

# Iowa Research Online

---

## Two-dimensional gas chromatography time-of-flight mass spectrometry (gcxgc-tof-ms) data analysis for identifying unknown persistent organic pollutants in milk samples

Flores, Allison

<https://iro.uiowa.edu/esploro/outputs/graduate/Two-dimensional-gas-chromatography-time-of-flight-mass-spectrometry/9984284950902771/filesAndLinks?index=0>

---

Flores, A. (2022). Two-dimensional gas chromatography time-of-flight mass spectrometry (gcxgc-tof-ms) data analysis for identifying unknown persistent organic pollutants in milk samples [University of Iowa]. <https://doi.org/10.25820/etd.006675>

---

<https://iro.uiowa.edu>  
Free to read and download  
Copyright 2022 Allison Flores  
Downloaded on 2024/11/05 16:52:14 -0600

---

Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GCxGC-ToF-MS)  
Data Analysis for Identifying Unknown Persistent Organic Pollutants in Milk Samples

by

Allison Flores

A thesis submitted in partial fulfillment  
of the requirements for the Master of Science  
degree in Electrical and Computer Engineering in the  
Graduate College of  
The University of Iowa

August 2022

Thesis Committee:

Ananya Sen Gupta, Thesis Supervisor  
Michael Wichman  
Anton Kruger

Copyright by

Allison Flores

2022

All Rights Reserved

## ACKNOWLEDGMENTS

I would firstly like to thank Dr. Ananya Sen Gupta for being my professor and thesis supervisor. She has been supportive and encouraging throughout this journey. I would also like to extend my thanks to Jeffrey Archer, the collaborator at the FDA for all the guidance, support, and outstanding feedback. I am also thankful for all the processing and information Daniel KamghemKom was able to provide for the target compounds. He was able to identify the target compounds in each of the samples utilizing relative retention times defined in section 4.1.1. I would also like to thank Dr. David Miles for his inspirational and motivating words. Finally, I am thankful for the support that I have had from my friends and family for encouraging me to finish this project. I would also like to thank the National Science Foundation (NSF) grant number 1808463 for supporting the entirety of this thesis research.



## ABSTRACT

Persistent organic pollutants (POPs) i.e., polychlorinated biphenyls (PCBs) and polybrominated diphenyl ethers (PBDEs), are targeted in chemometric analysis. Comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometer (GCxGC-ToF-MS) data is collected to identify the POPs that are toxic and have adverse effects to human health. The Food and Drug Administration (FDA) has collected data relevant to this research over the past ten years. While thousands of compounds are extracted in the analytical hardware, only twenty-three compounds, well-known to the chemists are targeted in these samples. This leaves a knowledge gap regarding the role and prominence of the other compounds, i.e, non-target analytes, in each sample. The purpose of this research is to analyze the instrumental data for non-target compounds to determine matches between samples with similar compound distributions, discover (potential) coeluting compounds across multiple samples, and find patterns in between the samples using the statistical method of principal component analysis (PCA). To account for variances in the primary and secondary retention time attained from the analytical hardware of a compound in each sample, relative retention times were introduced. Relative retention times are

n 1 in this thesis focuses on utilizing the mass spectra comparison algorithm by focusing on a pre-selected window of a GCxGC image, where an unknown compound has been identified between the PCB-101 and PCB-123 in collaboration with the FDA. The robust mass spectra comparison algorithm had a success rate of 93.5% when identifying matches. The remaining 6.5% can be reduced by only returning the best matching algorithm and limiting the percent change in the primary relative retention time to 3.5%. The PCA correlated the 51,143 unique relative retention time pairs of a GCxGC image to 71 principal components. The scores and coefficients of the principal components identify samples that differ substantially and samples with compounds in

the same relative retention time locations. In conclusion, the mass spectral data comparison was performed as intended. The PCA did not indicate any strict patterns but identified locations to be investigated in future work

## PUBLIC ABSTRACT

Over the past decade, the Food and Drug Administration (FDA) has been investigating toxic compounds in milk samples. When analyzing milk samples using a two-dimensional gas chromatography coupled to a time-of-flight mass spectrometer (GCxGC-ToF-MS), thousands of compounds may be identified. The FDA targets up to twenty-four compounds. The purpose of this research is to be able to sift through the data and identify non-target compounds that occur across multiple samples as well as identify any patterns that may emerge between the samples' GCxGC images. To compare compounds, an algorithm was developed that uses the mass spectra data collected to identify undiscovered patterns among the detected peaks, statistical methods were employed to assess data across seventy-two samples based on their GCxGC image. Compound locations are determined relative to the target compounds, being that, over time, the compound retention time locations may shift in the GCxGC image. Overall, the comparison algorithm compared a known non-target compound and identified 77 compounds with similar mass spectra as a 'match' with a success rate of 93.5%. Only one compound per sample can be a match. The remaining 6.5% accounted for repeated 'matches' from the same sample. When exploring the patterns in the sample, seven of the files featured abundances of non-target compounds that differentiated them from the other samples. Otherwise, based on the relative locations of the target analytes, the remaining samples were found to have little variance between one another. Further research is required to further quantify the profiles of each sample to improve the clustering, however, locations that are important for further investigation were identified and can provide more insight to the FDA of what compounds may potentially need to be monitored and possibly regulated.

## TABLE OF CONTENTS

LIST OF TABLES .....	viii
LIST OF FIGURES .....	x
CHAPTER 1: INTRODUCTION.....	1
1.1. Key Contributions.....	3
CHAPTER 2: BACKGROUND MOTIVATION .....	4
2.1. Persistent Organic Pollutants and Why They Are Investigated.....	4
2.2. Monitoring Persistent Organic Pollutants in Air and Water .....	6
2.3. Investigating Persistent Organic Pollutants in Milk Samples.....	7
2.4. Targeted Analysis Data Collection to Identify Persistent Organic Pollutants in Lipophilic Samples .....	8
2.5. Summary .....	10
CHAPTER 3: RELATED WORK.....	11
3.1. Overview of Past Research Analyzing Raw Instrument Signals for Target Analysis.....	11
CHAPTER 4: TECHNICAL APPROACH .....	15
4.1. How to Compare Two Analytes Together Using Their Mass Spectra Data.....	15
4.1.1. Introducing Relative Retention Times for Proper Relations Between Samples Collected Over the Years .....	16
4.1.2. Discussion of the Mass Spectra Comparison Algorithm .....	23
4.2. Patterns in the Gas Chromatograms Utilizing Statistical Methods.....	28
4.2.1. Performing the Principal Component Analysis Against Areas in the GCxGC Image.....	28
4.2.2. Method for Clustering Scores from the Principal Component Analysis .....	30
CHAPTER 5: RESULTS.....	33
5.1. Evaluation of Mass Spectra Comparison Algorithm .....	33

5.1.1. Analytes that Returned Base Coeluted from the Mass Spectra Comparison Algorithm for Unknown 3744 .....	36
5.1.2. Analytes that Returned Coelution from the Mass Spectra Comparison Algorithm for Unknown 3744 .....	40
5.1.3. Analytes that Returned a Match from the Mass Spectra Comparison Algorithm for Unknown 3744 .....	43
5.2. Exploring GCxGC Images to Recognize Patterns Throughout the Samples.....	52
5.2.1. Performing the Clustering of the Scores from the Principal Component Analysis.....	55
5.2.2. Investigating Scores of Principal Components and their Meaning.....	61
5.2.3. Investigating Coefficients of Principal Components and their Meaning.....	64
5.3. Further Discussion of the Principal Component Analysis and its Limitations .....	66
CHAPTER 6: CONCLUDING REMARKS .....	68
6.1. Key Contributions.....	70
APPENDIX A: MASS SPECTRA .....	72
APPENDIX B: GCxGC IMAGES.....	77
APPENDIX C: SCORES OF PRINCIPAL COMPONENTS .....	155
APPENDIX D: COMPONENT WEIGHTS.....	170
REFERENCES .....	191

## LIST OF TABLES

Table 4-1 A comprehensive list of each target compound and their corresponding target compound number. ....	18
Table 4-2 Statistics of the absolute and relative retention times of PCB-052 for 72 samples.....	19
Table 4-3 The coefficient of variation of the primary retention time (PRT) and secondary retention time (SRT) for each marker are listed for each target compound. The PRT has a coefficient of variation between 2-4.1% as the SRT has a coefficient of variation between 10-16.7%. It was expected for the PRT to not vary too much as each compound should leave the primary column around the same time.....	20
Table 4-4 The coefficient of variation of primary relative retention time (PRRT) and secondary relative retention time (SRRT) for each target compound are listed. PCB-028 does not have a value as the PRRT and SRRT are calculated relative to itself. The PRRT has a coefficient of variation between 0-2.5% as the SRRT has a coefficient of variation between 0-4.2%. The variation of the data is less than the variation in the absolute retention times, concluding that relative retention times are the preferred method. ....	21
Table 4-5 Result of algorithm based on the value $V_1$ and $V_2$ . The value $V$ can return a -1 if not all of the fingerprint mass to ion ratios are identified in the compared normalized signal, or if the difference in the normalized peak areas are greater than 0.3500. Otherwise, a certainty percentage will be outputted with the result of matching compounds.....	27
Table 5-1 The RRTs percent change is compared between the identified coeluted (potentially) compounds and matching compounds from the same sample based on the results of the comparison algorithm. Unknown 3744 where the Unknown 2632, Unknown 1427, Unknown 3500, Unknown 3526, and Unknown 3620 correspond to the coeluting compounds as Unknown 2421, Unknown 1357, Unknown 3219, Unknown 3296, and Unknown 3324 correspond to the matching compounds. Each row includes unknowns that belong to the same file .....	45
Table 5-2 The percent change of the PRRT and SRRT from the Unknown 3744 are compared between Unknown 2911, Unknown 3495, Unknown 2698, and Unknown 3230 as well as the certainty measurement. Unknown 2911 and Unknown 3495 both had a mass spectra mass of low intensity. Unknown 2698 and Unknown 3230 are from the same samples as the other unknowns respectively. The values that have low percent change in RRT. Another indicator is that the PRRT should be at least less than 4%. ....	46
Table 5-3 Unknown compounds from the same sample that were identified as either a coelution, match, or base coelution. The PRRT and SRRT percent change are both labeled with their matches that indicate the percent certainty. Those highlighted in green were labeled as a match visually by the chemist expert. ....	50
Table 5-4 The statistical measurements of the primary relative retention time percent change, secondary relative retention time percent change, and the certainty measurement are listed.	

