 Da was recommend as the tolerance which was twice the quoted mass accuracy (0.01 Da) of UPLC-HRMS instrument from Waters Inc.

Use relative retention time: This option allowed the software to automatically detect all markers with respect to the internal standard (ISTD) retention time. The retention time of a marker might vary in a tolerance range between the samples, but this method shifted all markers' retention times to the ISTD retention time, which located the markers relative positions more accurately. The ISTD (*m/z = 228.2033* ± *0.02*) retention time was set in a range of 6.7 ± 0.5 minutes.

Peak Width at 5% Height: This option allowed the software to automatically determine the peak width and integrate the peak area. The width in unit of seconds on retention time axis was initially set at 5% height of the peak. However, due to noise and poor smoothing of the peak shape, the automatically determined peak width sometimes varied significantly in replicate samples creating huge variations of integrated areas in the replicates. Therefore, an accurate peak width would be suggested to be manually entered for more accurate area integrations (Water Inc., 2010). However, no information was given to assess which part of each peak was integrated to determine if a peak was overestimated or underestimated during the integration process. As recommended by a technician from Waters Corporation (Milford, MA), the maximum peak width from the first automatic analysis run was applied for all remaining samples in order to obtain better integration results. Thus, as shown in Table 3-1, the automatic peak width determination option was turned off as indicated by a red cross mark, and the maximum peak width of 10.5 seconds entered in the value column would be applied for all markers' area integrations. It should be noted that the maximum peak width to integrate all peaks leads to the overestimation of areas for narrow peaks. However, by comparing the integration results with and without using maximum peak width, the differences for narrow markers were not very significant. The possible reason might be the overestimation would only include noises within maximum peak width, but the noise with intensities below critical values would be rejected, and noise areas counted could be negligible in magnitudes compared with the actual peak areas.

Peak-to-Peak Baseline Noise: Baseline noise between peaks on a typical extracted ion chromatogram directly impacts the number of markers detected. A higher value resulted in less detected markers because peaks with relatively low intensity approaching the baseline noise value were considered as noise rather than a peak. The parameter could be either manually set or automatically estimated by software. The green check mark in Table 3-1 indicated the value was automatically determined by software.

Apply Smoothing: This option specified if the software applied two iterations of mean smoothing function (three data points wide) during the peak detection. By comparing the results with and without applying smoothing, it was found that some markers (sometimes even the included internal standard) could not be detected in the samples by applying smoothing. Thus, no smoothing was applied for any analysis.

Marker Intensity threshold (counts): This parameter defined the minimum level of intensity threshold (in counts) for a spectral peak to be considered as a marker. A 100 count default value was recommended by Waters Inc., so the peaks with threshold less than 100 would be considered as noise.

Mass window: The value was in unit of daltons (Da) and specified the mass tolerance for a particular marker to be considered as the same marker across the different samples. A tolerance of 0.02 Da was used, because it was twice the quoted mass accuracy (0.01 Da) for a time-of-flight instrument as recommended (Waters Inc., 2010).

Retention time window: The value (in minutes) specified the retention time tolerance for a particular marker to be considered as the same marker across the different samples. A 0.2 minute window was recommended as an appropriate value (Waters Inc., 2010) and used for all analyses.

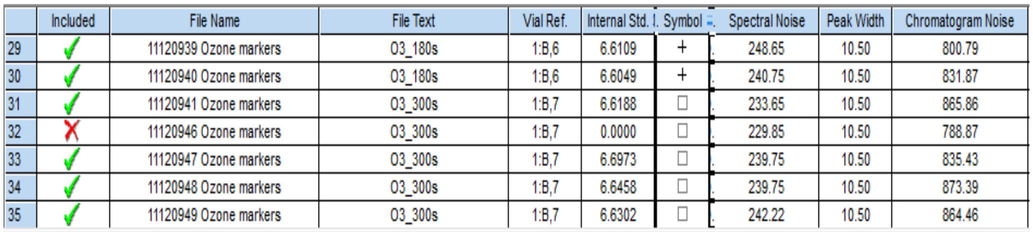
Noise elimination level: This option allowed the software to automatically eliminate the noise based on the value entered in the field during the peak detection. A larger value tended to discard more spectral peaks. Values in the range of 4 to 10 were suggested by Waters Inc. In Table 3-1, 6 times the standard deviation of the background noise was recommended for elimination by a field technician from Waters Corporation (Milford, MA) and used for all analyses.

Deisotope data: This option removed isotope signals to prevent the assignment of isotope peaks as markers.

Replicate % Minimum: Specified the minimum percentage of the total samples in the entire list that a marker had to be detected. Otherwise, the intensity of that marker was scaled to 0 for every sample in the list. This minimum percentage was another reason why replicates were necessary to give reliable markers detection. With a minimum of 10% of 35 samples in total, the markers had to be detected in at least 3 samples which was the minimum number of replicates suggested for each sample group. If the marker was not consistently present in a minimum of 10% of total samples, that marker would be discarded.

After modifying the method, the software was used to automatically process the samples and detect markers with the partial results shown in samples table (Table 3-2) and markers table (Table 3-3).

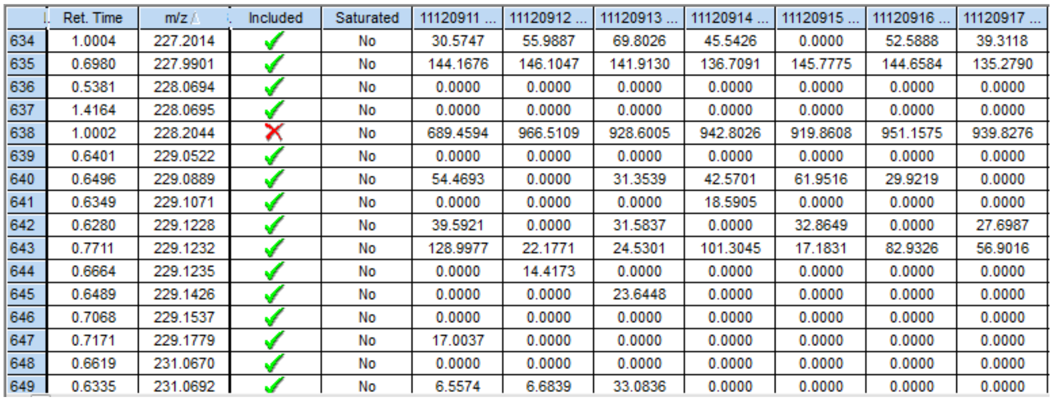
Table 3- 2: Portion of Typical Samples Table from Markerlynx Software (adapted from Waters Inc., 2010).



	Included	File Name	File Text	Vial Ref.	Internal Std. I.	Symbol	Spectral Noise	Peak Width	Chromatogram Noise
29	✓	11120939 Ozone markers	O3_180s	1:B,6	6.6109	+	248.65	10.50	800.79
30	✓	11120940 Ozone markers	O3_180s	1:B,6	6.6049	+	240.75	10.50	831.87
31	✓	11120941 Ozone markers	O3_300s	1:B,7	6.6188	□	233.65	10.50	865.86
32	✗	11120946 Ozone markers	O3_300s	1:B,7	0.0000	□	229.85	10.50	788.87
33	✓	11120947 Ozone markers	O3_300s	1:B,7	6.6973	□	239.75	10.50	835.43
34	✓	11120948 Ozone markers	O3_300s	1:B,7	6.6458	□	239.75	10.50	873.39
35	✓	11120949 Ozone markers	O3_300s	1:B,7	6.6302	□	242.22	10.50	864.46

In the samples table (Table 3-2) the first column counts the number of samples. In the second column, a green check mark indicates that the corresponding sample was involved in PCA. Otherwise, a red cross mark was displayed to indicate that the corresponding sample was excluded. In the next two columns, each sample was labeled by file text and file name. In the following columns, information of vial references for injections, ISTD retention time, sample symbol, spectral noise, peak width and chromatogram noise are illustrated. The spectra noise was calculated based on the noise elimination level set in the method, and this value determined the intensity below which was considered as noise. Increasing of the noise elimination level also increased the spectral noise value. The chromatogram noise showed the value of peak-to-peak baseline noise automatically determined by the software. In summary, information about ISTD retention time, spectral noise, peak width and chromatogram noise for each sample would be displayed in the sample table. These parameters are helpful to assess if the individual sample is an outlier which needs to be excluded for PCA. For instance, the sample #32 in Table 3-2 is excluded (red cross mark) because the software did not detect the ISTD as indicated by a retention time value of zero.

Table 3- 3: Portion of Typical Markers Table from Markerlynx Software (adapted from Waters Inc., 2010).



	Ret. Time	m/z	Included	Saturated	11120911 ...	11120912 ...	11120913 ...	11120914 ...	11120915 ...	11120916 ...	11120917 ...
634	1.0004	227.2014	✓	No	30.5747	55.9887	69.8026	45.5426	0.0000	52.5888	39.3118
635	0.6980	227.9901	✓	No	144.1676	146.1047	141.9130	136.7091	145.7775	144.6584	135.2790
636	0.5381	228.0694	✓	No	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
637	1.4164	228.0695	✓	No	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
638	1.0002	228.2044	✗	No	689.4594	966.5109	928.6005	942.8026	919.8608	951.1575	939.8276
639	0.6401	229.0522	✓	No	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
640	0.6496	229.0889	✓	No	54.4693	0.0000	31.3539	42.5701	61.9516	29.9219	0.0000
641	0.6349	229.1071	✓	No	0.0000	0.0000	0.0000	18.5905	0.0000	0.0000	0.0000
642	0.6280	229.1228	✓	No	39.5921	0.0000	31.5837	0.0000	32.8649	0.0000	27.6987
643	0.7711	229.1232	✓	No	128.9977	22.1771	24.5301	101.3045	17.1831	82.9326	56.9016
644	0.6664	229.1235	✓	No	0.0000	14.4173	0.0000	0.0000	0.0000	0.0000	0.0000
645	0.6489	229.1426	✓	No	0.0000	0.0000	23.6448	0.0000	0.0000	0.0000	0.0000
646	0.7068	229.1537	✓	No	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
647	0.7171	229.1779	✓	No	17.0037	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
648	0.6619	231.0670	✓	No	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
649	0.6335	231.0692	✓	No	6.5574	6.6839	33.0836	0.0000	0.0000	0.0000	0.0000

In the markers table (Table 3-3) the first column counted the number of markers. The second and third columns recorded the retention time and *m/z* values of the detected marker. The fourth column with a green check mark or a red cross mark indicates if the corresponding marker was included or excluded in further PCA study. For example, the internal standard (C13-Myristic acid, *m/z = 228.2044*) was excluded. The following columns show the area integration results for the marker in each sample. In addition, the chromatogram of a selected marker across the samples could be viewed in Markerlynx to help decide if the marker actually showed a peak or was only noise. Markerlynx software was mistaken in identifying noise chromatograms which had intensities higher than the noise intensity set in the method as peaks. The noise would be excluded for later PCA study. Examples of noise chromatograms are reviewed in Chapter 4.

3.3.2 Step 2: EZ Info Method Setup

The Markerlynx add-on statistical tool EZ Info (V2.0, Umetrics AB) was applied to analyze the data matrix with detected markers. In EZ Info, the PCA model was selected for the datasets analysis. Prior to PCA, data transformation methods such as log transformations and data scaling methods like centering and pareto scaling were available for data pre-treatment.

3.3.2.1 Data Pre-treatment Methods

PCA is sensitive to data scales and generally needs data pre-treatment prior to its use (Gao et al, 1999; Praveena et al, 2012; Van den Berg et al, 2006). Each data pre-treatment method has its own purpose, strengths and limitations (Van den Berg et al, 2006). However, there are no specific rules to define which method will give the best PCA results. The conditions vary depending on the properties of datasets, so the critical way to determine the appropriate data pre-treatment method in a specific situation is to compare the PCA results with and without different data pre-treatments (Van den Berg et al, 2006; Praveena et al, 2012; Reid and Spencer, 2009; Arruda et al, 2011; Gao et al, 1999).

The data pre-treatment methods that are commonly used include data scaling such as centering, auto scaling, range scaling, pareto scaling, vast scaling, level scaling; and data transformations such as log transformation and power transformation (Van den Berg et al, 2006). A general overview has been introduced in Chapter 2. However, in the EZ Info Software, only centering, pareto scaling as the data scaling method and log transformation method are available. The purposes, advantages and disadvantages of each available pre-treatment method are discussed below.

Centering:

$$\widetilde{\boldsymbol{X}}_{\boldsymbol{ij}} = \boldsymbol{X}_{\boldsymbol{ij}} - \overline{\boldsymbol{X}_{\boldsymbol{\iota}}} \quad \textbf{..............................Equation 3- 1}$$

Where X_{ij} is the value presents at ith row and jth column; $\bar{X}_{i}$ is the mean value of ith row; $\tilde{X}_{ij}$ represents the result after centering.

The purpose of centering is to focus the analysis on the differences rather than the similarities in the datasets. The method has the advantage of successfully removing the offset from datasets, but it is insufficient to treat datasets with heterosedastic properties that are defined as sub-populations having different variability from others, which leads to the failure in normal distribution (Van den Berg, 2006).

Pareto Scaling:

$$\widetilde{\boldsymbol{X}}_{\boldsymbol{ij}} = \frac{\boldsymbol{X}_{\boldsymbol{ij}} - \bar{\boldsymbol{X}}_{\boldsymbol{i}}}{\sqrt{\boldsymbol{S}_{\boldsymbol{i}}}} \quad \textbf{...Equation 3- 2}$$

Where S_i is the standard deviation of the ith row in datasets

The purpose of pareto scaling is to reduce the relative importance of variables with large values so that more correlations of the variables can be studied. The advantage is that it keeps part of the original data structure when it reduces the relative importance of the large variables, so it prevents creating variables (high variation and low variation) that are equally important. However, the limitation is that the scaling method is sensitive to the datasets with large fold (order of magnitude) changes.

Log Transformation:

$$\widetilde{\boldsymbol{X}}_{\boldsymbol{ij}} = \boldsymbol{log}(\boldsymbol{X}_{\boldsymbol{ij}})\boldsymbol{or}\ \widetilde{\boldsymbol{X}}_{\boldsymbol{ij}} = \boldsymbol{log}(\boldsymbol{X}_{\boldsymbol{ij}} + \mathbf{1}) \text{ ....Equation 3- 3}$$

The log transformation is the most common method applied to pre-treat datasets prior to PCA. The purpose is to correct the datasets with heteroscedasticity and transfer the datasets into log scale to satisfy the normal distribution with skew value < 0.5. However, such transformation is poor when dealing with datasets containing relatively large standard deviations and zeros (Van den Berg, 2006). Although log (X+1) transformation improves the situations of numerous zeros in the datasets, Reid and Spencer (2009) pointed out that the log (X+1) transformation shifts the skewness of datasets to a more negative side, so log (X+1) transformation will worsen the data distribution if it has already shown a negative skew. Moreover, Gao et al (1999) suggested that the log transformation compresses the upper end of the data on the scale and reduces the relative importance of variables. As a result, the transformation leads to a more balanced weight of variables, so the relative significance of the variables is reduced. Further, the reducing of extreme values leads to the reducing of the effects of outliers (Gao et al, 1999).

With the available data scaling method of centering and pareto scaling, and transformation method of log (X+1) as data pre-treatment methods in EZ Info software, six combinations of different data pre-treatments were considered below:

(1) No scaling & no transformation

(2) No scaling & log(x+1) transformation

(3) Pareto scaling & no transformation

(4) Pareto scaling & log(x+1) transformation

(5) Center scaling & no transformation

(6) Center scaling & log(x+1) transformation

All current datasets were first pre-treated by each method above. The PCA was applied on each modified dataset and the results were compared to choose the most appropriate data pre-treatment method to suit the data analysis. The critical evaluation was based on the more appropriate sample clustering and higher total variance of the original data that the principal components were able to recover.

3.3.2.2 Validation Method Setup

Validation is the technique to test if the model is well developed based on current datasets, and how well it performs to predict new datasets (Johnson and Wichern, 2007). Thus, the validation step usually estimates the uncertainty of the model prediction on new datasets and the model is considered to be valid if the uncertainty is in an acceptable range (CAMO, 201). Since the samples population (35 samples) were much smaller than the number of variables (>1,500 markers), cross validation which is designed to solve problems of limited sample populations (Johnson and Wichern, 2007; CAMO, 2011), was recommended to be applied in EZ Info Software (Waters Inc., 2010). The general review of cross validation was previously discussed in Chapter 2.

3.3.3 Step 3: Significant Markers Selection and Identification

Typical PCA score and loading plots were shown in Figure 1-3, and the basic interpretations of PCA plots were introduced in Chapter 1 as well. The distributions of the variables (i.e., markers) in the loading plot directly determine the clustering of samples in the score plot. However, it is not realistic to study over 1,500 individual markers, so it was necessary to extract significant markers which can be used to best describe the differences between samples. Theoretically, the markers in the loading plots are positively correlated to the samples in the corresponding regions in the score plots. Markers with higher loading values are more specific to corresponding samples, and markers near the origin of loading plot are common or similar in concentrations in all samples. Thus, the analysis first focused on markers with larger loadings located on the outside of the loading plot, and narrowed down toward the origin in loading plot. The areas of markers were plotted across the samples by Microsoft Excel. The significant markers showed a significant change in area across sample groups; in contrast, non-significant markers showed a relatively stable or constant trend. Comparing the trends in areas of markers, a relative significance boundary was defined to separate significant and non-significant markers. The results for this analysis are shown in detail in Chapter 4.

The *m/z* values of each significant marker were matched with exact masses of NAs and oxidized NAs (NA+Ox, where x = 1 to 4) shown in Table A1 to A5 in Appendix A. The marker was assumed to be NAs or NA+Ox if the exact mass was within the error of ± 0.01 Da. If the *m/z* value of the marker did not match any exact mass, that marker was considered as an unknown compound. As stated in previous chapters, this project would focus on the unknown markers' behaviours during ozonation in different conditions and biological treatment processes, so Microsoft Excel was used to plot the trends of significant unknown markers across the sample groups in terms of average peak areas automatically integrated by Markerlynx.

3.3.4 Step 4: Targetlynx Method Setup

Unlike Markerlynx which is a qualitative tool mainly used to detect the markers across the samples and integrates peak areas automatically, Targetlynx is a quantitative tool mainly used for quantification (Waters Inc., 2010), so the peaks of the significant markers were manually selected and adjusted in order to more accurately integrate the peak area in the extracted chromatogram window.

First of all, it was necessary to create a target list which included all significant markers with their *m/z* values and retention times with error tolerance set as ±0.02 minute. The peak responses were expressed as area integrations. The software automatically determined the peak-to-peak baseline noise and peak width at 5% height. The mean smoothing method with 2 iterations and 3 scans were applied during the area integrations. Targetlynx processed all selected samples automatically with the extracted ion chromatogram window displaying the peaks of each marker in the target list. Peaks could also be manually selected and adjusted in order to optimize the area integration. Manually adjusted and integrated significant markers peaks and integrated areas were exported into Microsoft Excel file for trends plots and comparison with the results based on the Markerlynx approach.

3.3.5 Step 5: Elemental Composition Calculator Setup

After selecting out the significant markers that described the main differences across the samples from PCA study and reviewing their trends plots, the next step was to study their possible elemental compositions and molecular structures. The Elemental Composition calculator in Masslynx software provided suggestions of elemental compositions and double bond equivalents for the significant markers based on *m/z* values. However, it was still necessary to setup the parameters in the calculator in order to get accurate and reliable suggestions.

The double bond equivalent (DBE) range was set from 0 to 50 which would cover most possible double bonds present in organic compounds and it was the default setting suggested by Waters Inc. DBE is defined as the number of H_2

molecules that have to be added to a molecule of the chemical compound to convert all *pi* bonds to single bonds, and all rings to acyclic structures (Clayden et al., 2001). DBE could be calculated based on the number of carbon, nitrogen and hydrogen molecules in the compound structure based on Equation 3-4:

$$\mathbf{DBE} = \boldsymbol{C} - \frac{\boldsymbol{H}}{\mathbf{2}} + \frac{\boldsymbol{N}}{\mathbf{2}} + \mathbf{1} \quad \textbf{........... Equation 3- 4}$$

(Adapted from Clayden et al., 2001)

Each time a ring or a double bond forms, two hydrogen atoms have to be lost. Generally, one DBE indicates one ring or one double bond. Two DBE may be two rings, or two double bounds, or one triple bond, or one ring plus one double bound.

Only carbon, hydrogen, nitrogen, oxygen and sulphur were selected as the possible elements of the compounds because they are the basic and common elements of organic compounds found in the tailing ponds (Allen, 2008; Corinee, 2010; Pourrezaei et al., 2011). In addition, the range for each element was set as 0-50 for carbon, 0-100 for hydrogen, 0-2 for nitrogen, 0-6 for oxygen and 0-2 for sulphur based on possible elements in organic compounds found in OSPW (Allen, 2008; Corinee, 2010). After entering the exact mass of the marker, the calculator gave all the possible elemental compositions which had the error less than the tolerance of 10 ppm. The mass error tolerance was set up to 10 ppm or accurate to 0.001 Da based on the mass differences between the exact mass of detected marker and the exact mass of the suggested compound, and it was calculated by:

$$\boldsymbol{ppm} = \frac{\textbf{detected}\ \boldsymbol{m/z} - \boldsymbol{calculated}\ \boldsymbol{m/z}}{\boldsymbol{calculated}\ \boldsymbol{m/z}} \times \mathbf{10^6} \quad \textbf{........... Equation 3- 5}$$

Since some markers had already been assumed as NAs or NA+Ox if their exact masses matched in the tolerance error of ±0.01 Da, the elemental compositions calculator was only applied to unknown compounds whose *m/z* values did not match with any exact masses.

4.0 Results and Discussion

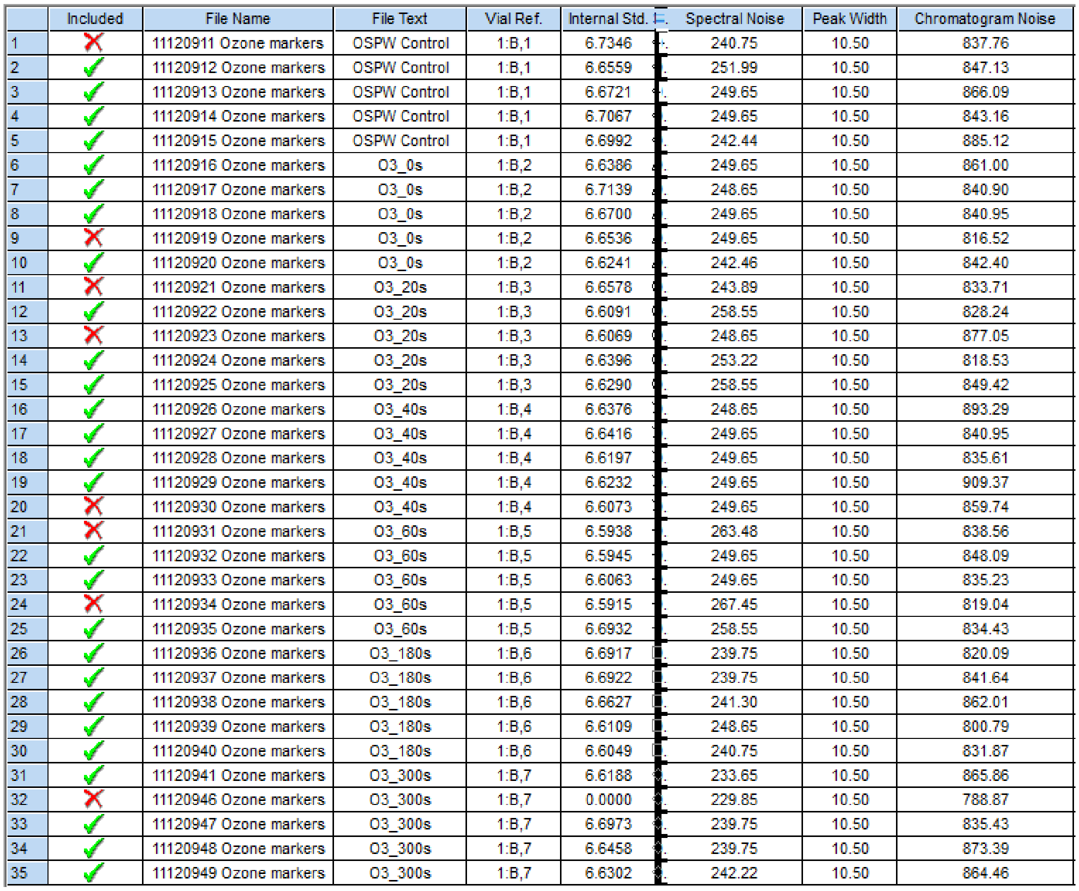
This chapter presents PCA results based on three experimental datasets measured using HRMS: (1) ozonation in different conditions experiments including O_3, O_3+CO_3^{2-}, O_3+TBA, O_3+TBA+CO_3^{2-}, O_3+TNM and O_3+Fe (II); (2) raw and ozonated OSPW biodegradation experiments; and (3) raw OSPW from different sites including Syncrude West in Pit, Suncor pond 7 and CNRL OSPW. Sections 4.1 to 4.8 include the details of O_3 results as an overall example for data analysis. Data pre-treatments, Markerlynx software validation, and the use of Targetlynx as complementary software to Markerlynx to more accurately monitor makers' trends across samples are discussed in ozonation in different conditions results.

PCA results based on raw and ozonated OSPW are reviewed in Section 4.9 by the same approaches as those illustrated in the ozonation in different conditions results. Additionally, raw OSPW from different sites are characterized in Section 4.10 using the presences or absences of unknown markers which were significantly changed during the previous ozonation in different conditions treatment processes or biological treatment processes. Key results were summarized in Section 4.11.

4.1 Markers Detection from Ozonation Datasets

After initial set up of the Markerlynx software methods in Chapter 3, the software was used to automatically process the sample list to detect markers across the samples. The results are shown in samples table and markers table in Table 4-1 and 4-2 respectively.

Table 4- 1: Samples Table Showing Samples Information in Ozonation Datasets Processed by Markerlynx Software.



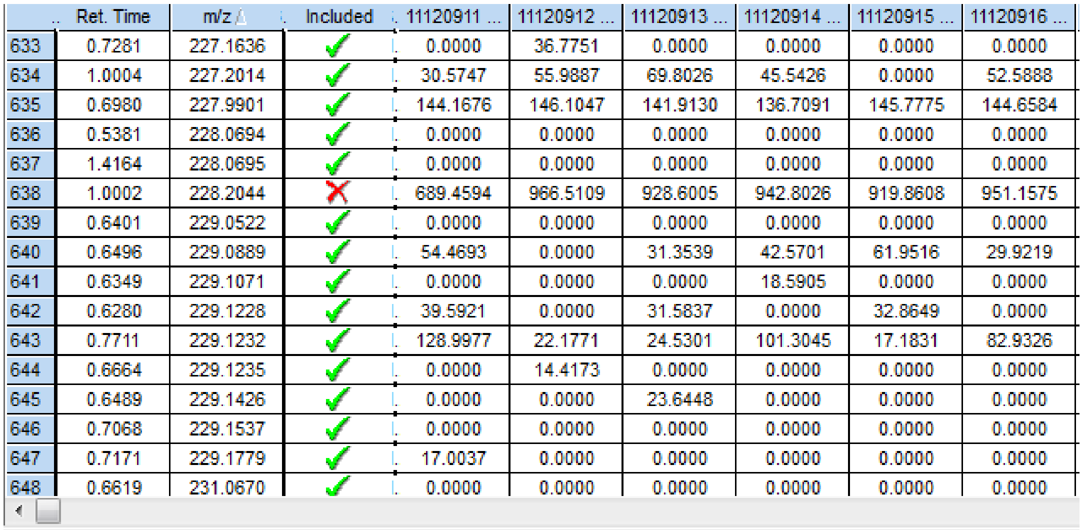
	Included	File Name	File Text	Vial Ref.	Internal Std.	Spectral Noise	Peak Width	Chromatogram Noise
1	✗	11120911 Ozone markers	OSPW Control	1:B,1	6.7346	240.75	10.50	837.76
2	✓	11120912 Ozone markers	OSPW Control	1:B,1	6.6559	251.99	10.50	847.13
3	✓	11120913 Ozone markers	OSPW Control	1:B,1	6.6721	249.65	10.50	866.09
4	✓	11120914 Ozone markers	OSPW Control	1:B,1	6.7067	249.65	10.50	843.16
5	✓	11120915 Ozone markers	OSPW Control	1:B,1	6.6992	242.44	10.50	885.12
6	✓	11120916 Ozone markers	O3_0s	1:B,2	6.6386	249.65	10.50	861.00
7	✓	11120917 Ozone markers	O3_0s	1:B,2	6.7139	248.65	10.50	840.90
8	✓	11120918 Ozone markers	O3_0s	1:B,2	6.6700	249.65	10.50	840.95
9	✗	11120919 Ozone markers	O3_0s	1:B,2	6.6536	249.65	10.50	816.52
10	✓	11120920 Ozone markers	O3_0s	1:B,2	6.6241	242.46	10.50	842.40
11	✗	11120921 Ozone markers	O3_20s	1:B,3	6.6578	243.89	10.50	833.71
12	✓	11120922 Ozone markers	O3_20s	1:B,3	6.6091	258.55	10.50	828.24
13	✗	11120923 Ozone markers	O3_20s	1:B,3	6.6069	248.65	10.50	877.05
14	✓	11120924 Ozone markers	O3_20s	1:B,3	6.6396	253.22	10.50	818.53
15	✓	11120925 Ozone markers	O3_20s	1:B,3	6.6290	258.55	10.50	849.42
16	✓	11120926 Ozone markers	O3_40s	1:B,4	6.6376	248.65	10.50	893.29
17	✓	11120927 Ozone markers	O3_40s	1:B,4	6.6416	249.65	10.50	840.95
18	✓	11120928 Ozone markers	O3_40s	1:B,4	6.6197	249.65	10.50	835.61
19	✓	11120929 Ozone markers	O3_40s	1:B,4	6.6232	249.65	10.50	909.37
20	✗	11120930 Ozone markers	O3_40s	1:B,4	6.6073	249.65	10.50	859.74
21	✗	11120931 Ozone markers	O3_60s	1:B,5	6.5938	263.48	10.50	838.56
22	✓	11120932 Ozone markers	O3_60s	1:B,5	6.5945	249.65	10.50	848.09
23	✓	11120933 Ozone markers	O3_60s	1:B,5	6.6063	249.65	10.50	835.23
24	✗	11120934 Ozone markers	O3_60s	1:B,5	6.5915	267.45	10.50	819.04
25	✓	11120935 Ozone markers	O3_60s	1:B,5	6.6932	258.55	10.50	834.43
26	✓	11120936 Ozone markers	O3_180s	1:B,6	6.6917	239.75	10.50	820.09
27	✓	11120937 Ozone markers	O3_180s	1:B,6	6.6922	239.75	10.50	841.64
28	✓	11120938 Ozone markers	O3_180s	1:B,6	6.6627	241.30	10.50	862.01
29	✓	11120939 Ozone markers	O3_180s	1:B,6	6.6109	248.65	10.50	800.79
30	✓	11120940 Ozone markers	O3_180s	1:B,6	6.6049	240.75	10.50	831.87
31	✓	11120941 Ozone markers	O3_300s	1:B,7	6.6188	233.65	10.50	865.86
32	✗	11120946 Ozone markers	O3_300s	1:B,7	0.0000	229.85	10.50	788.87
33	✓	11120947 Ozone markers	O3_300s	1:B,7	6.6973	239.75	10.50	835.43
34	✓	11120948 Ozone markers	O3_300s	1:B,7	6.6458	239.75	10.50	873.39
35	✓	11120949 Ozone markers	O3_300s	1:B,7	6.6302	242.22	10.50	864.46

As introduced previously in the software methods setup section in Chapter 3, the first column in the samples table (Table 4.1) counts the number of samples. In ozonation datasets, there were 5 replicates contained in each of 7 sample groups (OSPW control, 0s, 20s, 40s, 60s, 180s and 300s sample groups) for a total number of 35 samples analyzed. In the second column, a green check mark indicates that the corresponding sample was included in PCA. In contrast, a red cross mark indicates that the corresponding sample was excluded. In the next two columns, each sample was labelled by file text and file name which simply described samples. The vial references column shows replicates of repeated injections from the same sample vial. In the following columns, information of internal standard (ISTD) retention time, spectral noise, peak width and chromatogram noise are listed which were used to determine if the samples were potential outliers, and to provide information on markers detection and peaks integration. For example, one sample at 300s (sample #32) was excluded because the software did not detect the internal standard retention time which was incorrectly shown as 0 minute. The remaining samples shown in Table 4-1 had ISTD retention time of 6.7 ± 0.2 minutes which was in the tolerant range of 6.7 ± 0.5 minutes set in the method. The spectra noise was calculated based on the noise elimination level set in the method, and this value varied from 230 - 250 which indicated the minimum intensity below which the signal was considered as noise. The maximum peak width of 10.5 seconds was applied for all markers for optimal area integration (Water Inc., 2010). The chromatogram noise varied from 790 - 900 which directly indicated the peak-to-peak baseline noise automatically determined by software, and a higher value resulted in less detected markers. The values in the spectral noise column and chromatogram noise column had variations of approximately 10% across the whole sample list, which indicates the consistency of the markers detection method across samples during the analysis process.

Other than the one sample at 300s (sample #32) which was excluded as mentioned previously, six other samples including one OSPW control sample (sample #1), one 0s sample (sample #9), two 20s samples (sample #11, #13), one 40s sample (sample #20), and two 60s samples (sample #21, #24) were excluded in Table 4-1 due to being potential outliers. For example, although the software detected the ISTD of those outliers within a tolerant retention time, the integrated areas of ISTD (*m/z=228.2044*) and other markers were much different compared with areas in other replicates as shown in markers table (partially shown in Table

4-2). Thus, only 27 samples remained for the PCA study with each sample group containing at least 3 replicates which allowed for a reliable statistical analysis.

Table 4- 2: Portion of Markers Table Showing Markers Information in Ozonation Datasets Processed by Markerlynx Software.



..	Ret. Time	m/z	Included	11120911 ...	11120912 ...	11120913 ...	11120914 ...	11120915 ...	11120916 ...
633	0.7281	227.1636	✓	0.0000	36.7751	0.0000	0.0000	0.0000	0.0000
634	1.0004	227.2014	✓	30.5747	55.9887	69.8026	45.5426	0.0000	52.5888
635	0.6980	227.9901	✓	144.1676	146.1047	141.9130	136.7091	145.7775	144.6584
636	0.5381	228.0694	✓	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
637	1.4164	228.0695	✓	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
638	1.0002	228.2044	✗	689.4594	966.5109	928.6005	942.8026	919.8608	951.1575
639	0.6401	229.0522	✓	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
640	0.6496	229.0889	✓	54.4693	0.0000	31.3539	42.5701	61.9516	29.9219
641	0.6349	229.1071	✓	0.0000	0.0000	0.0000	18.5905	0.0000	0.0000
642	0.6280	229.1228	✓	39.5921	0.0000	31.5837	0.0000	32.8649	0.0000
643	0.7711	229.1232	✓	128.9977	22.1771	24.5301	101.3045	17.1831	82.9326
644	0.6664	229.1235	✓	0.0000	14.4173	0.0000	0.0000	0.0000	0.0000
645	0.6489	229.1426	✓	0.0000	0.0000	23.6448	0.0000	0.0000	0.0000
646	0.7068	229.1537	✓	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
647	0.7171	229.1779	✓	17.0037	0.0000	0.0000	0.0000	0.0000	0.0000
648	0.6619	231.0670	✓	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Similar to the samples table, the first column in the partial markers table in Table 4-2 counted the number of markers detected by the software Table 4-2. As defined in Chapter 3, markers were referred to as the ions analyzed by HRMS and they represent the organic compounds detected in the OSPW samples. The second and third column label markers based on their retention time relative to ISTD retention time, and the exact mass to charge ratio (*m/z*) which was equivalent to the exact mass (Z = 1). Table 4-2 only partially shows 16 markers (#633 to #648) with *m/z* values from *227.1636* to *231.0670*, as there were 1,525 markers detected in total with *m/z* values ranging from *101.2019* to *399.2647* (the range of *m/z=100* to *400* was set in the method) by Markerlynx software. The next column with a green check mark or red cross mark indicates if the marker was included or excluded for PCA study. The following columns labelled with file names illustrate the integrated areas of the markers in each sample. Note that marker #638 was removed since it is identified as the internal standard.

Examples of Noise Chromatograms

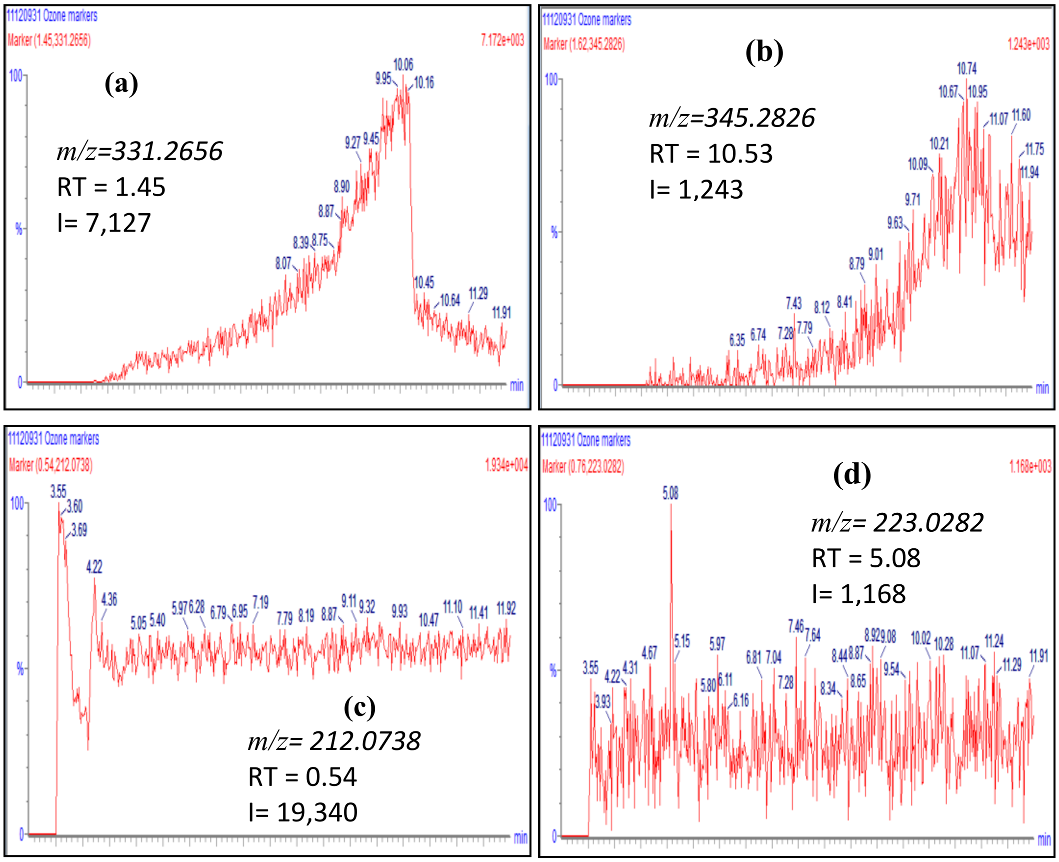


Figure 4- 1: Examples of Noise Chromatograms Extracted from Markerlynx Software. (a) noise chromatogram of *m/z=331.2656*; (b) noise chromatogram of *m/z=345.2826*; (c) noise chromatogram of *m/z=212.0738*; (d) noise chromatogram of *m/z=223.0282*.

Markerlynx software sometimes wrongly recognizes background noise as a marker if the noise had signal intensity greater than the minimum noise level set in the method (Figure 4-1 (a) to (d)). The inclusion of noises would lead to a further PCA study on noises rather than actual markers. Figure 4-1 (a) and (b) shows that the chromatograms have a peak shape so that the software determined it was a single peak with very high intensity. However, this type of chromatogram was not common, and these peaks were not considered as a normal markers, therefore two markers (*m/z=331.2656* and *m/z=345.2826*) with noise chromatograms shown in Figure 4-1 (a) and (b) were excluded in further analysis. In addition, Figure 4-1 (c) and (d) indicate that software considered the noise with high intensity as peaks. In plot (c), the incorrect marker (*m/z=331.2656*) had retention time of 3.61 minutes where there was high intensity noise found in the plot. Similarly, in Figure 4-1 (d), the noise with highest intensity is found at 5.08 minutes, which was the retention time of the incorrect marker (*m/z=223.0282*). The chromatograms shown in Figure 4-1 were extracted from a 60s sample, but similar chromatograms were found in other samples, they were concluded as noise across all samples. Thus, four markers with chromatograms shown in Figure 4-1 were excluded from the markers table to provide for a more reliable PCA study.

However, since it would be inefficient and time consuming to check for background noise determined as markers in all chromatograms considering more than 1,500 markers detected, it was determined that the PCA should first be applied to the datasets. The noise was only checked for the chromatograms of significant markers assigned with relatively large loading in the loading plot, since they were used to describe major differences in samples (Johnson and Wichern, 2007). The markers with potential noisy chromatograms were excluded, and the PCA was reprocessed on the remaining markers. The whole process was repeated several times until there was no noise with relatively large loading detected in the PCA loading plot. However, this approach only eliminated the noises with large PCA loadings, but noises with relatively small loadings were still included in the markers table. Fortunately, the small PCA loading values indicated that the remaining noises had negligible area variations across samples, so the corresponding impacts on the PCA model were negligible. Therefore, this approach saved a substantial amount of time on sorting out significant noise while minimizing the effects of non-significant noise on PCA results reliabilities.

With the approach to sort out significant noise introduced previously, in addition to the four noises shown in Figure 4-1, some other markers were found to have noisy chromatograms but with large intensities and integrated areas including: *m/z = 114-115, 148-149, 156-157, 184-185, 210-213, 223-224, 255.23-256, 283-284, 331-332,* and *345-346*. These noises were excluded before processing PCA; otherwise the PCA would have concluded that these noises were significant markers describing the differences across samples, which would have negatively impacted the PCA results and lead to a wrong direction of analysis.

4.2 Appropriate PCA Data Pre-treatment Selection

EZ Info (V2.0, Umetrics AB) was applied to process the PCA studies based on the datasets after initial processing of samples and markers tables. As discussed in Chapter 2 and 3, datasets may need to be pre-treated prior to PCA for optimum analysis. Six different combinations of data pre-treatment methods are available in EZ Info software:

(1) No scaling & no transformation

(2) No scaling & log(x+1) transformation

(3) Pareto scaling & no transformation

(4) Pareto scaling & log(x+1) transformation

(5) Center scaling & no transformation

(6) Center scaling & log(x+1) transformation

As reviewed in Chapter 2, the critical technique to determine the appropriate pre-treatment method was to compare the PCA results from different pre-treatments with the most appropriate treatment selected as the optimum data pre-treatment method.

(1) <u>No Scaling & no transformation</u>

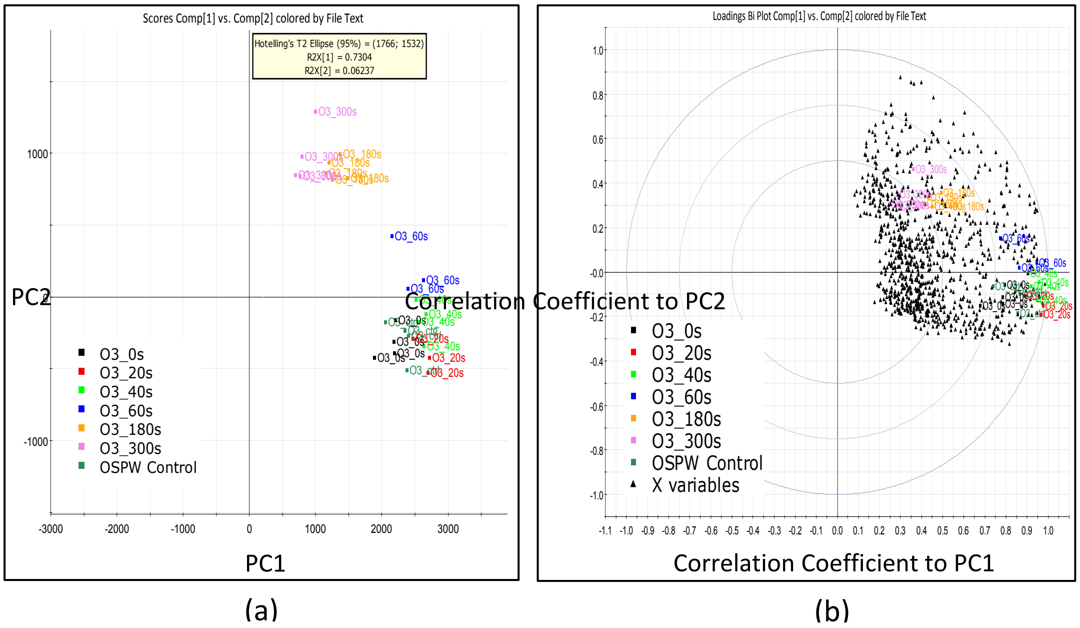


Figure 4- 2: PCA (a) Score plot; and (b) Bi-plot, based on Data without Pretreatment from EZ Info Software. OSPW control samples were in dark green; 0s samples were in black; 20s samples were in red; 40s samples were in light green; 60s samples were in blue; 180 sampls were in organe; 300s samples were in pink.

Figure 4-2(a) shows the PCA score plot based on the ozonation datasets without any data pre-treatment. In the score plot, all samples were plotted on the positive axis of PC1 which explained 73% of total variances. A total of 79% of original data variances were recovered by PC1 (73%) and PC2 (6%). PC3 would explain even less variance than PC2, so it was considered as not significant. Ideally, the replicates are close to each other so that samples can be clustered into groups, because the actual plot distance is equivalent to the differences between samples (Johnson and Wichern, 2007). However, in Figure 4-2(a), one 60s sample (blue) is separated from the sample group and there is significant overlapping of sample groups. For instance, samples in 0s (black), 20s (red), 40s (light green), and 60s (blue) were overlapped, and samples in 180s (orange) and 300s (pink) were overlapped as well. Although the overlapping reflected the similarities between samples, such observations also indicated poor separation of clusters of the individual samples, so the data might not be appropriately treated before applying PCA. The bi-plot as shown in Figure 4-2(b) is used to study the correlations between the samples and markers toward principal components, with the correlation coefficients read directly from the vertical and horizontal axis. The inner ellipse boundary has a correlation coefficient of 0.50, the middle ellipse boundary of 0.75, and the outer ellipse boundary has value of 1.00. Theoretically, a higher correlation coefficient indicates that samples or markers are highly correlated to PC1 or PC2 (Johnson and Wichern, 2007). Also, markers or samples located close together are positively correlated to each other. In contrast, markers and samples that are widely separated from each other are negatively correlated (CAMO, 2011). In Figure 4-2(b), all samples and markers are clustered together on the positive PC1 axis, so they were all positively correlated to PC1 with small differences between samples and markers. Such observations were not useful in determining differences between samples, which indicated that functions of PCA were not well developed to find the negative correlations between samples and markers, so the PCA results were not reliable. Thus, it was concluded that data had to be pretreated by scaling and/or transformation before applying PCA.