The average percent change was less than 1% in the relative retention times and the average certainty was greater than 90%. ..... 51

Table 5-5 The colors of each box indicated the sample numbers that are within the same color cluster of Figure 5-45. Samples 31 and 59 were expected to be the outliers of the  $L_2$  normalization score. .... 59

## LIST OF FIGURES

Figure 2-1 The GCxGC-ToF-MS instrument located in the FDA lab. Two arrows indicate the gas chromatograph oven and the time-of-flight mass spectrometer. Picture courtesy: Jeffrey Archer, FDA .....	8
Figure 2-2 The dual stage thermal modulator where there are cold and hot jets. The compounds from the primary column are entered into the modulate. The jets are then used to inject compounds into the secondary column. Picture courtesy: Jeffrey Archer, FDA. Artwork by Morgan Moore, FDA.....	9
Figure 4-1 Simplified Process of Primary and Secondary Columns Picture Courtesy: Jeffrey Archer, FDA. ....	17
Figure 4-2 An example of a peak true mass spectrum of peak 35. The x-axis shows the mass to charge ratio (m/z) and the y-axis shows the abundance of masses based on the area of the peak. The mass to charge ratio equates to the molecular mass of the compound. Picture courtesy: Jeffrey Archer, FDA.....	23
Figure 4-3 The mass spectrum representation of peak 35 is shown above. Each entry is the mass to charge ratio, or molecular weight, and the peak area associated with the mass as per Equation 4-7.....	24
Figure 4-4 The normalized peak information of peak 35 is shown above. Each of the areas from Figure 4-3 was divided by 1,376,601. The mass is separated by a colon with the normalized area of each mass. ....	25
Figure 4-5 The fingerprint of peak 35 with the corresponding normalized areas is shown above. Each normalized area is greater than 0.3000. The threshold value was determined based on collaboration with the FDA. ....	25
Figure 5-1 The caliper and peak true mass spectrum of Unknown 3744 is shown above. This compound is the 'base compound' used for comparisons. The red box indicates the masses that are in the fingerprint for Unknown 3744. The x-axis represents the mass to charge ratio as the y-axis represents the abundance of each mass.....	34
Figure 5-2 The mass spectrum data collected in the peak true of Unknown 3744. Each mass and their corresponding peak area are separated by a colon.....	34
Figure 5-3 The normalized peak information of Unknown 3744. Each mass is paired with their normalized area, separated by a colon. The normalized areas were found by dividing each peak area by 543,365. ....	34
Figure 5-4 Fingerprint of Unknown 3744 are the masses and their normalized area that are separated by a colon.....	35



Figure 5-5 The library search of Unknown 3744 performed by the FDA indicated that the compound structure is likely p-p'-DDE..... 35

Figure 5-6 The structure of p-p'-DDE compound [15] ..... 35

Figure 5-7 The mass spectrum of Unknown 3315 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. In the peak true signal, mass from the fingerprint of Unknown 3744 include 246, 248, 316, and 318.. 37

Figure 5-8 The structure of an o-p'-DDE [15]..... 37

Figure 5-9 The mass spectrum of Unknown 2584 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Peaks from the fingerprint of Unknown 3744 include 246, 248, and 316. .... 38

Figure 5-10 The mass spectrum of Unknown 4095 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Masses that are in the fingerprint of Unknown 3744 include 246 and 176. .... 38

Figure 5-11 The mass spectrum of Unknown 2725 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Masses from the fingerprint of Unknown 3744 are 246, 316, and 176. .... 38

Figure 5-12 The mass spectrum of Unknown 2606 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Compared to the fingerprint of Unknown 3744, only mass 246 is in the peak true. .... 39

Figure 5-13 The mass spectrum of Unknown 3560 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Compared to the fingerprint of Unknown 3744, mass 246, 248, 316, and 318 are included. .... 39

Figure 5-14 The mass spectrum of Unknown 3681 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Compared to the fingerprint of Unknown 3744, mass 246 and 318 appear in the peak true signal. .... 39

Figure 5-15 The mass spectrum of Unknown 3456 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Compared to the fingerprint of Unknown 3744, mass 246, 248, 176, 318 and 316 are found. Other masses that are not in Unknown 3744 include masses 212 and 165 so coelution was notated for this compound by the algorithm. .... 40

Figure 5-16 The mass spectrum of Unknown 3181 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Visually there seems to have a slight residual amount of Unknown 3744 as all the peaks within the fingerprint are accounted for..... 41

Figure 5-17 The mass Spectrum of Unknown 3270 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Visually there seems to have a slight residual amount of Unknown 3744 as all the peaks within the fingerprint are accounted for..... 41

Figure 5-18 The mass spectrum of Unknown 2632 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal visually represents a match to Unknown 3744 as each mass in the fingerprint is found and have proper ratios. However, the intensity of the peak true signal is low. .... 42

Figure 5-19 The mass spectrum of Unknown 1427 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal visually represents a match to Unknown 3744 as each mass in the fingerprint is found and have proper ratios. However, the intensity of the peak true signal is low. .... 42

Figure 5-20 The mass spectrum of Unknown 3500 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal visually represents a match to Unknown 3744 as each mass in the fingerprint is found and have proper ratios. However, the intensity of the peak true signal is low ..... 42

Figure 5-21 The mass spectrum of Unknown 3576 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal visually represents a match to Unknown 3744 as each mass in the fingerprint is found and have proper ratios. However, the intensity of the peak true signal is low ..... 42

Figure 5-22 The mass spectrum of Unknown 3620 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal visually represents a match to Unknown 3744 as each mass in the fingerprint is found and have proper ratios. However, the intensity of the peak true signal is low ..... 42

Figure 5-23 The mass spectrum of Unknown 2421 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal is verified to be a match with a high intensity to Unknown 3744..... 43

Figure 5-24 The mass spectrum of Unknown 1357 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal is verified to be a match with a high intensity to Unknown 3744..... 44

Figure 5-25 The mass spectrum of Unknown 3219 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal is verified to be a match with a high intensity to Unknown 3744..... 44

Figure 5-26 The mass spectrum of Unknown 3324 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal is verified to be a match with a high intensity to Unknown 3744..... 44

Figure 5-27 The mass spectrum of Unknown 2911 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The

peak true signal visually looks similar to Unknown 3744 at low intensities. The primary relative retention time of the compound is greater than 4% and determined not to be a match... 45

Figure 5-28 The mass spectrum of Unknown 3495 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal visually looks similar to Unknown 3744 at low intensities. The primary relative retention time of the compound is greater than 4% and determined not to be a match... 46

Figure 5-29 The mass spectrum of Unknown 2698 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The compound was confirmed to be a match with a certainty of 98.7%. .... 46

Figure 5-30 The mass spectrum of Unknown 3230 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The compound was confirmed to be a match with a certainty of 96.2% ..... 47

Figure 5-31 The mass spectrum of Unknown 3902 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The compound was not determined to be a match due to the relative retention times. .... 47

Figure 5-32 The mass spectrum of Unknown 3112 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The compound was not determined to be a match due to the relative retention times. .... 48

Figure 5-33 The mass spectrum of Unknown 3588 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The compound was not determined to be a match due to the relative retention times. .... 48

Figure 5-34 The mass spectrum of Unknown 3297 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The compound was the best match in the sample and has the p-p'-DDE structure..... 48

Figure 5-35 The mass spectrum of Unknown 3330 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The structure of the compound was o-p'-DDE and not verified to be a match for the sample. .... 49

Figure 5-36 The mass spectra of Unknown 2819 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The certainty measurement was 19.1% because mass 316 and 318 were not quantified in the fingerprint. .... 51

Figure 5-37 The Discrete GCxGC image of every unique relative retention time location (PRRT, SRRT) described over all the samples. There seem to be overlapping points that can be identified in the principal component analysis. Recall that there are 63,501 analytes stored identified in the GCxGC image of window five. There are 51,143 unique RRT locations found in all 72 samples. The PCA was performed on the 72 by 51,143 data matrix and resulted in a seventy-two by seventy-one data matrix. A combination of the 51,143 locations was reduced to seventy-one different PCs. The scree plot describes how the seventy-one PCs

describe the variance in Figure 5-37. The scree plot has an inflection point at PC 7. The PC 1-7 describe 55.4936% of the variance in the data set as PC 8-71 describes 44.4064%..... 53

Figure 5-38 Scree Plot of PCA performed over all GCxGC relative locations. Each component explains a percentage of the total variance within the original data matrix. The data matrix stored the normalized areas, based on PCB-123, at each RRT in Figure 5-37..... 54

Figure 5-39 The scree plot is zoomed in on principal components 8-71. Components 8-28 represent at least 1% of the total variance in the data matrix. Components 29-38 explain 0.5% to 1% of the total variance. The remaining compounds 39 to 71 account for less than 0.5%, with principal components 50 to 71 accounting for barely a percentage. When components explain a low percentage of the total data matrix, a near linear relationship is described based on the components. .... 55

Figure 5-40 The  $L_2$  normalization scores of all 72 samples based on all seventy-one principal components. The blue line indicates the seven samples that have high correlation to the first seven components. .... 56

Figure 5-41 The  $L_2$  normalization scores for the 72 samples of principal components 8-71. The red box indicates files 28, 30, 31, 47, 48, 49, and 59 that have low associations to principal components 8 to 71..... 56

Figure 5-42 The dendrogram of the Square Euclidean Distance. The cophenetic correlation for the hierarchical tree is 0.7360. The square Euclidean distance therefore is not the best representation of the data when using k-means. .... 57

Figure 5-43 The dendrogram of the sum of absolute distances referred to as Cityblock in MATLAB The right most tree defines the first seven samples into one cluster. The remaining samples seem to evenly belong to two separate clusters. The add cophenetic correlation coefficient is 0.8513 and is a better representation of the data. .... 58

Figure 5-44 The first clustering of all seventy-two samples based on the total  $L_2$  normalization score is shown. There are three clusters. The five samples in the yellow cluster identify the most variant samples. The blue and red clusters have samples that seem to be very close to one another. Clusters blue and red are re-clustered and shown in Figure 5-46. The x-axis is the sample numbers as the y-axis is the  $L_2$  normalization score..... 59

Figure 5-45 The second clustering of the  $L_2$  normalization scores not including sample 28, 30, 47, 48, and 49. The clustering representation has a cophenetic correlation coefficient of 0.9629 when clustering the data points above using the sum of absolute distance measurement. The x-axis is the sample numbers as the y-axis is the  $L_2$  normalization score..... 60

Figure 5-46 Scatter Plot of Files of the  $L_2$  normalization scores of PCs 1 to 7 vs PCs 8 to 71. The sample numbers that are shown have high associations with the 55.4936% of the total variance ..... 61

Figure 5-47 The  $L_2$  normalization score of principal component 1 where sample 49 is highly correlated with. Sample 28, 30, 47, and 48 each have at least one location that is in common with sample 49 since their value is not zero ..... 63

Figure 5-48 The normalized coefficients of principal component 1 indicates how relevant the locations of the discrete GCxGC image are for the principal component score..... 65

## CHAPTER 1: INTRODUCTION

Chemometrics are used to quantify toxic compounds in a sample. Comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry (GCxGC-ToF-MS) data has been collected over the past ten years that target twenty-three persistent organic pollutants in milk samples. The instrument processes thousands of compounds that co-exist in the sample extract. Only the twenty-three polychlorinated biphenyls (PCBs) and polybrominated diphenyl ethers (PBDEs) are identified and quantified in the sample extracts. The research presented in this thesis discusses methods to analyze the non-targeted compounds that are also collected during the target analysis.

A GCxGC-ToF-MS records data based on two axes, the primary retention time (PRT) and secondary retention time (SRT). An analyte will vary in the PRT and SRT. To account for the retention time variances across the samples analyzed, relative retention times (RRTs) are introduced. Each sample is divided into twenty-four windows defined by the target compounds identified in each sample. The RRTs are dependent on the target compounds primary and secondary retention time. Compounds that may be a match are found within 4% change of the primary relative retention time (PRRT) and secondary relative retention time (SRRT) of the analyte. A robust mass spectra comparison algorithm is presented. The algorithm normalizes the mass peak areas in the mass spectrum based on the largest peak. Each mass that has a normalized abundance greater than 0.3000 are considered the unique identifier of the compound called the fingerprint. The fingerprint is then quantified to a single value for the mass spectrum using robust mathematics. When comparing to a separate compound, the difference in peak areas must be within 0.3500. The differences in peak areas and the fingerprint are quantified to a single

value and divided to return a measurement value as discussed in section 4.1.2. Conclusions on whether two compounds are a match or (potentially) coeluted can be made based on the values.

Each of the GCxGC images separated into twenty-four windows using the RRTs as the two axes. To determine if there are any compounds that may be found in multiple samples, a data matrix was created. The data matrix represents the abundance of a compound at each RRT location in the GCxGC image for the seventy-two files. The principal component analysis (PCA) [1] is applied to the data matrix. From the PCA, a total score for each sample is calculated using the  $L_2$  normalization formula. The total scores are then clustered using k-means clustering [2] and optimized from the silhouette method [3]. Interpretation of the principal components generated is also discussed.

The research presented is relevant to the thousands of unknown compounds, some of which could be dangerous, that are processed when target analysis is completed. With mass spectra comparisons and statistical analysis, chemists can retroactively recognize patterns and evaluate samples with one another. One major outcome of this research is the ability to compare non-target compounds together based on the relative GCxGC location. Based on the variance of the total dataset, the PCA identifies locations for further analysis. Samples differed based on geographical locations and whether they were organic. This discrepancy between samples was considered when analyzing the clusters based on the PCA scores. Accordingly, an important tangible product of this thesis research is the potential to develop a MATLAB package that aids in the investigation of non-targeted compounds in milk samples for the FDA, that later can be applied to other data matrices.