(2) <u>No scaling & log(x+1) transformation</u>

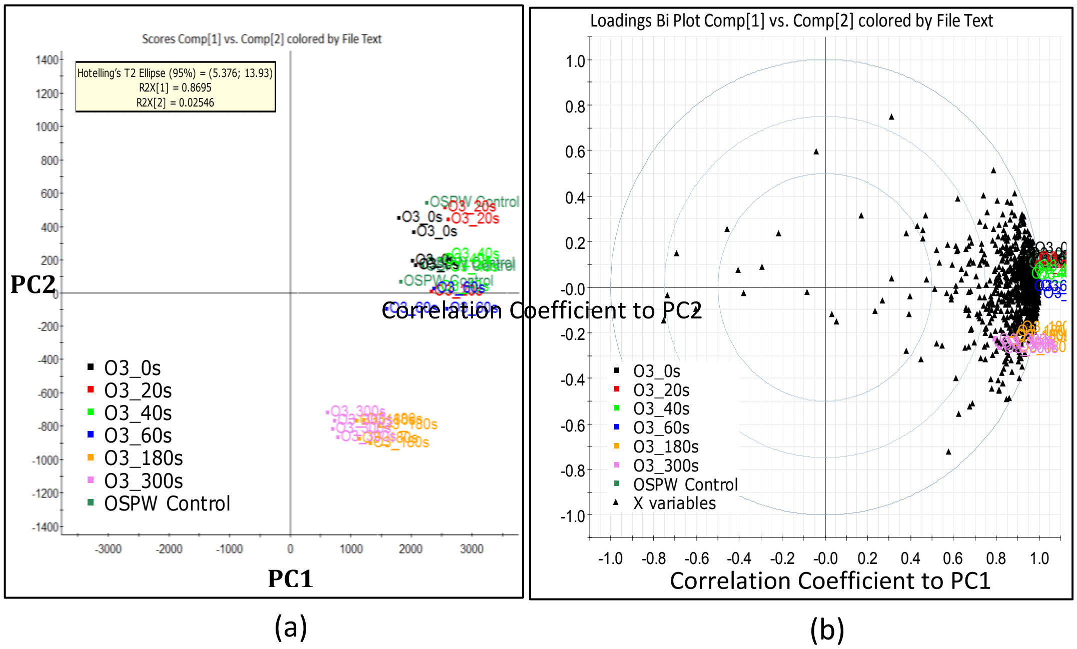
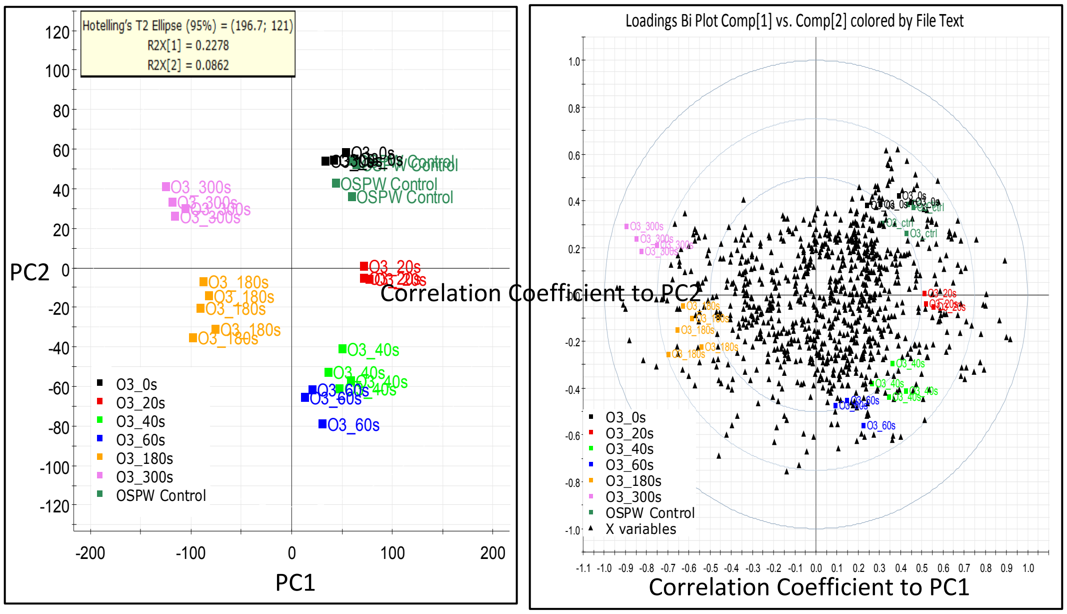


Figure 4- 3: PCA (a) Score Plot; and (b) Bi-plot, based on Data by log(X+1) Transformaiton from EZ Info Software. OSPW control samples were in dark green; 0s samples were in black; 20s samples were in red; 40s samples were in light green; 60s samples were in blue; 180 sampls were in organe; 300s samples were in pink.

Figure 4-3(a) shows the PCA score plot based on the ozonation data pretreated by log(X+1) transformation. In the clustering point of view, 300s (pink) and 180s (orange) sample groups were separated, but overlapping still happened in 0s (black), 20s (red), 40s (light green) and 60s (blue) samples. The PCA bi-plot shown in Figure 4-3(b) illustrates that although some markers were well distributed as expected on the negative axis along PC1 to show the negative correlations, samples and most of markers were still clustered together on the positive PC1 axis with correlation coefficients ranging from 0.75 to 1.00. Such clustering in Figure 4-5 shows that samples were very similar, so the function of PCA was not well developed to study the differences between samples. Although a total of 89% of the original variances were covered by PC1 (87%) and PC2

(2%), the loading values (not shown) were less than 0.1, which indicated that the components did not represent the markers. Thus, the low loading values of markers contradicted the high recovery (89%) of original data variances by principal components, which indicated problems within the PCA model. Thus, the log(X+1) transformation alone was not adequate to provide reliable PCA results.

(3) <u>Pareto scaling & no transformation</u>



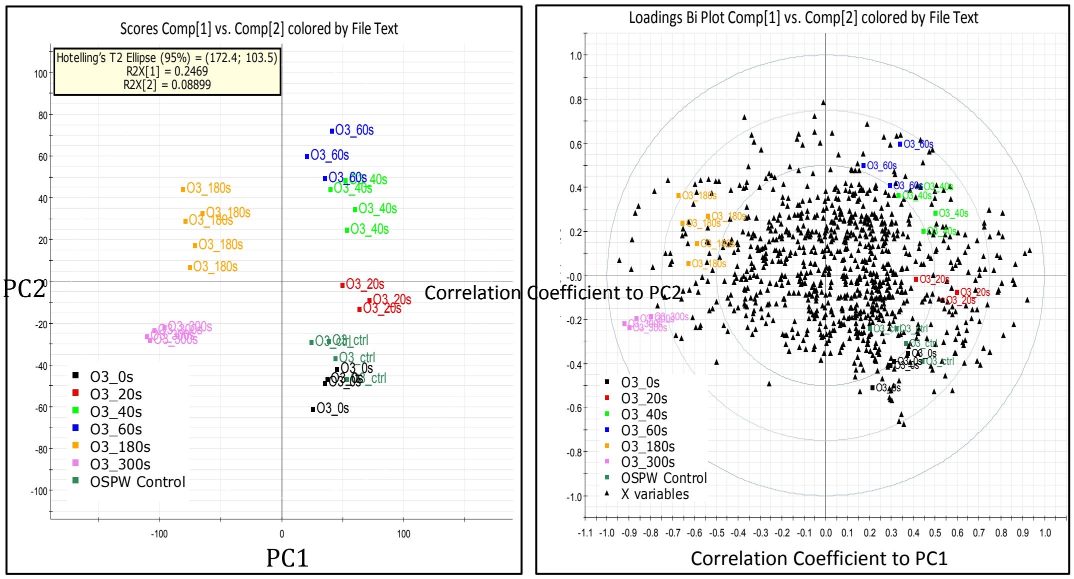
(a) (b)

Figure 4- 4: PCA (a) Score Plot; and (b) Bi-Plot, based on Data Pretreated by Pareto Scaling from EZ Info Software. OSPW control samples were in dark green; 0s samples were in black; 20s samples were in red; 40s samples were in light green; 60s samples were in blue; 180 sampls were in organe; 300s samples were in pink.

Since log(X+1) transformation alone did not improve PCA results, the next step was to test if a numerical scaling method would improve the PCA. In EZ Info software, two types of scaling methods were available including pareto and center scaling. Pareto scaling was first tested with the PCA score plot shown in Figure 4-4(a). The clustering of samples was much improved compared to the previous PCA score plots. Samples were well classified into groups, and the overlapping between the groups was reduced. Control samples (dark green) and 0s (black) samples were overlapped as expected, because those two groups of samples were collected from the beginning of the experiment, and should theoretically have the same number and type of markers. Although 40s (red) and 60s (blue) samples were closely clustered, they still have a noticeable separation and the small distances between them were possibly due to the presence of similar markers considering the samples were only 20 seconds apart from each other. However, PC1 and PC2 with 23% and 9%, respectively, only explained 32% of the total variance. The reduction of total explained variations indicated that the generated principal components lost some information from the original data. There are no specific rules to specify the minimum amount of original variances that an appropriate PCA model has to cover, but it is recommended that the outcomes will be significantly improved with a minimum sample population larger than five times the total number of analyzed variables (Osborne & Anna, 2004). With over 1,500 markers detected in each OSPW sample it would be impossible to prepare over 5,000 samples to analyze. However, PCA of similar sized datasets as in the current study are considered as acceptable to explain the linear relationships between markers and samples in recent literature (Fraser et al., 2013; Sleighter et al., 2010). For example, the PCA model generated by Sleighter et al. (2010) only covered 47% of original data variances by PC1 (28%) and PC2 (19%) due to 2,143 variables in datasets, but still provided sufficient information to indicate the linear relationship between variables. As well, the PC3 with additional 13% variance was rejected because it complicated analysis and provided little additional information. Similarly, the first two components currently were selected by EZ Info software to represent the PCA results, because

PC3 would explain less than the 8% of PC2, and would further complicate the analysis. The bi-plot shown in Figure 4-4(b) demonstrates that samples and variables were well separated and distributed along PC1 and PC2 axis, so that the correlations between principal components, samples and variables could be easily observed. Overall, the PCA plots generated based on the data pretreated by pareto scaling were considered acceptable for explaining the overall dataset.

(4) Pareto scaling & log (X+1) transformation



(a) (b)

Figure 4- 5: PCA (a) Score Plot; and (b) Bi-plot, based on Data Pretreated by Pareto Sacling & log(X+1) Transformation from EZ Info Software. OSPW control samples were in dark green; 0s samples were in black; 20s samples were in red; 40s samples were in light green; 60s samples were in blue; 180 sampls were in organe; 300s samples were in pink.

The pareto scaling alone successfully improved PCA samples clustering, but the loss of total variances of original datasets was a drawback, so the next analysis included pareto scaling with log(X+1) transformation to test if the log transformation could help increase the total variance. Theoretically, if logarithmic relationships are hidden in the original datasets, the log transformation is expected to expose them and transforms the data into a normal distribution to improve the PCA results (Van den Berg et al, 2006). Figure 4-5(a) and (b) show the PCA score plot and bi-plot for ozonation data pretreated by pareto scaling with log (X+1) transformation. A total of 34% of the variances were explained with PC1 at 25% and PC2 at 9%. Compared with 32% of total variances covered by PCA based on data pretreated by pareto scaling only, log(X+1) transformation only marginally improved the total variance of original data coverage. On the other hand, this transformation dispersed the replicate clusters with distances between replicates in 40s, 60s and 180s samples groups slightly increased, which indicated the differences between replicates in the same sample group. Thus, with little improvement in total original data recovery, the log(X+1) transformation was not recommended to be applied together with pareto scaling pre-treatment.

(5) <u>Center scaling & no transformation</u>

Pareto scaling improved the PCA results as shown previously, the center scaling method was tested if it could help better improve PCA results. Figure 4-6(a) shows the PCA score plot based on the ozonation data pretreated by center scaling. PC1 (37%) and PC2 (9%) explained approximately 46% of total variance, which was 14% higher than the data treated by pareto scaling alone. However, in comparison to pareto scaling, overlapping happened between 40s (light green) and 60s (blue) samples, and the distance between replicates in 60s and 180s samples (orange) groups increased, so the clustering was more dispersed. In the bi-plot shown in Figure 4-6(b), the samples and variables are well distributed, but the distances between replicates in 60s samples (blue) are considerably larger, which indicates that the clustering was rather dispersed, so pre-treating the data by center scaling did not to improve the PCA results.

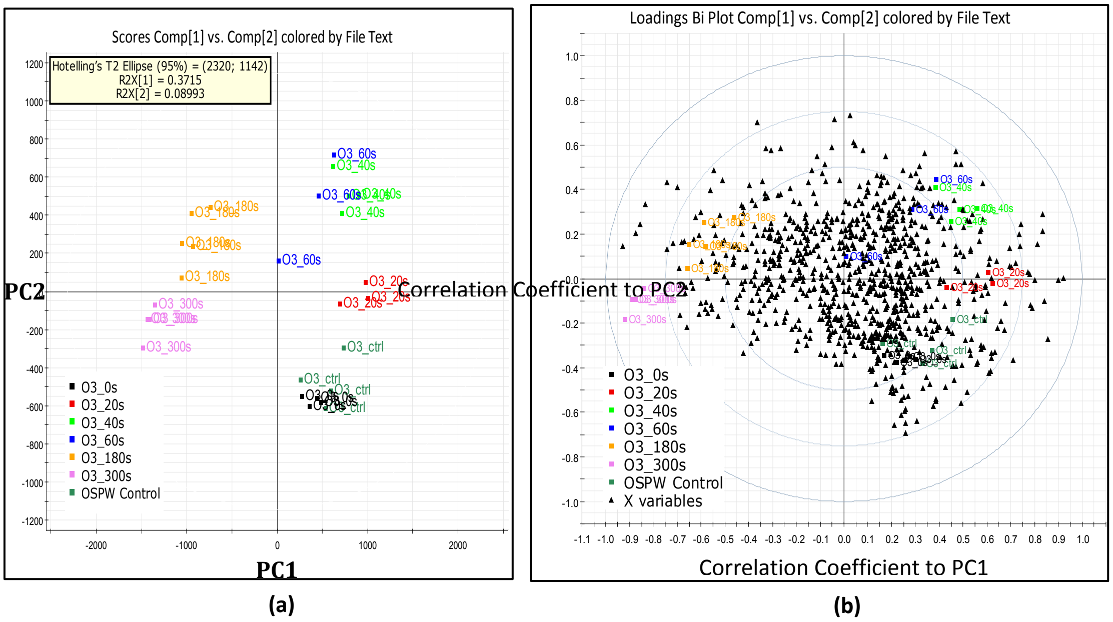
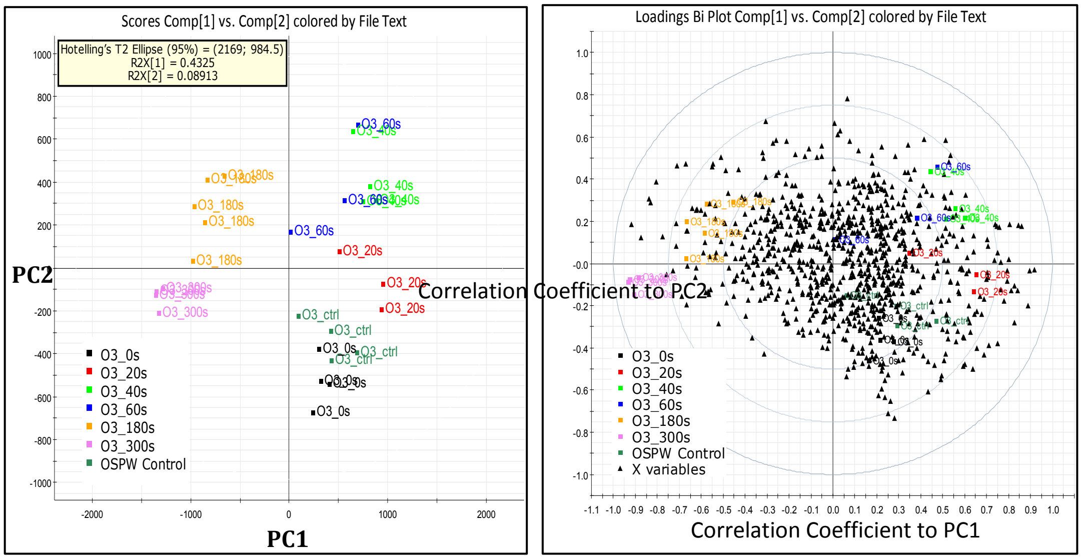


Figure 4- 6: PCA (a) Score Plot; and (b) Bi-Plot, based on Data Pretreated by Center Scaling from EZ Info Software. OSPW control samples were in dark green; 0s samples were in black; 20s samples were in red; 40s samples were in light green; 60s samples were in blue; 180 sampls were in organe; 300s samples were in pink.

(6) <u>Center scaling & log (X+1) transformation</u>

Figure 4-7(a) shows the PCA score plot of the data pre-treated by both center scaling and log(X+1) transformation. A total of approximately 50% variance with PC1 of 43% and PC2 of 9% was explained, which was similar to the data treated by center scaling alone. In both score plot and bi-plot, there was overlapping between 40s (light green) and 60s samples (blue), and the distances between replicates in 60s samples group showed large dispersion, so the center scaling and log(x+1) transformation together could not improve the overall PCA results.



(a)

(b)

Figure 4- 7: PCA (a) Score Plot; and (b) Bi-Plot, based on Data Pretreated by Center Scaling and log(X+1) Transformaiton from EZ Info Software. OSPW control samples were in dark green; 0s samples were in black; 20s samples were in red; 40s samples were in light green; 60s samples were in blue; 180 sampls were in organe; 300s samples were in pink.

PCA Pre-treatment Conclusions

Overall, after comparing the PCA results from different data pre-treatments shown above, it was concluded that data without any scaling treatment showed poor clustering results, because samples groups were overlapped, and distances between replicates in the individual groups were large. In bi-plots the samples were clustered together, so the negative correlations between samples and variables could not be studied, which indicated that the functions of PCA were not well developed.

It was observed that log(X+1) did not improve the total original data variances recovered by the PCA model, but actually hampered the clustering by increased distances between replicates and overlaps between sample groups. Reid and

Spencer (2009) indicated that the log transformation would shift the data to the negative side of skewness. The skew values of the ozonation datasets applied in this project varied from negative one to positive one, so the log(X+1) transformation would not be appropriate for the datasets with the negative skew values.

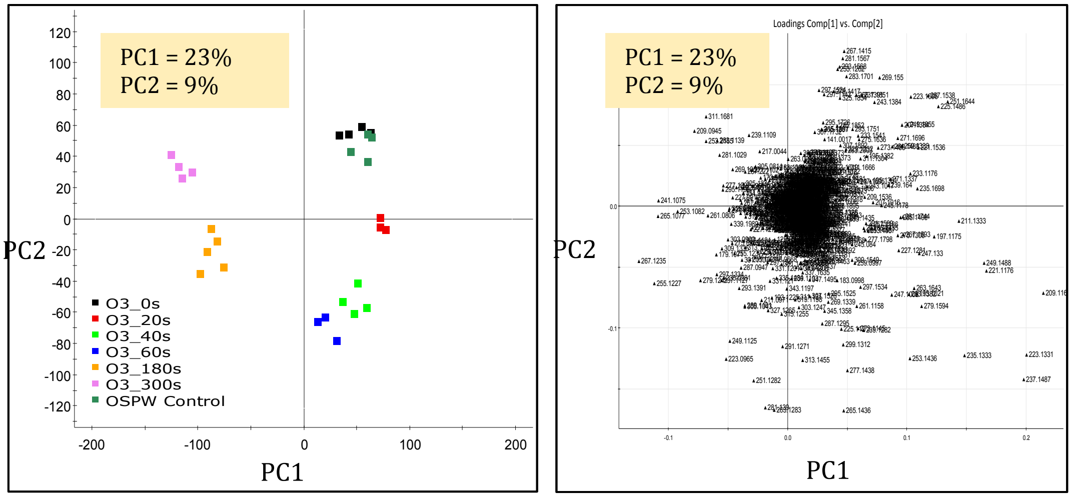
Pareto scaling and center scaling showed acceptable PCA score plots and bi-plots in the point view of clustering. However, by comparing the results from these two methods, center scaling still showed the overlapping of samples, especially between 40s and 60s samples, and relative large distances between replicates. Therefore, the pareto scaling alone showed the best clustering results with little overlapping and small distances between replicates as expected, which indicated those replicates had similar compounds given they were the repeated injections from same sample. Therefore, the pareto scaling method was considered as the appropriate data pre-treatment method for HRMS ozonation datasets before applying PCA.

Similar results were observed by applying PCA on the other datasets including CO_3^{2-}, TBA, TBA+CO_3^{2-}, TNM and Fe (II) (PCA plots generated with pareto scaling as optimum data pre-treatment were shown in Appendix B-1 to B-5). Data without any treatment showed dispersed clustering and log(X+1) transformation did not improve either the clustering or the recovery of original data variances. Compared with data treated by center scaling, data treated by pareto scaling showed tighter clustering results. In addition, the pareto scaling has the advantage of compensating the relative importance of variables with large values so that more correlations of the variables can be studied, while keeping partial original data structure without making all variables equally important as in other scaling processes (Van den Berg, 2006). Thus, the pareto scaling was selected as the appropriate data pre-treatment method before applying PCA on all ozonation in different conditions datasets.

4.3 PCA Advanced Interpretation

With pareto scaling as data pre-treatment, the PCA results based on ozonation datasets are shown in the PCA score and loading plots in Figure 4-8.

In the score plot shown in Figure 4-8(a), samples are clustered into groups. Since the distances between samples were proportional to the differences between samples, small distances between replicates within each sample group indicates their consistency. In addition, a total of 32% of original data variances were recovered by PC1 (23%) and PC2 (9%). Compared with the Sleighter et al. (2010) PCA model which explained 47% of variances in total by the first two PCs, it was assumed that the first two components shown in Figure 4-8 were sufficient to indicate the linear relationship between markers due to the large variables datasets (>1,500 markers in each OSPW sample). Moreover, it was observed that samples collected in the first minute were distributed vertically along PC2 axis, and samples taken after first minute were separated horizontally along PC1 axis (Figure 4-8(a)). Since PC1 (23%) explained more variances compared with PC2 (9%), the distance along PC1 axis indicated greater differences between samples compared with the distance along PC2 axis. Therefore, 180s and 300s samples which were horizontally separated were more dissimilar from OSPW control and 0s samples, compared to vertically separated samples collected in first minute.



(a) (b)

Figure 4- 8: PCA (a) Score Plot; and (b) Loading Plot, based on Ozonation Data from EZ Info Software. OSPW control samples were in dark green; 0s samples were in black; 20s samples were in red; 40s samples were in light green; 60s samples were in blue; 180 sampls were in organe; 300s samples were in pink.

In the loading plot shown in Figure 4-8(b), the markers which represented the organic compounds detected in the samples are labelled using their exact mass to charge ratio (*m/z*), and they are also assigned with loading values plotted on PC1 and PC2 axis. Markers with small loading values near the origin were very similar in concentrations across the samples, but markers with higher loading values (i.e., located at furthest from the origin of loading plot) would vary most in concentrations and describe the major differences in the samples. Thus, the overall analysis of the individual markers currently started from markers furthest from the origin, and narrowed down toward the origin. However, there were over 1,500 markers plotted in Figure 4-8(b), with many of them not being important due to their similar concentrations in all samples, so it was necessary to separate significant and non-significant markers prior to further interpretation.

Traditionally, the correlation coefficient between the variables and principal components is a critical way to separate significant and non-significant variables (Johnson and Wichern, 2007; CAMO, 2011). Figure 4-9 is a representative PCA Bi-Plot which was used to study the correlations between samples or markers between the PC1 and PC2 based on the correlation coefficients shown on each axis. The outer ellipse sets the boundary of maximum correlation coefficient value of 1.00, so all samples and markers were plotted within this boundary. The inner ellipse represents the correlation coefficient value of 0.75, and all markers and samples out of the middle ellipse have correlation coefficients greater than 0.75 and are considered as significantly correlated to the corresponding components (CAMO, 2011). All variables and samples within the inner ellipse (correlation coefficient value of 0.50) are considered as non-significant to the PCA model, because their correlation coefficients are less than 0.50 (Johnson and Wichern, 2007; CAMO, 2011). Therefore, all variables labelled by triangles within the outer and middle ellipse were highlighted with red squares, because they had correlation coefficients greater than 0.75 for both PC1 and PC2, and were considered as significant variables to PCA model (Figure 4-9). The large loading values of those highlighted markers found in the loading plot further proved their significance to the PCA model. However, there were only 27 highlighted variables, and only 300s samples group had correlation coefficients greater than 0.75 on PC1 axis, due to the low total variances of original data covered by PC1 and PC2. In order to explore more markers that could be used to determine differences between samples, it was necessary to study the markers with correlation coefficients of 0.50 - 0.75 between the middle and inner ellipse. Only markers within the inner ellipse (0.50) were defined as non-significant (CAMO, 2011), so those markers with correlation coefficients between 0.50 to 0.75 were considered as less significant than those greater than 0.75, but still useful to explore the data information.

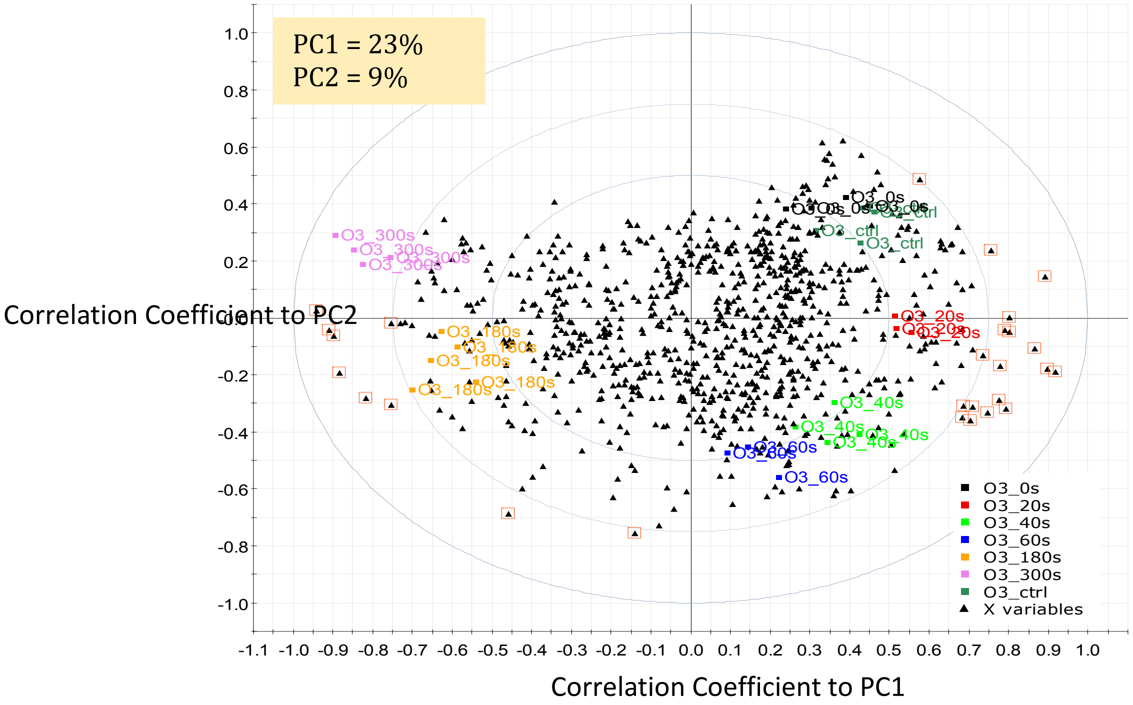


Figure 4- 9: PCA Bi-Plot based on Ozonation Data from EZ Info Software. Samples and variables within inner ellipse have correlation coefficients less than 0.50; samples and variables between middle and inner ellipse have correlation coefficients of 0.50-0.75; samples and variables between middle and outer ellipse have correlation coefficients of 0.75-1.00.

Since it is not reasonable or useful to study all the variables between the middle and inner ellipse (correlation coefficient = 0.50 – 0.75), a boundary was considered to separate the relative significant and non-significant markers in either bi-plot or loading plot. Due to the different scales of the bi-plot and loading plot, and the impacts of PC1 and PC2 loadings simultaneously on one marker, it was observed that the markers which were similar in the bi-plot (i.e., tightly clustered) might actually be dissimilar (i.e., more dispersed) in the loading plot. For example, although the 27 highlighted markers in bi-plot (Figure 4-9) were all on the outside of loading plot, not all of these markers were found on the outside of loading plot (Figure 4-10) as highlighted in the bi-plot. However, some of them were found between the outer and middle ellipse (0.50 to 0.75). In addition, the samples in the score plot are traditionally combined with the markers in the loading plot to interpret PCA results and highlight the markers for characterizing the samples (Johnson and Wichern, 2007). Therefore, the boundary to separate relative significant and non-significant markers was set in the loading plot as shown in Figure 4-10. The correlation coefficients of those significant markers out of the boundary were checked in the bi-plot to make sure they were between the middle and inner ellipse with correlation coefficient within range of 0.50 - 0.75.

In order to set the relative significance boundary to separate relatively significant and non-significant markers in loading plot (Figure 4-10), the analysis started with markers furthest from the origin and moved toward the origin in loading plot. It was expected that the markers on the outside were specific to certain samples, so they would have dramatic differences between samples, such as increasing or decreasing in their peak areas. However, when the analysis proceeded toward the origin, the markers would show similar concentrations in all samples, so their trends in terms of areas would tend to be constant across the samples. By comparing the trends in areas of markers across samples, the analysis was stopped when those approaching constant trends were observed, because the purpose of PCA was to identify the significant variables (markers) that help to describe the major differences between samples. The relative significance boundary in red line is shown in Figure 4-10 was drawn by connecting corresponding markers which were becoming constant between samples.

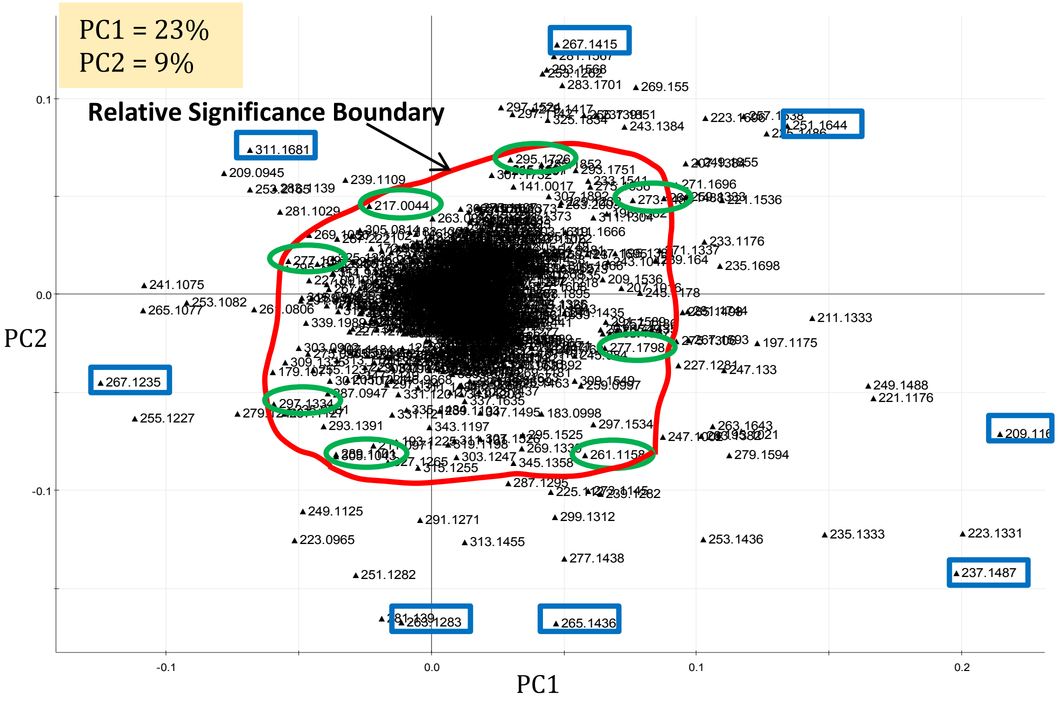


Figure 4- 10: Relative Significance Boundary Defined in PCA Loading Plot based on Ozonation Data from EZ Info Software.

In Figure 4-10, 8 markers in green circles were used to define the relative significance boundary within the red highlighted line, and their trends across samples (Figure 4-11) were relatively constant compared with the 8 markers in blue squares (Figure 4-12) which were furthest from the origin. Of note are the error bars (± 1 standard deviation) in Figure 4-11 and Figure 4-12 are approaching the marker area values, which indicated that the variations of areas between the replicates in each sample group were large. Discussion regarding such errors is included in Section 4.5. The trends for relatively significant and non-significant markers were reviewed based on average area values.

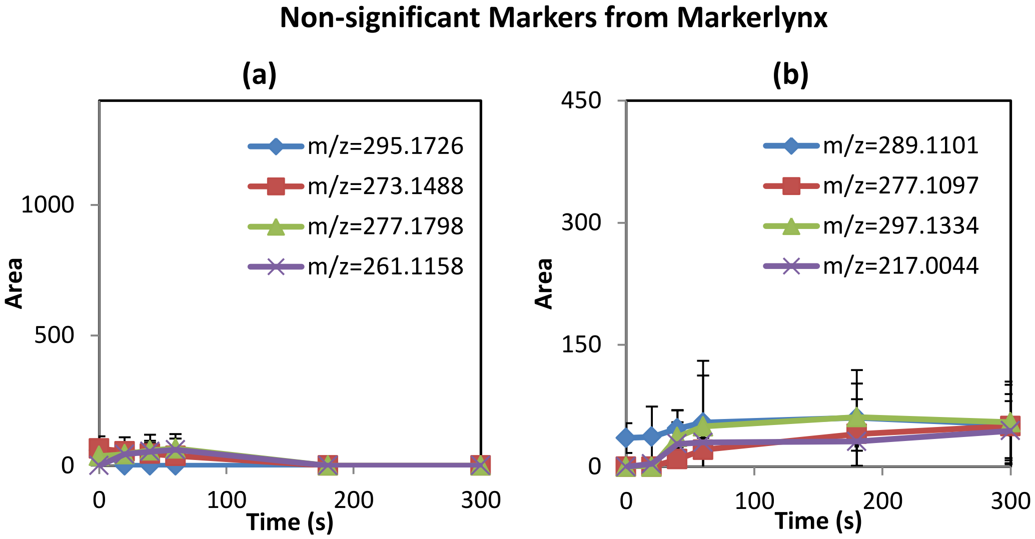


Figure 4- 11: Relative Non-significant Markers Selected from PCA Loading Plot based on Ozonation Datasetsby Markerlynx, where (a) shows markers on positive PC1 axis in loading plot; (b) shows markers on negative PC1 axis in loading plot.

Figure 4-11 shows the trends of the 8 non-significant markers within the relative significance boundary. Plot (a) shows the markers (*m/z=295.1726, 273.1488, 277.1798* and *261.1158*) on positive PC1 axis, and plot (b) shows markers (*m/z =289.1101, 277.1097, 297.1334* and *217.0044*) on negative PC1 axis. All markers were shown in the same vertical axis scale in which the significant markers were plotted (Figure 4-12) for consistent comparisons.

In Figure 4-12 (a), the trends of 5 significant markers (*m/z = 267.1415, 251.1644, 209.1167, 237.1487* and *265.1436*) with larger loadings on the positive axis of PC1 were plotted. It was found that *m/z=237.1487* varied most (range of 900 total area) across samples. Alternatively, *m/z = 267.1415* and *265.1436* showed lower variations of 350 total area across samples. Compared with the four markers with lower loadings on positive PC1 axis plotted in Figure 4-11(a), it was found that markers plotted in Figure 4-11(a) show relatively constant trends and the variations in areas were only about 60 across the samples, which were approximately 7% - 17% of the maximum area variation of 900 and minimum of 350 respectively in Figure 4-12(a) where four markers with larger loadings on positive PC1 axis were plotted.

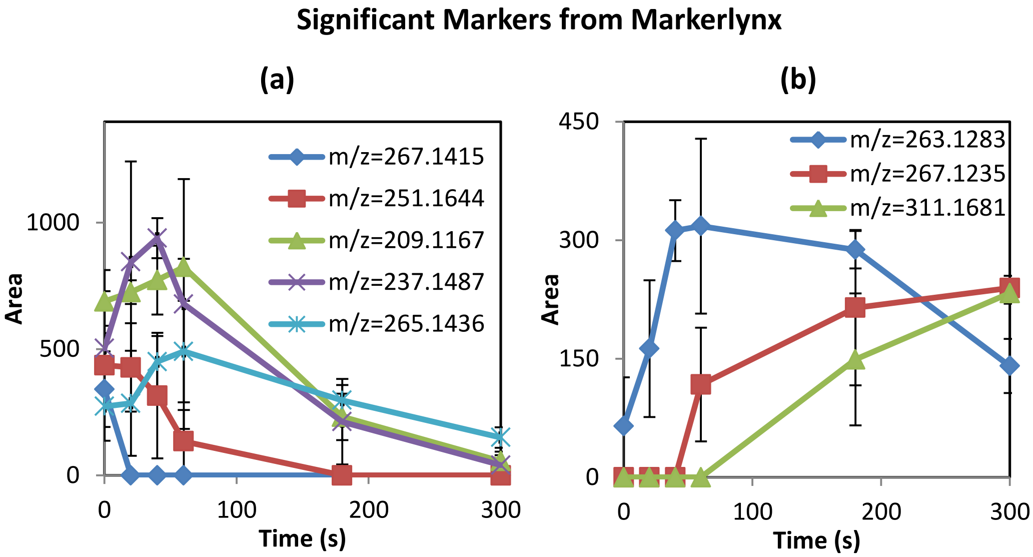


Figure 4- 12: Relative Non-significant Markers Selected from PCA Loading Plot based on Ozonation Datasets by Markerlynx, where (a) shows markers on positive PC1 axis in loading plot; (b) shows markers on negative PC1 axis in loading plot.

Similarly, compared with the trends of significant markers (*m/z = 263.1283, 267.1235* and *311.1681*) with larger loadings on negative PC1 axis plotted in Figure 4-12(b), the trends of markers with lower loadings on negative PC1 axis in Figure 4-11(b) are relatively constant. Therefore, the 8 markers with relatively constant trends shown in Figure 4-11(a) and (b) were considered as non-significant markers. Thus markers closer to the origin in the loading plot would be even less significant. All markers outside of the relative significance boundary were considered as significant markers which would be further analyzed, and all markers within the boundary were considered as the non-significant markers which would to be considered for further analysis in this project since they did not contribute to understanding of differences between samples.

4.4 Significant Markers Identification and Monitoring

In the first minute of ozonation, the direct reaction between molecular ozone and contaminants in OSPW dominates, as the molecular ozone takes time to decompose into hydroxyl radicals (Catalkaya and Kargi, 2009; Gamal El-Din et al., 2011). After the first minute, the non-selective and rapid reaction between hydroxyl radicals and contaminants becomes the dominant reaction (Catalkaya and Kargi, 2009; Gamal El-Din et al., 2011). These reaction mechanisms may be linked with the results of the PCA score and loading plots for a better interpretation of significant markers speciation. The PC2 axis, in which samples in first minute were vertically distributed, represented the direct reaction between markers and molecular ozone in the first minute where the molecular ozone preferentially reacted with markers with higher absolute PC2 loadings and lower absolute PC1 loadings. The PC1 axis, in which samples after the first minute were well separated, represented the non-selective reaction between markers and the hydroxyl radicals, where the markers with higher absolute PC1 loadings and lower absolute PC2 loadings were more reactive with hydroxyl radicals.

Significant markers defined previously were further identified by matching their exact mass to charge ratio (*m/z*) values to the known exact masses of NAs and oxidized NAs (NA+O, NA+2O, NA+3O and NA+4O) listed in Table A1-A5 in Appendix. The markers were assumed to be NAs or oxidized NAs if their exact masses were matched in a tolerance of ±0.01 Da. However, if the *m/z* values of the markers did not match with any exact mass of NA or oxidized NA, they were considered as unknown compounds. All significant markers are highlighted with different symbols in different colors in Figure 4-13.

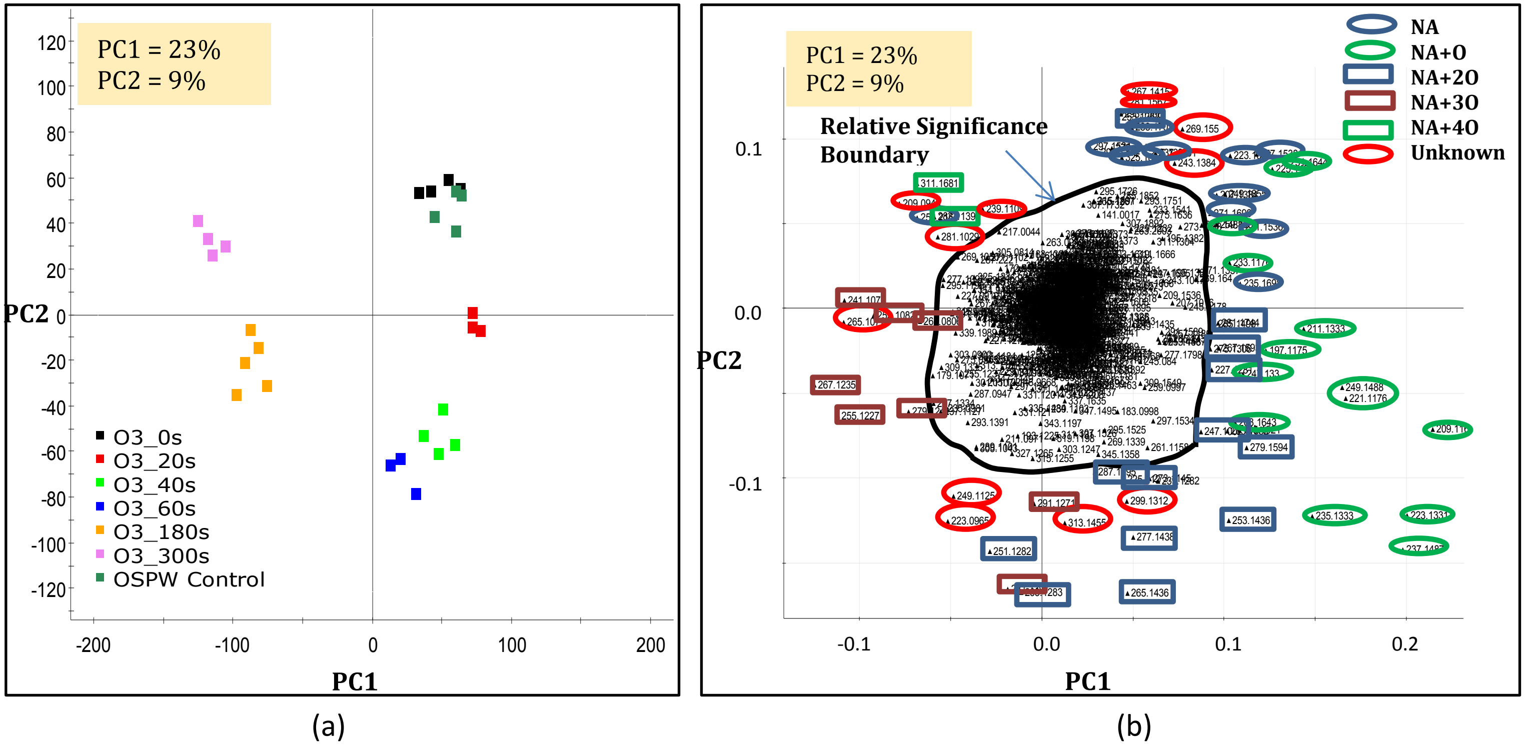


Figure 4- 13: PCA (a) Score Plot; and (b) Loading Plot, based on Ozonation Data, where blue circle=NA; green circle=NA+O, blue square=NA+2O; brown square=NA+3O; green square=NA+4O; red circle=unknown compounds.

The correlations between samples and markers were determined by combining the score and loading plots as shown in Figure 4-13. The samples in specific regions in the loading plot were positively correlated to the markers co-located in the corresponding regions in the score plot (CAMO, 2011; Jackson, 1991; Johnson and Wichern, 2007). For example, OSPW control and time 0s samples are on the upper right corner (positive PC1 and PC2) in score plot in Figure 4-13(a), so the markers on the upper right corner in loading plot Figure 4-13(b) were positively correlated to the OSPW control and time 0s samples, and those markers were mostly NAs and NA+O which were expected to have higher peak areas (proportional to higher concentrations) at the beginning of the experiment. Similarly, the markers on the negative PC2 axis would be correlated to 40s and 60s samples co-located in the score plot, so they were expected to have larger areas in samples collected at the first minute of the experiment, and the majority were oxidized NAs with the exact masses matching with NA+O, NA+2O and NA+3O. Markers on the negative axis of PC1 in the loading plot were correlated to 180s and 300s samples in the score plot, and those markers would have higher areas in samples collected at the end of the experiment, and the majority were oxidized NAs such as NA+3O and NA+4O based on their exact masses.

Therefore, samples at the beginning of the experiment had higher concentrations of NAs, NA+O and NA+2O, but through the process of ozonation, NAs and oxidized NAs were further oxidized into their higher oxidation states (NA+3O and NA+4O). These observations are in agreement with the findings reported in the literature (Gamal El-Din et al., 2011; Perez-Estrada et al., 2011; Scott et al., 2008). However, no study has previously reported on the fate of unknown compounds during ozonation process, so this project focused on tracking the behaviour of unknown markers (i.e., not identified as NAs nor oxidized NAs) during ozonation in different conditions.