Current research focuses on identifying associations between non-target and target compounds through raw signal analysis [4] [5] [6] [7]. Autonomous peak-cognizant techniques

have been established and graph-based visualizations can also identify non-target compounds [8] [4] [5] [6] based on the peak topography maps of the samples. The most recent research uses machine learning techniques to find the association where accumulation and degradation of compounds can be geographically analyzed [7]. Though the research presented does not focus on the raw signal of the instruments, non-target compounds can be analyzed utilizing the novel relative retention times that address the drift in PRT and SRT that typically occurs over time.

### **1.1. Key Contributions**

- (i) To account for the variance of absolute retention times between samples, relative retention times based on the already targeted analytes can be employed. Compounds that are within a threshold of percent change are considered for comparison of the mass spectra data.
- (ii) The interpretation and discussion of the statistics observed from the principal component analysis.
- (iii) A MATLAB package that can be employed on large data repositories for the Food and Drug Administration

Chemists can retroactively analyze the data for non-target compounds from their targeted analysis.



## CHAPTER 2: BACKGROUND MOTIVATION

In this chapter, we discuss the historical background and research motivation of the ideas and technical approach presented in this thesis. Persistent organic pollutants (POPs) have been a threat to human health and the environment due to their well-known toxicity causing a multitude of documented health problems [9]. Investigation of these compounds in our food is, therefore, imperative to public health interest, and of high priority to the Food and Drug Administration (FDA). This research is scoped out of a larger research project focused on computational monitoring POPs in air and water. The larger project involved both raw signal processing, chemometric analysis as well as quantitative informatic computations based on isolated compounds. The primary motivation of this thesis, in the context of the larger research effort, is to nail down two important computational issues, critical to precise quantitative analysis: (I) The issue of mapping a peak to the correct compound despite retention time variability, (II) Clustering compound peaks appropriately in a form that is interpretable by an experienced analyst. In this context, work presented in this thesis falls in the larger domain of chemical pattern recognition, and as such, we present the background motivation and related work in this broader context. We also discuss the data collection of the target analysis as it is used to retroactively find data patterns to identify non-targeted compounds in this research. Disclaimer: *Reference to any commercial materials, equipment, or process does not, in any way, constitute approval, endorsement, or recommendation by the US Food and Drug Administration.*

### **2.1. Persistent Organic Pollutants and Why They Are Investigated**

During the industrial boom after World War II, synthetic carbon-based chemicals were designed [9]. In the short term, these chemicals were beneficial in pest and disease control, crop

production, and had numerous uses for industry; however, in the long term, many of these synthetic chemicals had negative effects on human health and the environment. Many of these synthetic chemicals, including their unintentionally produced waste (dioxins), are referred to as persistent organic pollutants (POPs). POPs can be found in the air, water, soil, plants, fish, and other wildlife.

In 2001, the United States took part in the Stockholm Convention, where a global treaty was adopted by numerous nations to protect the health of all life and the environment [10]. POPs have been linked to cancer, damage the nervous system, cause reproductive disorders, and weaken the immune system. Since they are relatively stable, they can bioaccumulate in unhealthy concentrations in humans and animals, especially through consumption that occurs in the food chain.

The first 12 POPs initially listed under the Stockholm Convention contain forms of pesticides, industrial chemicals, and by-products. Specifically, as a by-product, polychlorinated biphenyls (PCBs) were listed [10]. The manufactured PCBs consist of carbon, hydrogen, and chlorine atoms where the location of the chlorine atoms in the chemical structure determines the physical and chemical properties of the PCB congener [11]. Toxicity varies between different PCBs. They were originally used for industrial purposes since they are non-flammable, chemically stable compounds, with a high boiling point and have electrical insulating properties.

Another group of compounds that were not listed in the Stockholm Convention are the polybrominated diphenyl ethers (PBDEs) [12]. The EPA continues to have concerns about these chemicals because they too are persistent, bioaccumulate, and are toxic to humans and the environment. They were utilized as flame retardants in textiles, plastics, wire insulation, and

automobiles. The levels of the PBDEs detected are increasing though their usage in the United States was phased out in 2004.

In summary, some POPs are known to cause adverse effects to human health and the environment. Chemists target several known chemicals to monitor their concentrations in the environment. This research study is a part of the larger effort of identifying and quantifying POPs funded by the National Science Foundation (NSF).

## **2.2. Monitoring Persistent Organic Pollutants in Air and Water**

There are instruments used to separate chemical compounds for identification and quantitation such as gas chromatography and mass spectrometry (GCxGC, GC-FID, GC-MS, GC-MS-MS) that produce a high-resolution signal [13]. This research has three aims; (1) raw signal analysis to capture co-eluting non-target compounds with target compounds, (2) pattern discovery to find the “partners in crime” that can be associated with how the contamination takes place, and (3) data integration where multiple types of data, such as chromatographic and mass spectrometry data can be combined to comprehensively interpret the data.

The broader impact of the proposed research can be applied to other environmental surveillance and toxin monitoring for the sake of public health. Multiple data sets are already being pursued with similar techniques, including data sets from oil spills and groundwater contamination.

**The research in this thesis does not focus on the raw interpretation of the signals, however, similar approaches of pattern discovery and data integration are pursued.** The approaches include principal component analysis (PCA) and k-means clustering. The data utilized for this research comes from the Food and Drug Administration (FDA) to explore what non-target chemicals may be present in milk samples included in the study.

### 2.3. Investigating Persistent Organic Pollutants in Milk Samples

The Persistent Organic Pollutants (POPs) Team at the FDA Office of Regulatory Affairs (ORA) and Office of Regulatory Science (ORS) Arkansas Laboratory (ORA/ARL) proposed the use of previous analytical data collected over the past 10 years to evaluate techniques that could be used to identify non-targeted, co-eluting analytes [14]. The manufacturer of the instrument used for the target analysis, LECO's two-dimensional gas chromatograph and time of flight mass spectrometer (GCxGC-ToF-MS), has developed a software package called "ChromaToF TILE" that can identify small differences in mass spectral results between samples. The Center for Food Safety and Applied Nutrition (CFSAN) works as a collaborator with the purpose of determining if other compounds should be incorporated in the targeted analysis by the conclusions of this research and the historical data associated with the new co-occurring compounds (potentially) identified. . This research will also be evaluated in comparison with these other techniques to try and learn more about non-targeted analytes.

Originally, it was proposed to use peak cognizant raw signal analysis techniques [13] to create a multi-dimensional evaluation of the data matrix. Then clustering the information created from unsupervised algorithms, and variable reduction from principal component analysis can aid in comparing full chromatographic runs from similar samples. However, the data that was used in collaboration was not the raw data but already processed data from the ORA/ARL. **Instead, the goal of this research is to provide aid in identifying compounds that are either (1) routinely found in a given matrix, (2) unique to a given matrix, and (3) found in a sample but not typically present in the matrix.** [14]

## 2.4. Targeted Analysis Data Collection to Identify Persistent Organic Pollutants in Lipophilic Samples

Historically the FDA used a HRGC-HRMS magnetic sector instrument to identify POPs. To find a more efficient production for analysis, the FDA began shifting away from the HRGC-HRMS as the instrument needs to have three analytical runs. Due to the instrument no longer being manufactured, the difficulty of maintaining instrument facility, the large instrument footprint, and the expense, the FDA changed to utilizing GCxGC-ToF-MS instrumentation in 2009. There are two portions of the GCxGC-ToF-MS instrument, the gas chromatograph (GC), and the mass spectrometer (MS).

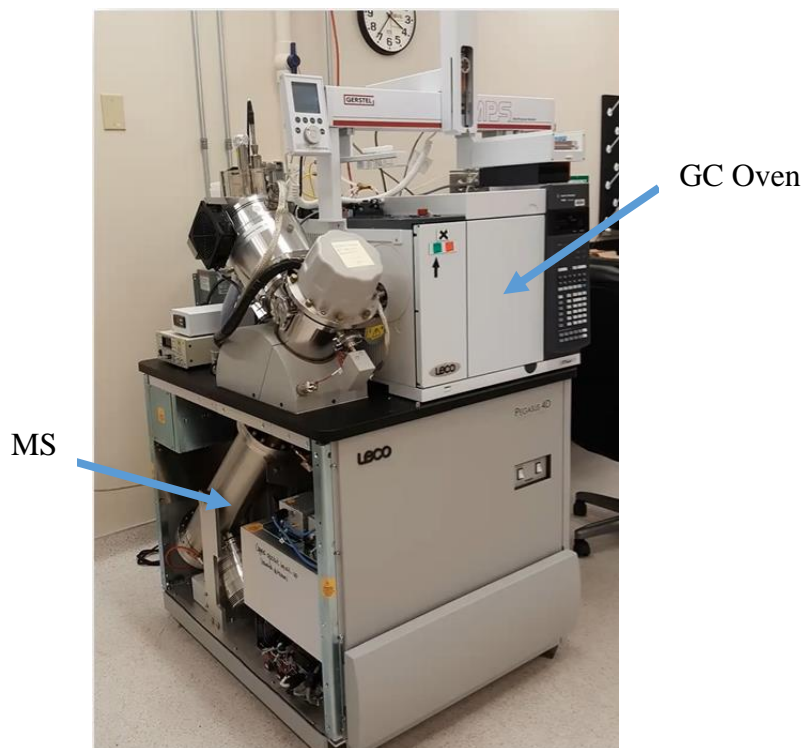


Figure 2-1 The GCxGC-ToF-MS instrument located in the FDA lab. Two arrows indicate the gas chromatograph oven and the time-of-flight mass spectrometer. Picture courtesy: Jeffrey Archer, FDA

The oven contains two columns, a primary column and secondary column. The milk sample extract is based upon approximately 95 grams of whole milk. The injection is approximately 20% of the total milk extract. 1-3 microliters extract of a sample is injected into a

precolumn that then feeds to the analytical primary column. Compounds are separated based on their polarity and then continue into a modulator. The modulator acts as a secondary injector. The sample goes into a dual stage thermal modulator as imaged in Figure 2-2. Cold temperature prevents the compounds from moving as the hot temperature allows the compounds to continue to the secondary column and into the mass spectrometer. Each section will blast cold and hot temperatures and the secondary column is packed with a different material that separates the PCBs based on the chlorine position relative to the ortho..

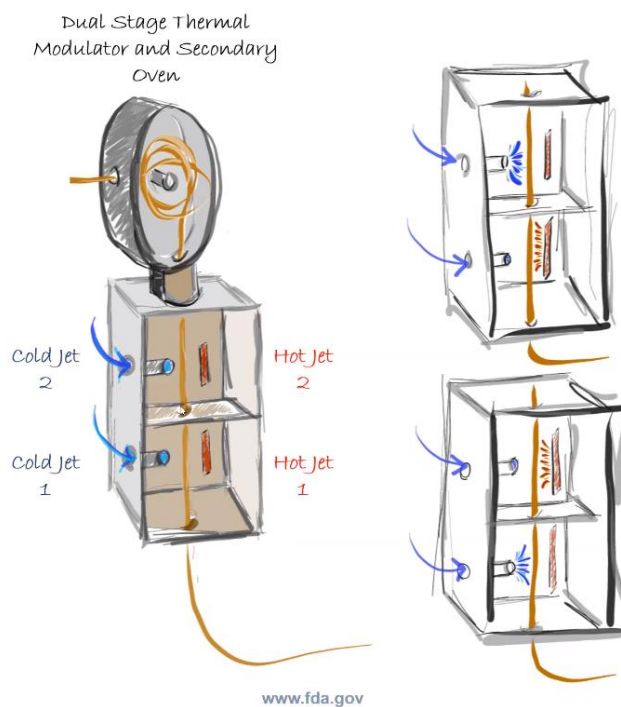


Figure 2-2 The dual stage thermal modulator where there are cold and hot jets. The compounds from the primary column are entered into the modulate. The jets are then used to inject compounds into the secondary column. Picture courtesy: Jeffrey Archer, FDA. Artwork by Morgan Moore, FDA.

To target analytes in the sample using this instrument, compounds are added that have the same chemical structure of the target compounds but instead each carbon is replaced with an isotopically labeled carbon-13 atom and serves as an internal standard. The compound has a higher mass; however, the retention time should be aligned with the naturally occurring carbon-

12 structure. Other processing occurs to quantify the amount of the POP found in the sample, however that is out of scope for this research. There is at least 10 years of GCxGC-ToF-MS data collected using this approach. Further investigation on how to use data to find matching compounds in the data matrix and determining similarities and differences between samples can be acquired when looking at 72 samples from different years and processed by different instruments.

## **2.5. Summary**

All in all, some persistent organic pollutants are investigated and monitored to protect the health of humans and the environment. Current research is focusing on how to process raw data signals from various instruments (GC-MS, LC-MS, etc.) to identify non-targeted compounds. Considering both the chromatographic and mass spectral data, patterns may emerge utilizing multiple methods. . Organizations, such as the FDA, are also interested in finding patterns through their extensive data set. Previous investigations on samples such as milk, eggs, and fish, were used to quantitate the concentration of POPs present. Using the same data collected over 10 years, exploration of the non-targeted compounds may demonstrate other compounds that may need to be monitored. Information of how the samples were collected, such as manufacturer, may also indicate if there are fertilizers or other contaminants that need to also be investigated. The goal of this research is to sort through their data to find matches, coelutions, and patterns of nontargeted analytes.

## CHAPTER 3: RELATED WORK

In this chapter, we discuss previous research analyzing target and non-target contaminants that are airborne or found crude oil. First petroleum forensics was investigated in "Interpreting comprehensive two-dimensional gas chromatography using peak topography maps with application to petroleum forensics". This paper used peak topography maps to identify target and non-target compounds. Next, "Peak-cognizant Signal Processing of Raw Instrument Signals to Quantify Environmental Weathering of Contaminants from the Deepwater Horizon Spill", investigated the graph-visualizations where compounds could be associated with one another. PCBs in airborne samples were discussed in "Signal Processing Methods to Interpret Polychlorinated Biphenyls in Airborne Samples", where an autonomous framework found associations of target and non-target compounds. Graph-visualizations were utilized and clustered for identifying the associations. Another research paper, "Raw signal processing and graph-based visualization to autonomously interpret large repositories of GC-MS data: applications to oil spill weathering studies," focused on large repositories of data utilizing similar techniques. Finally, machine learning was applied to the identify targets in each sample in "Employing and Interpreting a Machine Learning Target-Cognizant Technique for Analysis of Unknown Signals in Multiple Reaction Monitoring." Summaries of each paper and discussion of how this research differs is described below.

### **3.1. Overview of Past Research Analyzing Raw Instrument Signals for Target Analysis**

In 2016, peak topography maps (PTM) of two-dimensional gas chromatography (GCxGC) were utilized to identify target and non-targeted compounds when analyzing petroleum [8]. Without using training data, a robust quantitative measure for determining matches between samples was explored in the research. Comparison of PTM and statistical



methods, like PCA, were compared for robustness. PTM partitioning grouped peaks that are similar or dissimilar based on thresholds. The similar topography generates the cross-PTM score. The research focused on similarities and differences of oil samples based on targeted biomarkers topography.

Environmental weathering of contaminants from the Deepwater Horizon spill was researched using peak-cognizant quantification techniques [4]. Signal processing of the raw instrument signals would autonomously extract peak information from GC-MS signals. A graph-based quantitative computational framework relative to each peak represents the weathering that occurred. Before the signals were processed, the raw signal was normalized and aligned the retention time. The drift of the signals was applied to the weathered oil samples, which allowed compound associations to be discovered.

Next, signal processing methods were applied to PCBs in airborne samples [5]. A robust computational framework was used to autonomously analyze unknown associations between target and non-target industrial air pollutants. The autonomous framework utilized minimum mean-squared techniques to detect and separate coeluted peaks in the raw instrument signal. The amount of PCB in the raw signal was autonomously calculated by using a peak fitting technique based on  $L_2$  error minimization. Building off previous research, graph-based visualizations were utilized to find associations between target and non-target compounds through PCA. Clustering techniques like fuzzy c-means and k-means were implemented and compared. The relative contribution of PCB signals against ten potential source samples were evaluated utilizing parameter optimization techniques. Using 150 air samples, comparisons between target-only techniques that focus solely on target compounds versus the proposed techniques were based on efficiency.

Large repositories of oil spill weathering were analyzed using raw signal processing and graph-based visualizations [6]. Data were collected from GC/MS/MS instruments and focused on large-scale machine interpretations. Typical machine data interpretations use automated chemometric analysis that focuses on principal component analysis of thousands of compounds within a sample. Other manually driven interpretations are annotated by experts in the field. The analysis is typically costly in personnel time and is primarily focused on target compounds. Analysis of large data repositories are limited. To bridge the gap between expert analysis and instrument interpretation, automated peak-cognizant processing of the raw instrument signal and the graph-visualizations generated can provide opportunities to discover non-target compound peaks and quantify associations with target and non-target compounds.

The most recent publication employs instrument learning to analyze unknown signals [7]. The unknown signals are from multiple reaction monitoring mode. The autonomous instrument learning framework associates target along with non-target peaks present in raw GC/MS/MS instrument signals. Three instrument learning algorithms were evaluated based on the accuracy in training and testing. Peaks found at specific retention times are annotated and aligned using adaptive signal processing techniques. Discoveries of how target PCBs accumulate or degrade in certain locations can be found utilizing geographical topographical representations.

Overall, the approaches focused on target analysis of raw instrument signals where profiles of non-targeted compounds can be acquired through the analysis. Graph-based visualizations and statistical methods, such as principal component analysis, have been employed. Instead of accounting for the drift in retention times over the large data set, the research here introduces relative retention times that can be applied to a large-scale data repository. The analysis in this research focuses on retroactively identifying non-target

compounds utilizing the data collected from target analysis. Methods on how to investigate non-target compounds between target compounds are explored and new approaches with the data are proposed for future research.

## CHAPTER 4: TECHNICAL APPROACH

In this chapter, we will detail the overall technical approach as well as the nature of the dataset under consideration to justify the methodology chosen. Specifically, we provide detailed descriptions of what it means to compare two chemical analytes, i.e., compounds represented by peaks in the GCxGC image as well as the mass spectra. In particular, we introduce the concept of relative retention times that are determined by the target compounds. This allows us to robustly assess which compounds in the data matrix are worth comparing against the mass spectra comparison algorithm. We also use relative retention times to analyze the GCxGC image between two target compounds, referred to as a window. Statistical analysis techniques, i.e., principal component analysis, is used to determine the similarities and differences between milk samples based on the abundances of compounds in the GCxGC image. Finally, we discuss clustering techniques based on the  $L_2$  normalization scores of each sample from the PCA. The results are recorded in chapter 5 with further discussion on how the methodologies can be used for analysis.

### **4.1. How to Compare Two Analytes Together Using Their Mass Spectra Data**

The technical approach in this work is based on a diverse portfolio of analyzed milk samples provided by the Food and Drug Administration (FDA). The data set contains seventy-two milk samples that were processed using a LECO Pegasus 4D GCxGC-ToF-MS instrument, which provided a tiered resolution of each sample into GCxGC and mass spectrometric domains. A two-dimensional gas chromatogram image breaks down the sample into peaks based on their molecular weight and the abundance of the compound. Each GCxGC peak was paired with the mass spectral data, which provides another level of analytical breakdown of the compound(s) represented by the GCxGC peak. The seventy-two samples varied by location, specific

instruments used, and year collected. To compare the analytes together two factors were considered: (1) the relative primary and secondary retention time pairs and (2) the mass spectra.

#### **4.1.1. Introducing Relative Retention Times for Proper Relations Between Samples Collected Over the Years**

When the sample is processed by the LECO instrument, the output is a two-dimensional gas chromatogram. In the x-direction is the primary retention time (PRT) and the y-direction is the secondary retention time (SRT). The PRT has a range of 450-2600 seconds (s), where the SRT has a range of 0-4 seconds. Each represents how a compound moves through the primary and secondary columns of the LECO instrument as described in section 2.4.

In Figure 4-1 there are three rectangles. The first larger rectangle represents the primary column, the blue rectangle represents the modulator, and the last represents the secondary column. Compounds that are injected into the instrument are separated first based on their polarity and boiling point. Next, the modulator uses temperature variations to inject the compounds into the secondary column.

The secondary column, which is shorter, narrower, and thinner, has a different polarity and continues to separate the compounds. The time that the compound leaves the primary column corresponds to the PRT and the time that the compound leaves the secondary column corresponds to the SRT. The highest intensity of a compound is recorded into a file that was utilized for this research.

## GCxGC Simplified Separation

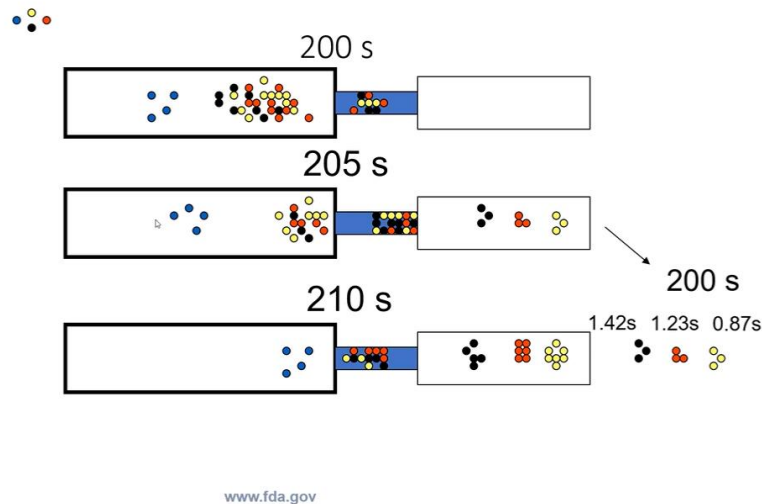


Figure 4-1 Simplified Process of Primary and Secondary Columns Picture Courtesy: Jeffrey Archer, FDA.

Each milk sample has twenty-four target compounds where twenty-three are typically found in the same general area in the chromatogram. A comprehensive list of the twenty-three compounds organized by order is found in Table 4-1. A retention time pair of the primary retention time and secondary retention time are used as a determinate for each compound found in the sample, however, it is known that the PRT of the same compound can vary between 3-15 seconds based on the instrument used and when the sample was processed. The SRT is also known to vary by hundredths of a second.

To account for the discrepancies in the PRT and SRT, relative retention times were assessed for each POP compound. Each sample has known target compounds that occur at a certain location. PCB-028 always appears before PCB-052. If the retention times of PCB-052 were found relative to PCB-028, the variance between each sample reduces. The primary relative retention time (PRRT) and secondary relative retention time (SRRT) are calculated based on the target compound number  $k$  — where PCB-028 represents target compound number one and

PCB-052 represents target compound number two that increments all the way to target compound number twenty-three which represents PBDE-153. The relative retention times are described in the equations 4-1 and 4-2. PCB-028 was calculated relative to itself and has a PRRT and SRRT of (1,1).

Table 4-1 A comprehensive list of each target compound and their corresponding target compound number.

1) PCB-028	6) PBDE-028	11) PCB-138	16) PBDE-047	21) PBDE-099
2) PCB-052	7) PCB-118	12) PCB-167	17) PCB-170	22) PBDE-154
3) PCB-070	8) PCB-114	13) PCB-156	18) PBDE-077	23) PBDE-153
4) PCB-101	9) PCB-153	14) PCB-157	19) PCB-189	
5) PCB-123	10) PCB-105	15) PCB-180	20) PBDE-100	

Equation 4-1 The primary retention time is dependent on the target compound found previous to the current target compound. For example, the primary retention time of target compound number one (PCB-028) is used to calculate the PRRT of target compound number two (PCB-052). PCB-028 has a PRRT of one.

$$PRRT_{k+1} = \frac{PRT_{k+1}}{PRT_k}$$

Equation 4-2 The secondary relative retention time is dependent on the target compound found previous to the current target compound. For example, the secondary retention time of target compound number one (PCB-028) is used to calculate the secondary relative retention time of the target compound number two (PCB-052). PCB-028 has a SRRT of one.

$$SRRT_{k+1} = \frac{SRT_{k+1}}{SRT_k}$$

Using PCB-052 as an example, the minimum, maximum, average, and standard deviation of the PCB-052's absolute PRT and SRT as well as the PRRT and SRRT was calculated and listed in Table 4-2. Note that the standard deviation of the relative retention times is much reduced when compared to the absolute retention times. The percent change using Equation 4-3

was calculated to be 15.3% and 82.4% respectively for PRT and SRT and 2.6% and 3.6% respectively for PRRT and SRRT. Clearly the PRRT and SRRT give a more precise location of the target compounds instead of the absolute retention times, and may result in trying to look for a compound in the wrong location.