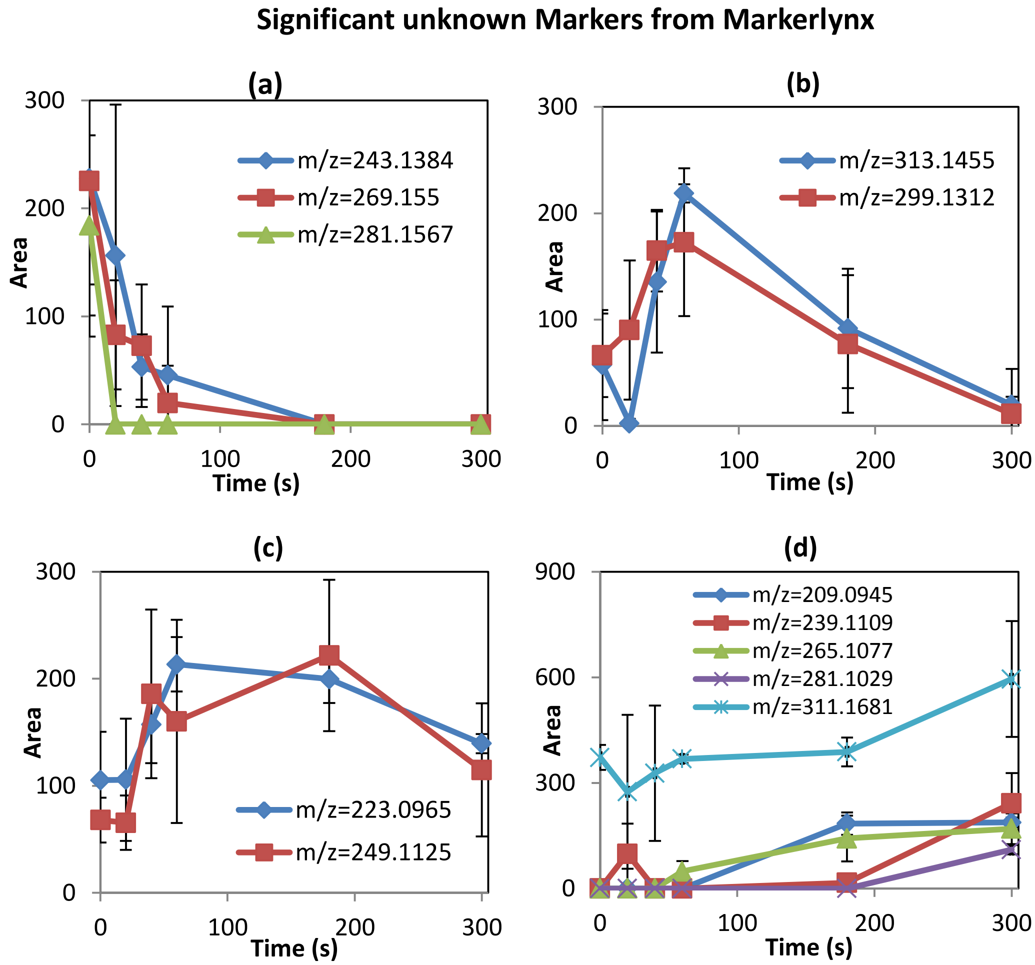


Figure 4- 14: Significant Unknown Markers Selected from PCA Results based on Ozoantion Datasets by Markerlynx. Plot (a) show markers with decreasing trends; Plot (b) and (c) show markers with increasing followed by decreasing trends, but at different rates; Plot (d) show markers with increasing trends.

Figure 4-14 shows the kinetic tendencies of unknown markers based on areas with respect to the ozonation time. The plots include the average value of replicates in each sample group, and the large error bars indicated the inconsistent peak area integration by Markerlynx which is discussed in detail in the following section. Figure 4-14 (a) shows the markers (*m/z = 243.1383, 269.1550* and *281.1567*) on the right upper corner (positive PC1 and PC2) in loading plot in Figure 4-13(b), which have a trend of decreasing area across the samples. This observation is in agreement with the prediction that those markers would have higher concentrations at the beginning of ozonation and are degraded in the ozonation process because they were positively correlated to OSPW control and 0s samples. In addition, the rates of degradation of each marker were different as shown in Figure 4-14(a). Marker *m/z=281.1567* decreased to zero area in the first 20s, which indicated that the marker was very sensitive to ozonation and could be easily degraded. However, *m/z=243.1384* and *269.1550* was more resistant to ozonation, because its degradation rate was much slower, and did not reach zero area until 180s.

Similarly, Figure 4-14 (b) and (c) show the trends of markers (*m/z = 313.1455* and *299.1312; 223.0955* and *249.1125* respectively) which were expected to be correlated to 60s samples on the negative PC2 axis. As predicted, those markers illustrated an increasing trend until 60s to reach the highest area and started to degrade after the first minute of ozonation. Both increasing and decreasing rates of markers (*m/z=313.1455* and *299.1312*) in Figure 4-14 (b) are faster than the rates of markers (*m/z=233.0965* and *249.1125*) shown in Figure 4-14 (c), which indicated that markers in plot (c) might be more resistant to ozonation.

Finally, Figure 4-14 (d) shows the trends of markers (*m/z = 209.0945, 239.1109, 265.1077, 281.1029* and *311.1681*) which were correlated to samples collected at end of the experiment. As expected, those five markers had their highest area in 180s or 300s samples. In addition, markers with *m/z=209.0945, 239.1109, 265.1077* and *281.1029* started with an initial area of zero, which indicated that those four markers were at negligible concentrations in raw OSPW but only formed later during the ozonation process. Thus, those markers are the by-products of ozonation. On the other hand, *m/z=311.1681* shows an increasing trend with initial area at beginning of ozonation in Figure 4-14(d), so it was reasonable to consider that slight oxidation had already occurred in the tailings ponds as suggested in the literature (Allen, 2008; Clemente, 2004), and the ozonation further increased its concentration. Also, it was found that unknown markers *m/z= 239.1109* and *281.1029* were formed after 180s, but unknown markers *m/z=209.0945* and *265.1077* formed after only 40s or 60s. The differences of formation time indicated that *m/z=209.0945* and *265.1077* could be formed from parent compounds that were more sensitive to ozone, but unknown compounds *m/z=239.1109* and *281.1029* might be formed from a parent compound that was more resistant to ozone (i.e., slower degradation rates) or further degradation of other by-products.

In conclusion, markers in the loading plot were correlated with the corresponding samples in the score plot, and such correlations could be reviewed in markers' trends in areas across samples with respect to ozonation time. Markers were assumed as NAs, oxidized NA or unknown compounds based on their *m/z* values. It was found that NAs or oxidized NAs were oxidized to their higher oxidation states during the ozonation process. By studying the trends of unknown compounds, it was found that some unknown markers were degraded at different rates during ozonation; some unknown markers increased at the beginning of experiments followed by degradation after the first minute of ozonation; and some unknown markers were formed or increased at different rates during the ozonation process. The markers decreasing over time indicated they were able to be removed from OSPW by ozonation, but markers increased or formed remained in OSPW after ozonation, and any unknown markers with increasing trends were the by-products from OSPW treated by ozonation in different conditions. The faster/slower degradation and formation/increasing rates indicated the markers' relative sensitivities or resistances to molecular ozone and the hydroxyl radical.

4.5 Markerlynx Validation

The area integrations by Markerlynx software showed relatively large variations between replicates. To track the variation, it was necessary to first review all integrated areas and chromatograms of the same marker across different samples. For example, Figure 4-15 shows the integrated areas of marker (*m/z=269.1550*) across all samples collected during ozonation and processed by PCA. Some sample groups only contained 3 or 4 samples after excluding the outliers. Each sample group was labelled with a different symbol across the horizontal axis. The vertical axis shows the area values of all samples, and the green and red lines set the boundary for two and three standard deviations determined using all sample values. It was observed the integrated area varied markedly in OSPW control samples labelled as blue squares and time 0s samples labelled as blue diamonds in Figure 4-15, because the variations exceeded the second standard deviation boundary.

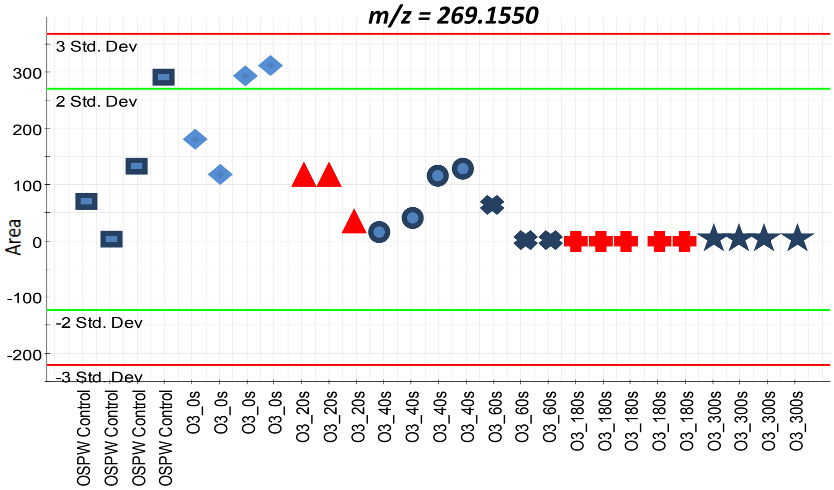


Figure 4- 15: Integrated Areas of *m/z=269.1550* across all Samples from Markerlynx Software.

To further track the high variability of the software for area integrations (e.g., area = 0 for the second OSPW control sample and area = 300 for the fourth OSPW control sample in Figure 4-15), it was necessary to review the chromatograms of the integrated peaks. Figure 4-16 shows the chromatograms of *m/z=269.1550* with four replicates in the OSPW control sample group. In Figure 4-16, the chromatograms of the extracted ion in each OSPW control sample look very similar. The intensity (I) and area (A) under the peaks were labelled with numerical values. The intensities among four replicates were 1,281, 1,556, 1,324 and 1,219 with a standard deviation of 147, which was about 10% relative standard deviation (RSD) with an average intensity of 1,345. However, the areas for four replicates were found to be 70, 0, 133 and 292. The standard deviation was 125 which was higher than the area of OSPW control sample 1, 2 and average value of 124 (RSD = 100%). Moreover, OSPW control sample 2 had the highest intensity of 1556, but the software did not recognize the peak. Thus, the problem observed above indicated that the Markerlynx software inconsistently integrated peak areas and sometimes did not recognize the peaks' presence.

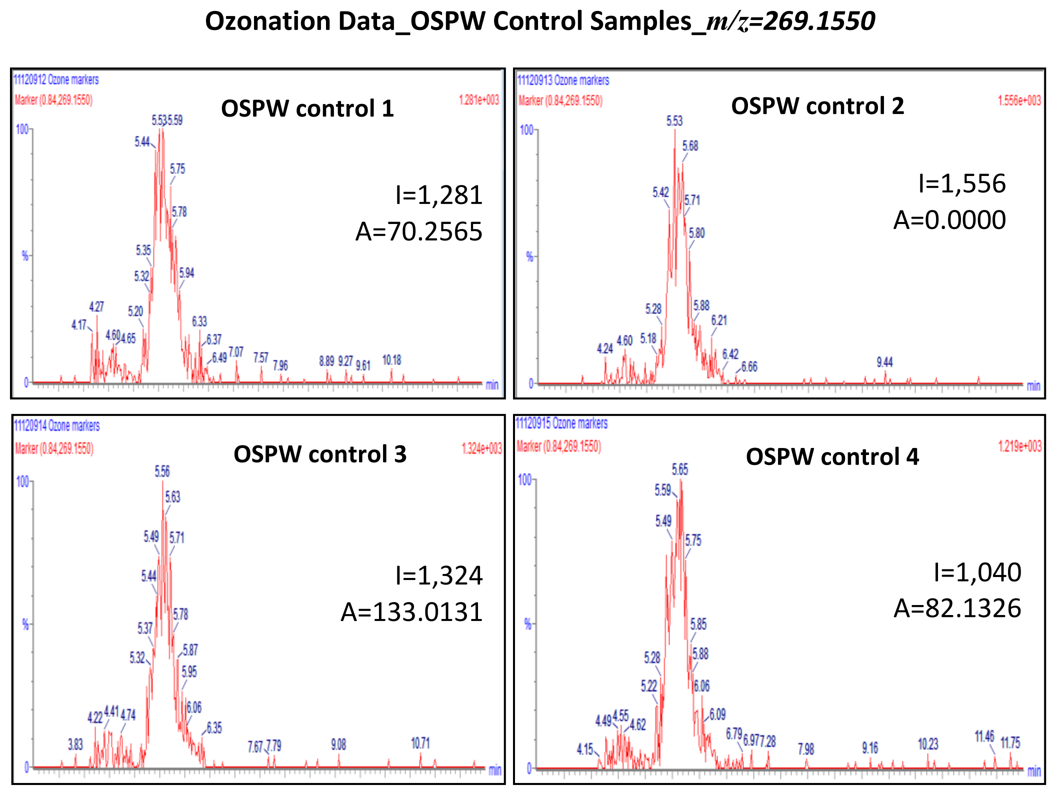

Ozonation Data_OSPW Control Samples_*m/z=269.1550*

Figure 4- 16: Extracted Ion Chromatograms for *m/z = 269.1550* in Four Injections of OSPW Control Sample.

Another example of integration issues is found in Figure 4-17 where Markerlynx software did not recognize peaks for *m/z=265.1077* which show the trend of formation in Figure 4-14(d). Since the chromatograms of replicates are similar, only one sample chromatogram from each sample group is shown in Figure 4-17 so that the change of marker's chromatograms through the entire ozonation processes could be viewed. It was found that peaks with intensities of 505, 521 and 543 actually existed in time 0s, 20s and 40s samples respectively, but the areas under those peaks were all integrated as 0. However, the peak areas in 60s, 180s and 300s samples were integrated as 82, 180 and 192 respectively, so it seemed the software could not recognize the peaks until the intensities reached

1,000. These problems indicated that the default parameters such as noise level, the minimum intensities and thresholds for peak detection used by the software needed adjustment.

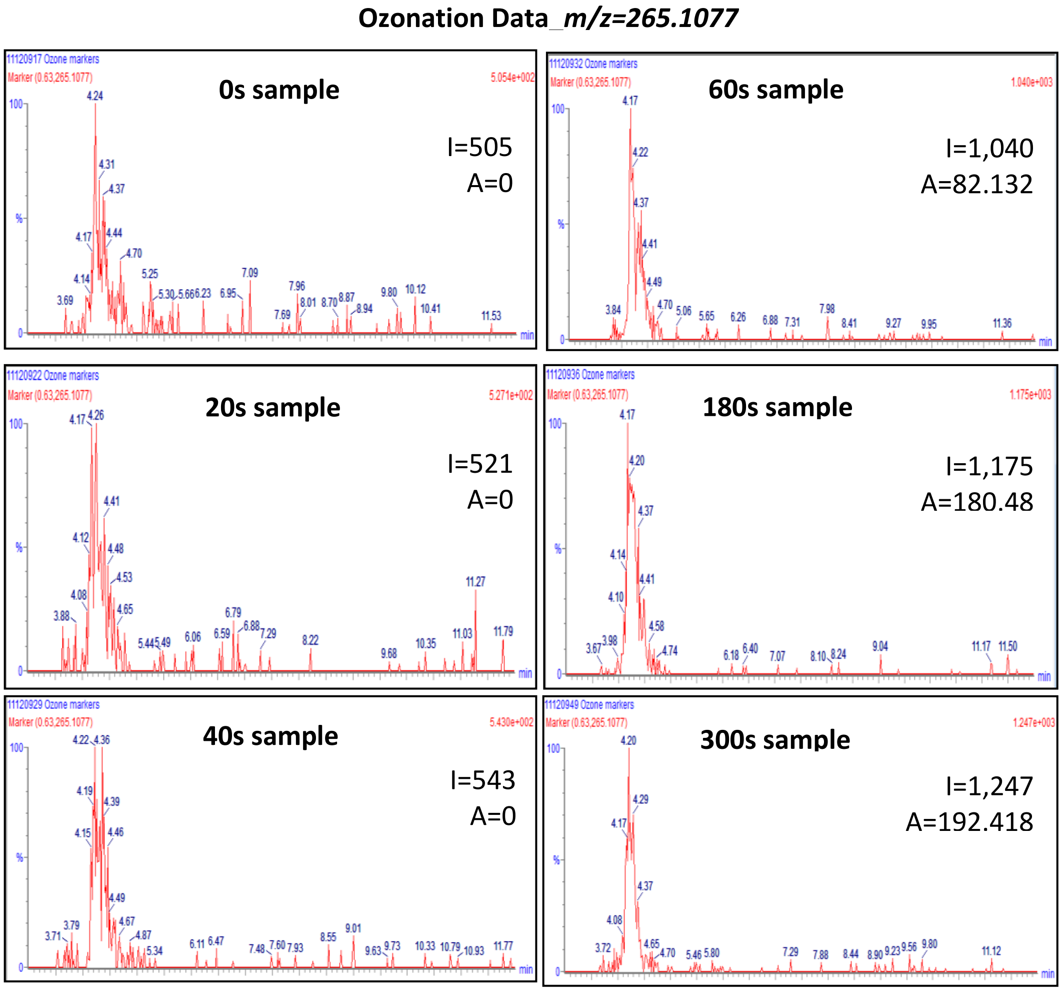


Figure 4- 17: Extracted Ion Chromatograms of *m/z = 265.1077* duirng Ozonation Processes.

The parameters set in the method for peak area integrations were discussed in Chapter 3 and shown in Table 3-1. The possible parameters that would affect the integration results were: (1) peak width, (2) peak-to-peak baseline noise, (3) smoothing, (4) marker intensity threshold (counts) and (5) noise elimination level.

The values of each parameters tested currently are shown in Table 4-3, and the effects after adjusting each parameter are discussed below.

Table 4- 3: Parameters in Method Which Potentially Influenced the Peak Area Integration

Parameters	Values			
Peak Width	10.5s	60s	100s	150s
Peak-to-peak Baseline Noise	automatic	500	300	100
Smoothing	Yes	No		
Intensity Threshold (counts)	100	50	30	10
Noise Elimination Level	6	4	2	0

(1) Peak Width

Initially, peak width was automatically determined at 5% height of the peak by Markerlynx software. Theoretically, peak width would vary from marker to marker based on the different markers' chromatograms. However, the integration results by automatic peak width showed large variations across samples. Based on Waters Inc. (2010)'s recommendations, using the maximum peak width would provide better integration results if the exact peak width was not known. Therefore the maximum peak width of 10.5 seconds automatically determined by software was applied across all samples, and all results shown previously were generated based on this maximum peak width value.

After reviewing a few markers' chromatograms, for example *m/z=269.1550* and *m/z=265.1077* in Figure 4-16 and 4-17, the peak widths were found to be about 60s which was much larger than the value of 10.5s applied previously. Therefore, a series of peak widths were manually entered including 60s, 100s and 150s to test if these changes could improve area integrations. However, after testing different peak width values, it was found that the increasing of peak width only resulted in differing areas for a few markers, but remained the same for the majority of markers. Table 4-4 shows the area results of *m/z=269.1550* and *m/z=265.1077* based on peak width of 10.5s, 60s, 100s and 150s. It was found that increasing peak width did not necessarily increase the integrated areas, because some of the areas reduced with extending peak width. For example, the integrated area of *m/z=269.1550* in 0s sample slightly increased when the peak width extended from 10.5s to 60s, and almost doubled in area when the peak width further extended to 100s, but the area dramatically decreased when the peak width extended to 150s. Also, the standard deviations of the peak areas within sample groups were still large when the peak width extended, so the area variations in the replicates were still high. In addition, increasing of peak width did not allow Markerlynx software to recognize the peaks of *m/z=265.1077* with areas still integrated as 0 in 0s to 40s samples in Table 4-4. The software did not provide output chromatographs to show which part of peak was integrated, so it was difficult to estimate the critical peak width and area values. Thus, it was not able to be concluded which area value shown in Table 4-3 was overestimated or underestimated. Therefore, increasing the peak width parameter did not help improve area integrations, so the maximum peak width of 10.5s automatically determined by Markerlynx software was kept for further analyses.

Table 4- 4: Integrated Areas for *m/z*=265.1077 and *m/z=269.1550* by applying Different Peak Widths.

Integrated Areas for *m/z = 265.1077*

Peak width	**10.5s**		**60s**		**100s**		**150s**	
Samples	average	SD	average	SD	average	SD	average	SD
ctrl	0	0	0	0	0	0	0	0
0s	0	0	0	0	0	0	0	0
20s	0	0	0	0	0	0	0	0
40s	0	0	0	0	0	0	0	0
60s	49	30	0	0	0	0	57	46
180s	142	65	135	36	208	51	179	48
300s	170	44	146	50	214	64	188	56

Where, SD = standard deviation

Integrated Areas for *m/z = 269.1550*

Peak width	**10.5s**		**60s**		**100s**		**150s**	
Samples	average	SD	average	SD	average	SD	average	SD
ctrl	124	125	168	113	465	260	202	150
0s	225	96	227	66	436	289	258	163
20s	83	50	99	51	156	155	101	37
40s	73	57	88	67	102	99	128	29
60s	20	34	0	0	67	49	0	0
180s	0	0	0	0	0	0	0	0
300s	0	0	0	0	0	0	0	0

(2) Peak-to-peak baseline noise

The baseline noise between peaks on extracted ion chromatograms were automatically determined by software, and the values varied from sample to sample with a range of 788 to 909 based on the analysis of ozonation datasets (Table 4-1). In order to allow the software to recognize more peaks with lower intensities, a series of peak-to-peak baseline noise values were tested at reduced values of 500, 300 and 100. However, the method was limited to a single baseline noise value applied for the whole sample list, so noise with higher intensities would be recognized as markers. After manually entering baseline noise, it was found that more potential noise was recognized as markers as the baseline noise was reduced. For instance, there were 1,525 markers detected with the automatic baseline and noise values ranging from 788 to 909. As the baseline noise was manually reduced to 500, there were 3,804 markers detected, and about 5,000 markers detected with baseline noise reduced to 300. Moreover, with baseline noise of 100, there were over 30,000 markers detected. However, despite the drawback of more noise with lower intensities being recognized as peaks by software, the integrated areas of *m/z=269.1550* and *m/z=265.1077* showed the same results as before, which indicated that software still could not properly recognize some peaks present in the chromatograms. Therefore, adjustment of the peak-to-peak baseline noise could not help the software to recognize peaks nor consistently integrate areas.

(3) Smoothing

Two iterations of a mean smoothing function (three data points wide) could be selected to be applied during the area integration process. However, the application of the smoothing function resulted in the Markerlynx software sometimes not detecting the internal standard. Also, this function did not solve the problem of recognizing peaks, because some peak areas of the examples markers *m/z=269.1550* and *m/z=265.1077* were still integrated as 0. Thus, smoothing was not applied since it did not improve the overall analyses.

(4) Marker intensity threshold (counts)

The intensity threshold (counts) defined the minimum intensity of a spectra peak that was recognized by software. This parameter was originally set as 100 as recommended by Waters Inc. (2010). All PCA results shown previously were based on the intensity threshold of 100, and there were 1,525 markers detected. After reducing the intensity threshold to 50, there were 1,536 markers detected, so the total markers slightly increased. However, the integrated areas were exactly the same for example markers *m/z=269.1550* and *m/z=265.1077*, and all other markers as well. In addition, similar results were observed by reducing the intensity threshold to 30 and 10, but realistically only more noises were being considered as markers. Therefore, the intensity threshold could not improve area integrations and was not applied further.

(5) Noise elimination level

Noise elimination level defined the number of standard deviations of the baseline noise to be eliminated. This parameter directly determined the magnitude of spectra noise shown in sample table (Table 4-1), and the spectra noise defined intensity threshold value below which the response was considered as noise (Waters Inc., 2010). The elimination level was originally set as 6, which is in the range of 4 to 10 recommended by Waters Inc. In order to have the software recognize more peaks, the elimination level was reduced to 4, 2 and 0. At the elimination level of 6, the spectra noises were about 250 consistently across all samples, and resulted in 1,525 markers detected. When the elimination level was reduced to 4, the spectra noises were about 200 across all samples; and 1,596 markers were detected. When the noise elimination level was set as 2, the spectra noise was reduced to about 70 and 1,728 markers were detected. As the elimination level was set to 0, the spectra noises became 0 for all samples and 1,888 markers were detected. However, regardless of the spectra noises reduction, the integrated areas for the example markers *m/z=269.1550* and *m/z=265.1077* remained unchanged. Thus, this parameter could not improve the area integrations.

Overall, with the combinations of different parameters tested previously shown in Table 4-3, the results still showed large standard deviations in integrated areas of replicates, and peaks in the chromatograms for markers *m/z=269.1550* and *m/z=265.1077* were integrated as 0 which indicated that the software was not able to recognize all actual peaks. Therefore, it was concluded that changing the parameters in the method could not help to improve the area integrations, and it was suggested that the software needed to be reassessed by the manufacturer. Unfortunately, there was no plot given to show which part of the peak had been integrated. Therefore, it was difficult to determine the critical value of correct area integration and to conclude which peak areas were being overestimated or underestimated.

4.6 Targetlynx

After discussion with Waters Inc. concerning the area integration issues associated with Markerlynx software, it was suggested that Targetlynx be applied for manually selecting peaks for more accurate area integrations. Targetlynx software is a quantitative tool to estimate the concentrations of organic compounds based on known concentrations of an appropriate internal standard, and the concentration estimation was expressed in Equation 4-1:

$$\text{Estimated Concentration} = \frac{\text{Peak Response in Area}}{\text{ISTD Response in Area}} \times \text{ISTD Concentration} \; \ldots. \; \textbf{Equation 4-1}$$

However, since many of the current markers considered were unknown compounds, the internal standard used for quantifications of NAs and oxidized NA was not suitable for estimating concentrations of the unknown compounds. Since the peak responses in the form of areas were positively proportional to the concentrations, the trends of the compounds concentrations changing during the treatment processes could be reviewed in the form of the peak areas (i.e., a semi-quantitative analysis).

The non-significant and significant markers shown in Figure 4-10 and 4-11 (processed by Markerlynx software) were re-processed by Targetlynx software, and their trend plots are shown in Figure 4-18 and 4-19.

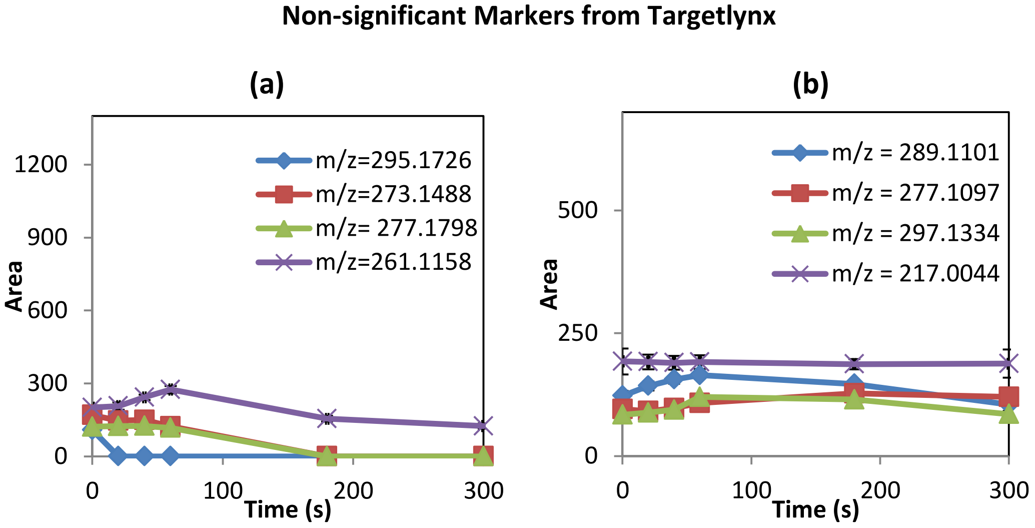


Figure 4- 18: Trends of Non-significant Markers by Targetlynx. (a) trends of markers highlighted with green circles on negative PC1 axis in Figure 4-10; (b) trends of markers highlighted with green circles on positive PC1 axis in Figure 4-10.

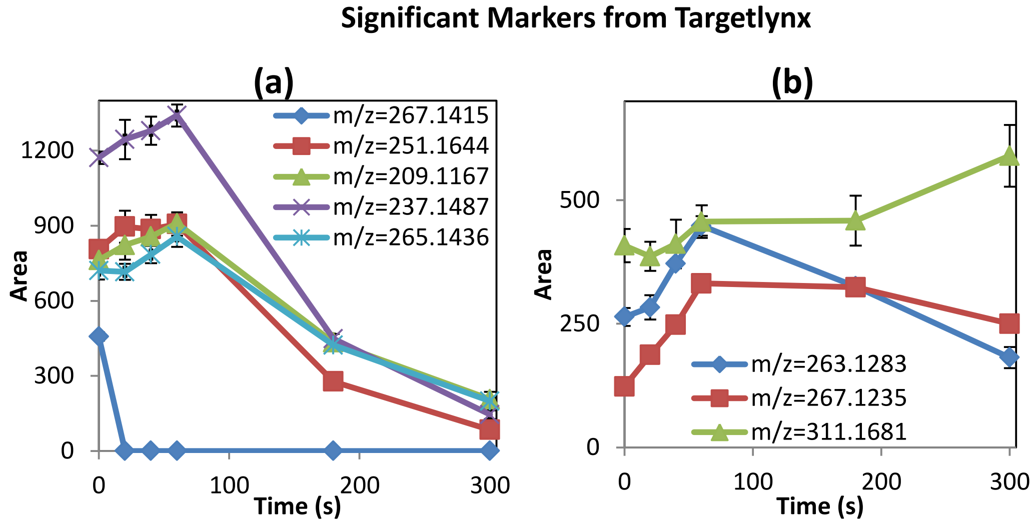


Figure 4- 19: Trends of Significant Markers by Targetlynx. (a) trends of markers highlighted with blue squares on negative PC1 axis in Figure 4-10; (b) trends of markers highlighted with blue squares on positive PC1 axis in Figure 4-10.

The areas of non-significant and significant markers integrated by Targetlynx were plotted using the same area and time scales for consistent comparisons to Markerlynx results. For example, four non-significant markers (*m/z=295.1726, 273.1488, 277.1798* and *261.1158*) that defined the positive PC1 axis of the relative significance boundary in Figure 4-10 were plotted in an area scale of 1400 in Figure 4-18(a). The five most significant markers (*m/z=267.1415, 251.1644, 209.1167, 237.1487* and *265.1436*) that had large positive PC1 or PC2 loadings (Figure 4-10) were plotted in Figure 4-19(a). It was clear to see that the trends shown in Figure 4-18(a) are relatively stable compared with the trends shown in Figure 4-19(a). For instance, the area variations of significant markers over time ranged from about 450-1,000 in area, but the area variations of non-significant markers were only about 40-200 in area, which was about 10%-20% of the variations of significant markers. Similarly, the trends of non-significant markers (on the negative PC1 axis in loading plot) shown in Figure 4-18(b) are stable as compared to the trends of significant markers (on the negative PC1 axis in loading plot) shown in Figure 4-19(b), because the area variations of non-significant markers over time were about 10-50 in area, which were about 5%-15% of the variations in areas of about 200-300 in area for significant markers. Therefore, the changes over time for the markers shown in Figure 4-18 were relatively non-significant compared with the 8 significant markers shown in Figure 4-19 based on their trends plots, so the relative insignificance boundary defined in Figure 4-10 by Markerlynx was acceptable based on the area integrations by Targetlynx.

Comparing the trends plots generated by Targetlynx (Figure 4-18 and 4-19) and Markerlynx (Figure 4-11 and 4-12), for the same marker, it could be observed that the average areas integrated by Targetlynx were larger than the average areas integrated by Markerlynx, and the standard deviations of areas in each sample group by Targetlynx were lower than those generated by Markerlynx. Such observations were expected since there should be more consistency of replicates in the same sample groups using a manually defined peak area. Generally, most markers showed similar overall trends by using the two different software approaches which validates the results considered in previous sections. However, there were some exceptions including markers *m/z=267.1235* and *311.1681*. By comparing Figure 4-12(b) and Figure 4-19(b) where these two markers were plotted by Markerlynx and Targetlynx, respectively, it was found that both *m/z=267.1235* and *311.1681* had no initial values until the first minute of the experiment by Markerlynx, but those two markers were assigned with initial values at the beginning of the experiment by Targetlynx. The reason for these initial values differences by the two different software approaches was due to Markerlynx not detecting peaks as the extracted ion chromatograms of *m/z=267.1235* shown in Figure 4-20, while the peak could be manually selected for integration in Targetlynx. The chromatogram shown in Figure 4-20(a) is from Markerlynx with the peak intensity of 653 and integrated area of 0. While, the lower chromatogram shown in Figure 4-25(b) is from Targetlynx, and the peak with intensity of 399 was integrated with an area of 143. Moreover, the signal to noise ratio of 32, which was greater than the critical value of 10 for quantification limit, indicated the peak integrated was not noise. The chromatograms shown in Markerlynx and Targetlynx are different in Figure 4-20, because a smoothing factor was applied in Targetlynx as recommended by a field technician from Waters Inc., which gave the peak an improvement in shape prior to integration.

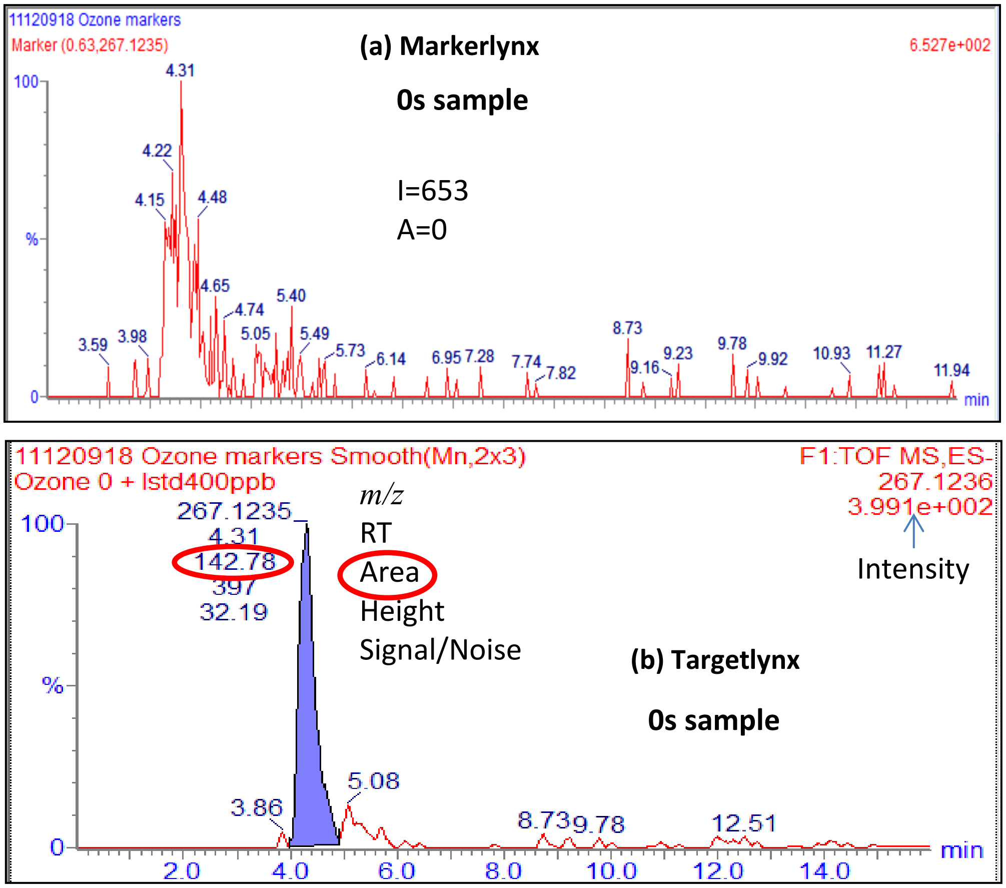


Figure 4- 20: Chormatogram of Significant Marker *m/z=267.1235* (a) in Markerlynx Software, and (b) in Targetlynx Software.

Not only did the initial area of *m/z=267.1235* differ between Markerlynx and Targetlynx, but the overall trend of the marker was different. The trend in Markerlynx (Figure 4-12(b) shows that the marker was formed after 40s, and the area increased until the end of experiment. However, the trend in Targetlynx (Figure 4-19(b) shows that the marker steeply increased in first minute, but slightly decreased after first minute. The differences between the trends shown in two different software approaches were due to several reasons. For instance, the trends were based on the average integrated areas, so the large variations of areas in replicates would lead to the average area to be overestimated or underestimated by Markerlynx. Also, the area integration algorithms between Targetlynx and Markerlynx were different (Waters Inc., 2010). The integrations in Targetlynx were more user-controllable contrary to the software-control of Markerlynx, therefore the area integrations from Targetlynx could generally be considered as more reliable.

Similarly, the significant unknown compounds shown in loading plot (Figure 4-13) were re-processed by Targetlynx, and their trends were plotted in Figure 4-21. Markers (*m/z=243.1384, 269.1550* and *281.1567*) in Figure 4-21(a) show a degradation trend during the ozonation process, so those markers were associated with OSPW control and 0s samples by combining PCA score and loading plots shown in Figure 4-13. While, markers (*m/z=313.1455, 299.1312, 223.0965* and *249.1125*) in Figure 4-26(b) and (c) show a trend of increasing in the first minute followed by decreasing at different rates after the first minute, so those markers were correlated with samples collected at the first minute of experiment. Thus, the observations from Targetlynx agreed with those from Markerlynx which were shown in Figure 4-14(a), (b) and (c). Although the integrated areas for markers (*m/z=243.1384, 269.1550, 281.1567, 313.1455, 299.1312, 223.0965* and *249.1125*) were not exactly the same between the two software approaches, the overall trends for those markers were similar. As for previous results, the areas integrated by Targetlynx were larger than the areas integrated by Markerlynx, and the standard deviations of areas in each sample group by Targetlynx were much smaller than the results generated by Markerlynx. Thus, Targetlynx helped to improve the consistent area integrations of replicates in each sample group as compared to Markerlynx.

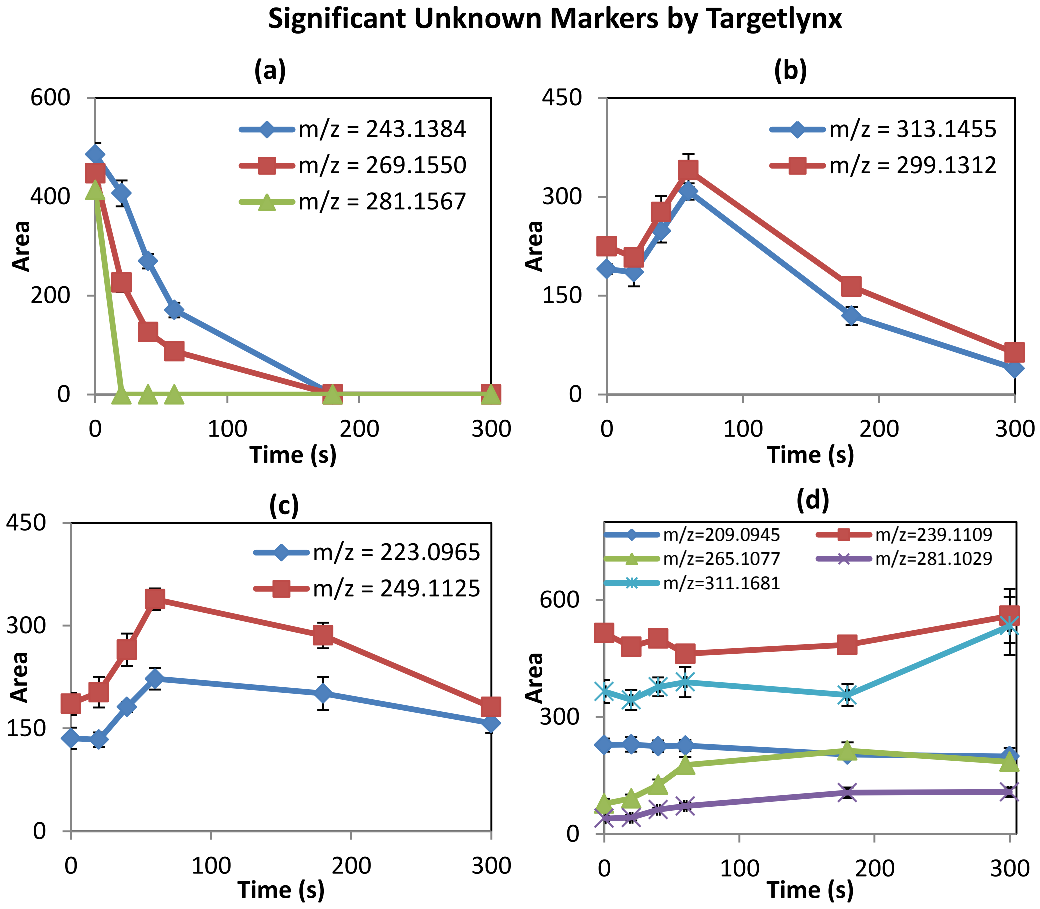


Figure 4- 21: Significant Unknown Markers Selected from PCA Results based on Ozonation Datasets by Targetlynx. Plot (a) shows markers with decreasing trends corresponded to Figure 4-14(a); Plot (b) and (c) show markers with increasing followed by decreasing trends corresponded to Figure 4-14(b) and (c), but at different rates; Plot (d) shows markers with increasing trends corresponded to Figure 4-14(d).

In addition, by comparing Figure 4-21(d) and 4-14(d) where the significant unknown markers with an increasing trend were plotted, it was found that the markers (*m/z=281.1029* and *265.1077*) had initial areas as integrated by Targetlynx, but those two markers showed no areas prior to the first minute in Markerlynx. In Figure 4-22, the extracted ion chromatograms of *m/z=265.1077* were obtained from the same sample but by both different software programs with Markerlynx in plot (a) and (b), and Targetlynx in plot (c) and (d). It was clear to visually see the peaks present in 0s sample and 40s sample in Figure 4-22(a) and (b), but both areas were integrated as 0 by Markerlynx. Alternatively, the areas were integrated as 79 in time 0s sample and 129 in time 40s sample by Targetlynx as highlighted by the red circles in Figure 4-22(c) and (d), and the signal to noise ratio of 25 and 27 indicated that the peaks were not noise.

As discussed previously, markers integrated with no initial area by Markerlynx were integrated with an initial area by Targetlynx. The inconsistent automatic area integration and peak detections indicated that Markerlynx validation had to be improved. Since PCA was processed based on the integrated areas from Markerlynx, it was recommended to use Markerlynx to review the statistical significance of PCA on data analysis, and to apply Targetlynx to monitor the actual changes of markers during the ozonation process.

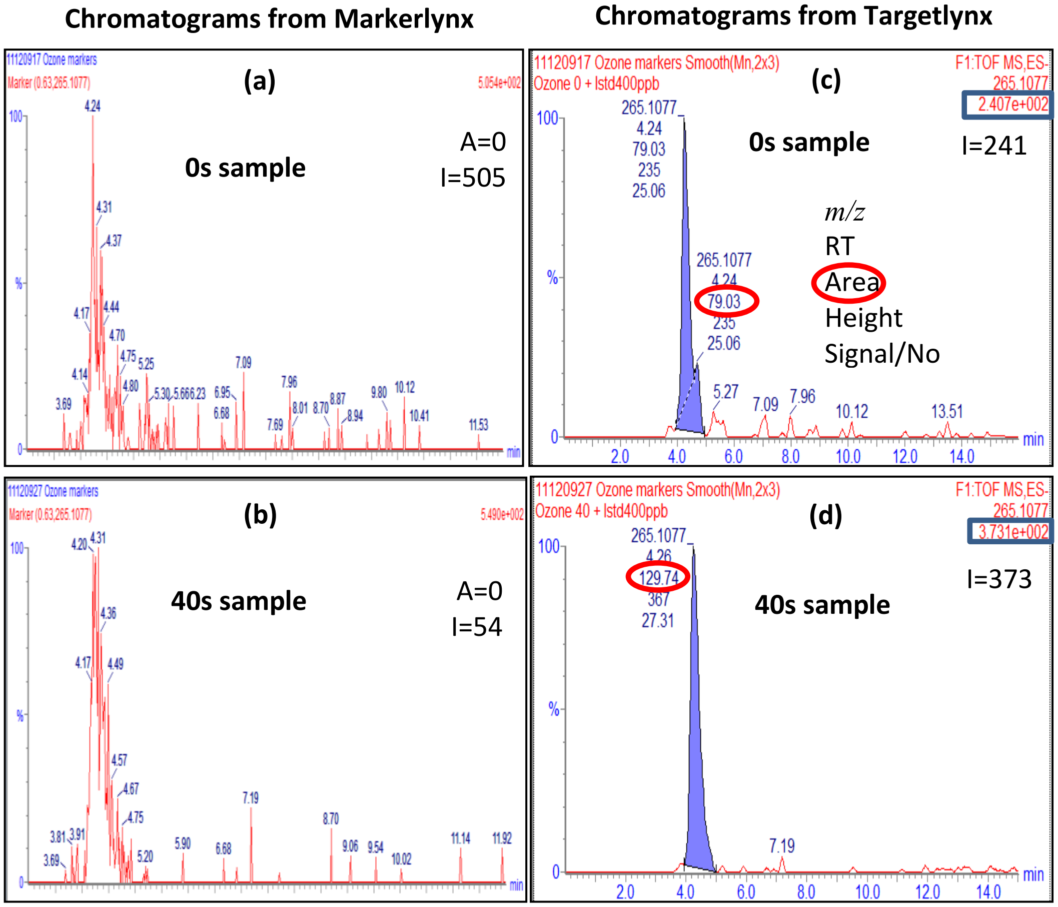


Figure 4- 22: Chromatograms of *m/z=265.1077* in (a) 0s sample from Markerlynx software; (b) 40s sample from Markerlynx software; (c) 0s sample from Targetlynx software; (d) 40s sample from Targetlynx software.

4.7 Possible Elemental Compositions

After determining the major unknown markers trends during the ozonation process, the next stage of analysis was to identify their elemental compositions. However, without further mass spectrometry experiments, the elemental compositions were only computed based on *m/z* values of markers in this project. The markers were assumed to be NAs or oxidized NAs if their exact masses were matched. However, for the unknown markers whose *m/z* values were not matched with any exact mass of NAs or oxidized NAs, the Masslynx Elemental Composition (V4.0, Water Inc.) calculator was applied to propose the possible elemental compositions based on their *m/z* values.

As introduced in Chapter 3, only elements including C (7-50), H (14-100), O (1-6), N (0-2) and S (0-2) were considered as possible compositions based on the previous study on the organic compounds in OSPW (Allen, 2008; Corinee, 2010; Pourrezaei et al., 2011). The possible elemental compositions given by the Masslynx Elemental Composition calculator for all significant unknown markers with red circles in the PCA loading plot (Figure 4-13) are listed in Table 4-5.

In Table 4-5, the first column shows the exact masses of detected unknown markers, and the second column shows the calculated exact masses based on suggested molecular formulas shown in the last column. The third and the fourth column represent the errors of exact masses between the detected mass and the calculated mass in mDa and PPM, respectively. Table 4-5 only lists the suggestions of elemental compositions within error tolerances less than 10 PPM. Negative values in errors indicated that the detected markers had exact masses lower than the suggested markers. The DBE column calculated the double bonds equivalents based on Equation 3-4. The DBE value suggests the number of double bonds, rings or triple bonds in the structure (Clayden et al., 2001). For example, one double bond or one ring would have a DBE value of one, and one triple bond have a DBE value of two.

Table 4- 5: Elemental Compositions for Unknown Markers from O_3 Datasets.

Detected Mass	Calculated Mass	Error mDa	Error PPM	DBE	Formula
209.0945	209.0960	-1.5	-7.2	0.5	C7 H17 N2 O3 S
	209.0926	1.9	9.1	5.5	C10 H13 N2 O3
	209.0966	-2.1	-10	9.5	C15 H13 O
223.0965	223.0970	-0.5	-2.2	5.5	C12 H15 O4
239.1109	239.1106	0.3	1.3	4.5	C13 H19 O2 S
243.1384	243.1385	-0.1	-0.4	7.5	C16 H19 O2
249.1125	249.1127	-0.2	-0.8	6.5	C14 H17 O4
265.1077	265.1076	0.1	0.4	6.5	C14 H17 O5
	265.1085	-0.8	-3	5.5	C15 H21 S2
	265.1051	2.6	9.8	10.5	C18 H17 S
	265.1103	-2.6	-9.8	11	C17 H15 N O2
269.155	269.1542	0.8	3	8.5	C18 H21 O2
	269.1575	-2.5	-9.3	3.5	C15 H25 O2 S
281.1029	281.1025	0.4	1.4	6.5	C14 H17 O6
	281.1034	-0.5	-1.8	5.5	C15 H21 O S2
	281.1052	-2.3	-8.2	11	C17 H15 N O3
281.1567	281.1575	-0.8	-2.8	4.5	C16 H25 O2 S
	281.1542	2.5	8.9	9.5	C19 H21 O2
299.1312	299.1310	0.2	0.7	14	C21 H17 N O
	299.1317	-0.5	-1.7	4.5	C15 H23 O4 S
	299.1283	2.9	9.7	9.5	C18 H19 O4
311.1681	311.1681	0	0	4.5	C17 H27 O3 S
	311.1674	0.7	2.2	14	C23 H21 N
	311.1708	-2.7	-8.7	9	C20 H25 N S
313.1455	313.1467	-1.2	-3.8	14	C22 H19 N O
	313.1440	1.5	4.8	9.5	C19 H21 O4
	313.1474	-1.9	-6.1	4.5	C16 H25 O4 S
	313.1433	2.2	7	0.5	C11 H25 N2 O6 S

According to Perez-Estrada et al. (2011), the AOPs mechanism of degradation is suggested as the reaction of the hydroxyl radical with organic compounds (e.g., NAs) by H atom abstraction. The markers with more rapid degradation during ozonation were suggested to have molecular structures of higher carbon number with more H atoms available for hydroxyl radical abstraction; more rings with more tertiary carbons in which the H atom is more reactive than that on primary and secondary carbon; and less quaternary carbons with no H atom available. The rapid degradation of the marker would lead to the rapid increasing in areas of another by-product marker. It was observed that significant unknown markers with higher carbon number or DBE values in their suggested formulas would tend to have faster rate in changes in Figure 4-14 by Markerlynx and in Figure 4-21 by Targetlynx. For example, *m/z=313.1455* ($C_{22}H_{19}NO$, DBE = 14, error = -3.8 PPM) and *m/z=299.1312* ($C_{21}H_{17}NO$, DBE = 14, error = 0.7 PPM) had similar rates of changes (increasing followed by decreasing) in Figure 4-14(b) and Figure 4-21(b), but their rates of change were faster compared to *m/z=223.0965* ($C_{12}H_{15}O_4$, DBE = 5.5, error = -2.2 PPM) and *m/z=249.1125* ($C_{14}H_{17}O_4$, DBE = 6.5, error = -0.8 PPM) in Figure 4-14(c) and Figure 4-21(c), because of their higher carbon number and DBE values which indicated that more rings with H atoms available for hydroxyl radical abstraction were potentially present in the molecular structure.

However, *m/z=281.1567* ($C_{16}H_{25}O_2S$, DBE = 4.5, error = -2.8 PPM) showed the highest degradation rate compared to *m/z=243.1384* ($C_{16}H_{19}O_2$, DBE = 7.5, error = -0.4 PPM) and *m/z=269.1550* ($C_{18}H_{21}O_2$, DBE = 7.5, error = 3 PPM) by both Markerlynx in Figure 4-14(a) and Targetlynx in Figure 4-21(a), but *m/z=281.1567* had lower carbon number and DBE values compared with *m/z=243.1384* and *m/z=269.1550*. Thus, the second possible formula with higher carbon number and DBE values for *m/z=281.1567* ($C_{19}H_{21}O_2$, DBE = 8.5, error =

8.9 PPM) might be more reliable based on the relative degradation rates. Overall, the analysis in this stage was only based on the possible elemental compositions given by software, which needs further experimental investigation.

4.8 Common Markers in Ozonation in Different Conditions

Markerlynx and Targetlynx analyses considered for O_3 datasets previously were applied to other ozonation in different conditions datasets including ozone with carbonate (O_3+CO_3^{2-}), ozone with *tert*-butyl-alcohol (O_3+TBA), ozone with carbonate and *tert*-butyl-alcohol (O_3+ CO_3^{2-}+TBA), ozone with tetranitromethane (O_3+TNM) and ozone with iron (O_3+Fe (II)). The significant unknown compounds were determined by PCA with all results including PCA score and loading plots, significant unknown compounds trends, and suggested elemental compositions for each significant unknown compound shown in Appendix B1 to B5.

Overall, 12 significant unknown markers were selected from the O_3 datasets; 20 from O_3+ CO_3^{2-} datasets; 21 from O_3+TBA datasets; 23 from O_3+ CO_3^{2-}+TBA datasets; 21 from O_3+TNM datasets; and 17 from O3+Fe (II) datasets. Comparing the results from each ozonation in different conditions dataset based on markers' *m/z* values, it was found that 21 unknown markers were present only in a specific individual ozonation condition experiment (shown in Appendix B-6), while 27 known markers (Figure 4-23 to Figure 4-27) were found more commonly in two or more ozonation conditions. The markers were assumed to be the same if their *m/z* values were in an error range of ±0.01 Da. However, the trends or behaviours of the common markers were specific to the different ozonation conditions. Figure 4-23 to Figure 4-27 show the 27 unknown markers found to be common in different ozonation conditions by Markerlynx. To facilitate a comparison between the software approaches as shown previously, the trends plots for the common markers by Targetlynx were plotted and included in Appendix B-7. Although the integrated areas were not exactly the same, the overall trends of these markers during ozonation in different conditions were found to be similar by both Markerlynx and Targetlynx software. Therefore comparisons are not further discussed.

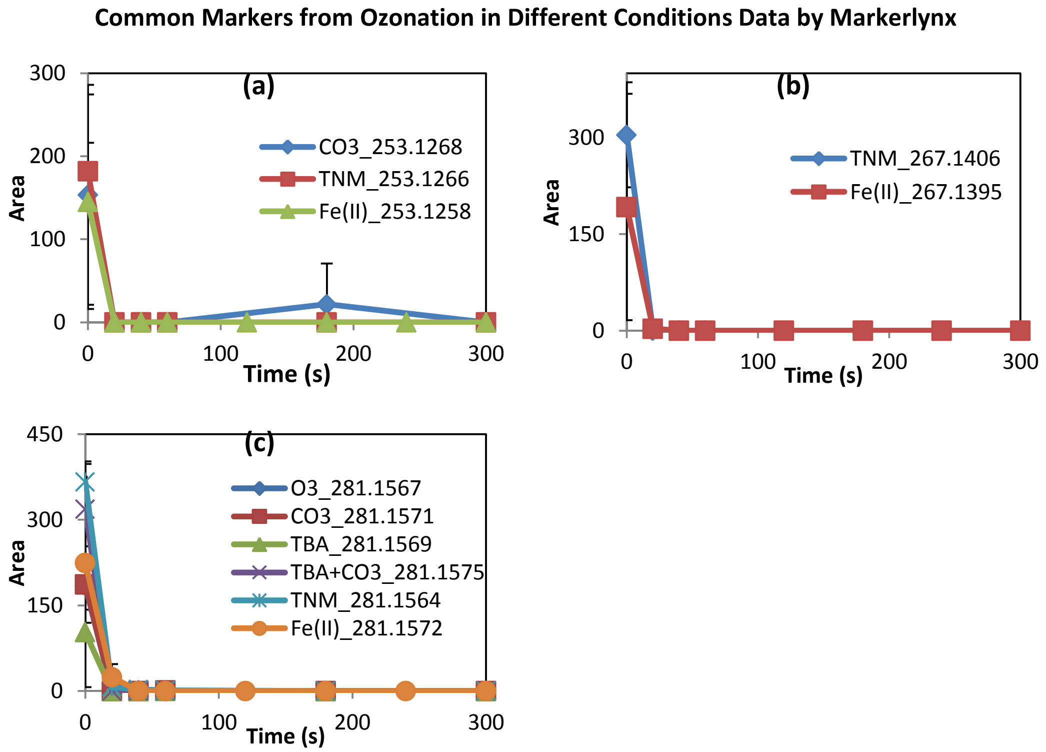


Figure 4- 23: Common Markers from Ozonation in Different Conditions Datasets by Markerlynx (part 1), where *m/z=253.12* in plot (a); *m/z=267.14* in plot (b); *m/z=281.15* in plot(c). Legend CO3_253.1268 indicated *m/z=253.1268* found in O_3 + CO_3^{2-} data.

Figure 4-23(a), (b) and (c) show that three markers *m/z=253.12, 267.14* and *281.15* were degraded to below the detection limit in the first 20s - 40s of ozonation in different conditions. The rapid degradation of three markers in Figure 4-23 indicated they were very sensitive to most ozonation in different conditions regardless of the addition of ozonation scavengers or catalysts.

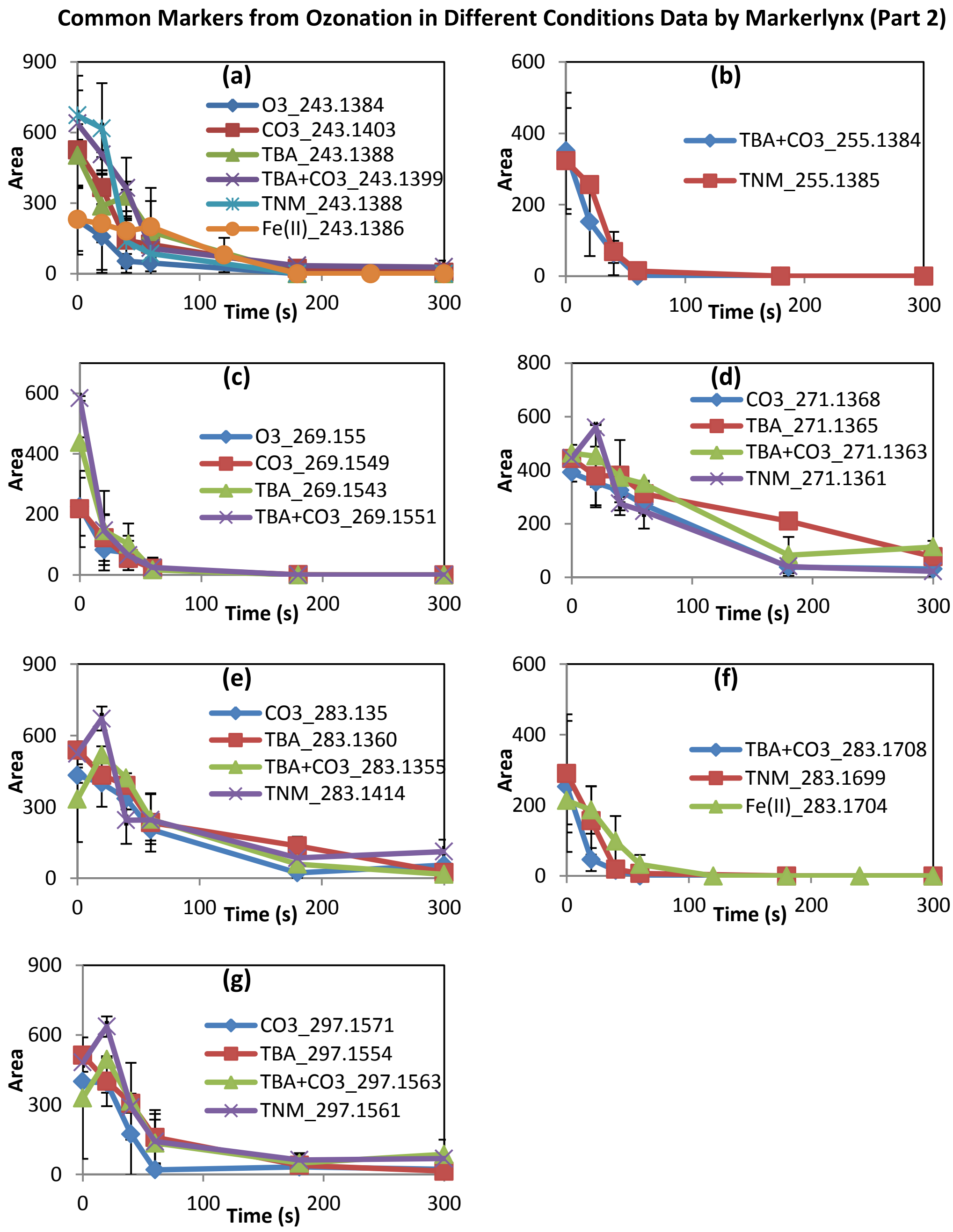


Figure 4- 24: Common Markers from Ozonation in Different Conditions Datasets by Markerlynx (part 2), where *m/z=243.13* in plot (a); *m/z=255.13* in plot (b); *m/z=269.15* in plot (c); *m/z=271.13* in plot (d); *m/z=283.13* in plot (e); *m/z=283.17* in plot (f); *m/z=297.15* in plot (g).

Figure 4-24 plots (a) to (g) show seven markers (*m/z=243.13; 255.13; 269.15; 271.13; 283.13; 283.17; 297.15*) that had degradation trends during ozonation in different conditions. However, compared with the markers rapidly removed shown in Figure 4-23, the areas of the seven markers shown in Figure 4-24 did not become zero until 1-5 minutes during ozonation in different conditions. Thus, the degradation rates of those markers were slower than those in Figure 4-23, so they were less sensitive to molecular ozone or hydroxyl radicals. In addition, it was expected that the addition of Fe (II) as a catalyst would enhance the hydroxyl radical production to accelerate the degradation rates of markers (Nawrocki and Kasprzylk-Hordern, 2010; Kishimoto and Ueno, 2012). However, the degradation rates of markers in the experiment with Fe (II) addition were slower than the rates found in other experiments with scavenger addition in plot (a) and (f). The reasons for these differences were not clear at this stage as the molecular structure of the markers and corresponding ozonation reaction mechanisms were not known.

Figure 4-25 show six markers with slower degradation rates (more resistant to ozonation in different conditions) compared with markers shown in Figure 4-24. For instance, marker *m/z=259.1327* in TBA in plot (a), *m/z=275.1294* in CO_3^{2-} in plot (c), *m/z=325.1834* in Fe (II) in plot (f) could not be reduced below detectable limits during the 5 minute experiments used for the ozonation in different conditions. Plot (e) and (f) show the same marker but at different scales, as area *m/z=325.1834* in Fe (II) data was more than 10 times larger than the areas in TBA and TBA+ CO_3^{2-} data. In addition, *m/z=273.1142* and *m/z=325.1835* in TBA+ CO_3^{2-} data in plot (b) and (e) respectively show constant trends across samples, which indicated they were quite stable during ozonation in different conditions treatment, so they would be considered as resistant to molecular ozone or hydroxyl radicals.

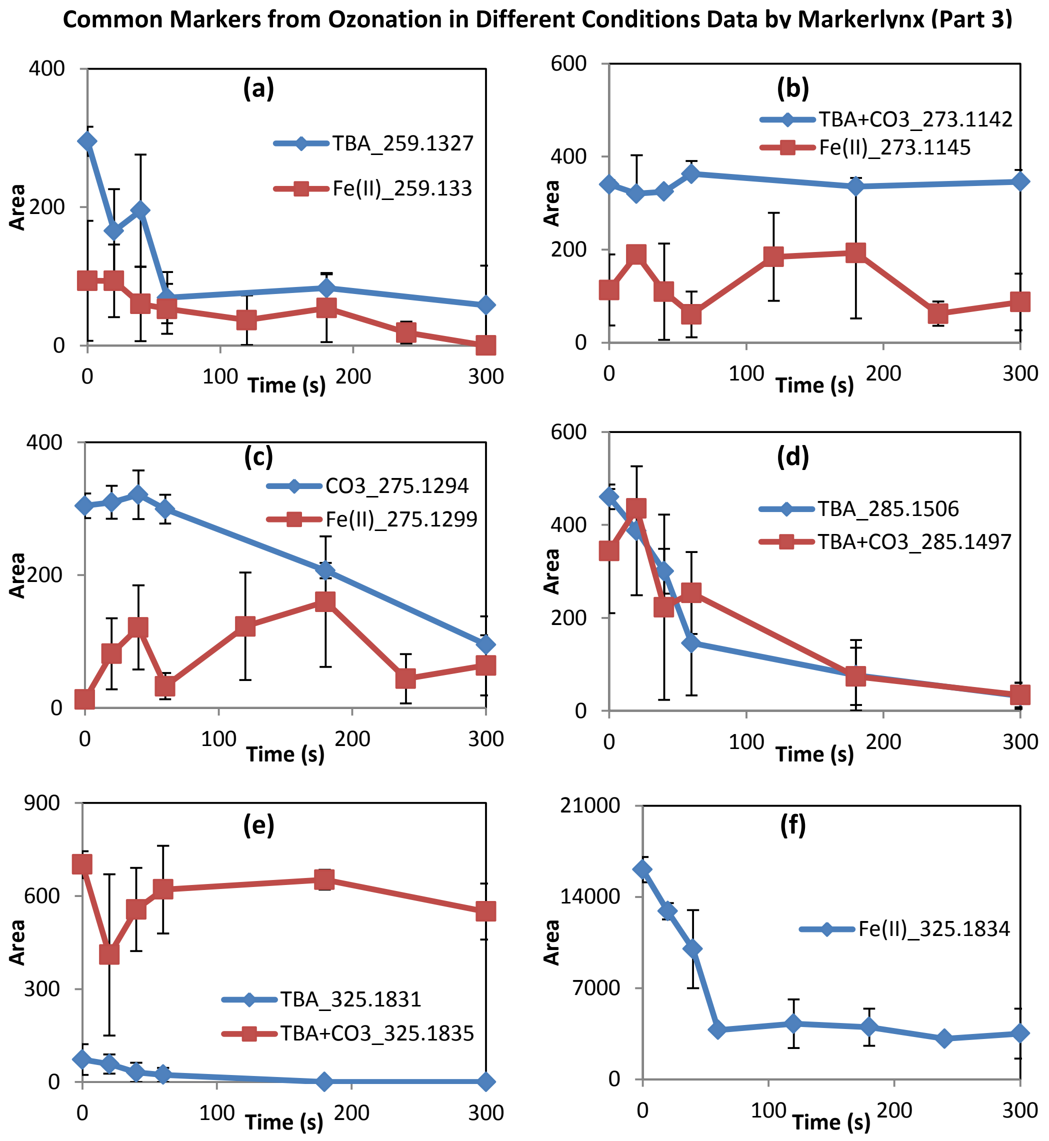


Figure 4- 25: Common Markers from Ozonation in Different Conditions Datasets by Markerlynx (part 3), where *m/z=259.13* in plot (a); *m/z=273.11* in plot (b); *m/z=275.12* in plot (c); *m/z=285.15* in plot (d); *m/z=325.18* in plot (c); *m/z=325.18* in both plot (e) and (f).

Figure 4-26 (a) to (g) show seven markers with varying trends in each ozonation condition. For instance, *m/z=223.0963* and *m/z=249.1123* in Fe (II) data showed increasing trends in plot (a) and (b) respectively, but their corresponding common markers (e.g., *m/z=223.0965* and *m/z=249.1125* in O_3 data) show trends of increasing followed by decreasing in plot (a) and (b) respectively. Thus, observations indicated the addition of Fe (II) might increase the efficiency of ozonation (Nawrocki and Kasprzylk-Hordern, 2010; Kishimoto and Ueno, 2012), as the continuing increase of the markers' areas indicated the continuing degradation of their parent compounds. However, the addition of Fe (II) did not always result in the increasing formation which would indicate the continuing degradation of another marker. For instance, *m/z=299.1307* and *m/z=313.1462* in Fe (II) data in plot (d) and (f) show the trends of increasing followed by decreasing, but their corresponding common compounds (e.g., *m/z=299.1326* and *m/z=313.1467* in TBA data) show increasing trends although TBA was added as a hydroxyl radical quencher to decrease the ozonation efficiency. The possible reasons were due to the ozonation in different conditions efficiencies and mechanisms which are directly related to the presence of scavengers or catalysts, and the molecular structures of target compounds.

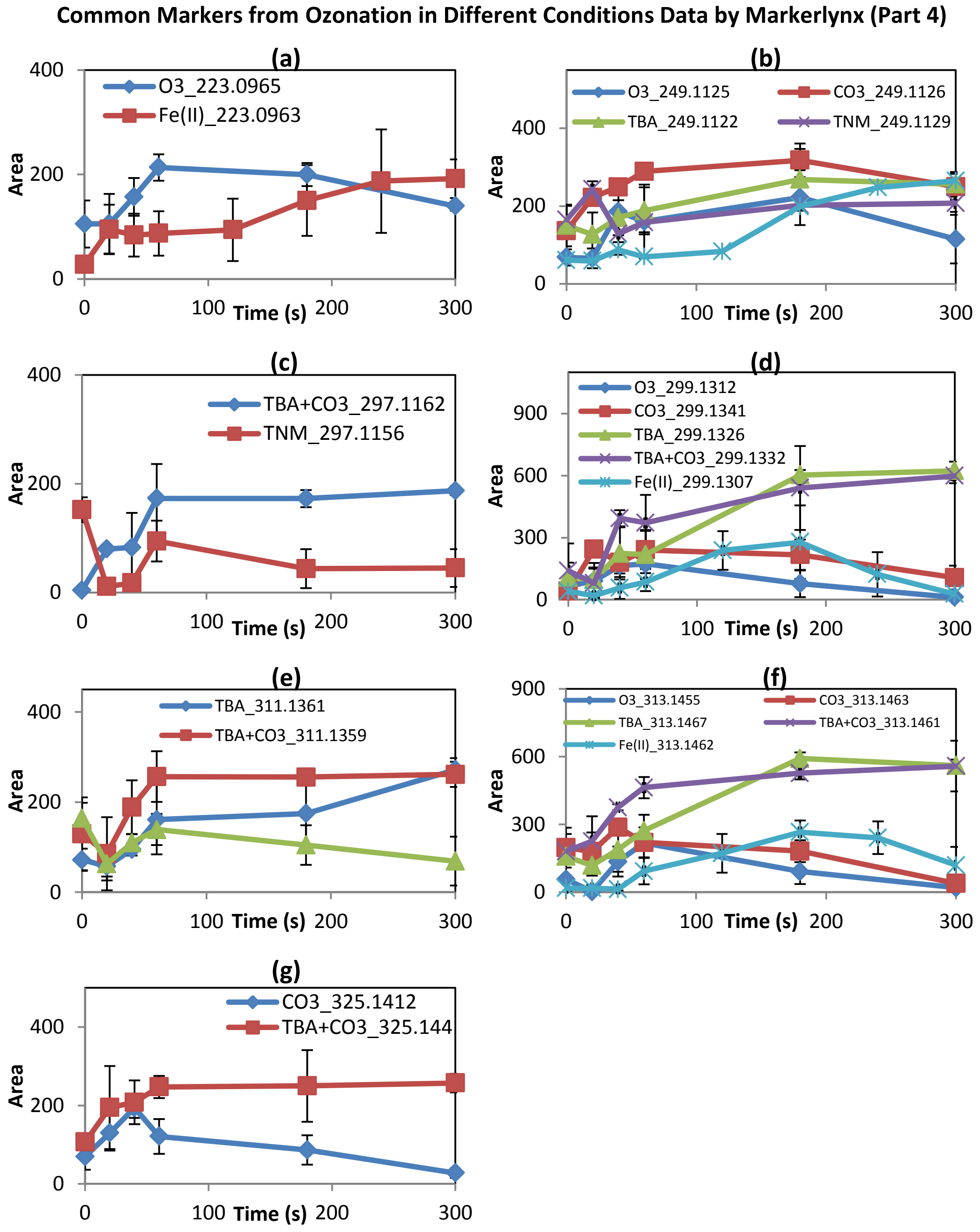


Figure 4- 26: Common Markers from Ozonation in Different Conditions Datasets by Markerlynx (part 4), where *m/z=223.09* in plot (a); *m/z=249.11* in plot (b); *m/z=297.11* in plot (c); *m/z=299.13* in plot (d); *m/z=311.13* in plot (e); *m/z=313.14* in plot (f); *m/z=325.14* in plot (g).

Figure 4-27 illustrates the common unknown markers with increasing trends in different ozonation conditions. Plot (a) shows the marker *m/z=265.10* started increasing after 0s in CO_3^{2-} data, but the marker was not formed until 40s in O_3 data. After the first minute for each ozonation condition, the increasing rates were similar. Plots (b) to (e) show the markers increasing at different rates during ozonation in different conditions with the addition of scavengers or catalysts. For example, in plot (d), the presence of Fe (II) accelerates the increasing rate of *m/z=313.1111* compared with its corresponding common marker *m/z=313.1116* in TNM data. The faster rate of increase might be a result of a faster rate of degradation of parent compounds, which would indicate the Fe (II) improved ozonation in different conditions efficiencies as expected (Catalkaya and Kargi, 2009). In addition, the increasing rate for *m/z=285.1173* in TBA+ CO_3^{2-} in plot (c) was much faster at the beginning as compared with the rate after the first minute, but the rate of *m/z=313.1116* in TNM in plot (d) was much faster after first minute of ozonation compared with the rate at beginning. Gamal El-Din et al. (2011) and Perez Estrada et al. (2011) suggested that the organic compounds rapidly degraded in the first minute of ozonation were more sensitive to molecular ozone, and the rapid degradation after the first minute indicated that the markers were mainly degraded by hydroxyl radicals. Further research would be necessary to identify the molecular structures of those significantly increased unknown markers, which would lead to better understanding of mechanisms of ozonation in different conditions. Overall, the markers shown in Figure 4-27 indicated that they were by-products of degradation for the majority of the current ozonation in different conditions regardless of the addition of catalysts or scavengers.

Common Markers from Ozonation in Different Conditions Data by Markerlynx (Part 5)

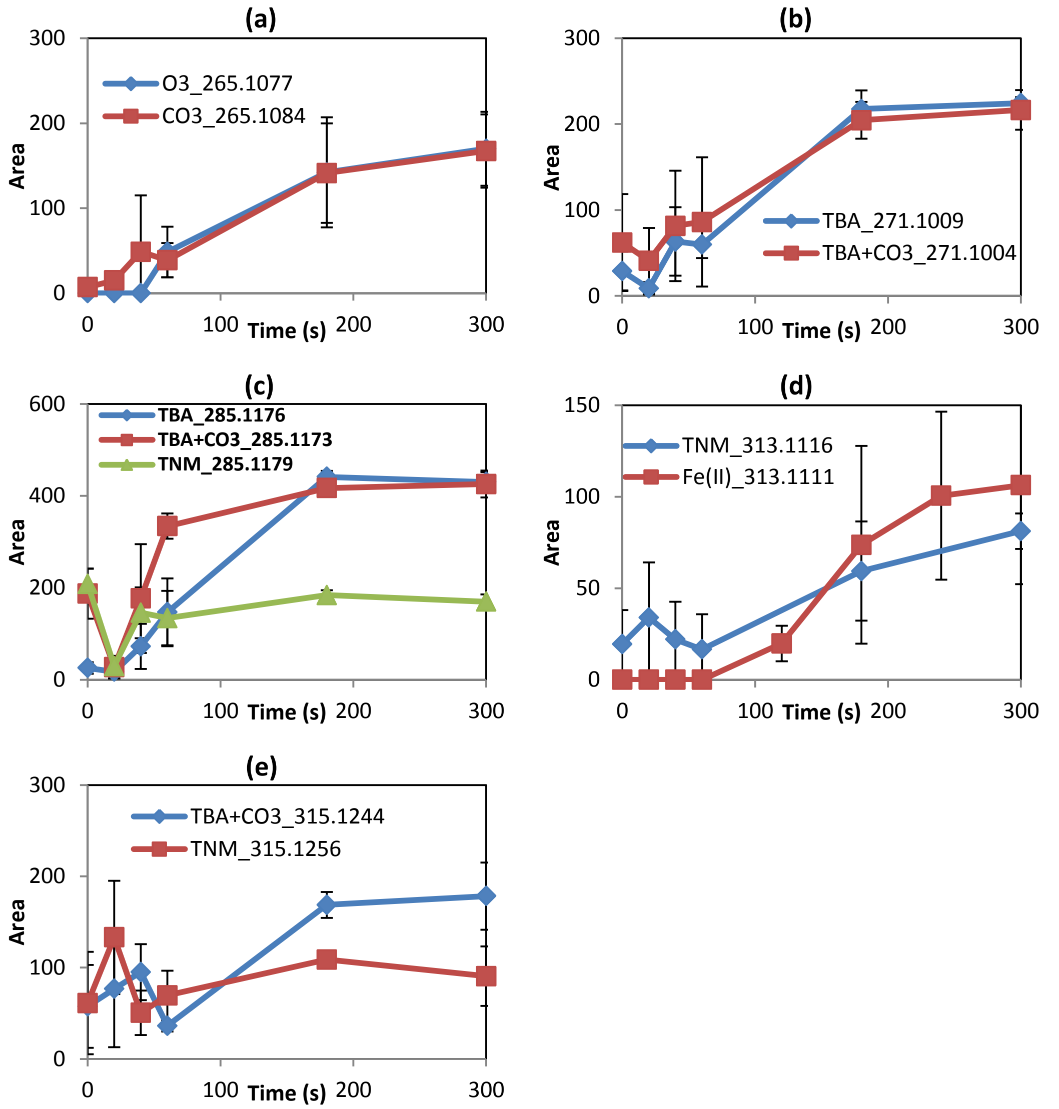


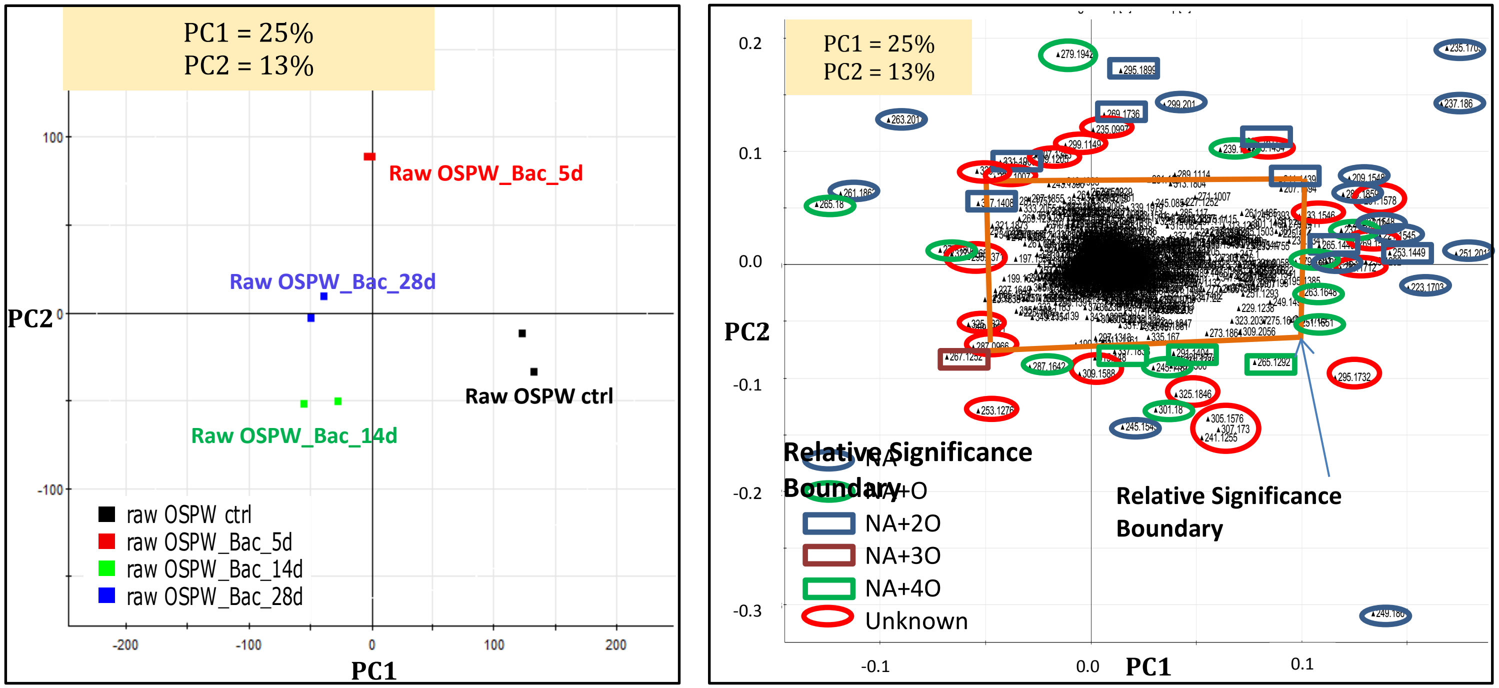
Figure 4- 27: Common Markers from Ozonation in Different Conditions Datasets by Markerlynx (part 5), where *m/z=265.10* in plot (a); *m/z=271.10* in plot (b); *m/z=285.11* in plot (c); *m/z=313.11* in plot (d); *m/z=315.12* in plot (e).

4.9 Biological Treatment Processes

4.9.1 Raw OSPW Biodegradation

Similarly to the approach used to study the ozonation in different conditions datasets, the PCA was applied to datasets for raw OSPW treated by biological treatment processes. Figure 4-28 shows the PCA score and loading plots based on datasets of raw OSPW biodegradation.

In the score plot in Figure 4-28(a), only duplicates were included in each sample group after removing outliers, although initially there were triplicates prepared for each sample group as mentioned in Chapter 3. As shown in Figure 4-28(a), the total variances of 38% are explained by PC1 of 25% and PC2 of 13%. Although the total variance was low because of the large variables datasets (>1,500 markers), two components were still sufficient to indicate the linear relationship as discussed previously (Sleighter et al., 2010). There was a weak tendency of samples distributed along PC1 axis with respect to the experiment time. For instance, raw OSPW control samples were on the positive PC1 axis, 5 days raw OSPW samples with bacteria addition were in the middle, and 14 and 28 days raw OSPW samples with bacteria addition were on the negative side of the PC1 axis. However, the score values for 14 and 28 days samples were similar in magnitude, so the differences in those two samples groups were small.



(a) (b)

Figure 4-28: PCA (a) Score, and (b) Loading Plots based on Datasets from Biodegradation of Raw OSPW.

In the loading plot in Figure 4-28(b), markers were identified as NAs, oxidized NAs and unknown compounds based on their *m/z* values, and were highlighted with different symbols. By combining the score and loading plots, it was observed that NAs correlated to control samples had higher loadings on the positive PC1 axis. These NAs had Z values of -4 to -8 and carbon number from 12 to 16 which agrees with the major group of NAs in raw OSPW reported in literature (Matthew, 2012). In contrast, those NAs with higher positive PC1 loadings were negatively correlated to 14 and 28 days samples with negative PC1 scores, which indicated that those NAs were biodegraded over time. The observations agreed with the literature that microorganisms preferred to degrade NAs with lower molecular weight and fewer branches (Han et al., 2008; Corinne, 2010). The NAs with Z = 0 and -2 were not selected as the most biodegraded species by PCA since the concentrations of those two groups of NAs were negligible in OSPW control samples.

The quantification results showed the total initial NAs concentrations of 17.1 mg/L in raw OSPW control decreased to 16.3 mg/L at day 5; 15.0 mg/L at day 14 and 14.9 mg/L at day 28. Thus, 14 day and 28 day samples had comparable NAs concentrations, which could further lead to the similar PC1 scores in the score plot in Figure 4-28(a). However, the total NAs degradation during the 28 day experimental period was only 2.2 mg/L which was 12% of initial NAs concentrations, so the biodegradation was limited. The possible reason for this minimal biodegradation could be the incubation period of 28 days being too short for microorganisms to adapt to degrading NAs and the fact that no other nutrition was provided which may be a limitation to microorganism degradation activity (Lai et al, 1996; Clemente et al., 2004). The limited biodegradation would lead to less potential by-product generation, which explained why more markers with higher loadings were found on positive PC1 axis than on the negative PC1 axis in

Figure 4-28(b). This indicated that more markers associated with samples at beginning of the experiment (control and 5 day) on the positive PC1 axis would be degraded, but fewer markers associated with samples at end (14 and 28 days) would increase during the biodegradation process.

The peak areas of relative significant unknown markers highlighted in Figure 4-28(b) were integrated by Markerlynx and Targetlynx, and the results are shown in Figure 4-29 and 4-30, respectively. The trends were plotted based on the average values and error bars were not included due to duplicates in each sample group not being sufficient for statistical analysis. Figure 4-29(a) shows the markers with large PC1 loadings decreased during the biodegradation process, so they were correlated to the OSPW control samples in the score plot as predicted previously. Figure 4-29(b) shows the markers with trends decreasing over the first 5 days followed by an increasing trend from 5 to 14 days, and becoming stable thereafter. Those markers had positive PC1 and negative PC2 loadings, so the positive PC1 loadings indicated the correlation with OSPW control samples as those markers degraded in first 5 days. However, the negative PC2 loadings suggested the anti-correlation to 5 days samples, so the areas of the markers dropped to the lowest point at 5 days, followed by an increasing trend afterward. Figure 4-29(c) shows the markers with increasing trends during biodegradation as those markers associated with 14 and 28 days samples due to their negative PC1 and PC2 loadings, and those increased markers were suggested as the by-products from raw OSPW biodegradation. Figure 4-29(d) shows the markers reached the highest area at 5 days followed with decreasing trends because they had small PC1 loadings and were located on the top of loading plot being correlated to 5 days samples. Overall, the decreasing trends indicated the biodegradation of markers, but the increasing trends indicated the potential formation of markers. For example, for *m/z=287.0966* in Figure 4-29(c), the marker was not detected in raw OSPW initially, but the area increased after 14 days which indicated the formation of this compound. However, it was important to note that, as discussed previously, Markerlynx had the issues with some of the area integrations, so the results shown in Figure 4-29 were confirmed by Targetlynx.

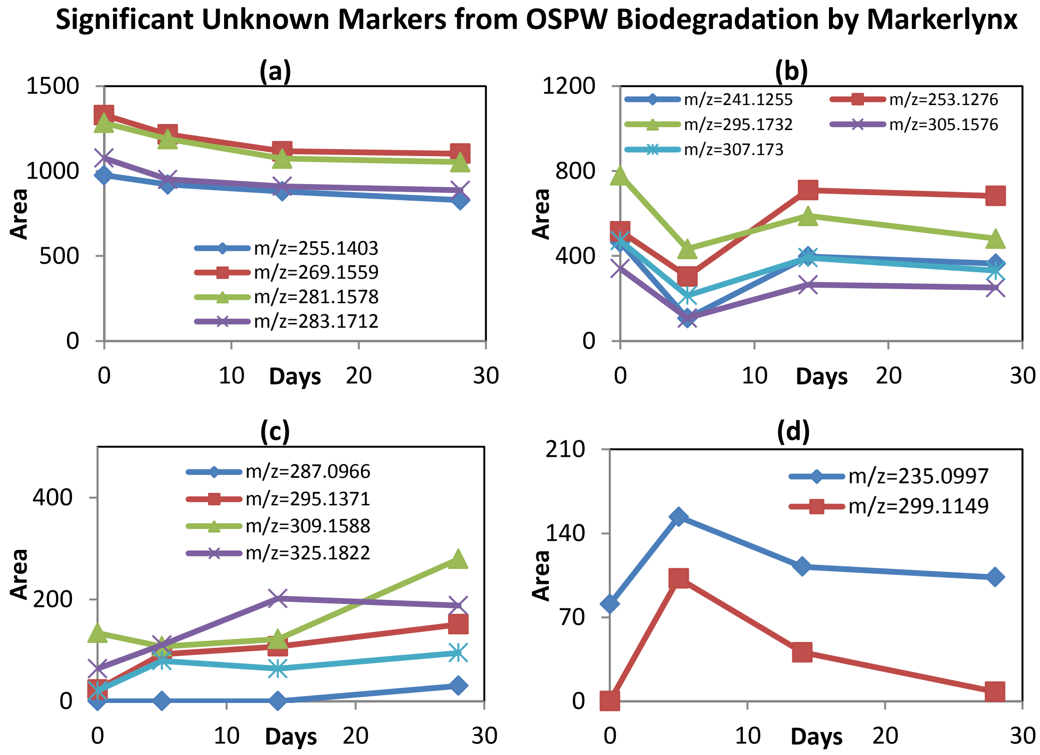


Figure 4- 29: Trends of Significant Unknown Markers from Raw OSPW Biodegradation by Markerlynx Software, where plot (a) shows unknown markers degraded along time; plot (b) shows unknown markers decreased in first 5 days followed by increasing trends; plot (c) shows unknown markers had increasing trends duirng biodegradation process; plot (d) shows markers increased in first 5 days followed by decreasing trends.

Figure 4-30 shows the same significant unknown markers as Figure 4-29 but integrated by Targetlynx. By comparing the trends in areas integrated by Markerlynx and Targetlynx, it was found that although Figure 4-30(a) shows the markers with similar intial areas degraded during the experiment, the decreasing rates by Targetlynx were much faster compared with Markerlynx. Figure 4-30(b) shows the markers with decreasing trends, so the observations were different than Figure 4-29(b) in which markers decreased in the first 5 days followed by increasing afterward. In addition, in Figure 4-30(c), *m/z=287.0966* and *295.1371* show marginal decreasing trends by Targetlynx (i.e., degrading), although Markerlynx showed a slight increasing trend (i.e., forming). The markers shown in Figure 4-36(d) had different trends by Targetlynx compared with the results by Markerlynx. In the Targetlynx, *m/z=299.1149* decreased during biodegradation, but it increased initially and decreased again after first 5 days by Markerlynx. The marker *m/z=235.0997* slightly increased in 14 days by Targetlynx, but increased initially and decreased after first 5 days by Markerlynx. Thus, the results found by the two software approaches showed high variability for these experiments.