Equation 4-3 The percent change of each primary retention time, secondary retention time, primary relative retention time, and secondary relative retention time was calculated using the minimum and maximum value for PCB-052 in Table 4-2.

$$\% = 100 * \frac{|\text{min} - \text{max}|}{\text{min}}$$

Table 4-2 Statistics of the absolute and relative retention times of PCB-052 for 72 samples

PCB-052	PRT	PRRT
MIN	746.4	1.096584
MAX	861.01	1.125097
AVG	799.1207	1.111418
SD	28.53849	0.005521
	SRT	SRRT
MIN	0.85	1
MAX	1.55	1.036541
AVD	1.25375	1.064409
SD	0.199433	0.008672

Next, consider the coefficient of variation (CV), which is the standard deviation divided by the average. This measurement describes the variability of a sample relative to its mean. The CV is expressed as a percentage and is used to compare the spread of data sets. The CVs for each of the absolute and relative retention time is calculated for each target compound and is expressed in Table 4-3 and Table 4-4.



Table 4-3 The coefficient of variation of the primary retention time (PRT) and secondary retention time (SRT) for each marker are listed for each target compound. The PRT has a coefficient of variation between 2-4.1% as the SRT has a coefficient of variation between 10-16.7%. It was expected for the PRT to not vary too much as each compound should leave the primary column around the same time.

PCB-028	4.00%	16.24%	PCB-153	2.52%	15.27%	PCB-170	2.17%	16.62%
PCB-052	3.60%	15.96%	PCB-105	2.58%	15.32%	PBDE-047	2.16%	14.03%
PCB-070	4.04%	16.01%	PCB-138	2.60%	15.16%	PCB-189	2.02%	13.62%
PCB-101	3.86%	15.61%	PCB-167	3.12%	15.60%	PBDE-100	2.08%	12.57%
PCB-123	2.65%	15.29%	PCB-156	2.96%	15.28%	PBDE-099	2.26%	12.08%
PBDE-028	2.71%	15.35%	PCB-157	2.52%	15.46%	PBDE-154	2.43%	10.87%
PCB-118	2.59%	15.27%	PCB-180	2.22%	15.18%	PBDE-153	3.14%	10.01%
PCB-114	2.64%	15.49%	PBDE-047	2.30%	15.00%			

Table 4-4 The coefficient of variation of primary relative retention time (PRRT) and secondary relative retention time (SRRT) for each target compound are listed. PCB-028 does not have a value as the PRRT and SRRT are calculated relative to itself. The PRRT has a coefficient of variation between 0-2.5% as the SRRT has a coefficient of variation between 0-4.2%. The variation of the data is less than the variation in the absolute retention times, concluding that relative retention times are the preferred method.

PCB-028	-	-	PCB-153	0.58%	1.88%	PCB-170	0.17%	0.83%
PCB-052	0.50%	0.86%	PCB-105	0.33%	1.28%	PBDE-047	0.11%	0.82%
PCB-070	2.48%	1.70%	PCB-138	0.45%	0.92%	PCB-189	0.28%	1.32%
PCB-101	2.38%	1.99%	PCB-167	2.08%	1.85%	PBDE-100	0.15%	1.47%
PCB-123	0.64%	1.35%	PCB-156	1.44%	2.40%	PBDE-099	0.85%	1.12%
PBDE-028	0.15%	0.88%	PCB-157	1.10%	1.54%	PBDE-154	1.90%	3.19%
PCB-118	0.21%	0.94%	PCB-180	1.01%	1.72%	PBDE-153	2.08%	4.20%
PCB-114	0.49%	1.94%	PBDE-047	0.20%	0.80%			

The variation of the data spread is smaller when using the relative retention time in both the primary and secondary locations. The PRT is expected to have a consistent or near consistent value, however with the variation of the PRRT being less than 1% in most cases, and 2.5% being the maximum variation spread, relative retention times provides a small window of where a compound may be found in the GCxGC image. When considering the secondary retention time, the spread of approximately 15% for each marker all reduced to a value less than 5%. This demonstrates that using a relative location for each sample will provide a more precise location of where the compound may be found.

Since relative retention times can easily identify where each marker compound is within each sample, the same idea can be applied to every non-targeted compound. A compound that is non-targeted must be within 5% change of their relative retention times. The non-targeted

compounds have a relative retention time based on which targeted analytes they are surrounded by. Each section divided by the targeted analytes are referred to as *windows*, where the compounds from 450s to the PRT of PCB-028 is referred to as window 1. Window 2 would be the unknown compounds between PCB-028 and PCB-052 with a PRRT and SRRT dependent on the location of PCB-052's PRT and SRT. The values are found similar to Equation 4-1 and 4-2, except the numerator is the non-targeted compound's PRT and SRT. For the last window 24, the RRTs are calculated based on PBDE-153.

Equation 4-4 Each non-target compound has a primary relative retention time based on the target compound *k*. Target compound *k* indicates the end of the window *k*. For window twenty-four, the target compound used in PBDE-153.

$$PRRT_{non-targeted} = \frac{PRT_{non-targeted}}{PRT_{target\ compound\ for\ window\ k}}$$

Equation 4-5 Each non-target compound has a secondary relative retention time based on the target compound *k*. The target compound *k* indicates the end of the window *k*. For window twenty-four, the target compound used in PBDE-153.

$$SRRT_{non-targeted} = \frac{SRT_{non-targeted}}{SRT_{window\ marker}}$$

As described earlier, the condition dictated in Equation 4-6 needs to be met for a compound to be considered a potential match. This condition speeds up the runtime for the algorithm to perform, as not every permutation of comparison needs to be done. Once locations are determined to be similar, then the algorithm described in 4.1.2 can be implemented to determine if there are any matches or coelutions in the data matrix.

Equation 4-6 The percent change of the primary relative retention time must be within 5% of the 'base compound', where the 'base compound' is the analyte that is being assessed if there are any matches within other milk samples.

$$100 * \frac{|PRRT_{compound\ to\ be\ compared} - PRRT_{base\ compound\ for\ comparison}|}{PRRT_{base\ compound\ for\ comparison}} < 5\%$$

Equation 4-7 The percent change of the secondary relative retention time must be within 5% of the ‘base compound’, where the ‘base compound’ is the analyte that is being assessed if there are any matches within other milk samples.

$$100 * \frac{|SRRT_{compound\ to\ be\ compared} - SRRT_{base\ compound\ for\ comparison}|}{SRRT_{base\ compound\ for\ comparison}} < 5\%$$

#### 4.1.2. Discussion of the Mass Spectra Comparison Algorithm

During each run of the LECO instrument process, compounds are separated and processed in a time-of-flight mass spectrometer. Sixty-seven spectra are captured every second. Mass spectrometry is a technique that analyzes the mass to charge weight of molecules. The raw signals are processed through software by the FDA and the resulting discrete signal is given in the data matrix. An example of a resulting mass spectrum is shown in Figure 4-2. The caliper signal is the total ion mass spectrum, and the peak true signal is the deconvoluted mass spectrum representation.

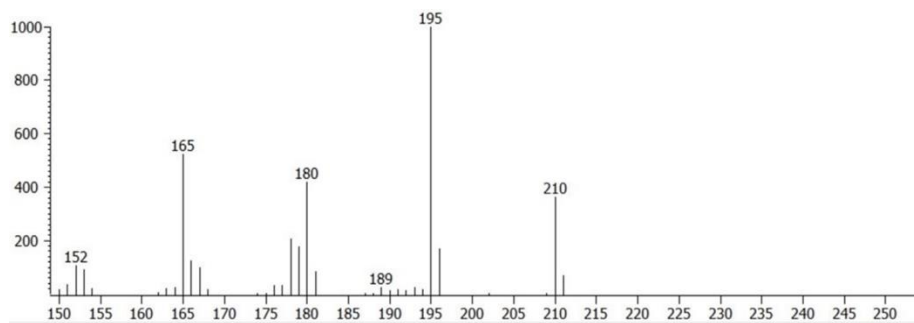


Figure 4-2 An example of a peak true mass spectrum of peak 35. The x-axis shows the mass to charge ratio (m/z) and the y-axis shows the abundance of masses based on the area of the peak. The mass to charge ratio equates to the molecular mass of the compound. Picture courtesy: Jeffrey Archer, FDA

Each of the peaks that are in the peak true signal have a corresponding peak area, or an abundance of a mass to charge ratio. To represent this data in the data matrix, the molecular weight is paired with the peak area and listed in an excel sheet. The representation of the mass

spectrum of Figure 4-2 is shown in Figure 4-3. The peak information of the mass spectrum is described in Equation 4-7. Note that the peak information is ordered from largest area to smallest area.

Equation 4-7 The peak information is stored in a two-element vector  $S$ . The two vectors are of the molecular weight, indicated as the mass of the ion, and peak area associated for all  $i$  peaks in the mass spectrum sorted from largest area to smallest area

$$S = (mass_i, area_i) \text{ for } i \text{ valued pairs in the mass spectrum}$$

195:1376601 165:722364 180:573602 210:501452 178:287023 179:242915 196:235082 166:170339
152:146764 167:136112 153:127566 181:114523 211:96974 151:50054 177:46094
176:41580 193:34176 189:33830 164:33714 154:30378 163:28668 191:24646 168:24231 194:23233
150:23158 190:19842 192:18364

Figure 4-3 The mass spectrum representation of peak 35 is shown above. Each entry is the mass to charge ratio, or molecular weight, and the peak area associated with the mass as per Equation 4-7.

To be able to compare two of these discrete data representations together, it is important to find the fingerprint of the peak information, which represents the unique identity of the GCxGC compound peak analyzed in the mass spectrum. The fingerprints are the molecules with large peak areas that can identify a compound. To do so, each peak area is normalized based on the largest peak area in the signal. In this case, Peak 35 would be normalized based on mass 195's peak area of 1,376,601 for each mass and is listed in Figure 4-5. A fingerprint is all of the peaks that have an normalize area greater than 0.3000 also referred to as being 30% of the normalized maximum peak. An example of the fingerprint is seen in Figure 4-6. The value of 30% was chosen based on consultations with the FDA expert regarding appropriate choice of a noise threshold for isolating peaks within the mass spectrometry profile. The normalized peak information  $N$  is listed as a two-element vector of mass and normalized area and is described in

Equation 4-8. The fingerprint only includes the masses that are greater than 0.3000 from the normalized peak information vector as described in Equation 4-9.

195:1.000	165:0.5247	180:0.4167	210:0.3543	178:0.2085	179:0.1765	196:0.1708	166:0.1237
152:0.1066	167:0.0989	153:0.0927	181:0.0832	211:0.0704	151:0.0364	177:0.0335	
176:0.0302	193:0.0248	189:0.0246	164:0.0245	154:0.0221	163:0.0208	191:0.0179	168:0.0176
194:0.0169	150:0.0168	190:0.0144	192:0.0133				

Figure 4-4 The normalized peak information of peak 35 is shown above. Each of the areas from Figure 4-3 was divided by 1,376,601. The mass is separated by a colon with the normalized area of each mass.

195:1.000	165:0.5247	180:0.4167	210:0.3543
-----------	------------	------------	------------

Figure 4-5 The fingerprint of peak 35 with the corresponding normalized areas is shown above. Each normalized area is greater than 0.3000. The threshold value was determined based on collaboration with the FDA.