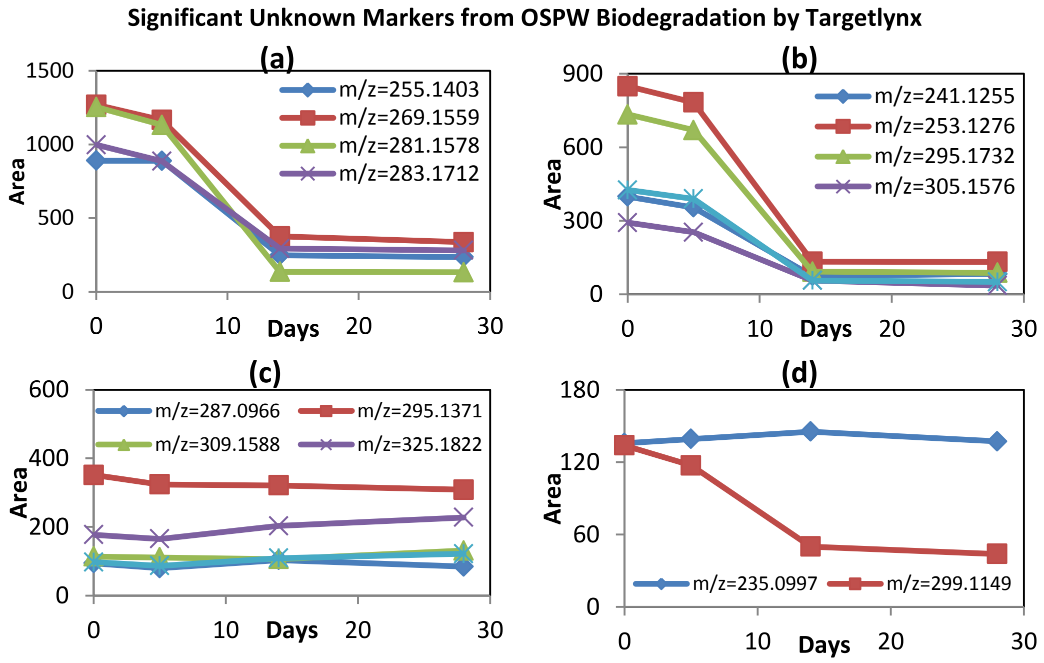
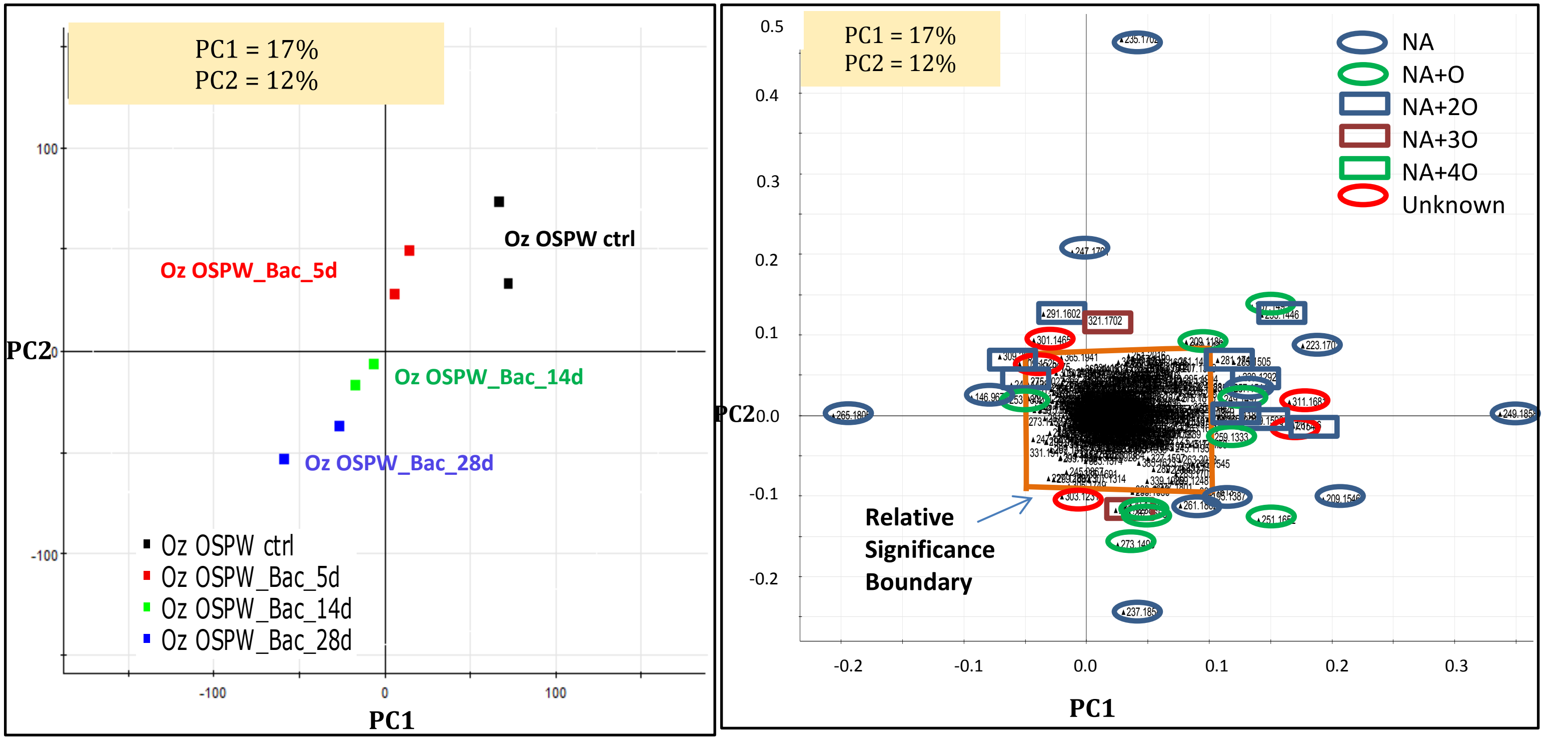


Figure 4- 30: Trends of Significant Unknown Markers from Raw OSPW Biodegradation by Targetlynx Software. Plot (a), (b), (c) and (d) are corresponded to the markers shown in Figure 4-29.

As discussed previously, Targetlynx allows manual peak selection to integrate peak areas, so the markers' trends from Targetlynx are suggested to more accurately demonstrate the actual changes of markers during biodegradation. However, the PCA was applied based on the integrated areas from Markerlynx, so the trends found using Markerlynx software demonstrate the statistical significance of PCA. Markers *m/z=287.0966* and *295.1371* were considered as potential by-products of raw OSPW biodegradation because they showed increasing trends by Markerlynx, but decreasing trends by Targetlynx. Ozonation in different conditions datasets with at least three replicates showed similar overall markers trends by Markerlynx and Targetlynx. However, due to the lack of sufficient samples in biodegradation datasets for reliable statistical analysis, the trends of markers plotted based on average values of duplicates showed inconsistencies between Markerlynx and Targetlynx. Therefore, for future studies it is highly recommended to increase the number of replicates (at least five) to improve the PCA study and minimize the errors in integrated areas by Markerlynx. In addition, the potential for improvement in the Markerlynx software by the manufacturer should be considered in future software updates.

4.9.2 Ozonated OSPW Biodegradation

Using the same analysis methods as raw OSPW, the PCA results for ozonated OSPW further treated by the biological treatment processes were shown in Figure 4-31. In the score plot, PC1 with 17% and PC2 with 12% explained a total variance of 29% of original data due to the large variables (>1,500 markers) matrix (Sleighter et al., 2010). The samples are distributed from positive PC1 and PC2 loadings to negative PC1 and PC2 loadings with respect to experimental duration from 0 day (control) to 28 days in Figure 4-31(a). In the loading plot in Figure 4-31(b), relative significant markers were identified as NAs, oxidized NAs and unknown compounds based on their *m/z* values. Most of the significant NAs identified had carbon number from 12 to 17 and Z number from -4 to -8, and the quantification results showed the concentrations of NAs with Z=0 and -2 were negligible. However, the NAs are distributed throughout the loading plot, rather than grouped in a specific region in Figure 4-31, so there was no specific correlation between NAs and sample groups at specific times as for raw OSPW Based on the previous results, the NAs were expected to be correlated to samples at the beginning of biodegradation. The possible reason for no correlations might be the lack of significant biodegradation since the total NAs decreased only marginally from 12.1 mg/L at 0 day to 10.5 mg/L at 28 days (10% degradation). As for the raw OSPW biodegradation experiment, the possible reason could be the incubation period of 28 days being too short for microorganisms to degrade NAs and the lack of any nutrition provided which may limit microorganism activity (Lai et al, 1996; Clemente et al., 2004). In addition, more markers correlated to ozonated OSPW control samples were found on the positive PC1 axis and fewer markers correlated to 28 days samples were found on negative PC1 axis. This suggests that more markers were degraded, and fewer would be formed during biodegradation processes.



(a)

(b)

Figure 4- 31: PCA (a) Score, and (b) Loading Plots based on Datasets from Biodegradation of Ozonated OSPW.

Trends plots of the relative significant unknown compounds were shown in Figure 4-32 and 4-33 by Markerlynx and Targetlynx, respectively. In Figure 4-32(a), two markers (*m/z=297.1543* and *311.1681*) with high loadings on the PC1 axis decrease in area during biodegradation, so they were correlated to ozonated OSPW control samples with the highest score values in the score plot. Markers (*m/z=301.1465* and *309.1525*) shown in Figure 4-32(b) reach their highest area at 5 days followed by a slight decrease in area afterward. Such observations were expected because those two markers were on the upper left side in loading plot (negative PC1 and positive PC2) which would be correlated to 5 days samples. Finally, Figure 4-32(c) shows the marker (*m/z=303.1231*) had an increasing trend prior to reaching its highest area at 14 days, but followed with a decreasing trend afterward. However, this marker was expected to reach its highest area at 28 days, because it was correlated to 28 days samples by its co-located region in loading plot. This anomaly could be due to the small PC1 loading value assigned to marker *m/z=303.1231* in Figure 4-31. The 28 days samples were expected to be correlated to markers with more negative PC1 loadings, because PC1 explained more total variance compared with PC2 (P1=17%; PC2=12%).

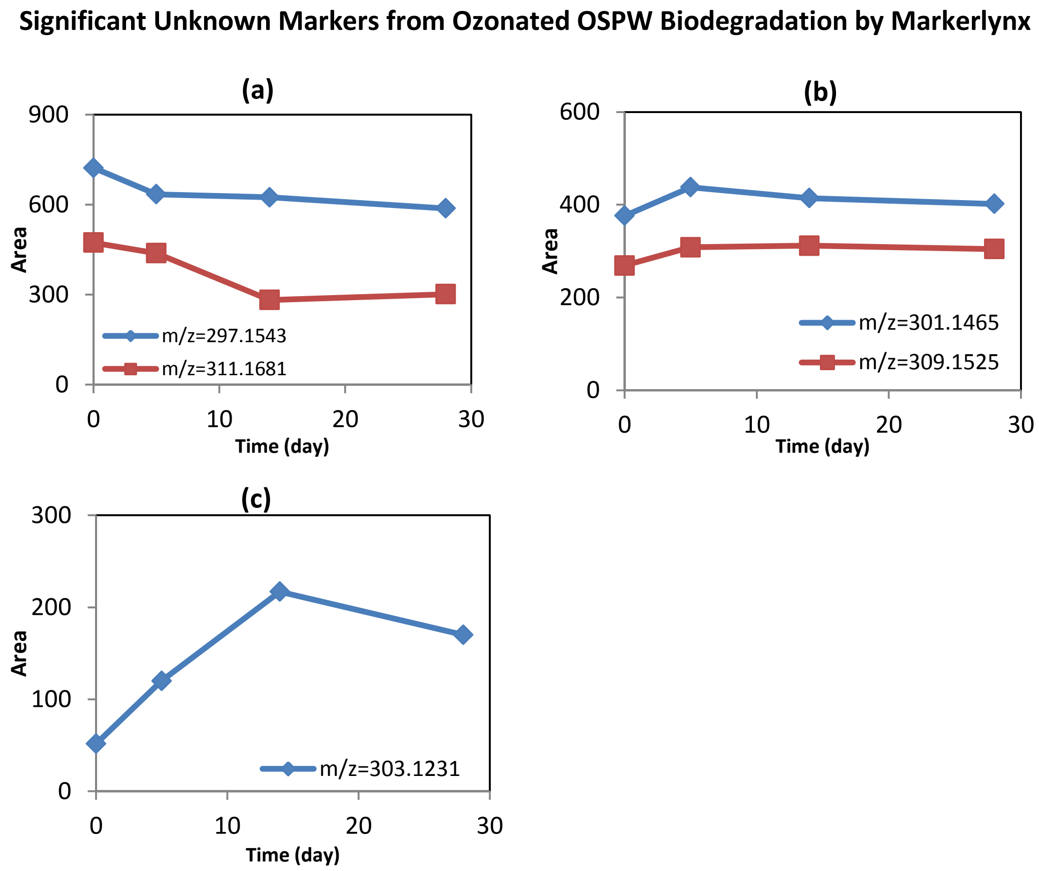


Figure 4-32: Trends of Significant Unknown Markers from Ozonated OSPW Biodegradation by Markerlynx Software. Plot (a) shows unknown markers with decreasing trends during biodegradation; plot (b) shows unknown markers with highest area in 5 days samples; plot (c) shows unknown markers with increasing trend in 14 days followed by decreasing trend.

The results from Markerlynx demonstrated the statistical significance of PCA as discussed previously. In order to more accurately monitor the changes of markers during biodegradation, Targetlynx was applied and the trends of the unknown markers during ozonated OSPW biodegradation were shown in Figure 4-33. Compared to the Markerlynx results in Figure 4-32(a) and Figure 4-32(b), unknown markers (*m/z=297.1543* and *311.1681* shown in Figure 4-33(a); *m/z=301.1465* and *309.1525*m shown in Figure 4-33(b)) by Targetlynx had similar decreasing trends and magnitudes in integrated areas. However, the marker (*m/z=303.1231*) in Figure 4-33(c) shows a decreasing trend which was different from Markerlynx which showed an increasing followed by decreasing trend (Figure 4-32(c)). The reason for this difference could be the inconsistent area integration by Markerlynx, and the lack of sufficient samples which could not decrease the overall errors in integrations.

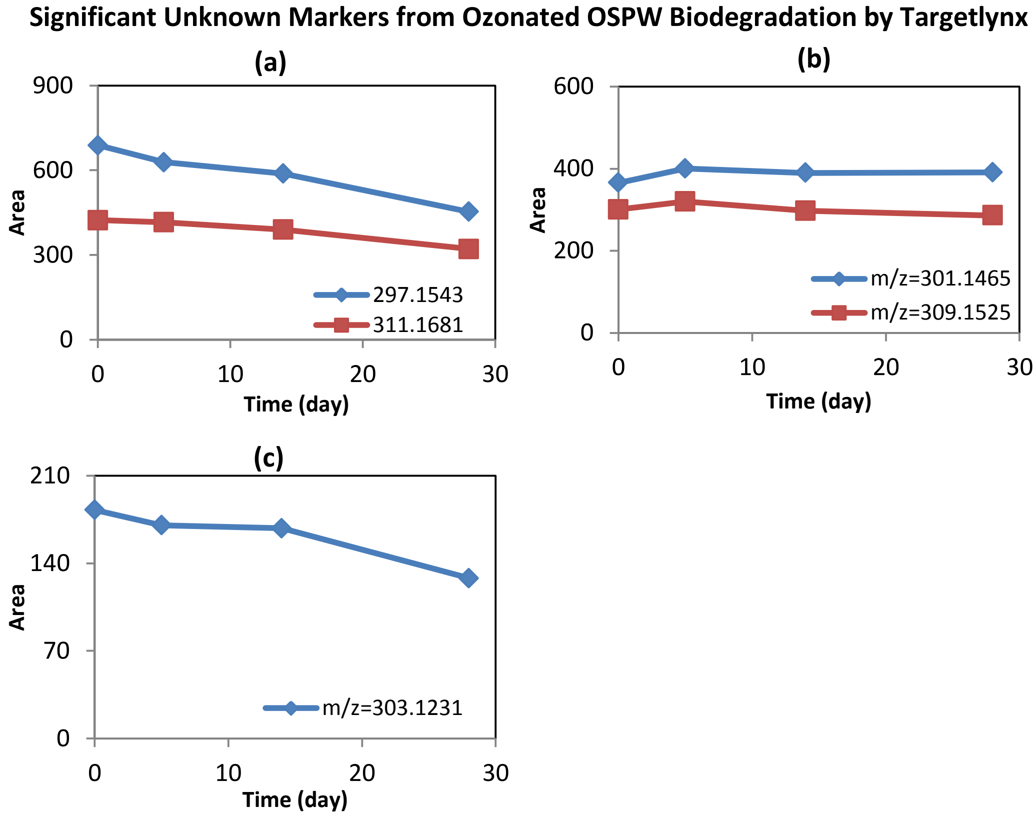


Figure 4-33: Trends of Significant Unknown Markers from Ozonated OSPW Biodegradation by Targetlynx Software. Plot (a), (b) and (c) are corresponded to the markers shown in Figure 4-32.

Overall, based on the significant unknown markers selected by PCA and their trends by both Markerlynx and Targetlynx in Figure 4-32 and 4-33, respectively, there was no detection of a possible by-product showing an increasing trend during the ozonated OSPW biodegradation processes. The reason could be attributed to limited biodegradation during the 28 day incubation period. However, it was expected that there would be greater biodegradation of ozonated OSPW than in raw OSPW since ozonation breaks down the larger molecules into more easily biodegradable smaller molecules (Martin et al., 2010). As discussed previously, Markerlynx sometimes did not recognize peaks. Thus, it was possible that some markers were not appropriately integrated so that the PCA was not able to identify them as significant markers. For future work, it is suggested to prepare more replicates (at least 5) to better define the outliers and minimize the inconsistent area integrations errors. As well, experiments with a longer incubation period and additional nutrition should be provided to allow for significant biodegradation of both raw and ozonated OSPW.

4.9.3 Common Markers from Raw and Ozonated OSPW Biodegradation

By PCA study, 16 significant unknown markers (Figure 4-29) were selected from raw OSPW biodegradation samples and 5 significant unknown markers (Figure 4-32) were selected from ozonated OSPW biodegradation samples. However, it was found that only *m/z=309.1588* in raw OSPW and *m/z=309.1525* in ozonated OSPW were possible common markers as their exact mass error was in a tolerant range of ±0.01 Da. The trends for each of them are shown in Figure 4-34. By using Markerlynx shown in Figure 4-34(a), it was found that the marker had a decreasing trend in the first 5 days followed by increasing from 5 days to 28 days in raw OSPW. In contrast, in ozonated OSPW, the marker increased in the first 5 days and slightly decreased afterward. However, since the initial area of marker in the ozonated OSPW was higher than the initial area in the raw OSPW, this result indicates that its concentration in OSPW was increased after the ozonation process so it is a by-product of ozonation. In raw OSPW biodegradation, this marker showed increasing trends and approached similar total areas found in ozonated OSPW after 28 days of biodegradation, which indicated that this common unknown marker would likely be oxidized products produced by either ozonation or biodegradation.

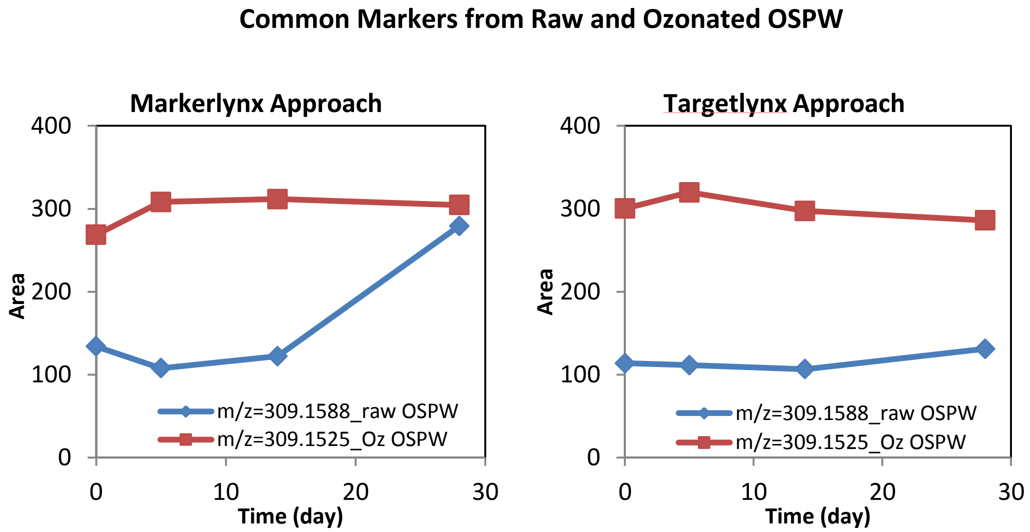


Figure 4-34: Common Markers from Raw and Ozonated OSPW, where (a) shows the trends of common marker in Markerlynx; (b) shows the trends of common marker in Targetlynx.

Figure 4-34(b) shows the same marker (*m/z=309.1588* in raw OSPW and *m/z=309.1525* in ozonated OSPW) integrated by Targetlynx. The marker was stable until a slight increase from 14 days to 28 days in raw OSPW, but the increasing rate was much slower compared to the Markerlynx plot. In ozonated OSPW biodegradation, the marker increased in the first 5 days followed by a decreasing trend. The initial area of the common unknown marker in ozonated OSPW was higher than that in raw OSPW by Markerlynx, so ozonation increased its concentration in OSPW. Thus, the overall trends shown in Targetlynx were considered to be similar to those of Markerlynx. However, only one marker was found to be common in both raw and ozonated OSPWs biodegradation, so the general target organic compounds for microorganism degradation in raw and ozonated OSPW would be different. However, previous discussions must be considered carefully given the lack of sufficient replicates and analytical errors due to the inconsistent area integration by Markerlynx software. Also, the biodegradation found in the period of 28 days was limited which resulted in limited markers showing significant increases or decreases. Further experiments with sufficient replicates and longer incubation periods with nutrient supplementation would be suggested for future work.

4.10 Raw OSPW Variations

4.10.1 Variations of Raw OSPW from Different Sites

All previous analyses including OSPW treated by ozonation in different conditions or biological treatment processes was from Syncrude West in Pit. However, the contents and concentrations in OSPW can vary within the same tailings pond (both spatially and temporally), different sites and various bitumen extraction processes (Allen, 2008). To study this variability, three different sources of raw OSPW were analyzed including Syncrude West in Pit, Suncor Pond 7, and CNRL.

Markerlynx was used to detect markers in OSPW samples from three different sites, and PCA was applied to determine the variations of detected markers in the samples. The PCA plots are shown in Figure 4-35. A total of 81% of the original data variances were covered by PC1 (59%) and PC2 (22%), as shown in the score plot in Figure 4-35(a), which was much higher than the comparable PCA models for ozonation in different conditions and biodegradation. Based on the *m/z* values, the markers were further identified as NAs, NA+O, NA+2O, NA+3O, NA+4O and unknown compounds as labelled in the loading plot in Figure 4-35(b). By combining the score and loading plots, it was found that most NAs highlighted in blue circles were distributed on the positive PC1 axis in the loading plot, so the CNRL OSPW with higher PC1 scores contained the highest NAs concentrations. Similarly, the Suncor OSPW had the highest concentrations of NA+2O with blue squares co-located on the left upper corner. NA+O and NA+4O species with smaller loading values highlighted in green circles and squares were close to the relative significance boundary, so they were less significant. The quantification results further agreed with the PCA analysis, as the highest NAs concentrations of 24 mg/L was determined for CNRL OSPW, followed by 14 mg/L in Suncor and 9 mg/L in Syncrude; and the highest NA+2O concentrations of 11 mg/L was found in Suncor OSPW, followed by 7 mg/L in both Syncrude and CNRL OSPW. Based on these observations, it was predicted that there may have been greater oxidation occurring in Suncor Pond 7.

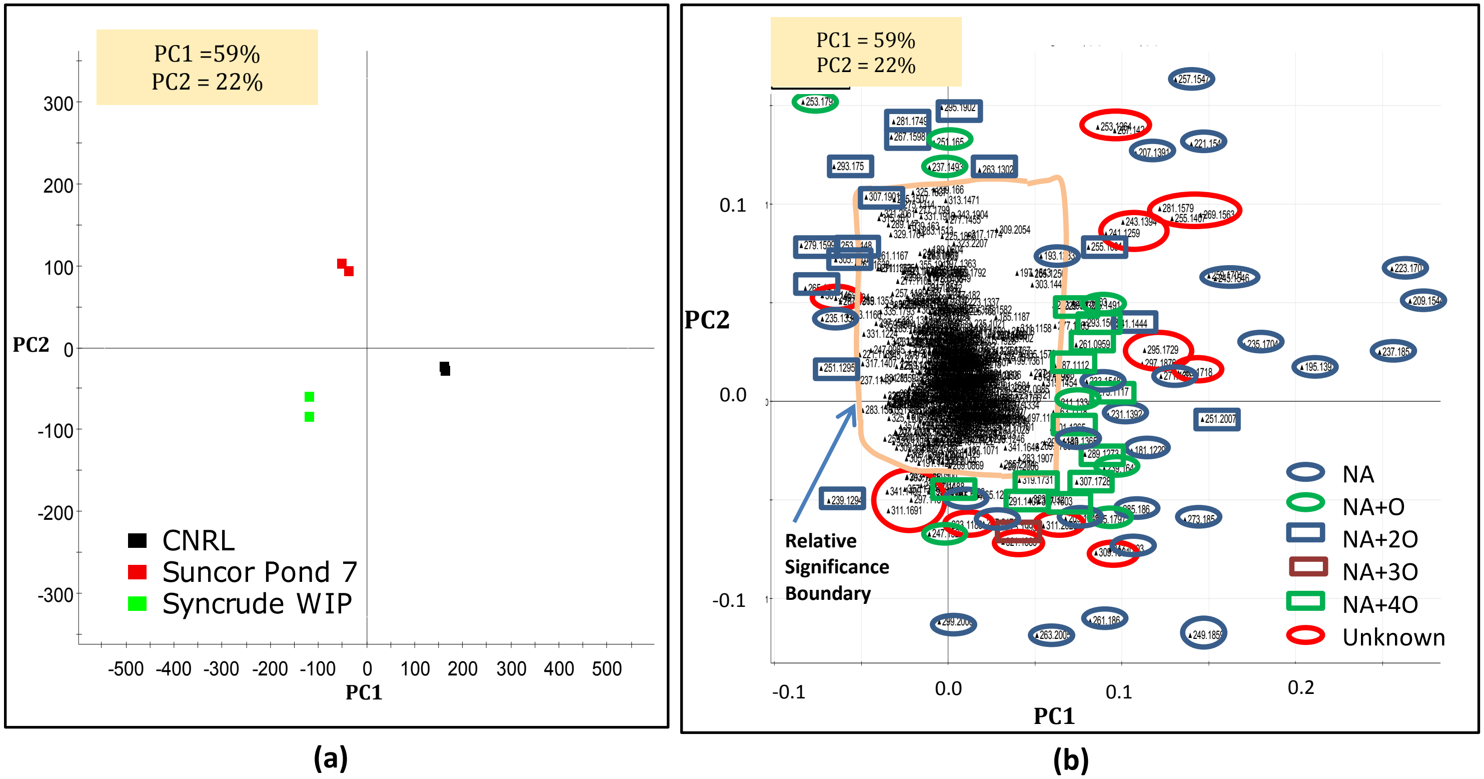


Figure 4- 35: PCA (a) Score, and (b)Loading Plots based on Datasets of OSPW from Suncor Pond 7, Syncrude West in Pit and CNRL.

In the loading plot in Figure 4-35(b), the unknown compounds are highlighted in red circles based on their *m/z* values. The peak areas in these markers' chromatograms were integrated by Markerlynx and Targetlynx, and the results are plotted in Figure 4-36 and 4-37, respectively. Since there were only two analyses for each sample group, the areas were plotted based on the average value of duplicates and no error bars or standard deviations were included.

In Figure 4-36(a), markers show the highest concentrations in CNRL OSPW and lowest concentrations in Syncrude OSPW, as those markers were associated with higher positive PC1 and PC2 loadings which were correlated to CNRL samples, but most anti-correlated to Syncrude samples. Figure 4-36(b) shows the markers with highest concentrations in CNRL but lowest concentrations in Suncor, since those markers had negative PC2 loadings which indicated the anti-correlation to Suncor OSPW. Further, it should be noted that the areas of four markers (*m/z=297.1876, 309.1881, 311.2028* and *321.1885*) were approaching zero in Suncor OSPW, so the concentrations of those four markers in Suncor OSPW were negligible. In addition, Figure 4-36(c) plots the markers with the highest concentrations in Syncrude OSPW but lowest concentrations in Suncor, because those markers were assigned with negative PC2 loadings. Moreover, the area of zero for *m/z= 311.1691* in CNRL and Suncor OSPW made this marker unique to Syncrude OSPW. Figure 4-36(d) shows markers with the lowest concentrations in CNRL, in particular *m/z=341.1409* which had an area of zero in CNRL OSPW. The presence and absence of NAs, oxidized NAs or unknown markers indicated the unique characteristic of OSPWs from different sites.

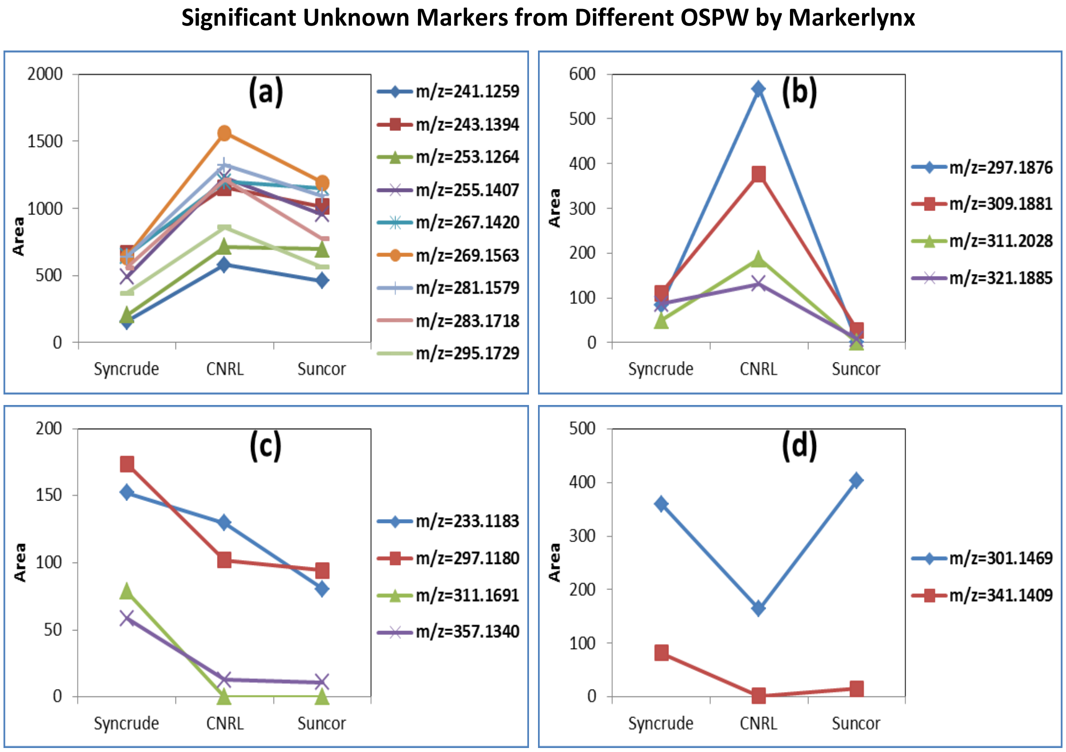


Figure 4- 36: Relative Responses of Significant Unknown Markers in Different OSPWs by Markerlynx Software, where plot (a) shows markers with highest area in CNRL and lowest in Syncrude; plot (b) shows markers with highest area in CNRL and lowest in Suncor; plot (c) shows markers with highest area in Syncrude; plot (d) shows markers with lowest area in CNRL.

By comparing the integrated areas by Targetlynx shown in Figure 4-37, it was found that the majority of markers had similar trends by Markerlynx. For example, Figure 4-37(a) shows the markers with highest concentrations in CNRL but lowest concentrations in Syncrude OSPW; (b) plots the markers with highest concentrations in CNRL but lowest concentrations in Suncor; (c) illustrates the markers with highest concentrations in Syncrude but lowest concentrations in Suncor; (d) demonstrates markers with the lowest concentrations in CNRL OSPW. However, there were differences between two software approaches shown in Figure 4-36 and Figure 4-37. For instance, the markers that had area of zero by Markerlynx were all assigned area values by Targetlynx. These areas indicated that the instrument detected the presence of those markers in either CNRL or Suncor OSPW, although the integrated areas were low.

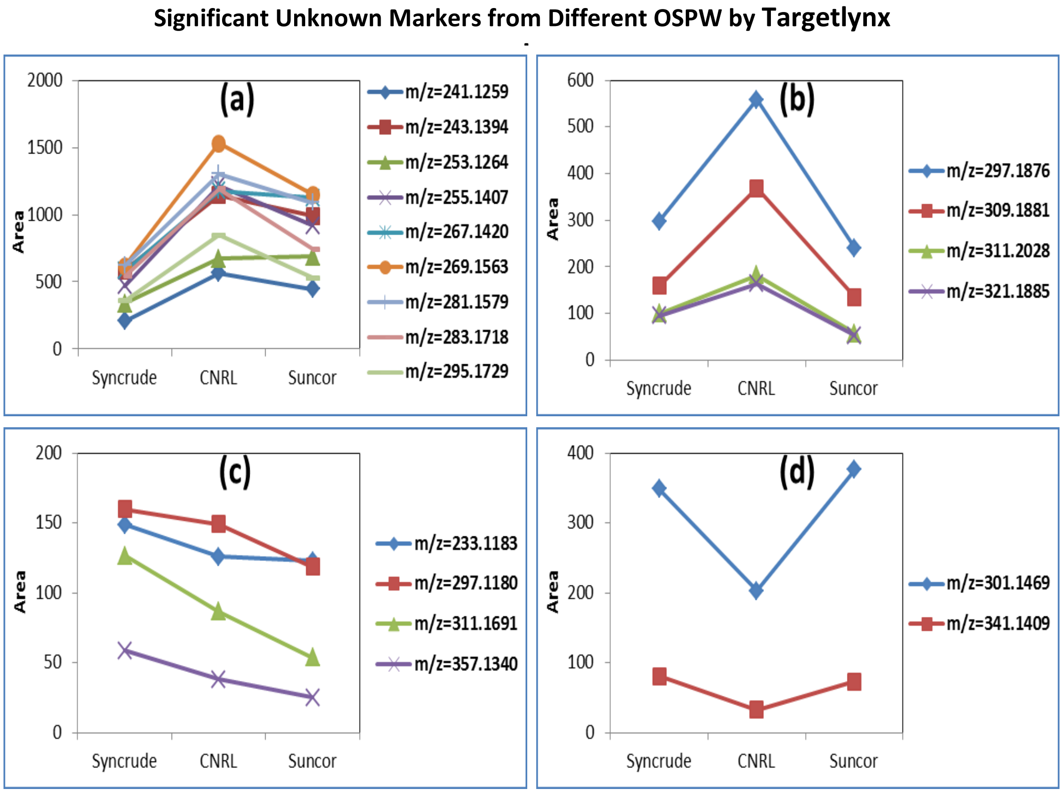


Figure 4- 37: Relative Responses of Significant Unknown Markers in Different OSPWs by Targetlynx Software.

4.10.2 Combined with Ozonation in Different Conditions Datasets

In order to predict the fates of unknown compounds or potential by-products from ozonation in different conditions in raw OSPWs from different sites, the significant unknown markers selected from different sites OSPWs were compared with significant unknown markers selected from the previous ozonation in different conditions datasets.

The common significant unknown markers selected from ozonation in different conditions and different sites OSPWs datasets are shown in Figure 4-38. Each of these markers showed a degradation trend during ozonation in different conditions. In addition, most of these markers had the lowest areas in Syncrude and highest areas in CNRL OSPW, except *m/z=297.1181* which had the highest areas in Syncrude OSPW. In the loading plot in Figure 4-35, these markers were clustered with NAs, so they were positively correlated to NAs and would be expected to be degraded during oxidation processes.

Unfortunately, none of significant unknown markers with increasing trends (defined as by-products) for the ozonation in different conditions experiments were selected as significant unknown markers from datasets of different sites OSPWs. Since the application of PCA was used to find significant markers that described the major differences between samples, those significant unknown markers with increasing trends for ozonation in different conditions treatments could be assumed to have similar concentrations in OSPWs from different sites if they were considered as non-significant by PCA. Thus, it was reasonable to assume that those significant unknown markers with increasing trends were associated with ozonation in different conditions, rather than naturally formed or generated from different bitumen extraction processes. However, this assumption needs to be further tested with more experimental evidence.

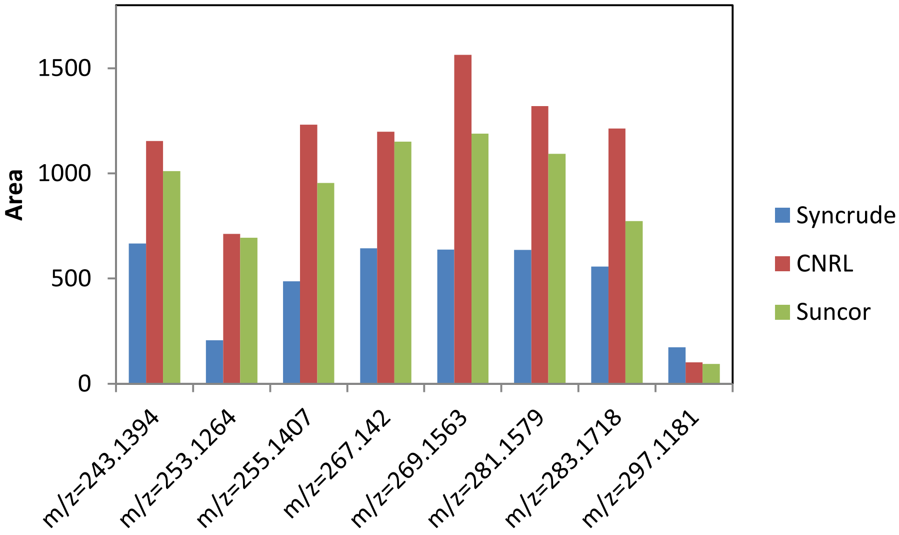


Figure 4- 38: Areas of Common Significant Unknown Markers from OSPWs from Different Sites Datasets and Ozonation in Different Conditions Data.

In order to track if the unknown markers which had increasing trends during ozonation in different conditions would be present in different sites OSPWs, all 24 increased or formed unknown markers shown in different ozonation condition experimental results were selected based on their *m/z* values from datasets of different sites OSPWs. Of these markers, 19 out of 24 were found in the different site OSPW chromatograms and their integrated areas were plotted in Figure 4-39. Since the purpose of the current analysis was to more accurately track the relative magnitudes of markers in different sites OSPWs, there was no PCA applied and the peak areas were only integrated by Targetlynx.

Figure 4-39(a) shows the markers with the highest areas in Syncrude and lowest areas in CNRL, and plot (b) shows the markers with the highest areas in Suncor and lowest areas in CNRL. However, the loading plot in Figure 4-35(b) shows most NAs had their highest concentrations in CNRL, which is in agreement with the quantification results with highest total NAs concentrations of 24 mg/L in CNRL, and 14 mg/L and 9 mg/L in Suncor and Syncrude OSPW, respectively. This suggests that the markers with increasing trends during ozonation in different conditions were negatively correlated to the NAs concentrations in raw OSPW. The higher concentrations of NAs, coupled with lower concentrations of unknown markers with increasing trends during ozonation in different conditions, indicate that there may be less oxidation in the CNRL OSPW. In contrast, lower NAs concentrations and higher concentrations of NA+2O in the Suncor OSPW indicated that Suncor OSPW was more oxidized. In addition, the unknown markers with increasing trends during ozonation in different conditions shown in Figure 4-39(a), (b) and (c) had higher areas in Suncor OSPW. Therefore, those unknown markers might be associated with oxidation by-products in raw OSPW. On the other hand, there were some exceptions as shown in Figure 4-39(d), in which some unknown markers have areas lower in Suncor and higher in CNRL. These markers could represent compounds associated with the degradation of other unknown compounds negatively correlated to NAs in raw OSPW.

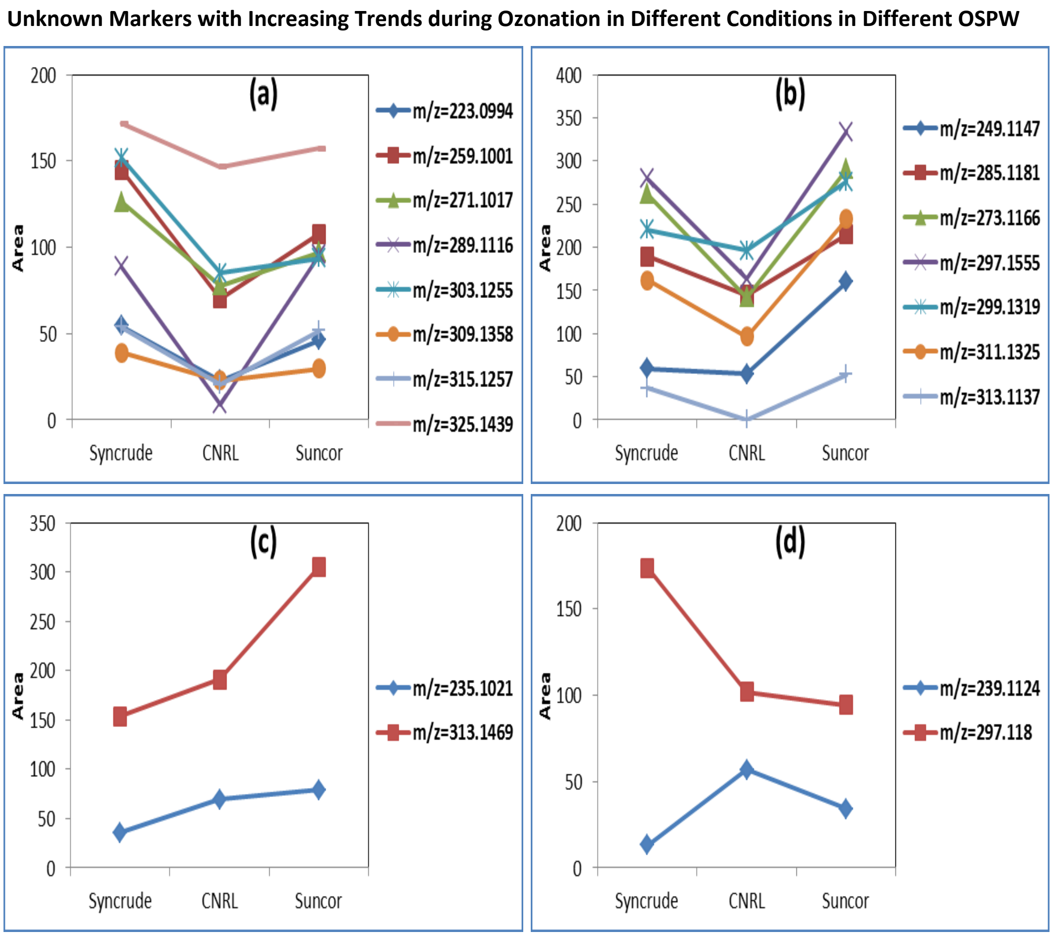


Figure 4- 39: The Relative Response of Markers with Increasing Trends duirng Ozonation in Different Conditions in Different OSPWS by Targetlynx, where plot (a) and (b) show markers with lowest areas in CNRL OSPW; plot (c) shows markers with highest areas in Suncor OSPW; plot (d) shows markers with highest areas in CNRL or Syncrude OSPW.

In addition the remaining 5 out of 24 of the significant unknown markers with increasing trends during ozonation in different conditions were not detected in OSPWs from different sites. For instance, unknown markers *m/z=209.0945, 265.1077* and *281.1029* in O_3 data, *m/z=265.1084* in O_3+ CO_3^{2-} data, and *m/z=287.0954* in O_3+TNM data were not found in other different site OSPWs. However, the peaks of markers were shown in the chromatograms in Syncrude OSPWs used for ozonation in different conditions experiments. Therefore, the concentrations of those unknown markers which increased during ozonation had to be below the detection limit in different sites OSPWs, and it was possible that those unknown markers were the by-products from other compounds oxidation rather than naturally existing in raw OSPW. However, the observations currently show the differences between the Syncrude raw OSPW from different barrels used for ozonation in different conditions experiment and the study on different sites OSPWs variations. All barrels of Syncrude OSPW were sent to Department of Civil and Environmental Engineering, University of Alberta on the same date in 2010. However, there was no information given if all barrels of Syncrude OSPW were sampled on the same date by same sampling procedures (e.g., same location, depth and etc.). Also, with the presence of iron species and scavengers like carbonates in raw OPSW (Allen, 2008), it is possible that partial oxidation may have occurred in raw OSPW during the storage period.

In summary, unknown markers with increasing trends during ozonation in different conditions were correlated to NAs oxidation based on their absence and relative areas in different sites OSPWs as shown in Figure 4-39(a), (b) and (c). Therefore, it was reasonable to predict those unknown markers were associated with the oxidation by-products of OSPW NAs. However, these suggestions are based only on the relative increasing and decreasing trends of NAs and unknown markers during ozonation in different conditions treatments. Further experiments focusing on the molecular structures to determine the formation of compounds and transformation products are recommended to corroborate the current study results.

4.10.3 Combined with Biological Treatment Processes Datasets

Significant unknown markers from raw OSPW biodegradation datasets were compared with the significant unknown markers from different sites OSPWs datasets, and seven common unknown markers were found including: (*m/z=241.1259, 253.1264, 255.1407, 269.1563, 281.1578, 283.1718,* and *295.1729*) (Figure 4-40). It was found that all common unknown markers were clustered with NAs in the loading plot in Figure 4-35(b), so they were positively correlated to NAs that had highest concentrations in CNRL OSPW, which is further corroborated their highest area in the CNRL OSPW in Figure 4-40. Also, with decreasing trends similar to NAs during the previous biological treatment processes, those common unknown markers were suggested as being positively correlated to NAs which were degraded by microorganisms rather than as biodegradation by-products.

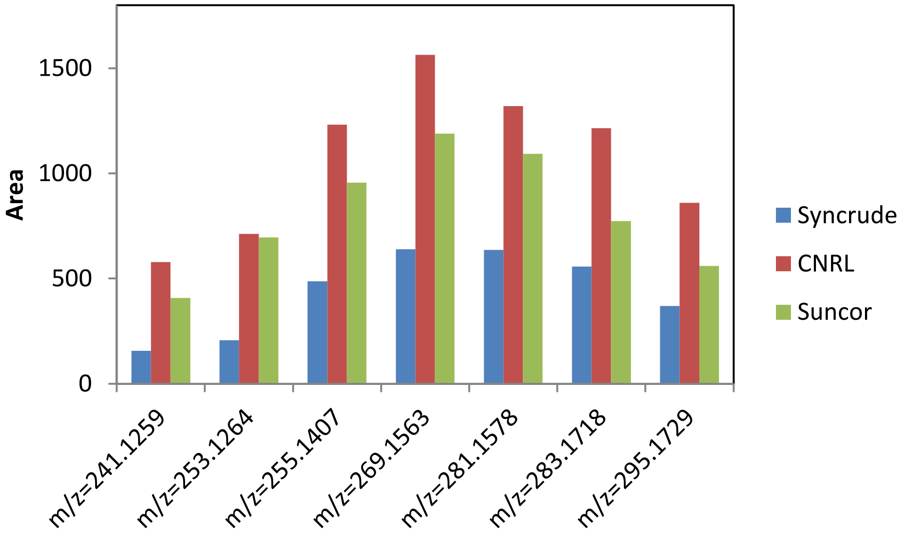


Figure 4- 40: Areas of Common Significant Unknown Markers from OSPWs from Different Sites Datasets and Raw OSPW Biodegradation Datasets.

The significant unknown markers with increasing trends during raw OSPW biodegradation were further tracked in different sites OSPWs, and their integrated areas by Targetlynx were plotted in Figure 4-41. It was found 4 out 5 markers (*m/z=295.1329, 309.1572, 325.1837* and *327.1227*) increased during Syncrude OSPW biodegradation had the lowest areas in CNRL OSPW, which is in contrast to the highest NAs concentrations in CNRL OSPW. Also, unknown markers in Figure 4-41 have highest areas in Suncor OSPW, which was positively correlated to the highest NA+2O concentrations in Suncor OSPW. Thus, combining these results with the increasing trends found for these markers during raw OSPW biodegradation process, the unknown markers shown in Figure 4-41 were suggested as being the by-products associated with NAs biodegradation in raw OSPW.

Additionally, marker *m/z=287.0966* (the 1 out of 5 marker that increased during Syncrude OSPW biodegradation by Markerlynx) with peak shown in chromatogram of Syncrude OSPW used for raw OSPW biodegradation experiment, was not detected in different sites OSPWs including the Syncrude OSPW from different barrels. Thus, *m/z=287.0966* marker can be considered as a by-products formed during biodegradation.

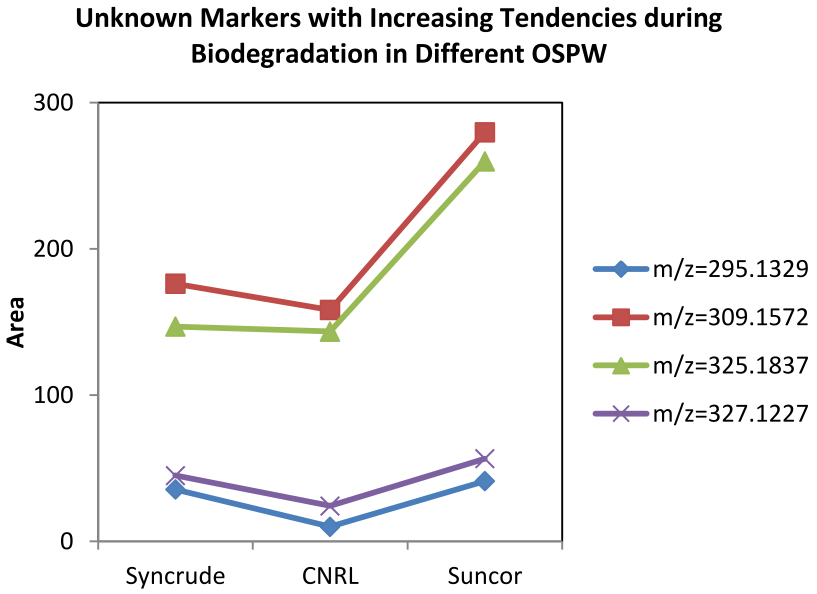


Figure 4- 41: The Relative Response of Markers with Increasing Trends during Biodegradation in Different OSPWS by Markerlynx.

4.11 Summary

In summary, HRMS detected over 1,500 markers from each OSPW sample for ozonation in different conditions and biological treatment processes. PCA with pareto scaling as the optimum data pre-treatment method successfully selected markers significantly changed during ozonation in different conditions and biological treatment processes (Figure 4-13; Figure 4-28; Figure 4-31). The significant markers selected by PCA were further defined as NAs (n = 7 to 22, Z = 0 to -12); oxidized NAs (NAs + Ox, x = 1, 2, 3, and 4); and unknown compounds based on their *m/z* values. The trends of significant unknown markers during treatment processes were reviewed in terms of areas integrated by Markerlynx software and Targetlynx software, and the results were used to evaluate the data analysis consistency and reliability.

PCA is processed based on the peak areas automatically integrated by Markerlynx, so the general trends of markers during treatment processes exhibit the statistical significance of PCA study. Targetlynx allows for manual peak selection where the area integrations are more accurate than Markerlynx, thus, it was applied as complementary software to more accurately review the actual markers changes during the treatment processes. For ozonation in different conditions datasets with at least 3 replicates after removing outliers, the significant unknown markers showed similar overall trends during ozonation in different conditions treatments using both Markerlynx and Targetlynx. In contrast, for biological treatment processes datasets with only duplicates after removing outliers, the overall trends of significant unknown markers showed relatively large differences between the two software approaches. Experiments using greater replication are suggested to allow for improved statistical power which is not available for only duplicate samples.

4.11.1 Ozonation in Different Conditions

PCA study verified that NAs were degraded into lower molecular compounds and/or higher oxidized forms during ozonation in different conditions, which is in agreement with previous studies (Gamal El-Din et al., 2011; Perez-Estrada et al., 2011). The unknown compounds found currently were reported previously to occur, however, have not been further identified in the literature (Scott et al., 2008; Drzewicz et al., 2010). Overall, 12 significant unknown markers were selected by PCA from the O_3 datasets (Figure 4-14); 18 from $O_3 + CO_3^{2-}$ datasets (Figure B-2); 21 from O_3+TBA datasets (Figure B-4); 23 from O_3+ CO_3^{2-} +TBA datasets (Figure B-8); 21 from O_3+TNM datasets (Figure B-11); and 15 from O_3 + Fe (II) datasets (Figure B-16). Generally, the significant unknown markers were found to have highly variable behaviour, both increasing and decreasing at various rates, throughout the ozonation in different conditions experiments. Based on the suggested elemental compositions (e.g., Table 4-5) and previous literature, significant unknown markers which increased at a higher rate were associated with a more rapid degradation of another marker (NAs, oxidized NAs or unknown compounds), and markers with higher rates of changes (e.g., decreased or increased) were more sensitive to ozonation (Gamal El-Din et al., 2011; Perez-Estrada et al., 2011). Based on *m/z* values, 26 significant unknown markers were detected in only one individual ozonation in different conditions experiment (Table B-6), and 27 significant unknown markers were present in at least two ozonation in different conditions experiments (defined as common markers). The 27 common markers had differing behaviours (Figure 4-23 to Figure 4-27) for each conditional ozonation treatment because of their varying molecular structures and the presence of scavengers and catalysts (iron salts) during ozonation in different conditions.

In addition, markers behaviours provided information on their fates in OPSW treated by ozonation in different conditions. All significant unknown markers which decreased over time were considered as degraded by ozonation in different conditions; while all significant unknown markers with increasing trends were concluded to be by-products. Overall, 24 by-products (Table B-7) from all ozonation in different conditions experiments showed similar results were selected by both Markerlynx and Targetlynx, with 12 by-products being common markers for all experiments (Figure 4-26 to Figure 4-27). In the Markerlynx, among the 24 by-products, 12 of them increased from zero initial area in raw OSPW sample, so they might be completely absent in raw OPSW and formed as by-products during ozonation in different conditions. However, in the Targetlynx, all 24 by-products showed increasing trends with an initial integrated area, so they were concluded as by-products which were originally present in raw OSPW, which indicates that there may be oxidation occurring in tailings ponds (Allen, 2008; Pourrezaei et al., 2011).

4.11.2 Biological Treatment Processes

By PCA study, 16 significant unknown markers (Figure 4-29) were selected from raw OSPW biodegradation datasets and 5 significant unknown markers (Figure 4-32) were selected from ozonated OSPW biodegradation datasets. All significant unknown markers showed variable behaviours (increasing and decreasing at various rates) during biological treatment processes. Among the 16 significant unknown markers from raw OSPW biodegradation datasets, 5 of them (Figure 4-29 (c)) were concluded as by-products based on their increasing trends with 3 increasing from an initial area of zero. However, in the Targetlynx approach, only 3 out of 16 significant unknown markers (Figure 4-30 (c)) were considered as by-products, and all of them increased from an initial area. There were 2 by-products (*m/z = 287.0996 and 295.1371*), which had decreasing trends by Targetlynx, while showed increasing trends by Markerlynx, due to the inconsistency of area integration in Markerlynx. Among the 5 significant unknown markers (Figure 4-32 and Figure 4-33) selected by PCA from the datasets of ozonated OSPW followed by the biological treatment processes, none of them were considered as by-products by either software program.

Overall, the PCA study showed NAs and oxidized NAs were degraded during the biological treatment processes of both raw and ozonated OSPW. However, only 5 by-products from Markerlynx and 3 by-products from Targetlynx were found from raw OSPW biodegradation datasets, and no markers were considered as by-products from ozonated OSPW biodegradation datasets. As compared to the number of by-products (24 in total) selected from datasets of raw OSPW treated by ozonation in different conditions, the number of by-products selected from datasets of both raw and ozonated OSPW treated by biological treatment processes were very limited. The possible reason for the limited by-products was the limited biodegradation of NAs in both raw and ozonated OSPW (~10% reduction in NAs concentrations). This result was unexpected since more biodegradation was predicted to occur for ozonated OSPW due to the increased biodegradability of the smaller compounds created by ozone degradation (Martin et al., 2010). In addition, only one significant unknown marker (Figure 4-34) was found as a common marker selected by PCA in both raw and ozonated OSPW biodegradation. Thus, the target compounds degraded by microorganisms were suggested to be different in raw and ozonated OSPW given the limit number of common markers.

4.11.3 Raw OSPW from Different Sites

By PCA study, it was found that NAs, oxidized NAs and 19 significant unknown markers were the major sources of variations of raw OSPW from Syncrude West in Pit, CNRL, and Suncor Pond 7. CNRL OPSW contained the highest concentrations of NAs but lowest concentrations of oxidized NAs, while Suncor OSPW had highest concentrations of oxidized NAs (especially NA+2O) but lower concentrations in NAs. 13 out of 19 significant unknown markers (Figure 4-36 and Figure 4-37) selected by PCA had their highest concentrations in CNRL and they were found to be degraded during ozonation in different conditions and/or biological treatment processes, so they were positively correlated to NAs concentrations.

In addition, 19 by-products selected (15 out of 24 by-products defined by both software programs from raw OSPW treated by ozonation in different conditions, and 4 out 5 by-products defined by Markerlynx from raw OSPW treated by biological treatment processes) had their lowest concentrations in CNRL OSPW and highest concentrations in Suncor OSPW (Figure 4-39 and Figure 4-41), which indicated their negative correlations to NAs concentrations and positive correlations to oxidized NAs concentrations. Moreover, 5 out of 24 by-products from ozonation in different conditions (*m/z=209.0945, 265.1077* and *281.1029* in O_3 datasets, *m/z=265.1084* in $O_3+CO_3^{2-}$ datasets, and *m/z=287.0954* in O_3+TNM datasets); and the remaining 1 out of 5 by-products (*m/z=287.0966*) from biological treatment processes on raw OSPW, were not detected in OSPWs from different sites, which indicated that their concentrations were negligible. Therefore, this further exhibits that those 5 by-products from ozonation in different conditions and one by-product from biological treatment processes were associated with degradation of NAs.

5.0 Conclusions

HRMS with Markerlynx software was able to detect over 1,500 markers in each OSPW samples collected during ozonation in different conditions and biological treatment processes, and raw OSPWs from different sites. Pareto scaling was found as the optimum pre-treatment method for these large datasets prior to PCA which successfully selected markers significantly changed during all treatment processes.

OSPW samples from different processes and sites were successfully characterized using significant markers selected by PCA, and those significant markers were considered to be in three groups of organic compounds including: classical naphthenic acids (NAs, n = 7 to 22, Z = 0 to -12); oxidized NAs (NAs + Ox, x = 1, 2, 3, and 4); and unknown compounds based on their *m/z* values.

PCA study verified that NAs were degraded into lower molecular compounds and/or higher oxidized NAs during ozonation in different conditions. In addition, 26 significant unknown markers (unknown compounds) were selected by PCA were present in only one individual ozonation condition experiment; 27 significant unknown markers selected by PCA were commonly present in at least two ozonation in different conditions experiments; and 24 significant unknown markers (included 12 common markers) with increasing trends were considered as by-products from all ozonation in different conditions. The differences in behaviour (e.g., decreasing and/or increasing) of the common markers for various ozonation conditions indicate the variability in the efficiency of each ozonation condition for the treatment of specific organic compounds. These differences can be used as a guide to characterize ozonation in different conditions and recommend the appropriate treatment for a specific organic compound target.

The biodegradation of NAs in both raw and ozonated OSPW were limited (~10% reduction) as shown in biological treatment processes datasets. There were only 5 out of 16 significant unknown markers selected by PCA were considered as by-products from raw OSPW biodegradation, while none of the 5 significant unknown markers selected by PCA were considered as by-products from ozonated OSPW biodegradation. In addition, only one significant unknown marker was found as the common marker selected by PCA in both raw and ozonated OSPW biodegradation samples. Thus, the conclusion was that the target compounds degraded by microorganisms were different in raw and ozonated OSPW given the limited number of common markers.

PCA study verified that raw OSPW from different sites (Syncrude West in Pit, CNRL, and Suncor Pond 7) had variable NAs concentrations with the highest concentrations in CNRL OSPW; oxidized NAs (especially NA+2O) had the highest concentrations in Suncor OSPW; and 19 unknown markers of which 15 had the highest concentrations positively correlated to the highest NAs concentrations in CNRL OSPW.

In addition, 15 by-products from ozonation in different conditions and 4 by-products from biological treatment processes had negative correlations to NAs concentrations and positive correlations to oxidized NAs (especially NA+2O) concentrations in raw OSPWs from different sites. Moreover, 5 by-products from ozonation in different conditions and 1 by-product from biological treatment processes were not detected in raw OSPWs from different sites, which indicated they were only formed during ozonation or biological treatment processes. Therefore, these further exhibits that the by-products selected by PCA from ozonation in different conditions and biological treatment processes were associated with degradation of NAs.

The use of PCA is advantageous to simplify complex data matrices, allowing for the rapid review of the potential sources of variations between OSPW samples. The PCA provided in the Markerlynx analysis software is also automated versus more labour intensive analyses such as ion mobility study or Excel plotting. With huge amount of unknown information provided by HRMS analysis, the use of a PCA is necessary to work as a filter to extract useful information from produced datasets allowing the focus on the further target analysis of significant markers.

PCA study with more replicates (at least 3) from ozonation in different conditions datasets showed more reliable results compared to the PCA study with duplicates only from biological treatment processes. Therefore, it is recommended to process at least 5 replicates for a more reliable PCA study in the future.

General trends of markers during treatment processes determined by Markerlynx exhibit the statistical significance of PCA study; while Targetlynx is recommended to be applied as a complementary approach to more accurately review the actual markers changes during treatment processes. The combination of the Markerlynx and Targetlynx approaches are currently the best method available from Waters Inc. to assess markers significance in particular treatments and to compare their behaviours in different treatment processes.

In summary, the manual analysis of all the compounds detected by HRMS in every OSPW sample is unviable. This project successfully demonstrates that the application of the PCA is a feasible approach to identify chemical markers that show significant changes (e.g., increase, decrease, or form) and potentially identify by-products from the OSPW treated by chemical (e.g., ozonation in different conditions) or biological treatment processes. The unknown by-products, other than the known NAs and oxidized NAs, defined in this study will lead the further research on the identification of their elemental compositions and molecular structures, which can contribute to filling the knowledge gap of reaction mechanisms in OSPW. As well, the markers most significantly changed, as determined in this study, can be used for monitoring the ozonation in different conditions and/or biological treatment processes. This monitoring may be useful to aid in the improvement of their efficiencies and potential use in OSPW treatments for the creation of environmentally safe final effluents. In addition, by linking the markers changed most at different times as defined by PCA to the toxicity data matrix, it may be possible to identify which markers can be attributed to the overall toxicity of different OSPW samples, and a more specific treatment process may be determined for reduction of these markers to reduce the toxicity of OSPW which will allow its release into receiving environments.

References

Abdelal, A., El-Enany, N., Belal, F. 2009. Simultaneous determination of sulpiride and its alkaline degradation product by second derivative synchronous fluorescence spectroscopy. *Talanta*, 80(2), 880-888.

Afzal, A., Drzewicz, P., Pérez-Estrada, L., Chen, Y., Martin, J.W., Gamal El-Din, M. 2012. Effect of Molecular Structure on the Degradation of Naphthenic Acids in the UV/H2O2 Advanced Oxidation Process. *Environmental Science & Technology*, 46(19), 10727-10734.

Allen, E.W. 2008. Process water treatment in Canada's oil sands industry: Target pollutants and treatment objectives. *Journal of Environmental Engineering and Science*, 7(2), 123–138.

Biryukova, O.V., Fedorak, P.M., Quideau, S.A. 2007. Biodegradation of naphthenic acids by rhizosphere microorganisms. *Chemosphere,* 67(10), 2058-2064.

Brient, J. A., Wessner, P. J., Doly, M. N. 1995. Naphthenic acids. *Encyclopedia of Chemical Technology*. New York: John Wiley and Sons.

CAMO. 2011. Webinars and Seminars. *Free webinars training*. Retrieved Jan 21, 2013, from http://www.camo.com/training/webinars-seminars.html

Catalkaya, E.C., Kargi, F. 2009. Dehalogenation, degradation and mineralization of diuron by peroxone (peroxide/ozone) treatment. *Journal of Environmental Science and Health Part A,* 44(6), 630-638.

Clayden, J., Greeves, N., Warren, S., Wothers, P. 2001. *Organic Chemistry*. New York: NY., Oxford University Press Inc.

Clemente, J.S., Mackinnon, M.D., Fedorak, P.M. 2004. Aerobic Biodegradation of two commerical naphthenic acids preparations. *Environmental Scicence & Technology*, 38(4), 1009-1016.

Corinne, W. 2010. Microbial Naphthenic Acid Degradation. *Advances in Applied Microbiology*, 70(1), 93-125.

Crawford, C.B., Ferguson, G.A. 1970. A general rotation criterion and its use in orthogonal rotation. *Psychometrika*, 35(3), 321-332.

Drzewicz, P., Afzal, A., Martin, J.W., Gamal El-Din, M. 2010. Degradation of a model naphthenic acid, cyclohexanoic acid, by vacuum-UV (172 nm) and UV (254 nm)/H2O2. *Journal of Physical Chemistry A*, 114(45), 12067–12074.

El-Aneed, A., Cohen, A., Banoub, J. 2009. Mass Spectrometry, Review of the Basics: Electrospray, MALDI, and Commonly Used Mass Analyzers. *Applied Spectroscopy Reviews*, 44(3), 210-230.

Energy Resources Conservation Board. 2010. Alberta's Energy Reserves 2010 and Supply/Demand Outlook 2011-2020; ST98-2011.

Frank, R.A., Kavanagh, R., Burnison, B.K., Headley, J.V., Peru, K.M., Van Der Krak, G., Solomon, K.R. 2006. Diethylaminoethyl-cellulose clean-up of a large volume naphthenic acid extract. *Chemosphere,* 64(8), 1346-1352.

Fraser, K., Lane, G.A., Otter, D.E., Hemar, Y., Quek, S., Harrison, S.J., Rasussen, S. 2013. Analysis of metabolic markers of tea origin by UHPLC and high resolution mass spectctrometry. *Food Research International,* 53(2), 827-835.

Gamal El-Din, M., Fu, H., Wang, N., Chelme-Ayala, P., Perez-Estrada, L., Drzewicz, P., Martin, J.W., Zubot, W., Smith, D.W. 2011. Naphthenic acids speciation and removal during petroleum-coke adsorption and ozonation of oil sands process-affected water. *Science of the Total Environment,* 409 (23), 5119-5125.

Gao, S., Zhang, Y., Meng, J., Shu, J. 2009. Online investigations on ozonation products of pyrene and benz[α]anthracene particles with a vacuum ultraviolet photoioinization aerosol time-of-flight mass spectrometer. *Atmospheric Environment.* 43(21), 3319-3325.

Gao, Y.; Williams, D.D., Williams, N.E. 1999. Data transformation and standardization in the multivariate analysis of river water quality. *Ecological*

Society of America, 9(2), 669-677.

Grebel, J., Pignatello, J.J., Mitch, W.A. 2010. Effects of halide ions and carbonates on organic contaminant degradation by hydroxyl radical-based advanced oxidation processes in saline waters. *Environmental Science and Technology,*. 44(17), 6822-6828.

Hadwin, A. K. M., Del Rio, L. F., Pinto, L. J., Painter, M., Routledge, R., Moore, M. 2006. Microbial communities in wetlands of the Athabasca oil sands: Genetic and metabolic characterization. *FEMS Microbial Ecology,* 55(1), 68–78.

Han, X., Mackinnon, M.D., Martin, J.W. 2009. Estimating the in situ Biodegradation of Naphthenic acids in oil sands process waters by HPLC/HRMS. *Chemosphere*, 76(1), 63-70.

Han, X., Scott, A.C., Fedorak, P.M., Bataineh, M., Martin, J.W. 2008. Influence of Molecular Structure on the Biodegradability of Naphthenic Acids. *Environemtnal Science & Technology*, 42(4), 1290-1295.

Head, I. M., Jones, M., Larter, S. 2003. Biological activity in the deep subsurface and the origin of heavy oil. *Nature*, 426(6964), 344–352.

Headley, J.V., Du, J.L., Peru, K., McMartin, D.W. 2009. Electrospray ionization mass spectrometry of the photodegradation of naphthenic acids mixtures irradiated with titanium dioxide. *Journal of Environmental Science and Health Part A –Toxic/Hazardous Substances & Environmental Engineering*, 44(6), 591–597.

Helena, B., Pardo, R., Vega, M., Barrado, E., Fernandez, J.M., Fernandez, L. 2000. Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. *Water Research,* 34(3), 807-816.

Herman, D.C., Fedorak, P.M., Mackinnon, M.D., Costerton. J.W. 1994. Biodegradation of naphthenic acids by microbial-populations indigenous to oil sands tailings. *Canadian Journal of Microbiology*, 40(6), 467–477.

Hiner, A.N.P., Hernandez-Ruiz, J., Williams, G.A., Arnao, M.B., Garcia-Canovas, F., Acosta, M. 2001. Catalyase-like oxygen production by horseradish peroxidase must predominatly be an enzyme-catalyzed reaction. *Archieves of Biochemistry and Biophysics,* 392(2), 295-302.

Holowenko, F.M., Mackinnon, M.D., Fedorak, P.M. 2001. Naphthenic Acids and Surrogate Naphthenic Acids in Methanogenic Microcosms. *Water Research*, 35(11), 2595-2606.

Holowenko, F.M., MacKinnon, M.D., Fedorak, P.M. 2002. Characterization of naphthenic acids in oil sands wastewaters by gas chromatography-mass spectrometry. *Water Research,* 36(22), 2843−2855.

Jackson, J.E. 1991. *A Users Guide to Principal Components*. New York: Wiley & Sons Inc.

Jeris J. S., McCarty P. L. 1965. The biochemistry of methane fermentation using 14C tracers. *Journal of Water Pollutant Control Fed*, 37(3), 178–192.

Johnson, R.A., Wichern, D.W. 2007. *Applied Multivariate Analysis*. New Jersey, NJ: Pearson Prentice Hall.

Jolliffe, I.T. 2002. *Principal Component Analysis*. New York, NY: Springer.

Jung, Y.J., Oh, B.S., Kim, K.S., Koga, M., Shinohara, R., Kang, J. 2010. The degradation of diethyl phthalate (DEP) during ozonation: oxidation by-products study. *Journal of Water & Health,* 8(2). 290-298.

Kavanagh, R.J., Burnison, B.K., Frank, R.A., Solomon, K.R., Van Der Kraak, G. 2009. Detecting oil sands process-affected waters in the Alberta oil sands region using synchronous fluorescence spectroscopy. *Chemosphere,* 76(1), 120-126.

Kenl, J. 2003. *Analytical Chemistry for Technicians.* Boca Raton: Lewis Publishers.

Kishimoto, N., Ueno, S. 2012. Catalytic effects of several iron species on ozonation. *Journal of Water and Environmental Technology,* 10(2), 205-215.

Lai, J.W.S., Pinto, L.J., Kiehlmann, E., Bendell-Young, L.I., Moore, M.M. 1996. Factors That Affect the Degradation of Naphthenic Acids in Oil Sands Wastewater by Indigenous Microbial Communities. *Enivronmental Toxicology and Chemistry*, 15(9), 1482-1491.

Liang, X., Zhu, X., Butler, E.C. 2011. Comparison of four advanced oxidation processes for the removal of naphthenic acids from model oil sands process water. *Journal of Hazardous Materials*, 190(1), 168-176.

Liu, J., Xu, Z., Masliyah, M. 2005. Processability of oil sand ores in Alberta. *Energy and Fuels*, 19(5), 2056–2063.

Lo, C.C., Brownlee, B.G., Bunce, N.J. 2006. Mass spectrometric and toxicological assays of Athabasca oil sands naphthenic acids. *Water Research*, 40(4) 655–664.

Mahdavi, H., Ulrich, A.C., Liu, Y. 2010. Metal removal from tailings pond water using indigenous micro-alga. In: Second International Oil Sands Tailing Conference, December 5-8, 2010, Edmonton, Alberta, Canada, pp. 279−283.

Martin, J.W., Barri, T., Han, X., Fedorak, P.M., El-Din, M.G., Perez, L., Scott, A.C. & Jiang, J.T. 2010. Ozonation of oil sands process-affected water accelerates microbial bioremediation. *Environmental Science & Technology,* 44 (21), 8350-8356.

Matthew, S.R., Pereira, A.d.S, Fennell, J., Devies, M., Johnson, J., Sliva, L., Martin, J.W. 2012. Quantitative and qualitative analysis of naphthenic acids in natural waters surrounding the Canadian oil sands industry. *Environmental Science & Technology*, 46(23), 12796-12805.

McMartin, D.W., Headley, J.V., Friesen, D.A., Peru, K.M., Gillies, J.A. 2004. Photolysis of naphthenic acids in natural surface water. *Journal of Environmental Science and Health. Part A, Toxic/Hazardous Substances & Environmental Engineering*, 39(6), 1361−1383.

Mishra, S., Meda, V., Dalai, A.K., McMartin, D.W., Headley, J.V., Peru, K.M. 2010. Photocatalysis of naphthenic acids in water. *Journal of Water Resource and Protection*, 2(7), 644−650.

National Energy Board. 2006. Canada's Oil Sands Opportunities and Challenges to 2015: An Update. The Publications Office National Energy Board, Calgary, Alberta, Canada.

Nawrocki, J., Kasprzyk-Hordern, B. 2010. The Efficiency and Mechanisms of Catalytic Ozonation. *Applied Catalysis B, Environmental*, 99(1), 27-42.

Olsen, R.L., Chappell, R.W., Loftis, J.C. 2012. Water quality sample collection, data treatment and results presentation for principal components analysis-literature review and Illinois River watershed case study. *Water Research*, 46(9), 3110-3122.

Osborne, J., Anna, B. 2004. Sample size and subject to item ratio in principal component analysis. *Practical Assessment, Research & Evaluation*, 9(11), 111-119.

Parinet, B., Lhote, A., Legube, B. 2004. Principal component analysis: an appropriate tool for water quality evaluation and management—application to a tropical lake system. *Ecological Modeling,* 178(3). 295-311.

Peldszus, S., Halle, C., Peiris, R.H., Hamouda, M., Jin, X., Legge, R.L., Budman, H., Moresoli, C., Huck, P. 2011. Reversible and irreversible low-pressure membrane foulants in drinking water treatment: Identification by principal component analysis of fluorescence EEM and mitigation by biofiltration pretreatment. *Water Research,* 45(16), 5161-5170.

Perez-Estrada, L.A., Han, X., Drzewicz, P., Gamal El-Din, M., Fedorak, P.M., Martin, J.W. 2011. Structure-reactivity of naphthenic acids in the ozonation process. *Environmental Science & Technology*, 45 (17), 7431-7437.

Pharr, D.Y., Mckenzie, J.K., Hickman, A.B. 1992. Fingerprinting petroleum contamination using synchronous scanning fluorescence spectroscopy. *Ground Water*, 30(4), 484-489.

Pourrezaei, P., Drzewicz, P., Wang, Y., Gamal El-Din, M., Perez-Estrada, L.A., Martin, J.W., Anderson, J., Wiseman, S., Liber, K., Giesy, J.P. 2011. The impact of metallic coagulants on the removal of organic compounds from oil sands process-affected water. *Environmental Science & Technology,* 45(19), 8452-8459.

Qi, F.; Xu, B. Chen, Z., Ma, J., Sun, D., Zhang, L. 2009. Influence of aluminum oxides Surface properties on catalyzed ozonation of 2,4,6-trichloroanisole. *Seperation and Purification Technology*, 66(2), 405-410.

Reid, M.K., Spencer, K.L. 2009. Use of principal components analysis on estuarine sediment datasets: The effect of data pre-treatment. *Environmental Pollution*, 157(8-9). 2275-2281.

Ricardo, A.R., Velizarov, S., Crespo, J.G., Reis, M.A.M. 2011. Multivariate analysis of the transport in an Ion Exchange Membrane Bioreactor for removal of anionic micropollutants from drinking water. *Water Science & Technology*, 63(10), 2207-2212.

Rowland, S.J., West, C.E., Jones, D., Scarlettt, A.G., Frank, R.A., Hewitt, L.M. 2011. Steroidal Aromatic 'Naphthenic Acids' in Oil Sands Process-Affected Water: Structural Comparisons with Environmental Estrogens. *Environmental Science & Technology*, 45(22). 9806-9815.

Sanguansat, P. 2012. *Principal Component Analysis*. Winchester: InTech.

SAS. 2012. Training Course. *Training & Books*. Retrieved August 10, 2013, from http://support.sas.com/training/index.html

Scott, A.C., Mackinnon, M.D., Fedorak, P.M. 2005. Naphthenic acids in Athabasca oil sands tailings waters are less biodegradable than commercial naphthenic acids. *Environmental Science & Technology*, 39(21), 8388−8394.

Scott, A.C., Zubot, W., Mackinnon, M.D., Smith, D.W., Fedorak, P.M. 2008. Ozonation of oil sands process water removes naphthenic acids and toxicity. *Chemosphere,* 71 (1), 156-160.

Skoog, D.A., Holer, F.J., Nieman, T.A. 1998. *Principals of Instrumental Analysis.* Ontario: Nelson Thomson Learning.

Sleighter, R., Liu, Z., Xue, J., Hatcher, P.G. 2010. Multivariate statistical approaches for the characterization of dissolved organic matter analyzed by ultrahigh resolution mass spectrometry. *Environmental Science and Technology*, 44(19), 7576-7582.

Smith, L.I. 2002. A tutorial on Principal Component Analysis. *Spring Conference on Computer Graphics*. Retrieved July 12, 2013, from http://www.sccg.sk/~haladova/principal_components.pdf

Sohn, J., Amy, G., Cho, J., Lee, Y., Yoon, Y. 2004. Disinfectant decay and disinfection by-products formation model development: chlorination and ozonation by-products. *Water Research*, 38(10). 2461-2478.

Stevens, J. 1986. *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K., van der Werf, M.J. 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *Biomed Central Genomic*, 7(1), 142-158.

Waters Inc. 2010. Masslynx Manuals. *Software Manual*. Retrieved Jan 20, 2014, from http://www.waters.com/waters/home.htm?locale=en_US

Webster, R. 2001. Statistics to support soil research and their presentation. *European Journal of Soil Science*, 52(2), 331-340.

Whitby, C. 2010. Microbial naphthenic acid degradation. A*dvances in Applied Microbiology*, 70(7), 93–125.

Appendix A

Exact Masses of Naphthenic Acids (NAs) & Oxidized NAs

Table A- 1: Exact Masses of Classical NAs ($C_nH_{2n+Z}O_2$)

Z	0	-2	-4	-6	-8	-10	-12
n							
7	129.0921	127.0765					
8	143.1078	141.0921					
9	157.1234	155.1078	153.0921				
10	171.1391	169.1234	167.1078				
11	185.1547	183.1391	181.1234	179.1078			
12	199.1704	197.1547	195.1391	193.1234			
13	213.1860	211.1704	209.1547	207.1391	205.1234		
14	227.2017	225.1860	223.1704	221.1547	219.1391		
15	241.2173	239.2017	237.1860	235.1704	233.1547	231.1391	
16	255.2330	253.2173	251.2017	249.1860	247.1704	245.1547	
17	269.2486	267.2330	265.2173	263.2017	261.1860	259.1704	257.1547
18	283.2643	281.2486	279.2330	277.2173	275.2017	273.1860	271.1704
19	297.2799	295.2643	293.2486	291.2330	289.2173	287.2017	285.1860
20	311.2956	309.2799	307.2643	305.2486	303.2330	301.2173	299.2017
21	325.3112	323.2956	321.2799	319.2643	317.2486	315.2330	313.2173
22	339.3269	337.3112	335.2956	333.2799	331.2643	329.2486	327.2330

Table A- 2: Exact Masses of NA+O ($C_nH_{2n+Z}O_3$)

Z	0	-2	-4	-6	-8	-10	-12
n							
7	145.0870	143.0714					
8	159.1027	157.0870					
9	173.1183	171.1027	169.0870				
10	187.1340	185.1183	183.1027				
11	201.1496	199.1340	197.1183	195.1027			
12	215.1653	213.1496	211.1340	209.1183			
13	229.1809	227.1653	225.1496	223.1340	221.1183		
14	243.1966	241.1809	239.1653	237.1496	235.1340		
15	257.2122	255.1966	253.1809	251.1653	249.1496	247.1340	
16	271.2279	269.2122	267.1966	265.1809	263.1653	261.1496	
17	285.2435	283.2279	281.2122	279.1966	277.1809	275.1653	273.1496
18	299.2592	297.2435	295.2279	293.2122	291.1966	289.1809	287.1653
19	313.2748	311.2592	309.2435	307.2279	305.2122	303.1966	301.1809
20	327.2905	325.2748	323.2592	321.2435	319.2279	317.2122	315.1966
21	341.3061	339.2905	337.2748	335.2592	333.2435	331.2279	329.2122
22	355.3218	353.3061	351.2905	349.2748	347.2592	345.2435	343.2279

Table A- 3: Exact Masses of NA+2O ($C_nH_{2n+Z}O_4$)

Z	0	-2	-4	-6	-8	-10	-12
n							
7	161.0819	159.0663					
8	175.0976	173.0819					
9	189.1132	187.0976	185.0819				
10	203.1289	201.1132	199.0976				
11	217.1445	215.1289	213.1132	211.0976			
12	231.1602	229.1445	227.1289	225.1132			
13	245.1758	243.1602	241.1445	239.1289	237.1132		
14	259.1915	257.1758	255.1602	253.1445	251.1289		
15	273.2071	271.1915	269.1758	267.1602	265.1445	263.1289	
16	287.2228	285.2071	283.1915	281.1758	279.1602	277.1445	
17	301.2384	299.2228	297.2071	295.1915	293.1758	291.1602	289.1445
18	315.2541	313.2384	311.2228	309.2071	307.1915	305.1758	303.1602
19	329.2697	327.2541	325.2384	323.2228	321.2071	319.1915	317.1758
20	343.2854	341.2697	339.2541	337.2384	335.2228	333.2071	331.1915
21	357.3010	355.2854	353.2697	351.2541	349.2384	347.2228	345.2071
22	371.3167	369.3010	367.2854	365.2697	363.2541	361.2384	359.2228

Table A- 4: Exact Masses of NA+3O ($C_nH_{2n+Z}O_5$)

Z	0	-2	-4	-6	-8	-10	-12
n							
7	177.0768	175.0612					
8	191.0925	189.0768					
9	205.1081	203.0925	201.0768				
10	219.1238	217.1081	215.0925				
11	233.1394	231.1238	229.1081	227.0925			
12	247.1551	245.1394	243.1238	241.1081			
13	261.1707	259.1551	257.1394	255.1238	253.1081		
14	275.1864	273.1707	271.1551	269.1394	267.1238		
15	289.2020	287.1864	285.1707	283.1551	281.1394	279.1238	
16	303.2177	301.2020	299.1864	297.1707	295.1551	293.1394	
17	317.2333	315.2177	313.2020	311.1864	309.1707	307.1551	305.1394
18	331.2490	329.2333	327.2177	325.2020	323.1864	321.1707	319.1551
19	345.2646	343.2490	341.2333	339.2177	337.2020	335.1864	333.1707
20	359.2803	357.2646	355.2490	353.2333	351.2177	349.2020	347.1864
21	373.2959	371.2803	369.2646	367.2490	365.2333	363.2177	361.2020
22	387.3116	385.2959	383.2803	381.2646	379.2490	377.2333	375.2177

Table A- 5: Exact Masses of NA+4O ($C_nH_{2n+Z}O_6$)

Z	0	-2	-4	-6	-8	-10	-12
n							
7	193.0718	191.0561					
8	207.0874	205.0718					
9	221.1031	219.0874	217.0718				
10	235.1187	233.1031	231.0874				
11	249.1344	247.1187	245.1031	243.0874			
12	263.1500	261.1344	259.1187	257.1031			
13	277.1657	275.1500	273.1344	271.1187	269.1031		
14	291.1813	289.1657	287.1500	285.1344	283.1187		
15	305.1970	303.1813	301.1657	299.1500	297.1344	295.1187	
16	319.2126	317.1970	315.1813	313.1657	311.1500	309.1344	
17	333.2283	331.2126	329.1970	327.1813	325.1657	323.1500	321.1344
18	347.2439	345.2283	343.2126	341.1970	339.1813	337.1657	335.1500
19	361.2596	359.2439	357.2283	355.2126	353.1970	351.1813	349.1657
20	375.2752	373.2596	371.2439	369.2283	367.2126	365.1970	363.1813
21	389.2909	387.2752	385.2596	383.2439	381.2283	379.2126	377.1970
22	403.3065	401.2909	399.2752	397.2596	395.2439	393.2283	391.2126

Appendix B

Supplementary Information for Ozonation in Different Conditions

B-1: O_3+CO_3^{2-} Datasets Analysis

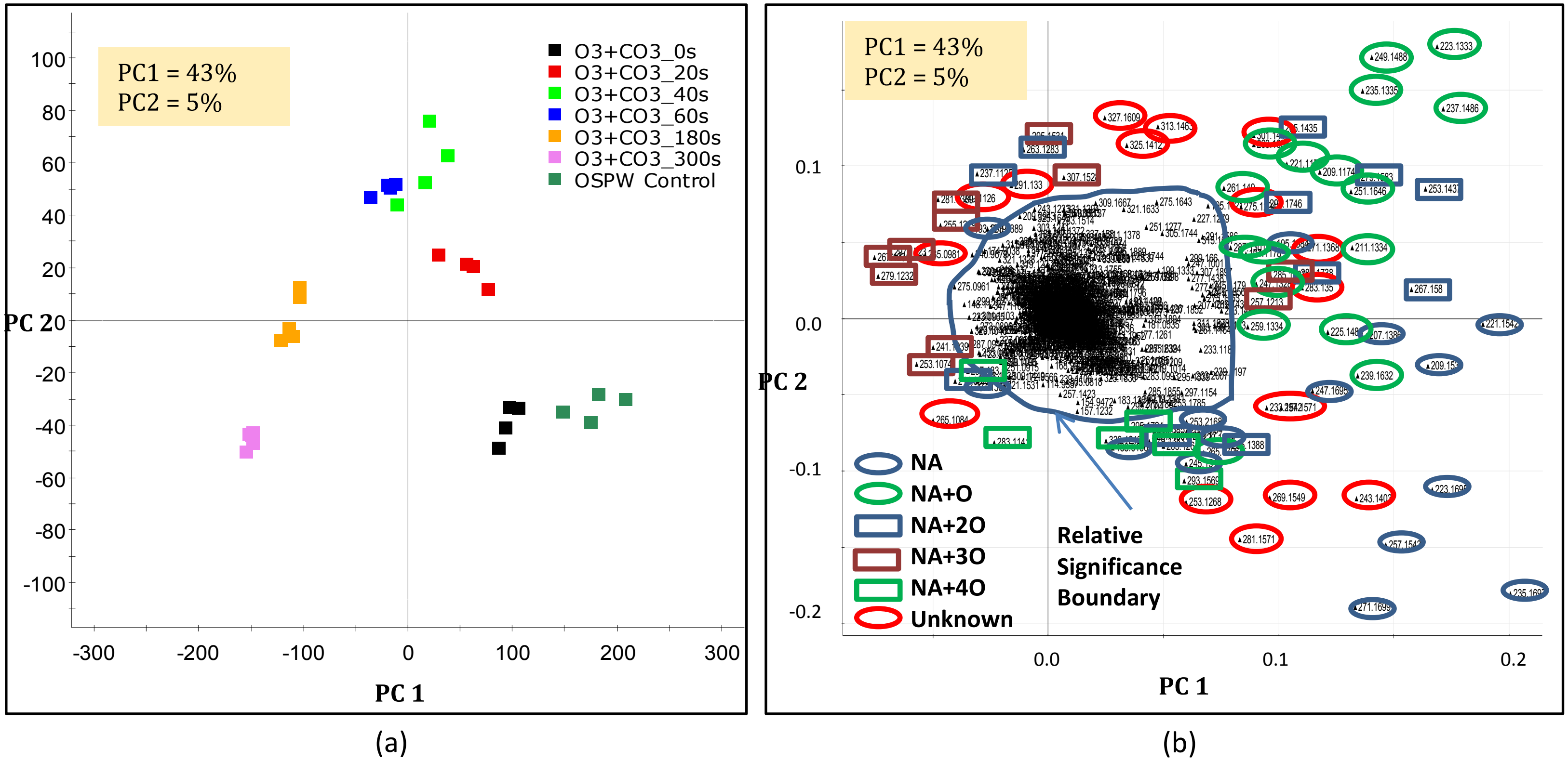


Figure B- 1: PCA (a) Score Plot; and (b) Loading Plot, based on O_3+CO_3^{2-} Datasets.

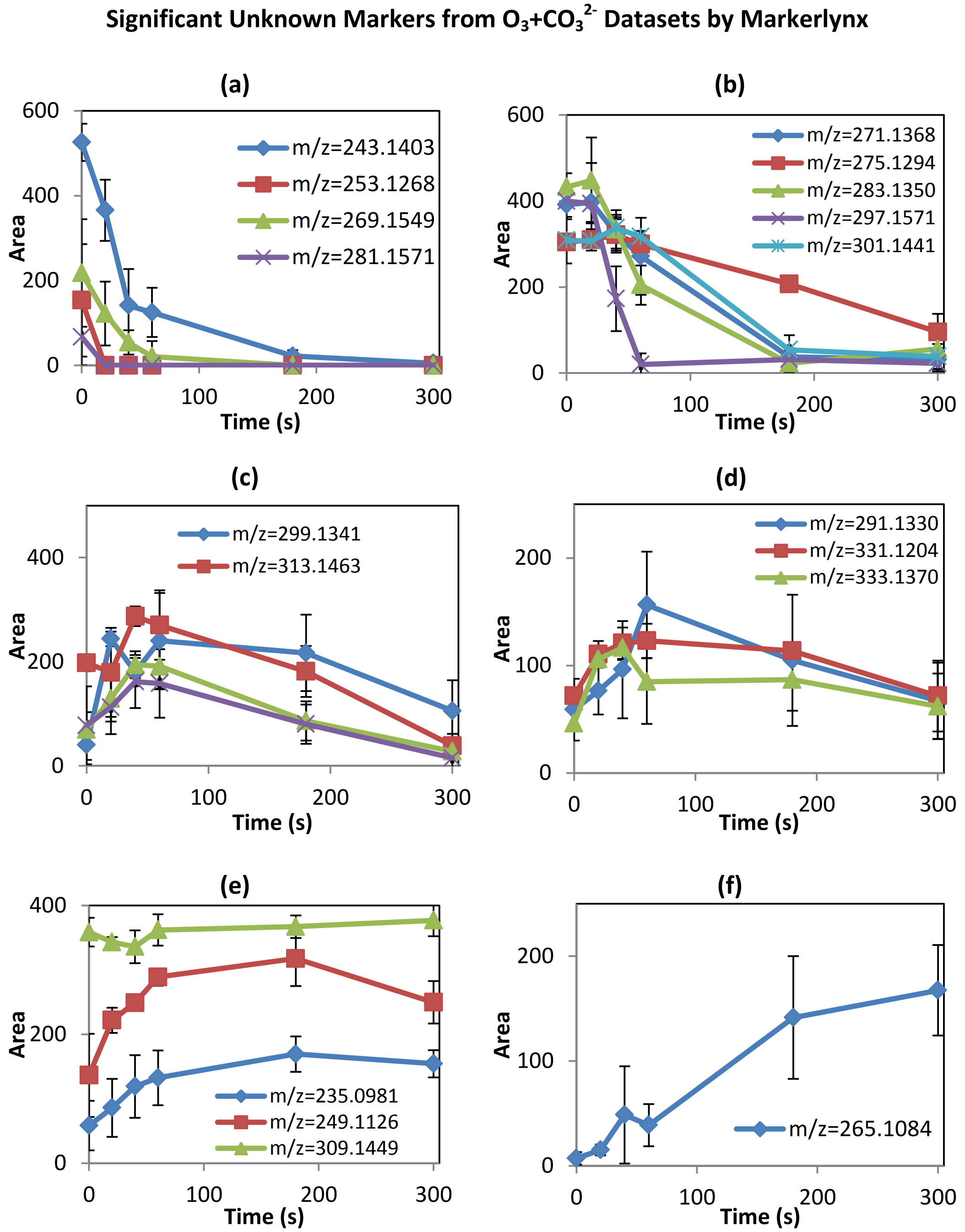


Figure B- 2: Trends of Significant Unknown Markers Selected from PCA Results based on O_3 + CO_3^{2-} Datasets by Markerlynx. Plot (a) and (b) show markers with decreasing trends; plot (c) and (d) show markers with increasing followed by decreasing trends; plot (e) and (f) show markers with increasing trends.