Equation 4-8 The normalized peak information is a two-element vector of mass and normalized area for all the  $i$  masses. The normalized area is found by dividing the largest area in the mass spectrum.

$$N = \left( mass, \frac{area_i}{\max area_1} \right) \text{ for } i \text{ valued pairs in the mass spectrum}$$

Equation 4-9 The fingerprint is a two-element vector of the peak information from the normalized peak information where the normalized area is greater than 0.3000. The mass and the corresponding normalized peak area are stored pairwise in the vector.

$$F = \{N_x | N_y > 0.3000\}$$

To perform the analysis, consider two mass spectrometry signals  $S_1$  and  $S_2$  that represent the information described in Figure 4-3 and Equation 4-7. The maximum peak areas in both  $S_1$  and  $S_2$  are the first valued pairs in the signal. The normalized signals  $N_1$  and  $N_2$  are found using Equation 4-8 as well as the fingerprints  $F_1$  and  $F_2$  using Equation 4-9. To quantify each fingerprint, the amount of normalized area for each fingerprint mass is squared and summed together. Equation 4-10 describes the quantification of the fingerprint that contains  $k$  elements.

Equation 4-10 To quantify the fingerprint of a compound, the magnitude of the normalized peak area is squared and summed for each  $k$  mass in the fingerprint.

$$E = \sum_{j=1}^k |F_y|^2$$

Now that the normalized signals, fingerprints, and a value to quantify the fingerprint are established, comparisons between the signals can be made. There are two comparisons that are to be done. One comparison is between  $S_1$  to  $S_2$ . The other comparison is between  $S_2$  to  $S_1$ . First,  $S_1$  is the “base” signal and is compared to  $S_2$ . For  $S_2$  to be considered a match to  $S_1$ ,  $N_2$  must contain all the masses in  $F_1$ . If each mass is found in the two vectors, then the difference of the normalized areas for each element is calculated. The difference has to be less than or equal to 0.3500 based on consultation with the FDA chemist expert. If the peak area difference is greater than 0.35, then  $S_2$  is not considered a match and returns -1. To quantify the differences of the two mass spectra, the difference of the two is squared and summed for all the  $k$  elements in  $F_1$  as shown in Equation 4-11.

Equation 4-11 Each of the masses in the fingerprint of  $S_1$  is compared with the normalized peak information of  $S_2$ . Each mass of  $F_1$  **must** be in  $N_2$ . If every mass is found, then the absolute difference between the normalized peak areas is calculated, squared, and summed for all  $k$  elements in  $F_1$ . This is to quantify the peak area ratios in each mass spectrum.

$$D = \sum_{j=1}^k |\{d_j | F_{y,k} - N_y < 0.35\}|^2$$

Finally, to quantify the total peak information the sum of differences from Equation 4-11 is divided by the quantification of the fingerprint.

Equation 4-12 To quantify the peak information, the values from Equation 4-10 and Equation 4-11 are divided to get the value  $V$ . The value  $V$  is used to determine the result of the compared mass spectra. The value returned is also used for the certainty of two matching compounds as listed in Table 4-5.

$$V = \frac{D}{E}$$

The same procedure is done except  $S_2$  is compared to  $S_1$ . There are three outcomes that can occur; (1) both  $V_{1to2}$  and  $V_{2to1}$  are positive values ( $\mathfrak{R}^+$  which includes zero) where the certainty is calculated in Equation 4-13, (2)  $V_1$  returns a value and  $V_2$  is -1, and (3)  $V_1$  is -1 and  $V_2$  is a value.

The table below dictates the results of the two signals  $S_1$  and  $S_2$ .

Equation 4-13 The certainty percentage measures how similar the mass spectra are with one another. It is calculated by dividing the minimum value  $V$  by the maximum value  $V$  generated when comparing the mass spectra together. When the fingerprints are the same in both spectra, the certainty will return a high percentage, otherwise, if there are spectra that have different fingerprints but still have the masses within them, then the certainty percentage would be lower.

$$Certainty(\%) = 100 * \frac{\min V_1, V_2}{\max V_1, V_2}$$

Table 4-5 Result of algorithm based on the value  $V_1$  and  $V_2$ . The value  $V$  can return a -1 if not all of the fingerprint mass to ion ratios are identified in the compared normalized signal, or if the difference in the normalized peak areas are greater than 0.3500. Otherwise, a certainty percentage will be outputted with the result of matching compounds.

$V_{1to2}$	$V_{2to1}$	Result
$\mathfrak{R}^+$	$\mathfrak{R}^+$	Match with some certainty percentage (%)
$\mathfrak{R}^+$	-1	Coelution
-1	$\mathfrak{R}^+$	Base Coelution

A real positive value indicates that each mass that was in  $F$  can be found in  $N$ , and that the peak area's difference was less than 0.35. When a -1 is returned, it means that either not all the masses of  $F$  were found in  $N$  or that the difference between the fingerprints mass and the normalized signal was greater than 0.35. When two real values are returned, that indicates that both fingerprints were found in both signals. If there is a -1 paired with a real number, there is a



possibility that there may be coeluting compounds. The fingerprint could be found in the mass spectrum in one direction but not in the other.

Using the relative retention times discussed in 4.1.1, the algorithm compares signals that are within 5% change of the PRRT and SRRT of the “base” compound or  $S_I$ . Outcomes of the algorithm and other observations are discussed in Chapter 5 under section 5.1.

How to Find

## **4.2. Patterns in the Gas Chromatograms Utilizing Statistical Methods**

An important goal of this research is not only to be able to identify matching compounds, but to also find patterns across the entirety of the data matrix. There are twenty-four windows across the data as explained in section 4.1.1 where the same process introduced here can be performed for each window. The relative retention times of each location of the GCxGC image is calculated. Each RRT corresponds to a compound. The goal is to use unsupervised techniques to find unbiased relations in the data set. Principal component analysis is employed to find the variance between each GCxGC image corresponding to a sample.

### **4.2.1. Performing the Principal Component Analysis Against Areas in the GCxGC Image**

To see the variance between the images of each sample, the principal component analysis was performed for a window as it is the current state of the art method that prioritizes the variance between observations. The PCA is a method that reduces the dimensionality of a data set that contains a large number of variables while maintaining the variance of the original data set. The new axes that represent the data are called principal components (PCs).

Say that there are  $p$  random variables and that the covariances or correlation between the variables are of interest [1]. Instead of looking at  $\frac{1}{2}p(p - 1)$  correlations or covariances, few derived variables can be calculated that preserve the information. This is done by finding the

eigen values and vectors of the data matrix. The largest eigenvalue corresponds to the first PC, the second largest the second PC, and so on until all of the variance is accounted for. It is expected for a large value of  $p$  that the number of components  $m$  is much smaller,  $m \ll p$ .

MATLAB software is used to calculate the PCA of the GCxGC image of a window [1]. Each unique PRRT and SRRT location in the window has an area associated with it. Each area is normalized based on the window target POP's abundance. After this normalization, the PCA produces two matrices. One matrix is the coefficients of each PC that translate the  $p$  variables to  $m$  PCs. The other is the score matrix where a sample is described by each  $m$  variables. The percentage of variance PC represents for the entire data matrix is also outputted.

The scree describes how each PC contributes to the total variance. According to Principal Component Analysis, Second Edition, if a set of variables have a substantial correlation among them, then the first few components will account for most of the variation of the original variables. The last few PCs identify directions in which there is very little or near-constant linear relationships among the original variables.

The scores are the translation of values based on the PCs. The coefficients describe how much weight a variable  $p$  contributes in describing  $m$  variables. The score of the file is based on the PCs. The coefficients multiplied based on each variable calculates to the score of a sample file.

Equation 4-14 The GCxGC image is composed of compound abundances. Each indicated compound from the data file are normalized based on the target compound  $k$  that defines the end of a window, similar to the relative retention times.

$$\text{normalized abundance} = \frac{\text{each compound abundance listed for a file}}{\text{target compound abundance for window } k}$$

Equation 4-15 The score of a sample  $f$  is calculated from the dot product of the coefficient vector and the normalized abundance of the sample. The coefficient of variable location  $p$  is multiplied by the normalized abundance found at that location to return the  $m$  score of sample  $f$ .

$$s_f = \sum_{j=1}^m \text{coeff}_{p,f} * \text{normalized abundance}_{p,f}$$

Since the PCA is performed on the GCxGC image, the coefficients dictate which locations are most prominent in describing the PC. If the coefficients are normalized for each PC based on the maximum coefficient value, the weights of each location can be calculated and graphed. Understanding which locations play a role in the describing the variance can indicate compound locations that have a large abundance that are not being targeted, compounds that have an area that is not heavily weighted as they are all similar, and other patterns can be explored.

Equation 4-16 The normalized coefficient vector, that has  $p$  elements, is calculated by taking each  $p$  coefficient dividing by the maximum coefficient for a principal component.

$$nc_p = \frac{\text{coeff}_p}{\max \text{coeff}_p}$$

#### 4.2.2. Method for Clustering Scores from the Principal Component Analysis

The scores matrix of the PCA relates the sample files to the PCs. To associate one value to the sample, the  $L_2$  normalization of each sample for each score was taken. This is finding the distance away from the origin to the score of the file based on all the  $n$  components. Each score is represented for sample  $i$  and component  $k$  by  $x_{i,k}$  and is summed and square rooted to the resulting score of  $X_i$ .

Equation 4-17 For  $m$  principal components found from the PCA, each  $m$  score is squared. The sum of each squared score is then square rooted to return a value for sample number  $i$ . This formula is referred to as the  $L_2$  normalization.

$$X_i = \sqrt{\sum_{k=1}^m |x_{i,k}|^2}$$

An unsupervised clustering technique commonly used is *K-Means Clustering* [2]. It is an iterative algorithm that sorts each observation into exactly one cluster. Each cluster is defined by a centroid. The number of clusters must be given before the data-partitioning can occur. The algorithm randomly chooses the centroid location for each cluster. Next, the distance of each point to the cluster centroids is computed. Each observation is then assigned to a cluster with the minimal distance. The average of the observations in each cluster is calculated. Each centroid is then placed in their clusters' average location. The process continues with the new centroid location except each observation can be individually assigned to a different centroid if the sum of distances would be reduced. Each cluster centroid is then recomputed after the reassignments. At the end of the process, data points are grouped together based on how similar their observations are.

To find the optimal value of clusters, the *Silhouette Method* [3] is employed. The silhouette method calculates a silhouette value for each point and determines similar points are to each other in the same cluster. If the points have a high silhouette value, then the clustering solution is appropriate otherwise there may be too many or too few clusters indicated. To calculate the silhouette value, two variables are attained. First the average distance  $a_i$ , which is calculated from the  $i$ th point to each point in the same cluster. Then the minimum average distance  $b_i$  is calculated. The average distance from the  $i$ th point to each point in a cluster is minimized depending on the clusters.

Equation 4-18 The silhouette value for point  $i$  is calculated by taking the difference of the smallest average distance of the  $i$ th point from all the points in a singular cluster ( $b_i$ ) and the average distance of the  $i$ th point to each point in the same cluster ( $a_i$ ), divided by the maximum of the two averages. The silhouette value can range from -1 to 1 where a high value near 1 indicates that the points are clustered appropriately.

$$SV_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}$$

An evaluation tool utilizes this method and returns the optimal number of clusters for K-means based on how many points have silhouette values. Since a random location is used for the centroids in the first iteration, it is important to repeat the clustering and use the most usual form. MATLAB allows ‘replicates’ as an option so it can be done internally.

The distance measurement described for the K-Means clustering is referred to as the Squared Euclidean distance, where each centroid is the mean of the points in the cluster. Another option is the sum of absolute differences where the centroid is the component-wise median of the points in that cluster. Others include cosine, correlation, and hamming distances [2].

The squared Euclidean distance was evaluated against distance measurements. To represent how well the distance represented the data, a hierarchical cluster tree was created. Each tree had a corresponding cophenetic correlation coefficient, which measures the quality of the solution. When the magnitude of the value is close to one, then the solution is of high-quality. The value is calculated by MATLAB’s ‘cophenet’ function. The best representation of the data is using the sum of absolute differences as the distance measurement.