Significant Unknown Markers from $O_3 + CO_3^{2-}$ Data by Targetlynx

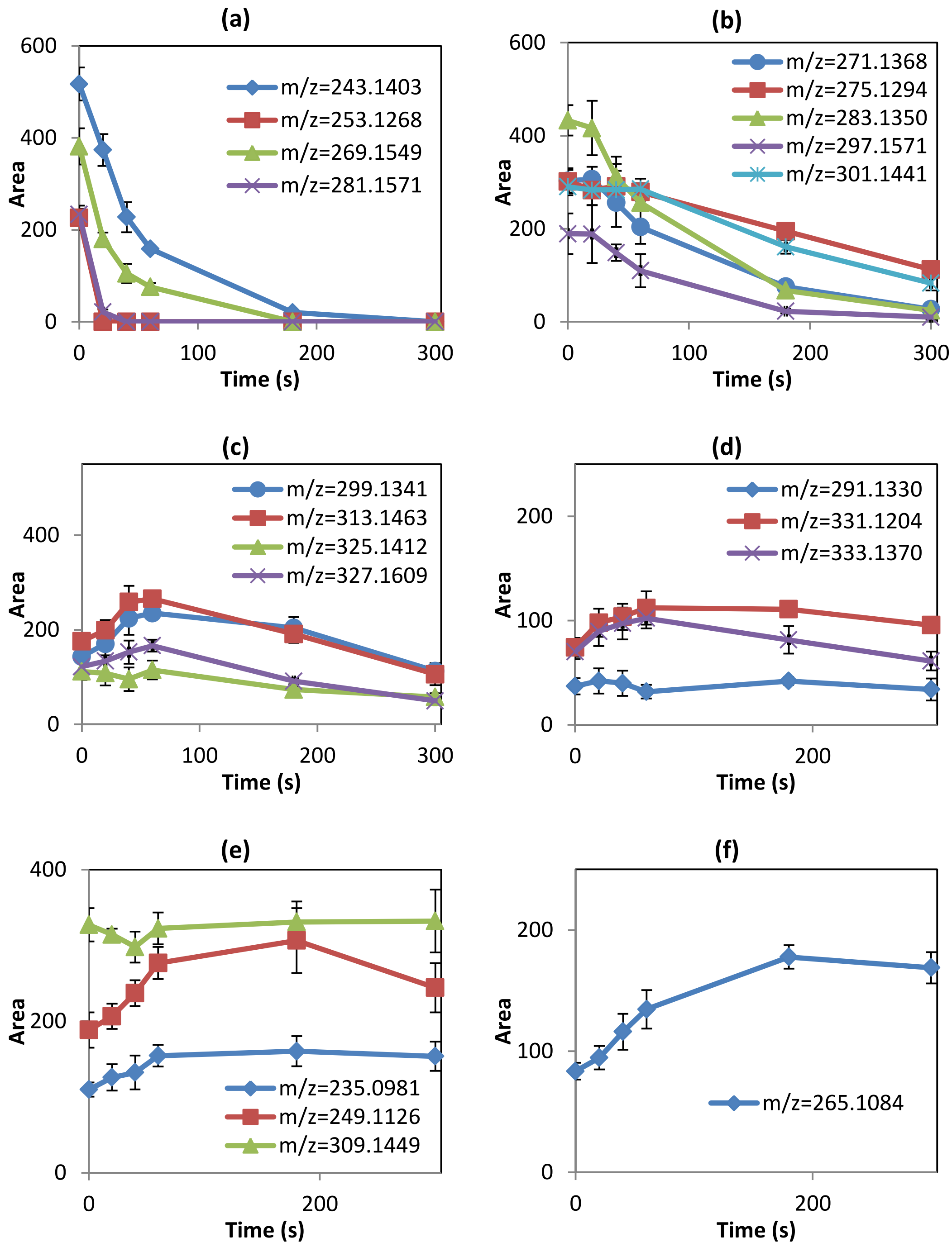


Figure B- 3: Trends of Significant Unknown Markers Selected from PCA Results based on $O_3 + CO_3^{2-}$ Datasets by Targetlynx. Plot (a) to (f) show the trends of markers corresponding to the markers in Figure B-2 (a) to (f).

Table B- 1: Elemental Compositions for Significant Unknown Markers from $O_3 + CO_3^{2-}$ Datasets.

Detected Mass	Calculated Mass	Error PPM	DBE	Formula	Detected Mass	Calculated Mass	Error PPM	DBE	Formula	Detected Mass	Calculated Mass	Error PPM	DBE	Formula
235.0981	235.0970	4.7	6.5	C13 H15 O4	***275.1294***	275.1283	4	7.5	C16 H19 O4	***309.1449***	309.1450	-0.3	6.5	C15 H21 N2 O5
	235.0997	-6.8	11	C16 H13 N O		275.1310	-5.8	12	C19 H17 N O		309.1459	-3.2	5.5	C16 H25 N2 S2
	235.1004	-9.8	1.5	C10 H19 O4 S		275.1317	-8.4	2.5	C13 H23 O4 S		309.1432	5.5	1	C13 H27 N O3 S2
243.1403	243.1419	-6.6	2.5	C13 H23 O2 S	***281.1571***	281.1575	-1.4	4.5	C16 H25 O2 S		309.1425	7.8	10.5	C19 H21 N2 S
	243.1385	7.4	7.5	C16 H19 O2	***283.1350***	283.1361	-3.9	14	C21 H17 N	***313.1463***	313.1467	-1.3	14	C22 H19 N O
249.1126	249.1127	-0.4	6.5	C14 H17 O4		283.1334	5.7	9.5	C18 H19 O3		313.1474	-3.5	4.5	C16 H25 O4 S
253.1268	253.1262	2.4	4.5	C14 H21 O2 S		283.1368	-6.4	4.5	C15 H23 O3 S		313.1440	7.3	9.5	C19 H21 O4
	255.0877	5.9	4.5	C13 H19 O S2		283.1328	7.8	0.5	C10 H23 N2 O5 S	***325.1412***	325.1408	1.2	5.5	C16 H25 N2 O S2
265.1084	265.1085	-0.4	5.5	C15 H21 S2	***291.1330***	291.1327	1	2	C13 H25 N O2 S2		325.1440	-8.6	10.5	C20 H21 O4
	265.1076	3	6.5	C14 H17 O5		291.1345	-5.2	7.5	C15 H19 N2 O4		325.1381	9.5	1	C13 H27 N O4 S2
	265.1103	-7.2	11	C17 H15 N O2	***297.1571***	297.1576	-1.7	5	C15 H23 N O5	***327.1609***	327.1596	4	9.5	C20 H23 O4
	265.1110	-9.8	1.5	C11 H21 O5 S		297.1585	-4.7	4	C16 H27 N S2		327.1623	-4.3	14	C23 H21 N O
269.1549	269.1542	2.6	8.5	C18 H21 O2		297.1551	6.7	9	C19 H23 N S		327.1630	-6.4	4.5	C17 H27 O4 S
	269.1575	-9.7	3.5	C15 H25 O2 S	***299.1341***	299.1344	-1	9	C18 H21 N O S	***331.1204***	331.1208	-1.2	14	C21 H17 N O3
271.1368	271.1368	0	3.5	C14 H23 O3 S		299.1317	8	4.5	C15 H23 O4 S		331.1190	4.2	8.5	C19 H23 O S2
	271.1361	2.6	13	C20 H17 N	***301.1441***	301.1440	0.3	8.5	C18 H21 O4		331.1235	-9.4	18.5	C24 H15 N2
	271.1395	-10	8	C17 H21 N S		301.1467	-8.6	13	C21 H19 N O	***333.1370***	333.1365	1.5	13	C21 H19 N O3
											333.1392	-6.6	17.5	C24 H17 N2
											333.1347	6.9	7.5	C19 H25 O S2
											333.1399	-8.7	8	C18 H23 N O3 S

B-2: O_3 + TBA Datasets Analysis

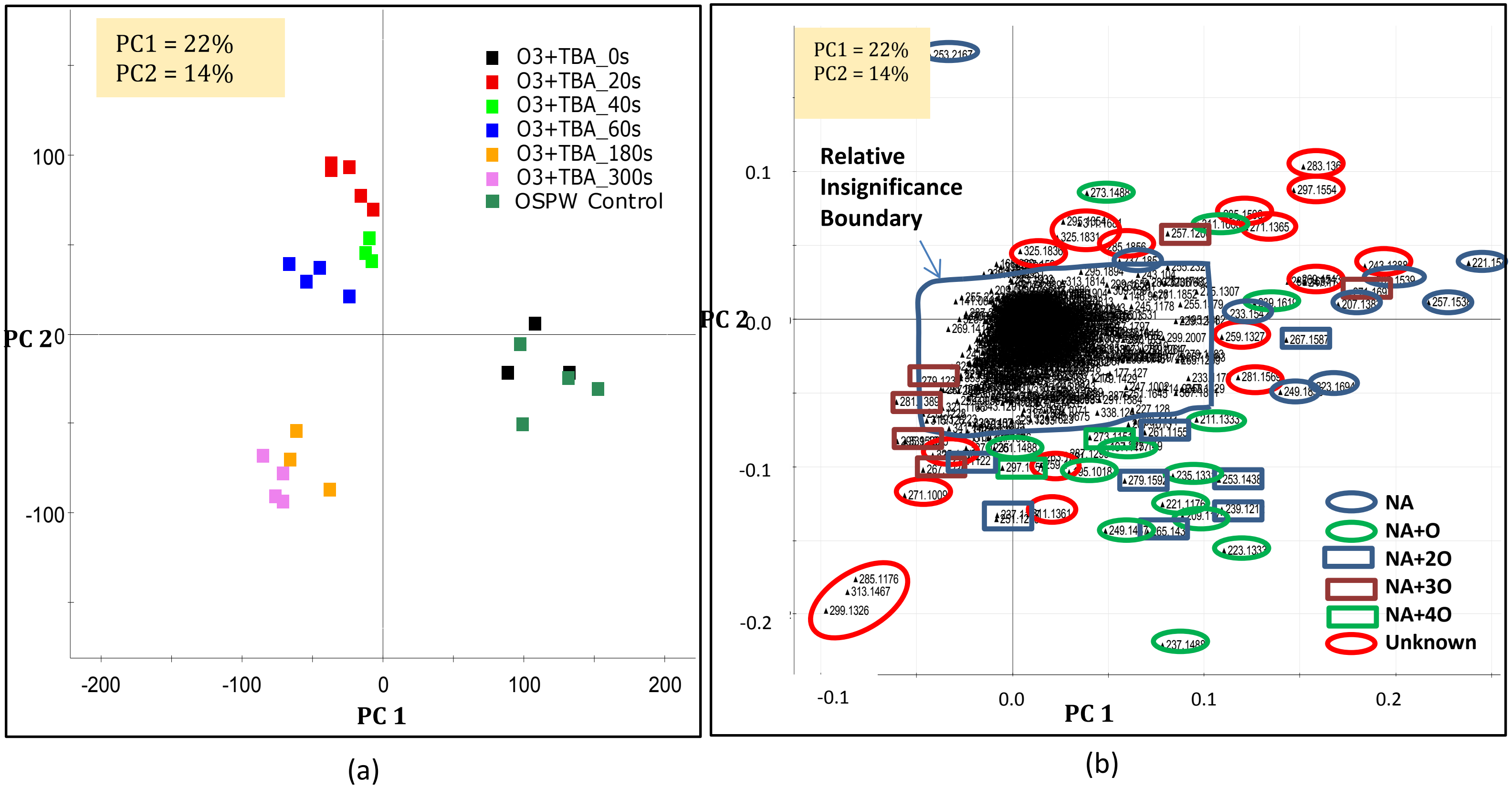
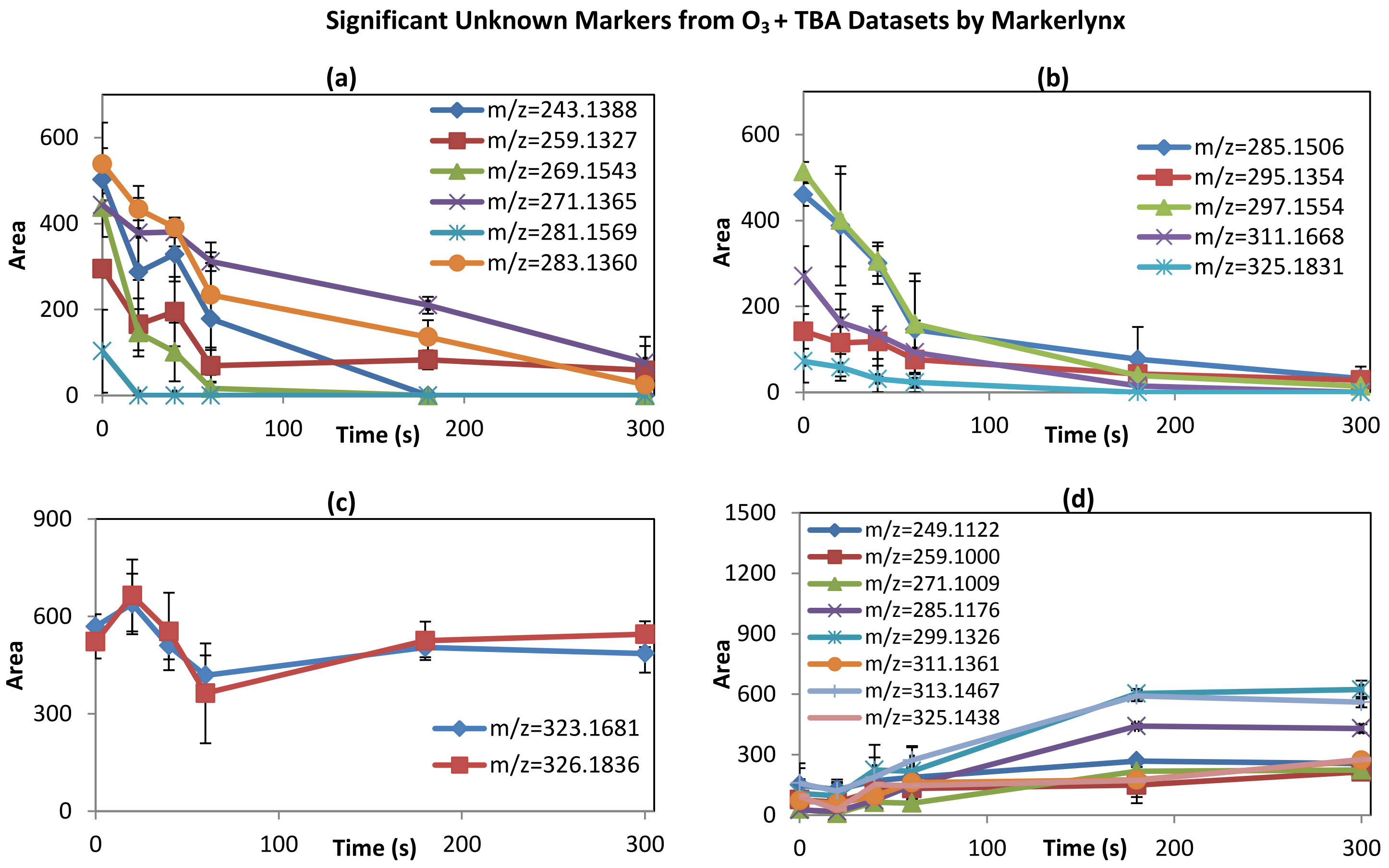


Figure B- 4: PCA (a) Score Plot; and (b) Loading Plot, based on O_3+TBA Datasets.


Significant Unknown Markers from O_3 + TBA Datasets by Markerlynx

(a)

(b)

(c)

(d)

Figure B- 5: Trends of Significant Unknown Markers Selected from PCA Results based on O_3 + TBA Datasets by Markerlynx. Plot (a) and (b) show markers with decreasing trends; plot (c) shows markers with increasing followed by decreasing trends; plot (d) shows markers with increasing trends.

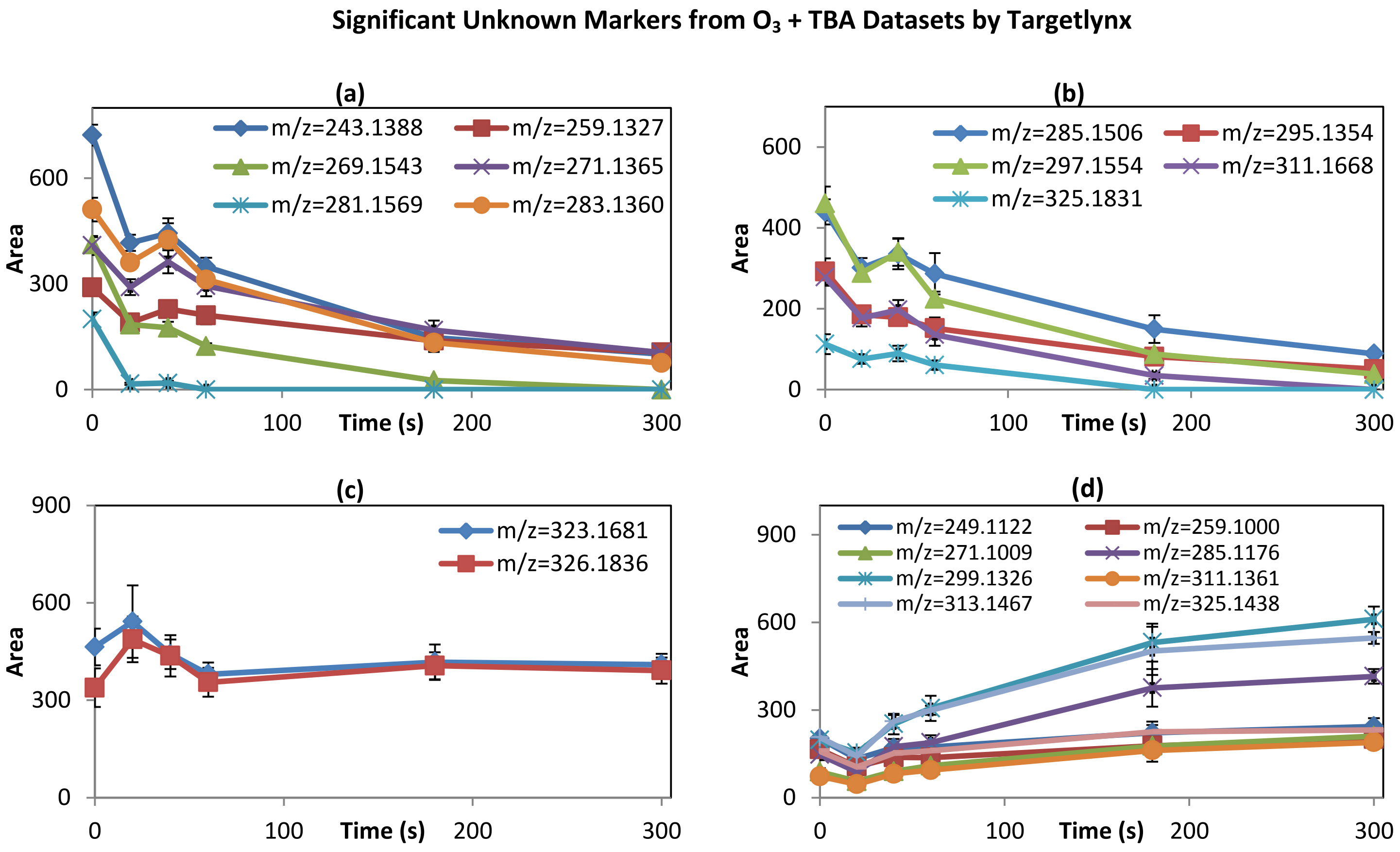


Figure B- 6: Trends of Significant Unknown Markers Selected from PCA Results based on O_3 + TBA Datasets by Targetlynx. Plot (a) to (d) show the trends of markers corresponding to the markers in Figure B-5 (a) to (d).

Table B- 2: Elemental Compositions for Significant Unknown Markers from O_3 + TBA Datasets.

Detected Mass	Calculated Mass	Error PPM	DBE	Formula	Detected Mass	Calculated Mass	Error PPM	DBE	Formula	Detected Mass	Calculated Mass	Error PPM	DBE	Formula
243.1388	243.1385	1.2	7.5	C16 H19 O2	***285.1506***	285.1517	-3.9	13	C21 H19 N	***323.1681***	323.1681	0	5.5	C18 H27 O3 S
249.1122	249.1127	-2	6.5	C14 H17 O4		285.1491	5.3	8.5	C18 H21 O3		323.1674	2.2	15	C24 H21 N
259.1000	259.0997	1.2	13	C18 H13 N O		285.1524	-6.3	3.5	C15 H25 O3 S		323.1708	-8.4	10	C21 H25 N S
	259.1004	-1.5	3.5	C12 H19 O4 S	***295.1354***	295.1361	-2.4	15	C22 H17 N	***313.1467***	313.1467	0	14	C22 H19 N O
259.1327	259.1334	-2.7	7.5	C16 H19 O3		295.1368	-4.7	5.5	C16 H23 O3 S		313.1474	-2.2	4.5	C16 H25 O4 S
	259.1303	9.3	2.5	C12 H23 N2 S2		295.1334	6.8	10.5	C19 H19 O3		313.1440	8.6	9.5	C19 H21 O4
269.1543	269.1542	0.4	8.5	C18 H21 O2		295.1328	8.8	1.5	C11 H23 N2 O5 S	***325.1438***	325.1440	-0.6	10.5	C20 H21 O4
271.1009	271.1004	1.8	4.5	C13 H19 O4 S	***297.1554***	297.1551	1	9	C19 H23 N S		325.1467	-8.9	15	C23 H19 N O
	271.0997	4.4	14	C19 H13 N O		297.1576	-7.4	5	C15 H23 N O5		325.1408	9.2	5.5	C16 H25 N2 O S2
	271.1031	-8.1	9	C16 H17 N O S	***299.1326***	299.1317	3	4.5	C15 H23 O4 S	***325.1831***	325.1830	0.3	14	C24 H23 N
271.1365	271.1368	-1.1	3.5	C14 H23 O3 S		299.131	5.3	14	C21 H17 N O		325.1837	-1.8	4.5	C18 H29 O3 S
	271.1361	1.5	13	C20 H17 N		299.1344	-6	9	C18 H21 N O S		325.1804	8.3	9.5	C21 H25 O3
281.1569	281.1575	-2.1	4.5	C16 H25 O2 S	***311.1361***	311.1351	3.2	0.5	C13 H27 O4 S2	***326.1836***	326.1842	-1.8	5	C16 H26 N2 O5
	281.1542	9.6	9.5	C19 H21 O2		311.1378	-5.5	5	C16 H25 N O S2		326.1850	-4.3	4	C17 H30 N2 S2
283.1360	283.1361	-0.4	14	C21 H17 N		311.1344	5.5	10	C19 H21 N O S		326.1817	5.8	9	C20 H26 N2 S
	283.1368	-2.8	4.5	C15 H23 O3 S	***311.1668***	311.1674	-1.9	14	C23 H21 N					
	283.1334	9.2	9.5	C18 H19 O3		311.1681	-4.2	4.5	C17 H27 O3 S					
285.1176	285.1187	-3.9	9	C17 H19 N O S		311.1647	6.7	9.5	C20 H23 O3					
	285.1161	5.3	4.5	C14 H21 O4 S		311.1641	8.7	0.5	C12 H27 N2 O5 S					
	285.1154	7.7	14	C20 H15 N O										

B-3: $O_3 + TBA + CO_3^{2-}$ Data Analysis

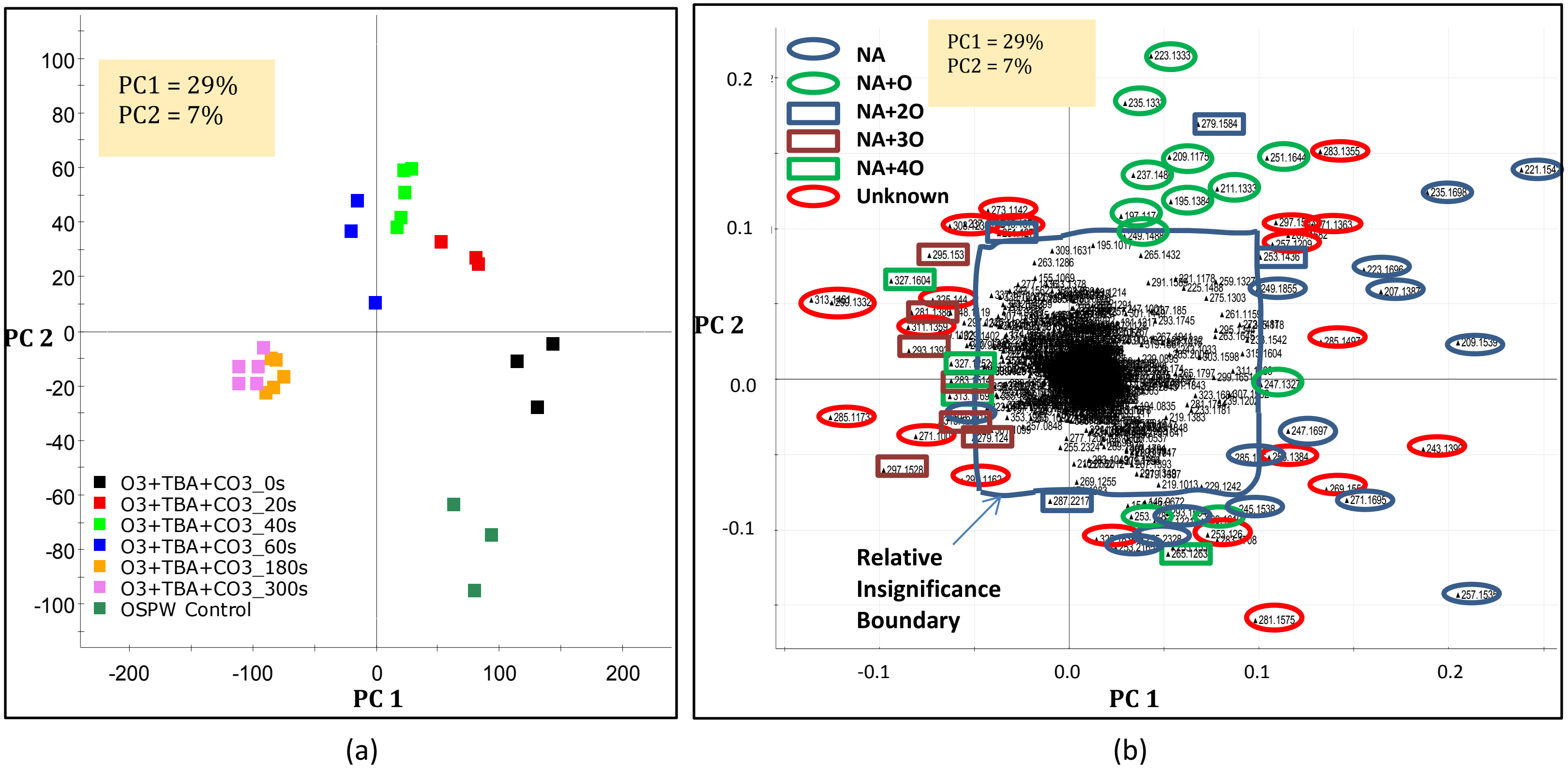


Figure B- 7: PCA (a) Score Plot; and (b) Loading Plot, based on $O_3 + TBA + CO_3^{2-}$ Datasets.

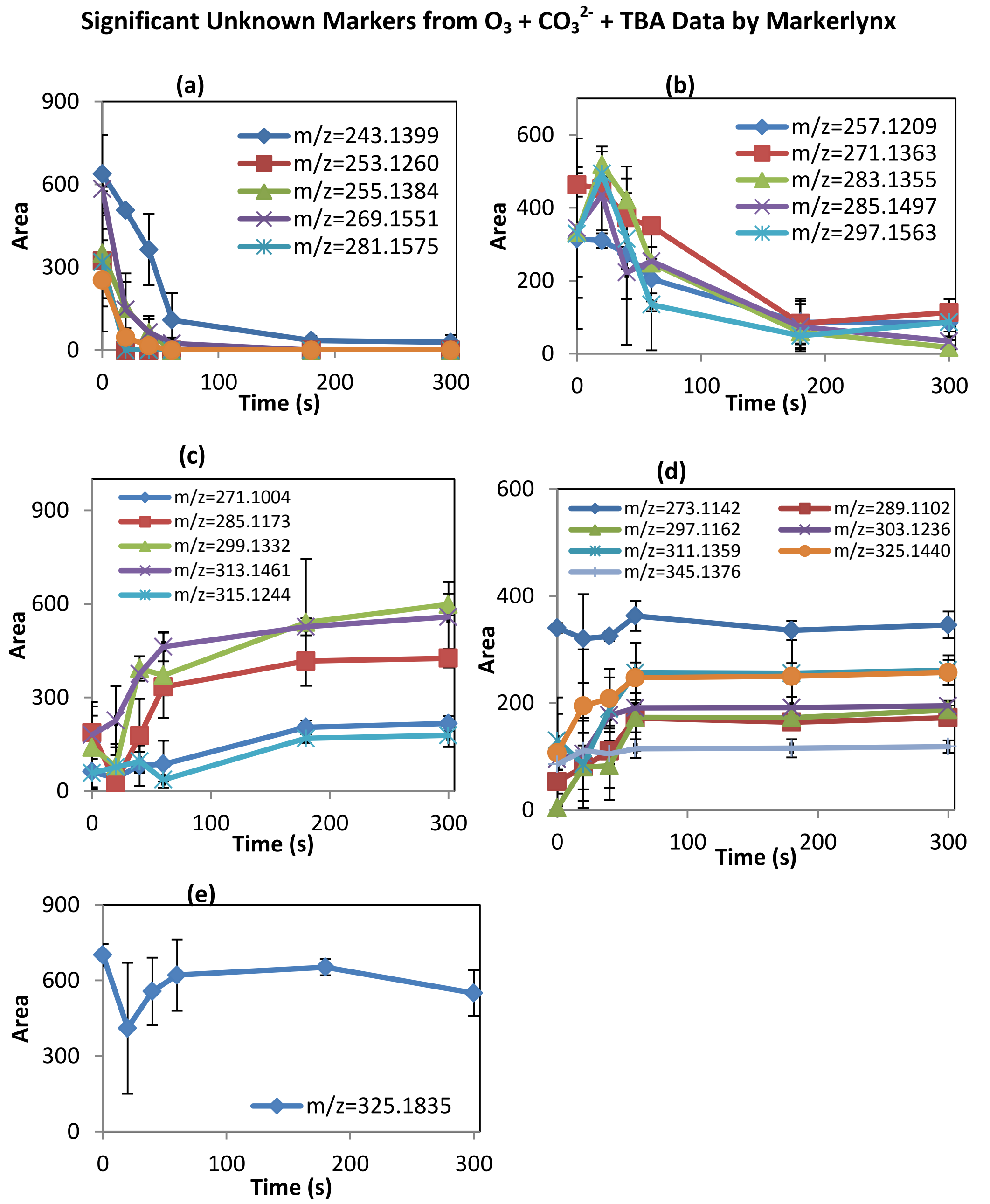


Figure B- 8: Trends of Significant Unknown Markers from PCA Results based on O_3 + CO_3^{2-} +TBA Data by Markerlynx. Plot (a) and (b) show markers with decreasing trends; plot (c) and (d) show markers with increasing trends; plot (e) shows marker with decreasing followed by increasing and decreasing trend.

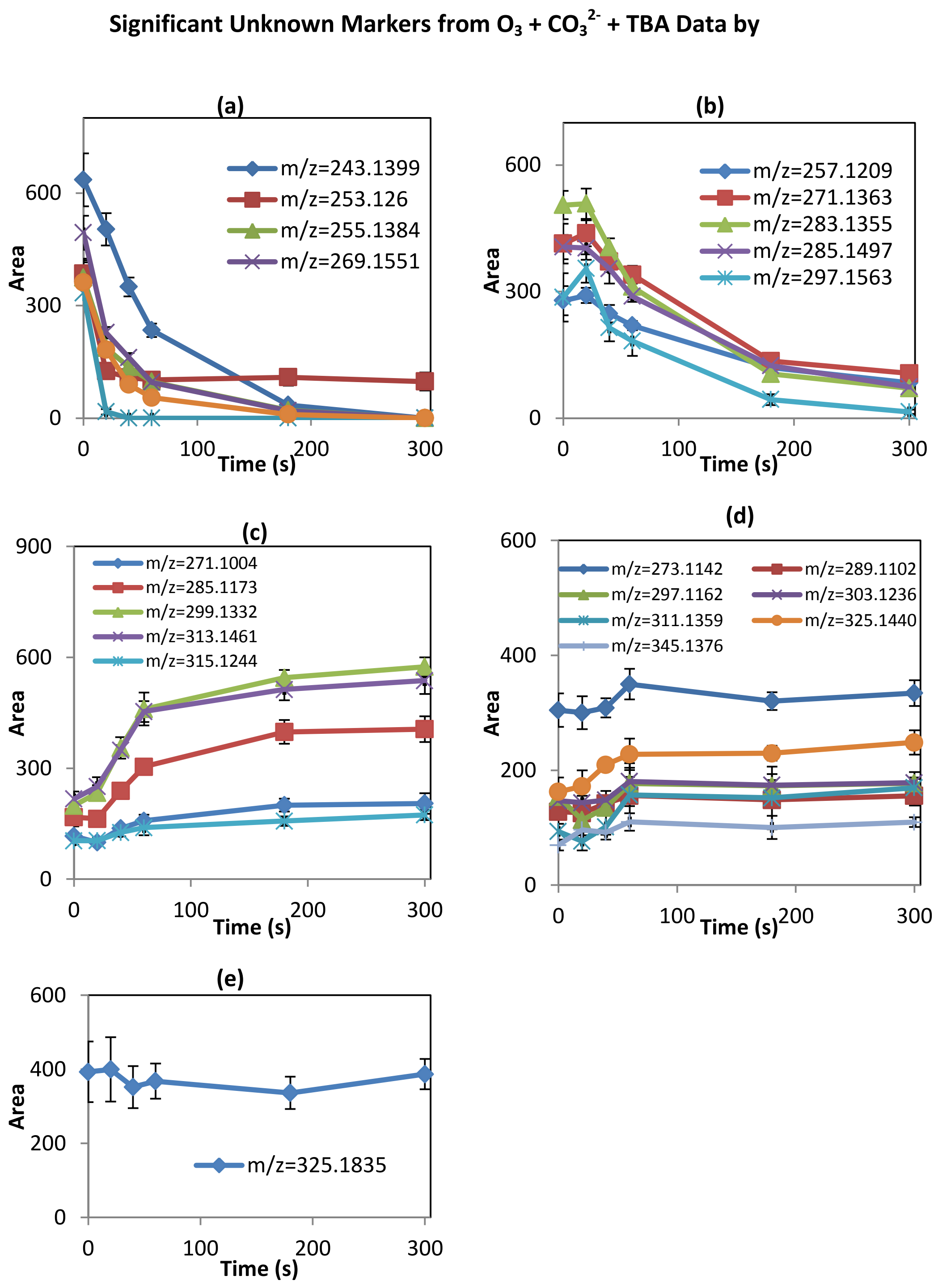


Figure B- 9: Trends of Significant Unknown Markers Selected from PCA Results based on O_3 + CO_3^{2-} + TBA Data by Targetlynx. Plot (a) to (e) show the trends of markers corresponding to the markers in Figure B-8 (a) to (e).

Table B- 3: Elemental Compositions for Significant Unknown Markers from O_3 + CO_3^{2-} + TBA Datasets.

Detected Mass	Calculated Mass	Error PPM	DBE	Formula	Detected Mass	Calculated Mass	Error PPM	DBE	Formula	Detected Mass	Calculated Mass	Error PPM	DBE	Formula
243.1399	243.1385	5.8	7.5	C16 H19 O2	***285.1497***	285.1491	2.1	8.5	C18 H21 O3	***311.1359***	311.1351	2.6	0.5	C13 H27 O4 S2
	243.1419	-8.2	2.5	C13 H23 O2 S		285.1517	-7	13	C21 H19 N		311.1344	4.8	10	C19 H21 N O S
253.1260	253.1262	-0.8	4.5	C14 H21 O2 S		285.1524	-9.5	3.5	C15 H25 O3 S		311.1378	-6.1	5	C16 H25 N O S2
255.1384	255.1385	-0.4	8.5	C17 H19 O2	***289.1102***	289.1103	-0.3	13	C19 H15 N O2	***313.1461***	313.1467	-1.9	14	C22 H19 N O
257.1209	257.1211	-0.8	3.5	C13 H21 O3 S		289.1110	-2.8	3.5	C13 H21 O5 S		313.1474	-4.2	4.5	C16 H25 O4 S
	257.1204	1.9	13	C19 H15 N		289.1085	5.9	7.5	C17 H21 S2		313.1440	6.7	9.5	C19 H21 O4
269.1551	269.1542	3.3	8.5	C18 H21 O2		289.1076	9	8.5	C16 H17 O5	***315.1244***	315.1241	1	8.5	C19 H23 S2
	269.1575	-8.9	3.5	C15 H25 O2 S	***297.1162***	297.1161	0.3	5.5	C15 H21 O4 S		315.1232	3.8	9.5	C18 H19 O5
271.1004	271.1004	0	4.5	C13 H19 O4 S		297.1154	2.7	15	C21 H15 N O		315.1259	-4.8	14	C21 H17 N O2
	271.0997	2.6	14	C19 H13 N O		297.1187	-8.4	10	C18 H19 N O S		315.1266	-7	4.5	C15 H23 O5 S
271.1363	271.1361	0.7	13	C20 H17 N	***297.1528***	297.1524	1.3	4.5	C16 H25 O3 S	***325.1440***	325.1440	0	10.5	C20 H21 O4
	271.1368	-1.8	3.5	C14 H23 O3 S		297.1517	3.7	14	C22 H19 N		325.1467	-8.3	15	C23 H19 N O
273.1142	273.1154	-4.4	13	C19 H15 N O		297.1551	-7.7	9	C19 H23 N S		325.1408	9.8	5.5	C16 H25 N2 O S2
	273.1127	5.5	8.5	C16 H17 O4	***297.1563***	297.1551	4	9	C19 H23 N S	***325.1835***	325.1837	-0.6	4.5	C18 H29 O3 S
	273.1161	-7	3.5	C13 H21 O4 S		297.1576	-4.4	5	C15 H23 N O5		325.1830	1.5	14	C24 H23 N
281.1575	281.1575	0	4.5	C16 H25 O2 S		297.1585	-7.4	4	C16 H27 N S2		325.1864	-8.9	9	C21 H27 N S
283.1355	283.1361	-2.1	14	C21 H17 N	***299.1332***	299.1344	-4	9	C18 H21 N O S		325.1804	9.5	9.5	C21 H25 O3
	283.1368	-4.6	4.5	C15 H23 O3 S		299.1317	5	4.5	C15 H23 O4 S	***345.1376***	345.1365	3.2	14	C22 H19 N O3
	283.1334	7.4	9.5	C18 H19 O3		299.1310	7.4	14	C21 H17 N O		345.1392	-4.6	18.5	C25 H17 N2
	283.1328	9.5	0.5	C10 H23 N2 O5 S	***303.1236***	303.1232	1.3	8.5	C17 H19 O5		345.1399	-6.7	9	C19 H23 N O3 S
285.1173	285.1161	4.2	4.5	C14 H21 O4 S		303.1241	-1.6	7.5	C18 H23 S2		345.1347	8.4	8.5	C20 H25 O S2
	285.1187	-4.9	9	C17 H19 N O S		303.1259	-7.6	13	C20 H17 N O2	***283.1708***	283.1698	3.5	8.5	C19 H23 O2
	285.1154	6.7	14	C20 H15 N O		303.1266	-9.9	3.5	C14 H23 O5 S		283.1732	-8.5	3.5	C16 H27 O2 S

B-4: O_3 + TNM Datasets Analysis

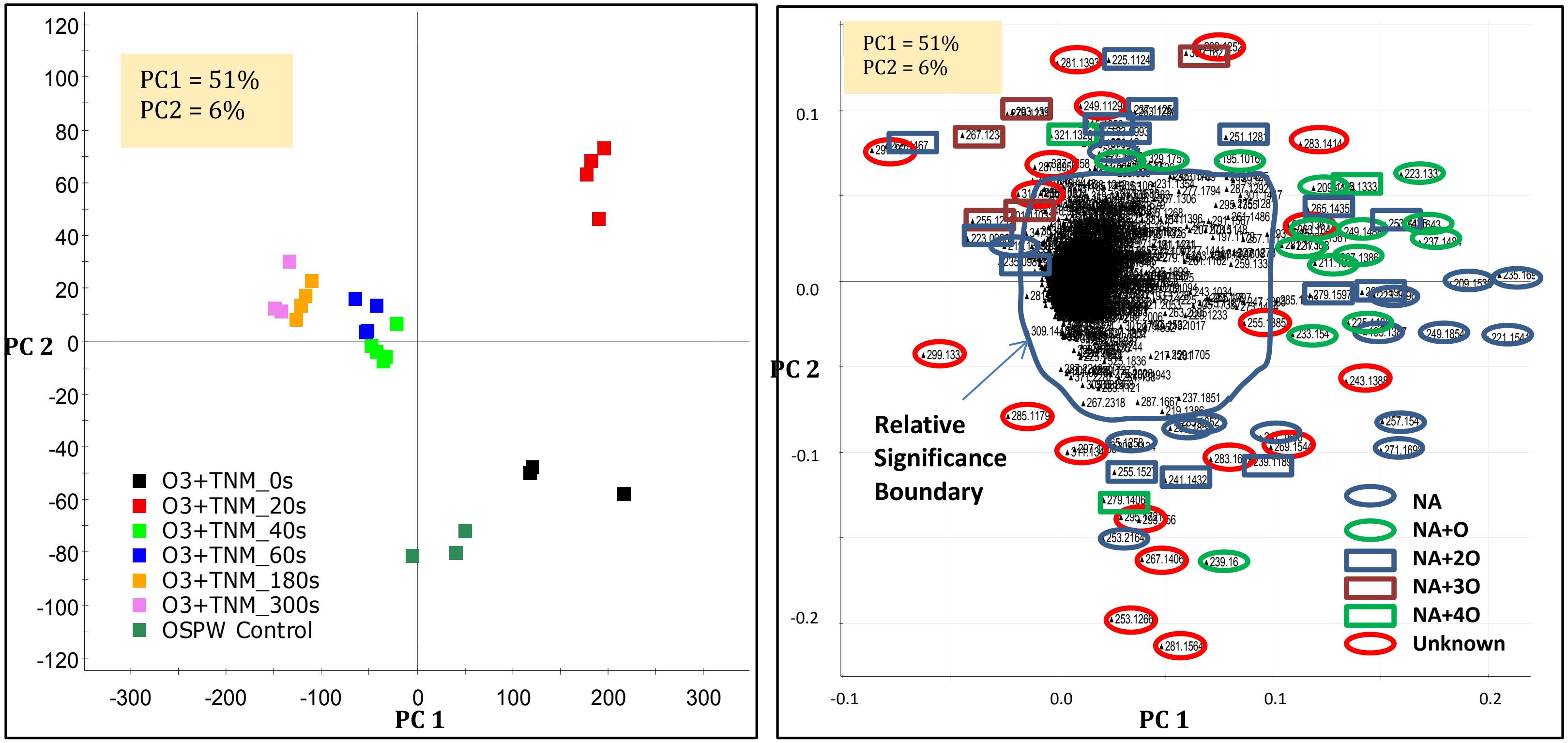


Figure B- 10: PCA (a) Score Plot; and (b) Loading Plot, based on O_3 + TNM Datasets.

Significant Unknown Markers from O_3 + TNM Data by Markerlynx

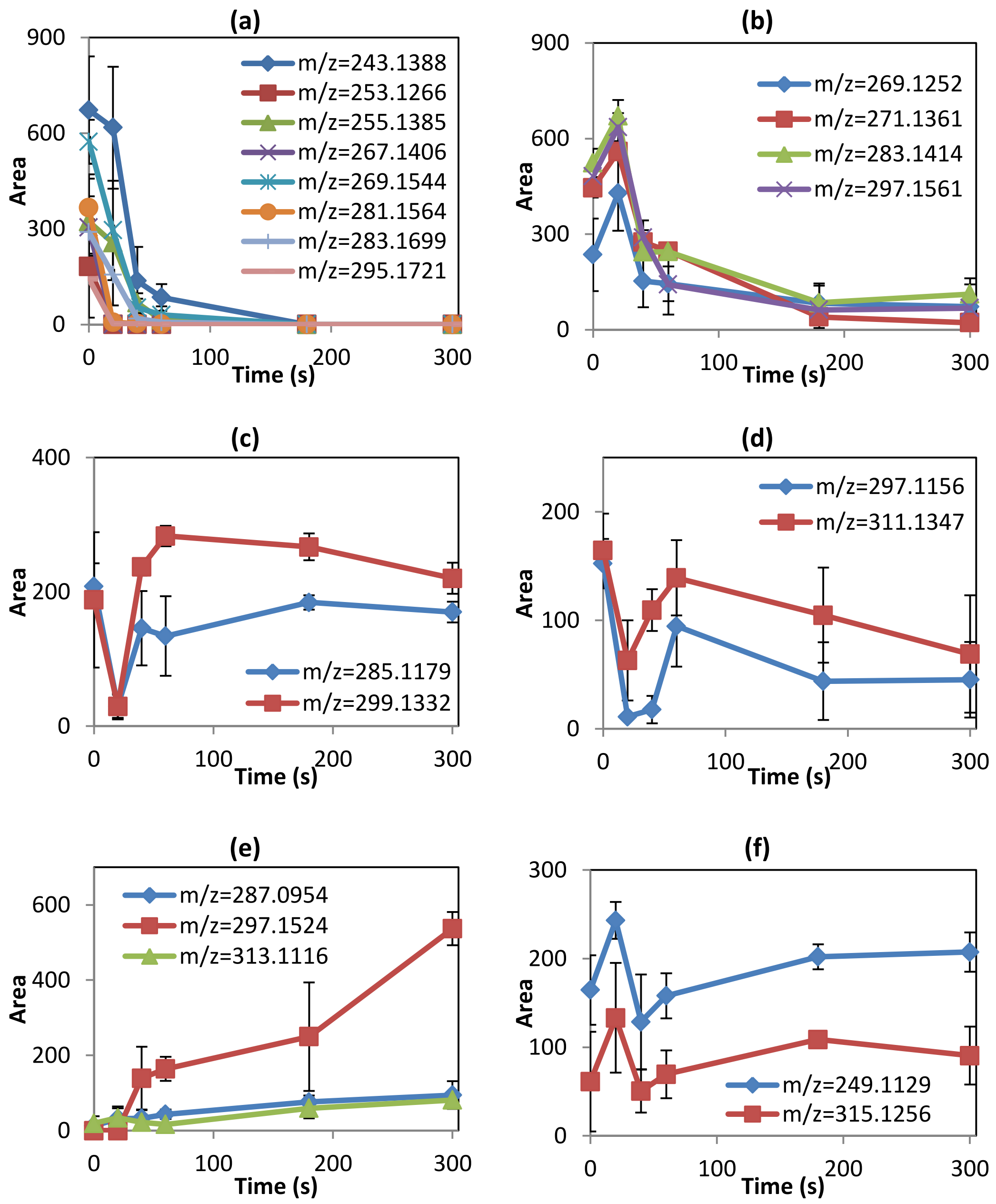


Figure B- 11: Trends of Significant Unknown Markers Selected from PCA Results based on O_3 + TNM Datasets by Markerlynx. Plot (a) and (b) show markers with decreasing trends; plot (c) and (d) show markers with decreasing followed by increasing and decreasing trends; plot (e) shows markers with increasing trends; plot (f) shows markers with increasing followed by decreasing and increasing trends.