Once the clusters are established, each sample can then be compared to other samples within the same clusters. To understand how the PCA was able to give a value to each sample can determine if there are similarities or differences between each file. Explorations on the clusters and determinations of methods that can utilize the results will be discussed in Chapter 5.

## CHAPTER 5: RESULTS

First, we evaluated the mass spectral data in chapter 4, and only one window was utilized for analysis. In this case, *Window 5*, which is between PCB-101 and PCB-123 was the focus of this study. This case study was determined based on consultation with the FDA as a representative window where dominant peaks, target and (potentially) non-target, would be present. Out of the seventy-two samples included in this study, there are 63,501 compounds identified in this window. The window was selected in part because there is an already known contaminant that has been identified in the window. This analyte, dichlorodiphenyldichloroethylene (DDE), provides a test case for the algorithm to determine if something is found correctly within the evaluation, as the compound has been verified in by the FDA as being present. The applications for this window can be applied to other windows in future work. Secondly, we discuss the trends and methodologies utilizing the principal component analysis (PCA) data generated from the GCxGC images of window 5 for each sample. The  $L_2$  normalized score of each sample is clustered. The coefficients and scores from each principal component (PC) are investigated and a function was created to pull relative retention time locations that may be of interest to the FDA to analyze data across study samples . All the locations are defined by relative retention times. Finally, discussion of potential future work is assessed at the end of the chapter.

### **5.1. Evaluation of Mass Spectra Comparison Algorithm**

The algorithm discussed in 4.1.2 was evaluated and applied to all the compounds identified in window 5. Many matches, coelution, and base coelutions were noted as a result. The results that occurred when compound Unknown 3744 was used as the ‘base compound’ was analyzed and verified in consultation with the FDA. Unknown 3744 was chosen as the ‘base compound’ because it was established to be the non-targeted DDE compound present.

The mass spectrum of the compound is found in Figure 5-1. The masses and the corresponding peak areas are listed in Figure 5-2. The fingerprint of Unknown 3744 is also presented in Figure 5-3. Recall the fingerprint is where each of the normalized peak heights are greater than 0.3000. The red box in Figure 5-1 outlines the threshold for the fingerprint masses.

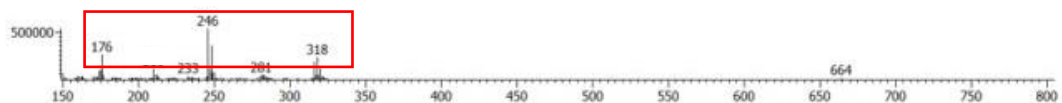


Figure 5-1 The mass spectrum of Unknown 3744 is shown above. This compound is the ‘base compound’ used for comparisons. The red box indicates the masses that are in the fingerprint of Unknown 3744. The x-axis represents the mass to charge ratio as the y-axis represents the abundance of each mass.

246:543365 248:351744 176:261545 318:226714 316:175286 320:111661 247:101330
210:101018 175:95305 174:84550 249:60057 250:59847 150:53046 212:42635 281:37743 283:36483
177:35867 211:34530 282:33649 319:31988 317:29196 170:27845 172:25401 160:24443 280:24022
322:23051 245:21652 163:21099 233:20386 321:16412 173:15753 162:15409 284:14547 285:14047
184:13960 235:13495 161:12884 213:12745 186:9702 151:9191 251:8463 209:8157 185:8095
199:7211 164:7144 171:6197

Figure 5-2 The mass spectrum data collected in the peak true of Unknown 3744. Each mass and their corresponding peak area are separated by a colon.

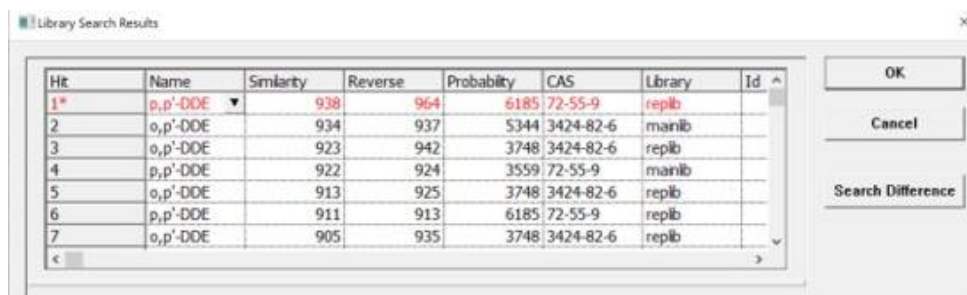
246:1.000 248:0.6473 176:0.4813 318:0.4172 316:0.3226 320:0.2055 247:0.1865 210:0.1859
175:0.1754 174:0.1556 249:0.1105 250:0.1101 150:0.0976 212:0.0785 281:0.0695 283:0.0671
177:0.0660 211:0.0635 282:0.0619 319:0.0589 317:0.0537 170:0.0512 172:0.0467 160:0.0450
280:0.0442 322:0.0424 245:0.0398 163:0.0388 233:0.0375 321:0.0302 173:0.0290 162:0.0284
284:0.0268 285:0.0259 184:0.0257 235:0.0248 161:0.0237 213:0.0235 186:0.0179 151:0.0169
251:0.0156 209:0.0150 185:0.0149 199:0.0133 164:0.0131 171:0.0114

Figure 5-3 The normalized peak information of Unknown 3744. Each mass is paired with their normalized area, separated by a colon. The normalized areas were found by dividing each peak area by 543,365.

246: 1.000 248: 0.6473 176: 0.4813 318: 0.4172 316: 0.3226

Figure 5-4 Fingerprint of Unknown 3744 are the masses and their normalized area that are separated by a colon.

The library search indicated that the compound is a p-p'-DDE. The p-p'-DDE has a specific structure as shown in Figure 5-6. The structure of the compound has a high certainty and likely is a p-p'-DDE according to the FDA. Two standards were purchased to verify the structure for Unknown 3744.



Hit	Name	Similarity	Reverse	Probability	CAS	Library	Id
1*	p,p'-DDE	938	964	6185	72-55-9	replib	
2	o,p'-DDE	934	937	5344	3424-82-6	manib	
3	o,p'-DDE	923	942	3748	3424-82-6	replib	
4	p,p'-DDE	922	924	3559	72-55-9	manib	
5	o,p'-DDE	913	925	3748	3424-82-6	replib	
6	p,p'-DDE	911	913	6185	72-55-9	replib	
7	o,p'-DDE	905	935	3748	3424-82-6	replib	

Figure 5-5 The library search of Unknown 3744 performed by the FDA indicated that the compound structure is likely p-p'-DDE

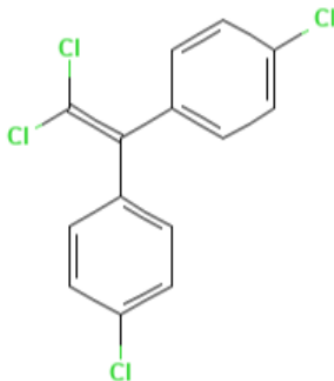


Figure 5-6 The structure of p-p'-DDE compound [15]

To assess the performance of the algorithm, the experienced analyst from the FDA evaluated each of the comparisons manually. When the algorithm was processed, both the PRRT and SRRT had to be 5% or less from PRRT and SRRT of Unknown 3744. The peak area



differences were set to a threshold of 0.3500. This value was set in consultation with the FDA. If the difference was greater than 0.3500, the compounds were not considered a match and the algorithm returned a -1. Comparing Unknown 3744 as the base compound, 104 other compounds from different files were identified as either a match, coeluted, or base coeluted.

Through collaboration with the FDA, the mass spectra were first visually compared. The compounds break into the same fragments when processed so the spectra are expected to contain similar peaks and similar ratios between peaks. If the spectra were considered to be a chemical match, then the relative retention times were verified to be within the proper range. After this was confirmed, the predicted structure of the compound was considered based on a library search for the compound.

### **5.1.1. Analytes that Returned Base Coeluted from the Mass Spectra Comparison**

#### **Algorithm for Unknown 3744**

First, the performance of the base coeluted labeled compounds was analyzed. There were eight compounds classified as base coeluted. Recall that base coelution is when the  $V_{2to1}$  returned a real value however  $V_{1to2}$  returned -1. Here two unknowns, 3315 and 3244 were identified as base coeluted. At first glance, both were seen as a match by the analyst from the FDA. The data had been reprocessed by the time the evaluation of the algorithm was done. Unfortunately, the mass spectrum of Unknown 3244 could not be located. For verification purposes, only Unknown 3315 will be presented. The mass spectrum of Unknown 3315 is shown below.

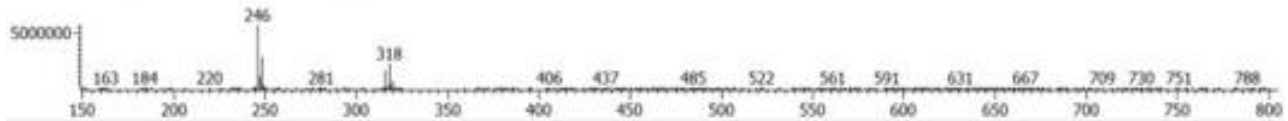


Figure 5-7 The mass spectrum of Unknown 3315 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. In the peak true signal, mass from the fingerprint of Unknown 3744 include 246, 248, 316, and 318.

Unknown 3315 has 4 out of the 5 masses from the fingerprint of Unknown 3744. Masses 246, 248, 316, and 318, are in the peak true spectrum of Unknown 3315, however, mass 176 is not. Therefore, it is possible that Unknown 3744 was coeluted with a different compound that has a molecular weight of 176 and Unknown 3315. The compound was considered a potential match based on visual analysis of the mass spectra. Next the RRT were verified by finding the percent change of the RRT from the base compound. For unknown 3315, the PRRT had -0.083% change while the SRRT had -1.22%. The RRT were within range. Next a library search of the compound's structure was performed. The compound was determined to have a chemical structure of an o-p'-DDE instead of a p-p'-DDE. The difference is the location of a chlorine atom as shown in para position in Figure 5-6 and ortho position in Figure 5-8.

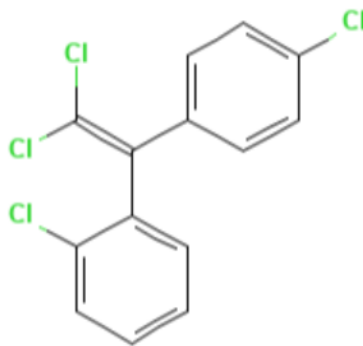


Figure 5-8 The structure of an o-p'-DDE [15]

Both structures have the same molecular weight. Typically, the retention time of the o-p'-DDE occurs after the p-p'-DDE. The results of the percent change correspond with this common

occurrence. To verify that Unknown 3315 is a match, a compound standard would need to be obtained and processed through the GCxGC-ToF-MS instrument.

The other six compounds identified were not considered a match. Two compounds, Unknown 2584 and Unknown 4095, visually were similar as there was likely a slight residual match with an interference. This residual amount does not confirm coelution, however presence is notable for further investigation by the lead chemist in consultation.

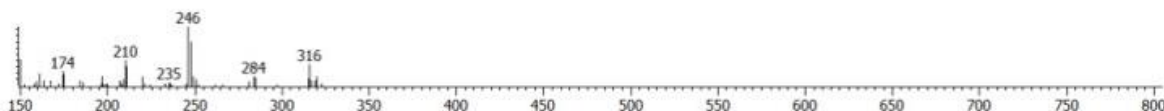


Figure 5-9 The mass spectrum of Unknown 2584 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Peaks from the fingerprint of Unknown 3744 include 246, 248, and 316.

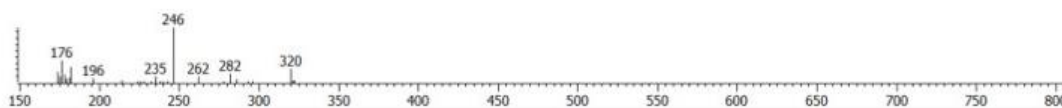


Figure 5-10 The mass spectrum of Unknown 4095 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Masses that are in the fingerprint of Unknown 3744 include 246 and 176.