Significant Unknown Markers from O_3 + TNM Data by Targetlynx

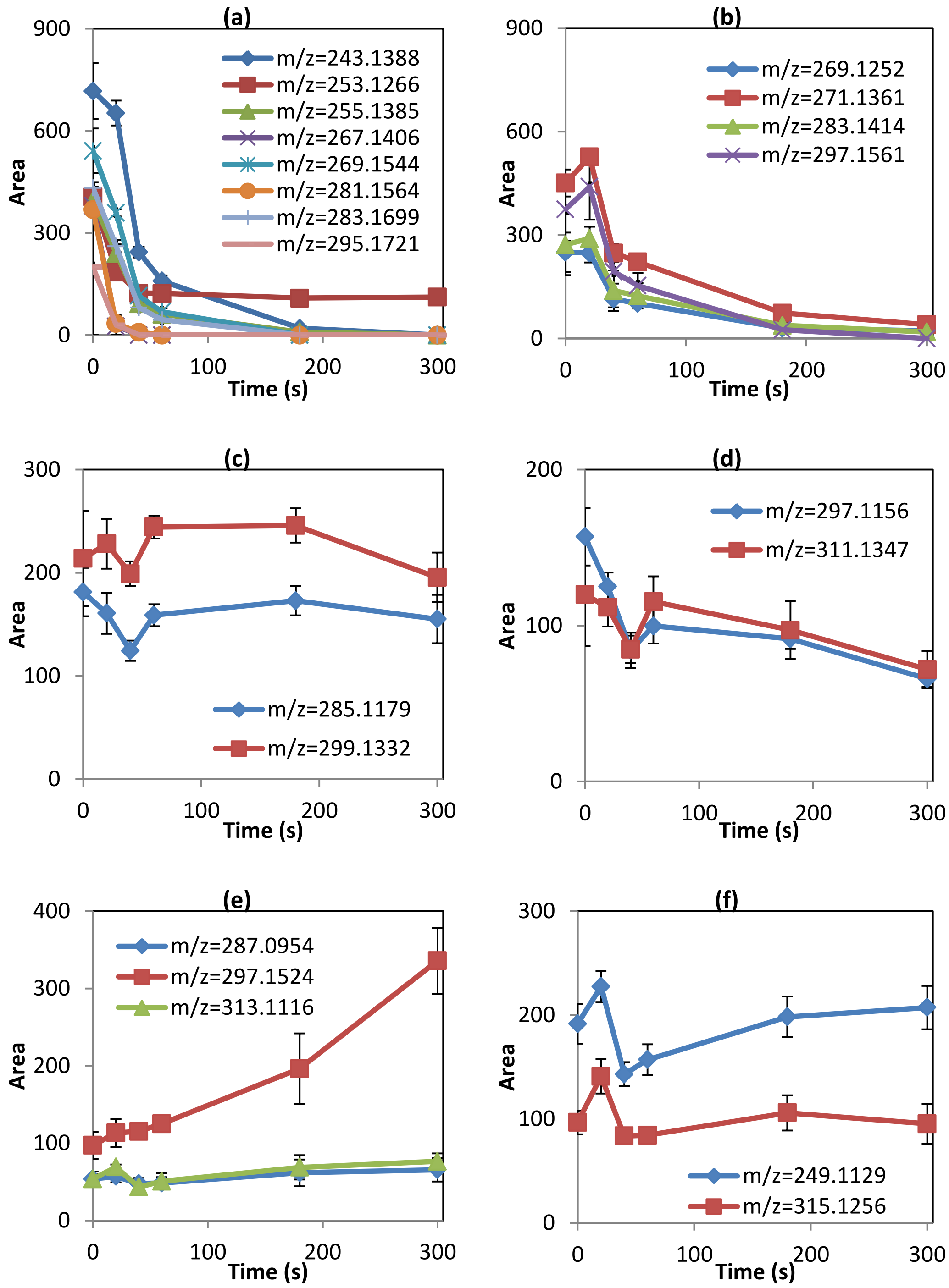


Figure B- 12: Trends of Significant Unknown Markers Selected from PCA Results based on O_3 + TNM Data by Targetlynx. Plot (a) to (f) show the trends of markers corresponding to the markers in Figure B-11 (a) to (f).

Table B- 4: Elemental Compositions for Significant Unknown Markers from O_3 + TNM Datasets.

Detected Mass	Calculated Mass	Error PPM	DBE	Formula	Detected Mass	Calculated Mass	Error PPM	DBE	Formula	Detected Mass	Calculated Mass	Error PPM	DBE	Formula
243.1388	243.1385	1.2	7.5	C16 H19 O2	***283.1699***	283.1698	0.4	8.5	C19 H23 O2	***297.1561***	297.1551	3.4	9	C19 H23 N S
249.1129	249.1127	0.8	6.5	C14 H17 O4	***285.1179***	285.1187	-2.8	9	C17 H19 N O S		297.1576	-5	5	C15 H23 N O5
253.1266	253.1262	1.6	4.5	C14 H21 O2 S		285.1161	6.3	4.5	C14 H21 O4 S		297.1585	-8.1	4	C16 H27 N S2
255.1385	255.1385	0	8.5	C17 H19 O2		285.1154	8.8	14	C20 H15 N O	***299.1332***	299.1344	-4	9	C18 H21 N O S
267.1406	267.1419	-4.9	4.5	C15 H23 O2 S	***287.0954***	287.0953	0.3	4.5	C13 H19 O5 S		299.1317	5	4.5	C15 H23 O4 S
	267.1385	7.9	9.5	C18 H19 O2		287.0946	2.8	14	C19 H13 N O2		299.1310	7.4	14	C21 H17 N O
269.1252	269.1263	-4.1	5	C13 H19 N O5		287.0928	9.1	8.5	C17 H19 S2	***311.1347***	311.1344	1	10	C19 H21 N O S
	269.1238	5.2	9	C17 H19 N S		287.0980	-9.1	9	C16 H17 N O2 S		311.1351	-1.3	0.5	C13 H27 O4 S2
	269.1272	-7.4	4	C14 H23 N S2	***295.1721***	295.1732	-3.7	4.5	C17 H27 O2 S		311.1317	9.6	5.5	C16 H23 O4 S
269.1544	269.1542	0.7	8.5	C18 H21 O2		295.1698	7.8	9.5	C20 H23 O2	***313.1116***	313.1110	1.9	5.5	C15 H21 O5 S
271.1361	271.1361	0	13	C20 H17 N		295.1692	9.8	0.5	C12 H27 N2 O4 S		313.1103	4.2	15	C21 H15 N O2
	271.1368	-2.6	3.5	C14 H23 O3 S	***297.1156***	297.1154	0.7	15	C21 H15 N O		313.1136	-6.4	10	C18 H19 N O2 S
281.1564	281.1575	-3.9	4.5	C16 H25 O2 S		297.1161	-1.7	5.5	C15 H21 O4 S		313.1143	-8.6	0.5	C12 H25 O5 S2
	281.1542	7.8	9.5	C19 H21 O2		297.1127	9.8	10.5	C18 H17 O4		313.1085	9.9	9.5	C19 H21 S2
283.1414	283.1420	-2.1	5	C14 H21 N O5	***297.1524***	297.1524	0	4.5	C16 H25 O3 S	***315.1256***	315.1259	-1	14	C21 H17 N O2
	283.1428	-4.9	4	C15 H25 N S2		297.1517	2.4	14	C22 H19 N		315.1266	-3.2	4.5	C15 H23 O5 S
	283.1395	6.7	9	C18 H21 N S		297.1551	-9.1	9	C19 H23 N S		315.1241	4.8	8.5	C19 H23 S2
											315.1232	7.6	9.5	C18 H19 O5

B-5: O_3 + Fe (II) Datasets Analysis

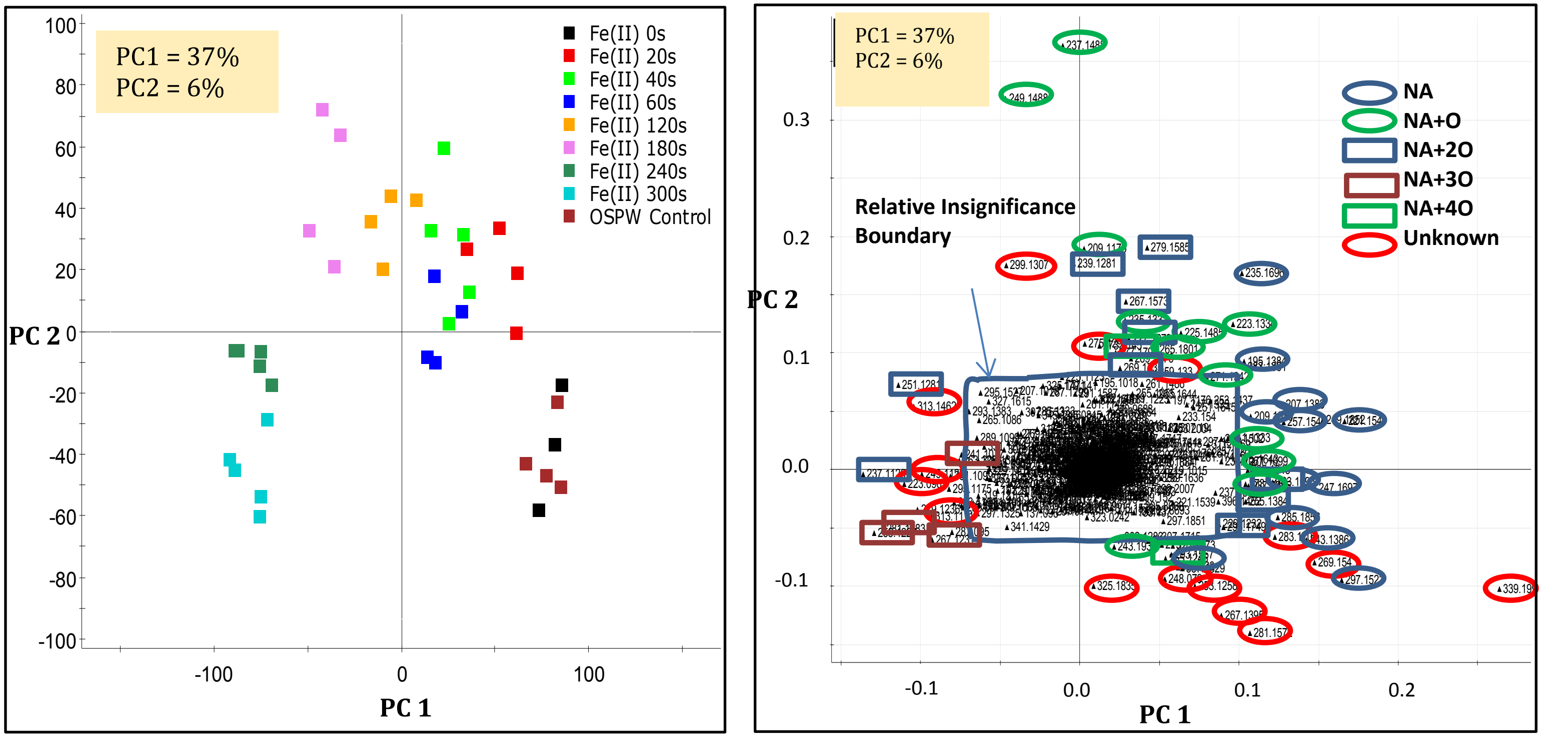


Figure B- 13: PCA (a) Score Plot; and (b) Loading Plot, based on O_3 + Fe (II) Datasets.

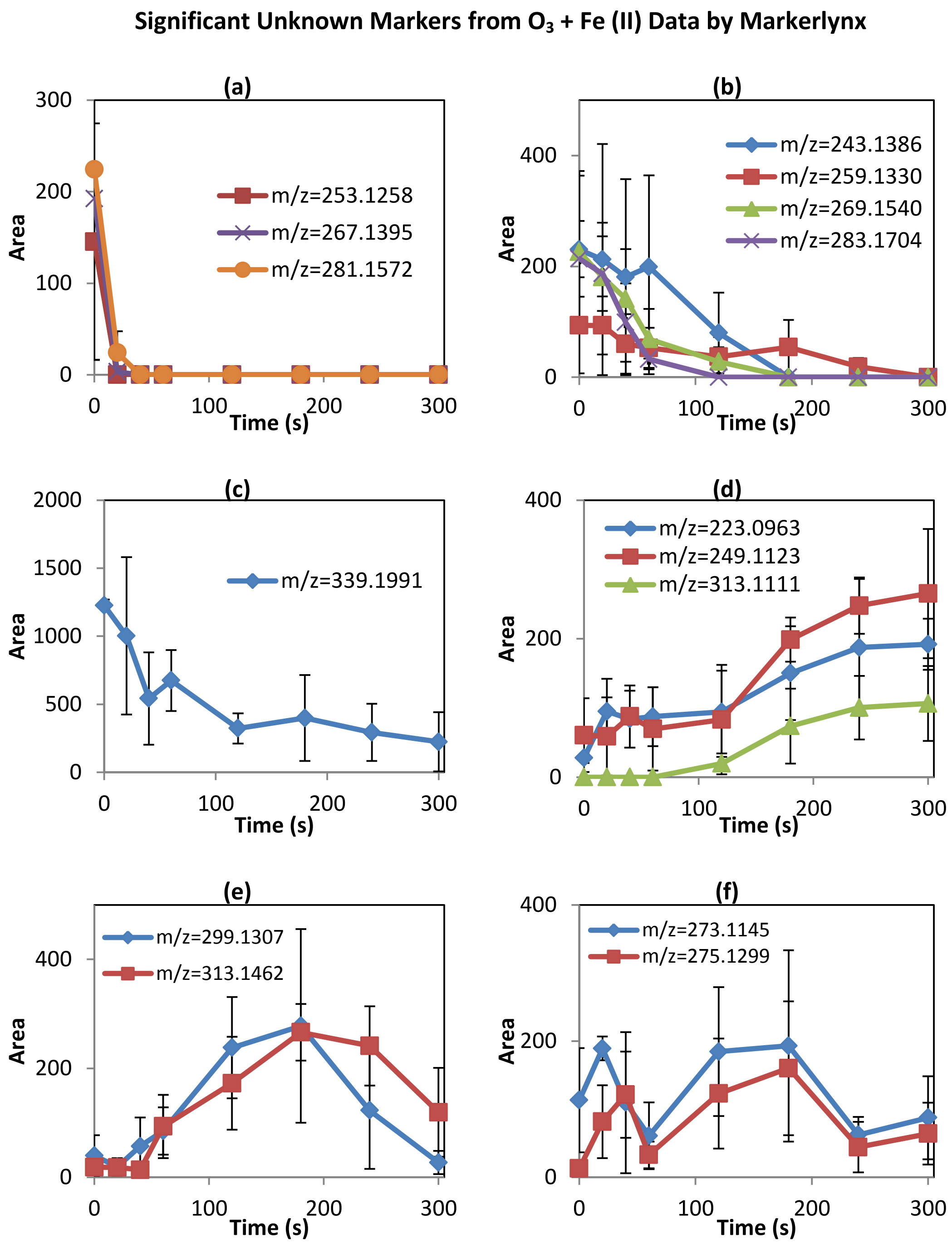


Figure B- 14: Trends of Significant Unknown Markers Selected from PCA Results based on O_3 + Fe (II) Datasets by Markerlynx. Plot (a) and (c) show markers with decreasing trends; plot (d) show markers with increasing trends; plot (e) shows markers with increasing followed by decreasing trends; plot (f) shows markers with variable trends.

Significant Unknown Markers from O_3 + Fe (II) Data by Targetlynx

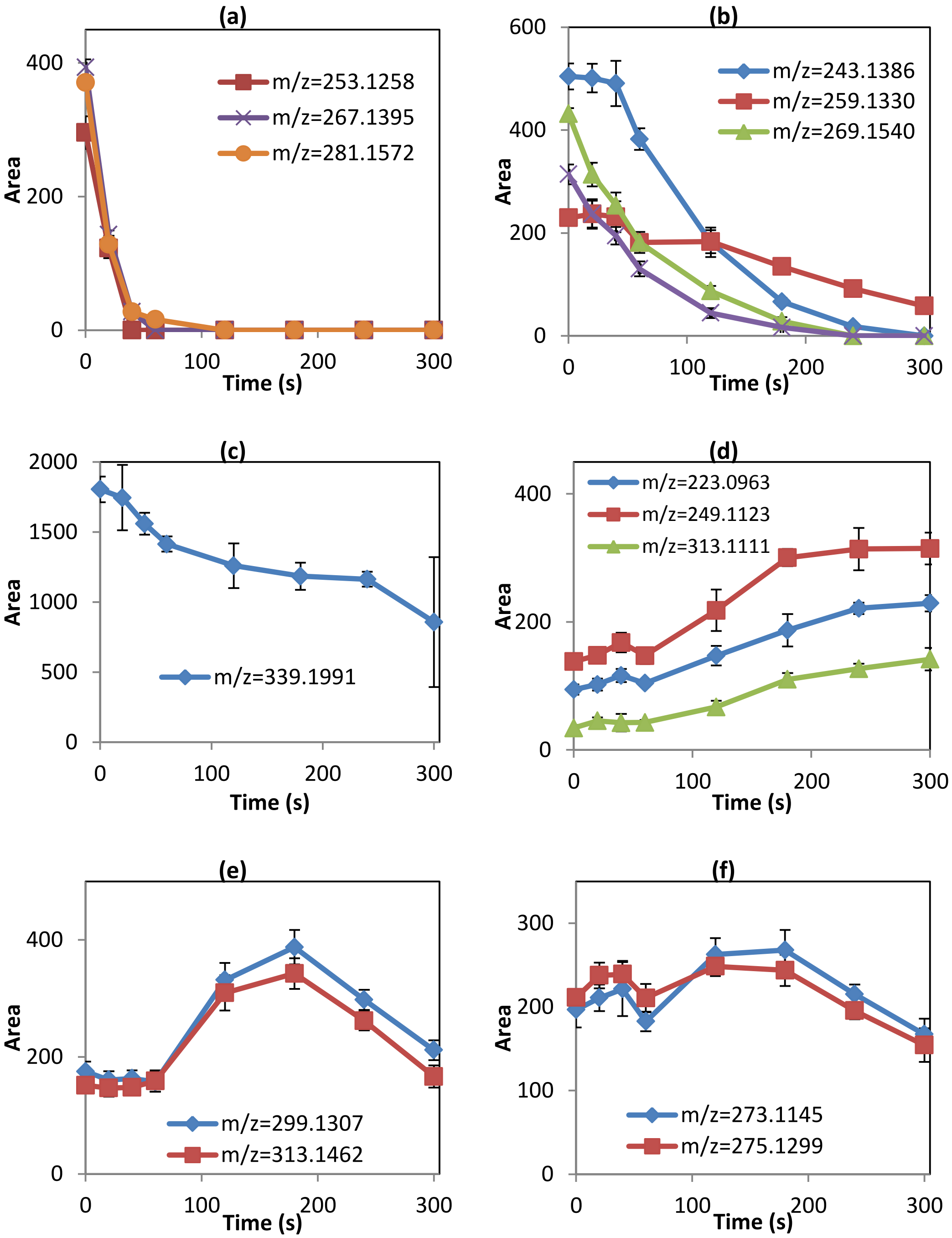


Figure B- 15: Trends of Significant Unknown Markers Selected from PCA Results based on O_3 + Fe (II) Datasets by Targetlynx. Plot (a) to (f) show the trends of markers corresponding to the markers in Figure B-16 (a) to (f).

Table B- 5: Elemental Compositions for Significant Unknown Markers from O$_3$ + Fe (II) Datasets.

Detected Mass	Calculated Mass	Error PPM	DBE	Formula
223.0963	223.0970	-3.1	5.5	C12 H15 O4
243.1386	243.1385	0.4	7.5	C16 H19 O2
248.0790	248.0797	-2.8	8	C12 H12 N2 O4
	248.0779	4.4	2.5	C10 H18 N O2 S2
249.1123	249.1127	-1.6	6.5	C14 H17 O4
253.1258	253.1262	-1.6	4.5	C14 H21 O2 S
259.1330	259.1334	-1.5	7.5	C16 H19 O3
267.1395	267.1385	3.7	9.5	C18 H19 O2
	267.1379	6	0.5	C10 H23 N2 O4 S
	267.1419	-9	4.5	C15 H23 O2 S
269.1540	269.1542	-0.7	8.5	C18 H21 O2
273.1145	273.1154	-3.3	13	C19 H15 N O
	273.1161	-5.9	3.5	C13 H21 O4 S
	273.1127	6.6	8.5	C16 H17 O4
275.1299	275.1310	-4	12	C19 H17 N O
	275.1283	5.8	7.5	C16 H19 O4
	275.1317	-6.5	2.5	C13 H23 O4 S
281.1572	281.1575	-1.1	4.5	C16 H25 O2 S
283.1704	283.1698	2.1	8.5	C19 H23 O2
	283.1732	-9.9	3.5	C16 H27 O2 S
297.1523	297.1524	-0.3	4.5	C16 H25 O3 S
	297.1517	2	14	C22 H19 N
	297.1551	-9.4	9	C19 H23 N S
299.1307	299.1310	-1	14	C21 H17 N O
	299.1317	-3.3	4.5	C15 H23 O4 S
	299.1283	8	9.5	C18 H19 O4
311.1678	311.1681	-1	4.5	C17 H27 O3 S
	311.1674	1.3	14	C23 H21 N
	311.1708	-9.6	9	C20 H25 N S
313.1111	313.1110	0.3	5.5	C15 H21 O5 S
	313.1103	2.6	15	C21 H15 N O2
	313.1136	-8	10	C18 H19 N O2 S
	313.1085	8.3	9.5	C19 H21 S2
313.1462	313.1467	-1.6	14	C22 H19 N O
	313.1474	-3.8	4.5	C16 H25 O4 S
	313.1440	7	9.5	C19 H21 O4
325.1834	325.1837	-0.9	4.5	C18 H29 O3 S
	325.1830	1.2	14	C24 H23 N
	325.1804	9.2	9.5	C21 H25 O3
	325.1864	-9.2	9	C21 H27 N S
339.1991	339.1994	-0.9	4.5	C19 H31 O3 S
	339.1987	1.2	14	C25 H25 N
	339.2021	-8.8	9	C22 H29 N S
	339.1960	9.1	9.5	C22 H27 O3

B-6: Specific and Common Markers from Ozonation in Different Conditions Datasets

Table B- 6: Specific Significant Unknown Markers Only Presented in One Individual Ozonation Condition

Ozonation Condition	Specific unknown markers in individual Ozonation (*m/z*)						
O_3	*209.09*	*239.11*	*281.10*	*311.17*			
O_3+CO_3^{2-}	*301.14*	*327.16*	*291.13*	*331.12*	*333.14*	*235.10*	*309.14*
O_3+TBA	*295.14*	*323.17*	*325.18*	*259.10*			
O_3+TBA+CO_3^{2-}	*257.12*	*289.10*	*303.12*	*273.14*	*303.12*	*345.14*	
O_3+TNM	*295.17*	*287.10*	*297.15*				
O_3+Fe (II)	*339.20*	*223.09*					

Table B- 7: Common and Specific By-products from All Ozonation in Different Conditions

Marker (*m/z*)	O_3	O3+CO_3^{2-}	O3+TBA	O3+TBA+CO_3^{2-}	O3+TNM	O3+Fe(II)
209.09	✔					
223.09						✔
235.09	Δ	✔				
239.11	✔					
249.11	Δ	✔	✔		Δ	✔
259.10			✔			
265.11	✔	✔				
271.10			✔	✔		
273.14				✔		
281.10	✔					
285.12			✔	✔	Δ	
287.10					✔	
289.10				✔		
297.12				✔	Δ	
297.15					✔	
299.13	Δ	Δ	✔	✔		Δ
303.12				✔		
311.13			✔	Δ	Δ	
311.17	✔					
313.11					✔	✔
313.15	Δ	Δ	✔	✔		Δ
315.12				✔	Δ	
325.14		Δ		✔		
345.14				✔		

Note: ✔= markers defined as by-products in corresponding ozonation condition; Δ = markers (not by-products) showing various behaviours in different ozonation conditions; blank = not detected. Markers highlighted in red are common markers.

B-7: Trends of Common Markers in Ozonation in Different Conditions by Targetlynx

Common Markers from Ozonation in Different Conditions Data by Targetlynx (Part 1)

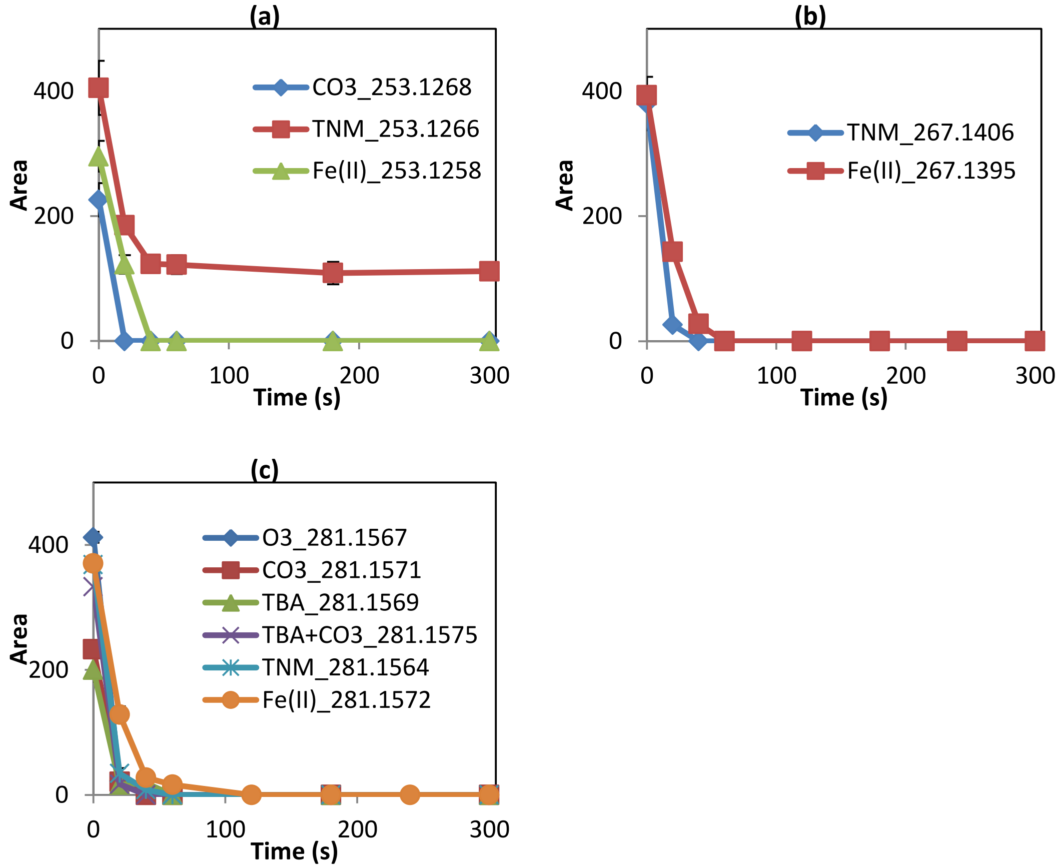


Figure B- 16: Common Markers from Ozonation in Different Conditions Datasets by Targetlynx (Part 1). Plot (a) to (c) showed trends of markers corresponding to the markers shown in Figure 4-23 (a) to (c).

Common Markers from Ozonation in Different Conditions Data by Targetlynx (Part 2)

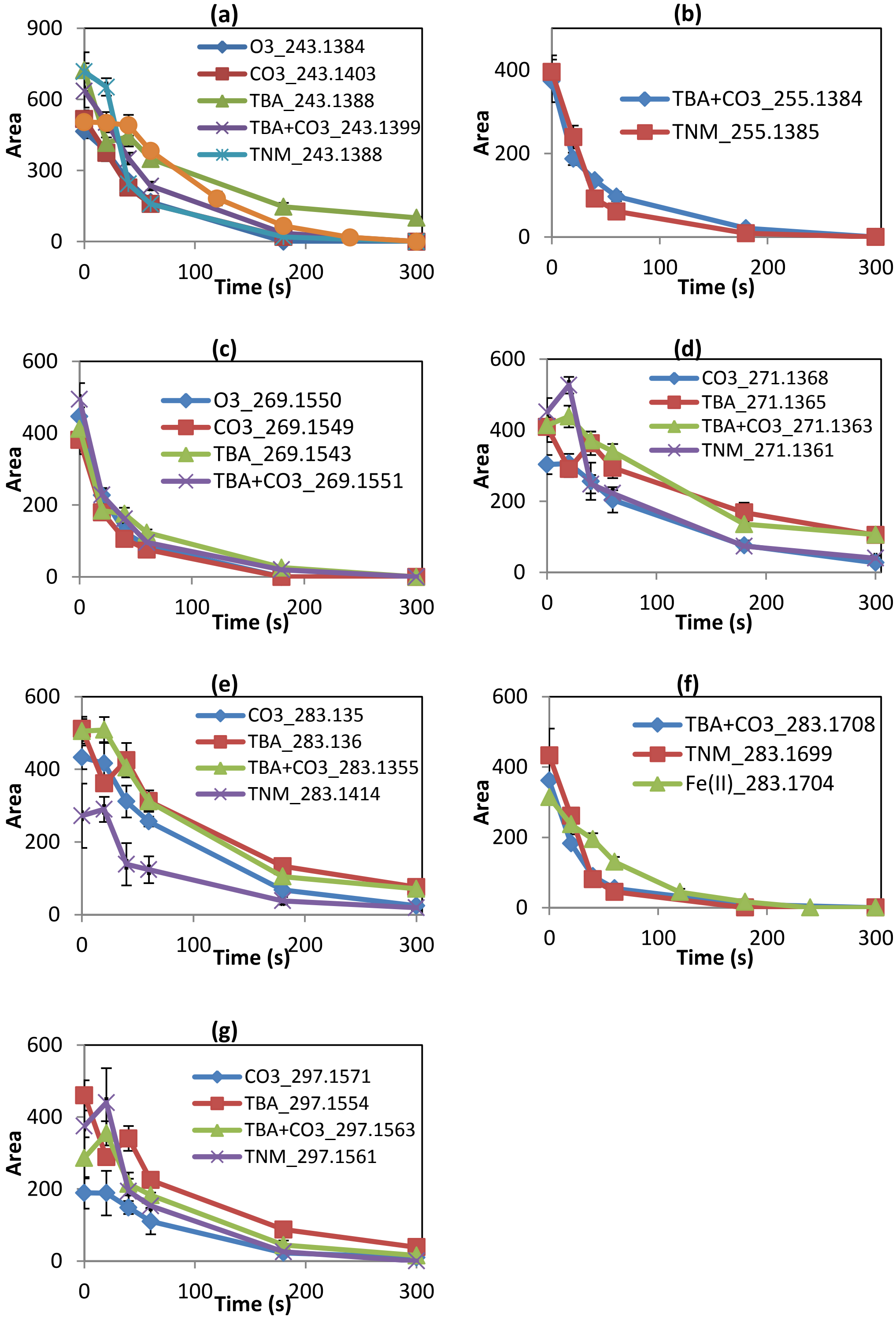
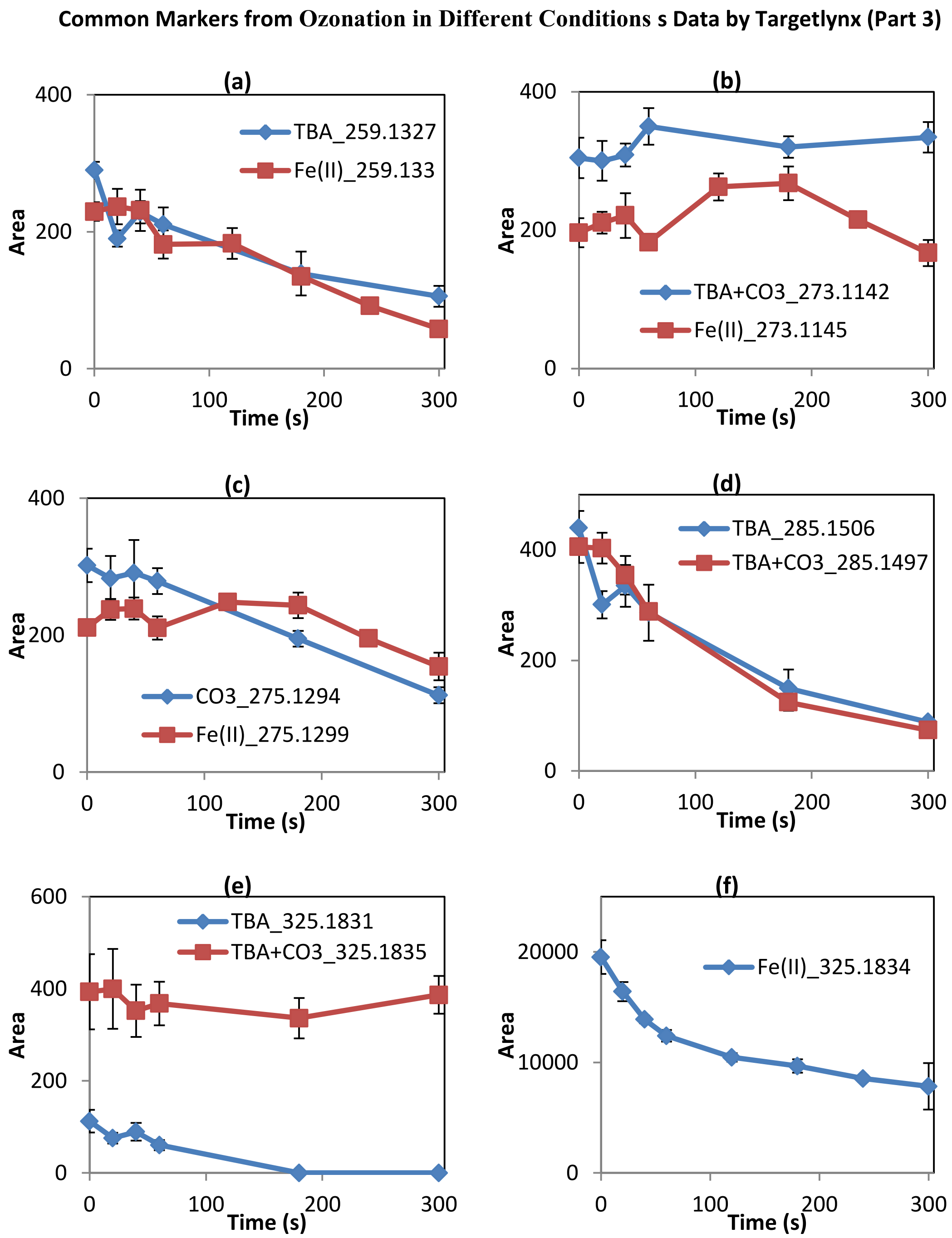


Figure B- 17: Common Markers from Ozonation in Different Conditions Datasets by Targetlynx (Part 2). Plot (a) to (g) showed trends of markers corresponding to the markers shown in Figure 4-24 (a) to (g).



Common Markers from Ozonation in Different Conditions s Data by Targetlynx (Part 3)

Figure B- 18: Common Markers from Ozonation in Different Conditions Datasets by Targetlynx (Part 3). Plot (a) to (f) showed trends of markers corresponding to the markers shown in Figure 4-25 (a) to (f).

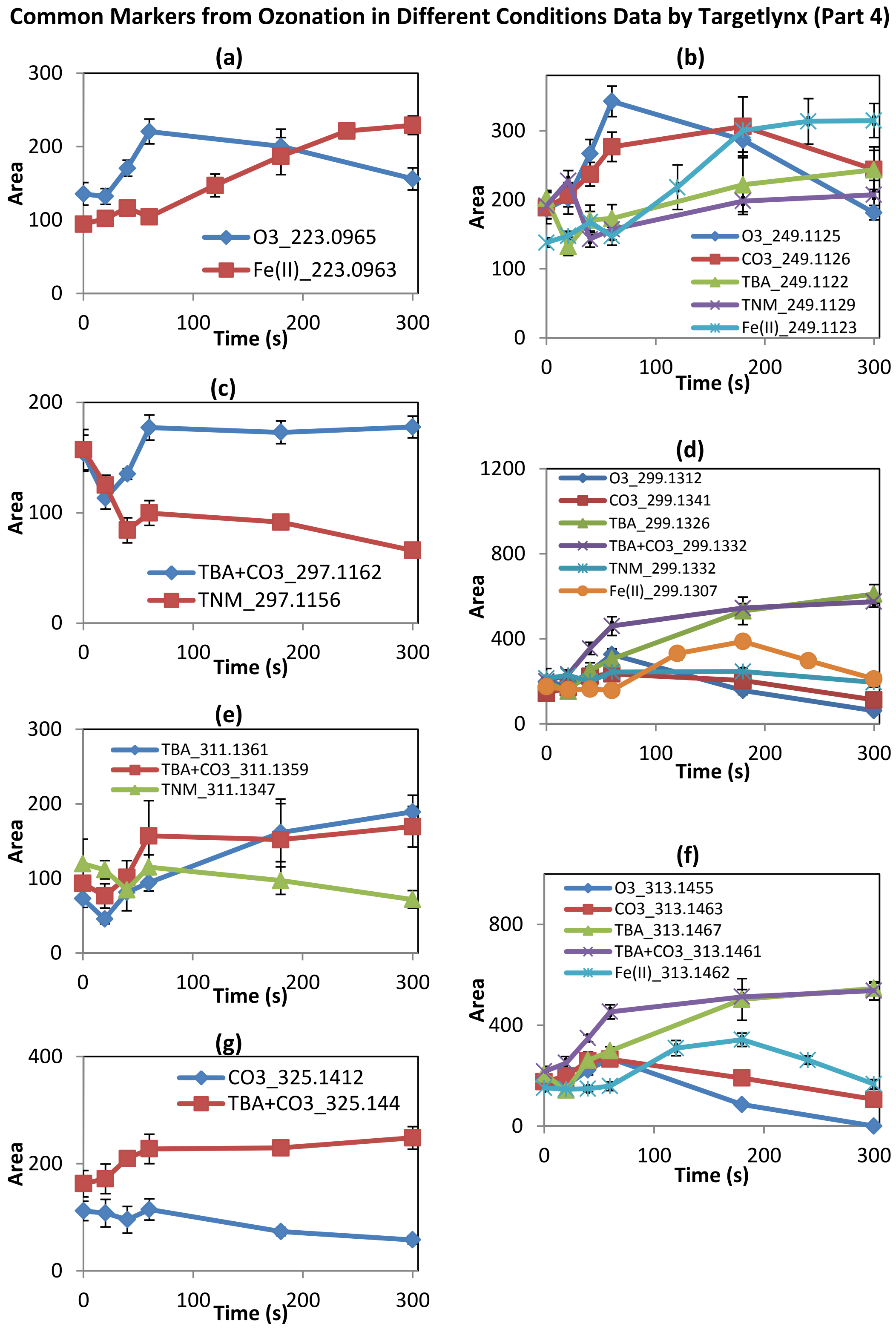


Figure B- 19: Common Markers from Ozonation in Different Conditions Datasets by Targetlynx (Part 4). Plot (a) to (g) showed trends of markers corresponding to the markers shown in Figure 4-26 (a) to (g).

Common Markers from Ozonation in Different Conditions Data by Targetlynx (Part 5)

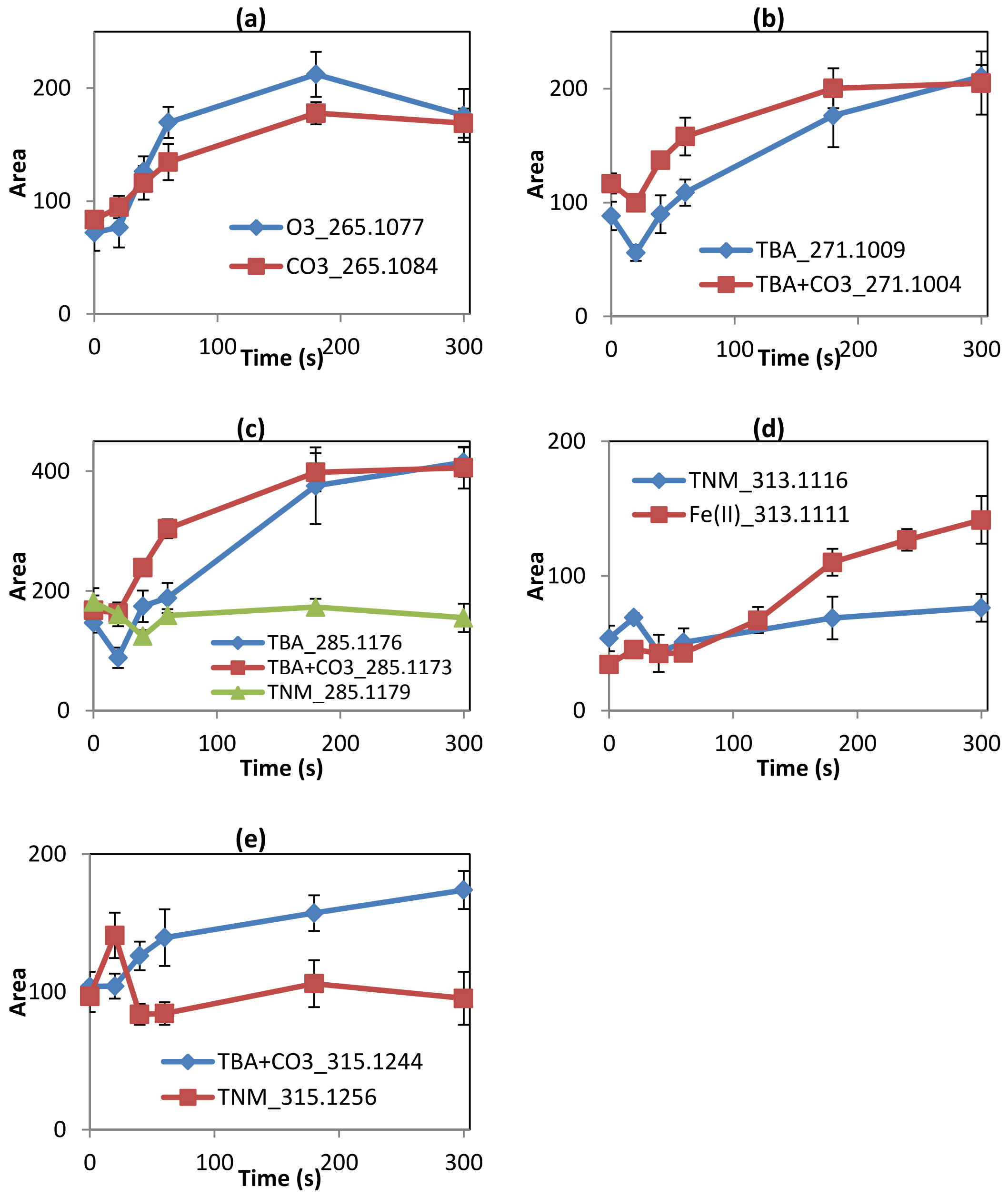


Figure B- 20: Common Markers from Ozonation in Different Conditions Datasets by Targetlynx (Part 5). Plots (a) to (e) showed trends of markers corresponding to the markers shown in Figure 4-27 (a) to (e).

Appendix C

Supplementary Information for Biological Treatment Processes

C-1: Raw OSPW Biodegradation

Table C- 1: Elemental Compositions for Significant Unknown Markers from Raw OSPW Biodegradation Datasets

Detected Mass	Calculated Mass	Error PPM	DBE	Formula	Detected Mass	Calculated Mass	Error PPM	DBE	Formula	Detected Mass	Calculated Mass	Error PPM	DBE	Formula
235.0997	235.0997	0	11	C16 H13 N O	***287.0966***	287.0953	4.5	4.5	C13 H19 O5 S	***309.1588***	309.1585	1	5	C17 H27 N S2
	235.1004	-3	1.5	C10 H19 O4 S		287.0980	-4.9	9	C16 H17 N O2 S		309.1576	3.9	6	C16 H23 N O5
241.1255	241.1262	-2.9	3.5	C13 H21 O2 S		287.0946	7	14	C19 H13 N O2		309.1603	-4.9	10.5	C19 H21 N2 O2
253.1276	253.1262	5.5	4.5	C14 H21 O2 S	***295.1371***	295.1368	1	5.5	C16 H23 O3 S		309.1610	-7.1	1	C13 H27 N O5 S
255.1403	255.1419	-6.3	3.5	C14 H23 O2 S		295.1361	3.4	15	C22 H17 N		309.1558	9.7	0.5	C14 H29 O3 S2
	255.1385	7.1	8.5	C17 H19 O2		295.1395	-8.1	10	C19 H21 N S	***325.1822***	325.1830	-2.5	14	C24 H23 N
269.1559	269.1575	-5.9	3.5	C15 H25 O2 S	***295.1732***	295.1732	0	4.5	C17 H27 O2 S		325.1837	-4.6	4.5	C18 H29 O3 S
	269.1542	6.3	8.5	C18 H21 O2	***299.1149***	299.1158	-3	10	C17 H17 N O4		325.1804	5.5	9.5	C21 H25 O3
281.1578	281.1575	1.1	4.5	C16 H25 O2 S		299.1139	3.3	4.5	C15 H23 O2 S2		325.1797	7.7	0.5	C13 H29 N2 O5 S
283.1712	283.1698	4.9	8.5	C19 H23 O2	***305.1576***	305.1575	0.3	6.5	C18 H25 O2 S	***327.1268***	327.1266	0.6	5.5	C16 H23 O5 S
	283.1732	-7.1	3.5	C16 H27 O2 S	***307.1730***	307.1732	-0.7	5.5	C18 H27 O2 S		327.1259	2.8	15	C22 H17 N O2
											327.1293	-7.6	10	C19 H21 N O2 S
											327.1241	8.3	9.5	C20 H23 S2
											327.1300	-9.8	0.5	C13 H27 O5 S2

Table C- 2: Elemental Compositions for Significant Unknown Markers from Ozonated OSPW Biodegradation Datasets

Detected Mass	Calculated Mass	Error PPM	DBE	Formula
297.1543	297.1551	-2.7	9	C19 H23 N S
	297.1524	6.4	4.5	C16 H25 O3 S
	297.1517	8.7	14	C22 H19 N
311.1681	311.1681	0	4.5	C17 H27 O3 S
	311.1674	2.2	14	C23 H21 N
	311.1708	-8.7	9	C20 H25 N S
301.1465	301.1467	-0.7	13	C21 H19 N O
	301.1474	-3	3.5	C15 H25 O4 S
	301.1440	8.3	8.5	C18 H21 O4
309.1525	309.1524	0.3	5.5	C17 H25 O3 S
	309.1517	2.6	15	C23 H19 N
	309.1551	-8.4	10	C20 H23 N S
303.1231	303.1232	-0.3	8.5	C17 H19 O5
	303.1241	-3.3	7.5	C18 H23 S2
	303.1207	7.9	12.5	C21 H19 S
	303.1259	-9.2	13	C20 H17 N O2
	303.1201	9.9	3.5	C13 H23 N2 O2 S2