The remaining four compounds were not determined to be a match by the FDA chemist. They are Unknowns 2725, 2606, 3560, and 3681. Each of the signals contained mass 246, however other fingerprint masses were not included and therefore compounds were determined to not be a match.

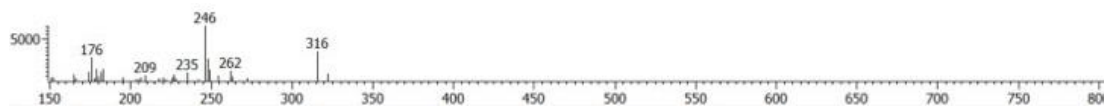


Figure 5-11 The mass spectrum of Unknown 2725 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Masses from the fingerprint of Unknown 3744 are 246, 316, and 176.

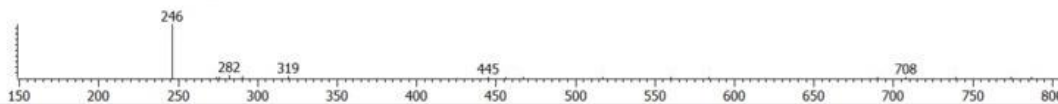


Figure 5-12 The mass spectrum of Unknown 2606 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Compared to the fingerprint of Unknown 3744, only mass 246 is in the peak true.

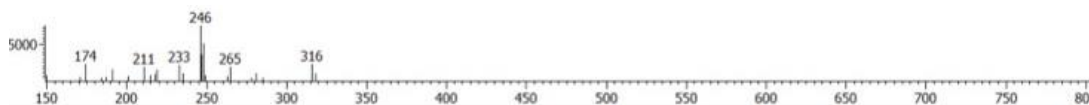


Figure 5-13 The mass spectrum of Unknown 3560 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Compared to the fingerprint of Unknown 3744, mass 246, 248, 316, and 318 are included.

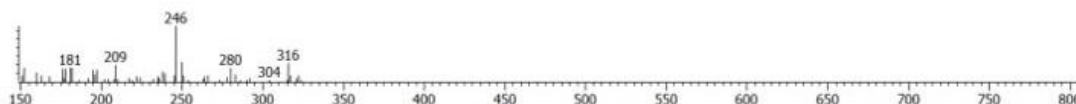


Figure 5-14 The mass spectrum of Unknown 3681 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Compared to the fingerprint of Unknown 3744, mass 246 and 318 appear in the peak true signal.

Overall, the base coelution results were not conclusive to finding coelutions. A coelution is when two or more compounds have the same retention time or co-elute from the analytical column at the same time. Two compounds had slight residuals in the mass spectrum. Unknown 3315 was similar enough to be considered a match, though it did not have the correct structure. Out of the 104 compounds identified in comparison to Unknown 3744, only 7.7% were labeled as base coelution. The low amount of base coelution results is to be expected as the base compound was known to be DDE. There is little to no interference of the base compound mass spectrum signal. The algorithm was successful in targeting mass spectra that have similar retention times that also have some, but not all the masses in the fingerprint. The algorithm performed as intended, however further investigation is necessary to verify coeluting compounds.

### 5.1.2. Analytes that Returned Coelution from the Mass Spectra Comparison Algorithm for Unknown 3744

Like base coelutions, coelutions are labeled when  $V_{1to2}$  returns a real value and  $V_{2to1}$  returns -1. The algorithm found nineteen coeluted compounds in window 5. After consultation with the FDA, only one was identified to be a match. Unknown 3456 came from the SPIKE file that is used for quality control purposes in the analyses of milk samples. The SPIKE refers to the blank matrix where the targeted analytes were added. This is used for quality control purposes to verify extraction efficiency in the target analysis.

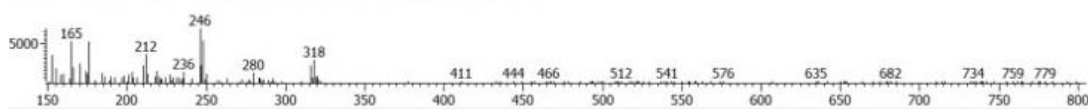


Figure 5-15 The mass spectrum of Unknown 3456 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Compared to the fingerprint of Unknown 3744, mass 246, 248, 176, 318 and 316 are found. Other masses that are not in Unknown 3744 include masses 212 and 165 so coelution was notated for this compound by the algorithm.

The peak true of Unknown 3456 shows each mass in the fingerprint for Unknown 3744. Examples of masses included in the fingerprint for Unknown 3744 that also appear in Unknown 3456 include 165 and 212. Through consultation with the FDA, it was concluded that Unknown 3744 was present at a residual amount. Next, the RRT percent change was investigated. The percent change for the PRRT was 0.194% and the SRRT was -1.46%. However, the library search indicated the o-p'-DDE structure. Recall that further investigation is necessary to determine differences between the o-p'-DDE structure and p-p'-DDE structure.

The remaining nineteen were concluded to not be a match. There were eleven that had no additional comments by the experienced analyst; Unknown 3381, Unknown 2723, Unknown 2741, Unknown 2760, Unknown 3855, Unknown 3456, Unknown 3476, Unknown 2466,

Unknown 3914, Unknown 3102, Unknown 3199, and a twelfth Unknown 3115. The mass spectrum of 3115 was not accessible due to reprocessing of the GCxGC data. The remaining mass spectra can be found in Appendix A.

From the remaining seven compounds identified as coeluting, Unknown 3181 and 3270 seemed to have a slight residual of Unknown 3744 with an interference of another compound. They were not confirmed as the same compound manually. The compounds have some masses from the fingerprint in Unknown 3744, however, there are clearly more masses in the mass spectra of Unknown 3181 and Unknown 3270

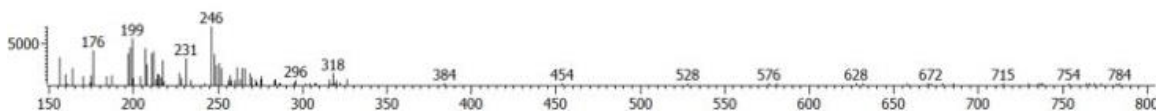


Figure 5-16 The mass spectrum of Unknown 3181 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Visually there seems to have a slight residual amount of Unknown 3744 as all the peaks within the fingerprint are accounted for.

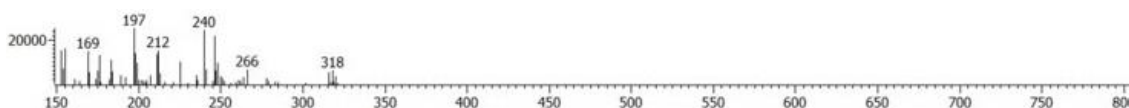


Figure 5-17 The mass Spectrum of Unknown 3270 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. Visually there seems to have a slight residual amount of Unknown 3744 as all the peaks within the fingerprint are accounted for.

The remaining five compounds were identified to have mass spectra matches with exceptionally low intensity. The PRRTs percent change was near the 5% threshold, being at least 4% or more. A match from the same sample was also identified by the algorithm. The mass spectra of Unknown 2632, Unknown 1427, Unknown 3500, Unknown 3526, and Unknown 3620 are shown below.

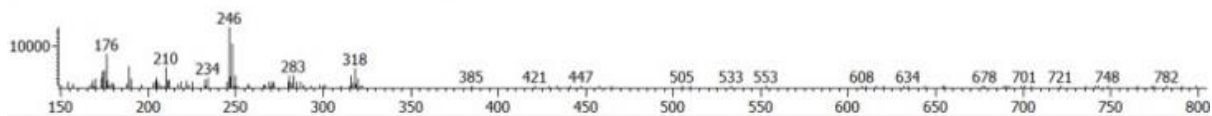


Figure 5-18 The mass spectrum of Unknown 2632 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal visually represents a match to Unknown 3744 as each mass in the fingerprint is found and have proper ratios. However, the intensity of the peak true signal is low.

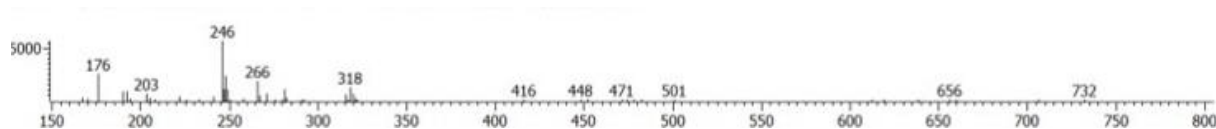


Figure 5-19 The mass spectrum of Unknown 1427 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal visually represents a match to Unknown 3744 as each mass in the fingerprint is found and have proper ratios. However, the intensity of the peak true signal is low.

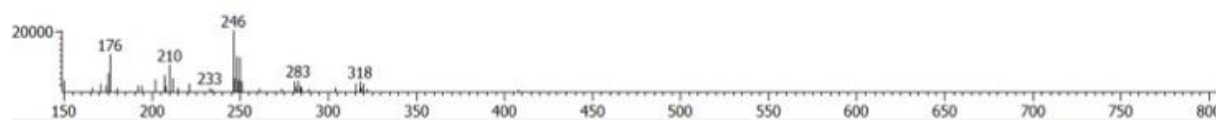


Figure 5-20 The mass spectrum of Unknown 3500 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal visually represents a match to Unknown 3744 as each mass in the fingerprint is found and have proper ratios. However, the intensity of the peak true signal is low

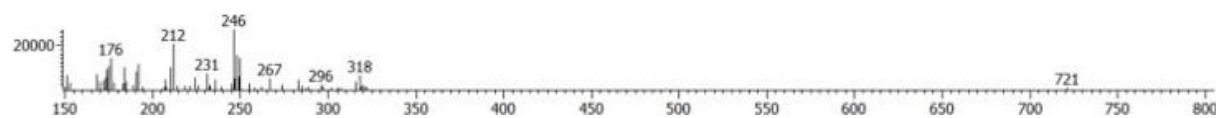


Figure 5-21 The mass spectrum of Unknown 3576 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal visually represents a match to Unknown 3744 as each mass in the fingerprint is found and have proper ratios. However, the intensity of the peak true signal is low

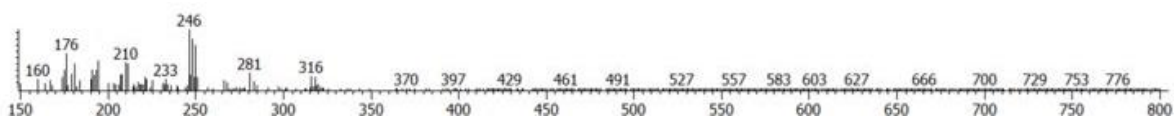


Figure 5-22 The mass spectrum of Unknown 3620 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal visually represents a match to Unknown 3744 as each mass in the fingerprint is found and have proper ratios. However, the intensity of the peak true signal is low

Overall, the coelutions were successful in targeting compounds that contained the fingerprint of Unknown 3744 in their mass spectra. The fingerprint of coeluted compounds had more masses than Unknown 3744. At least one of the masses was not found in the fingerprint of Unknown 3744. Further verification needs to be performed to verify if coelution occurred. The last five compounds each came from a sample that already had a match confirmed by the algorithm. The residue in the mass spectra had a low intensity. Discussion of the identified matches will be presented in the next section.

### **5.1.3. Analytes that Returned a Match from the Mass Spectra Comparison Algorithm for Unknown 3744**

There were seventy-seven matches identified by the algorithm. Recall that Unknown 2632, Unknown 1427, Unknown 3500, Unknown 3526, and Unknown 3620 were indicated to be coeluting compounds by the algorithm. Each unknown had a match previously identified in the same sample. They are Unknown 2421, Unknown 1357, Unknown 3219, Unknown 3296, and Unknown 3324. There were multiple iterations of the dataset. The mass spectrum of Unknown 3296 was lost as the data was reprocessed and therefore is not represented below, however, the remaining four mass spectra are shown below.

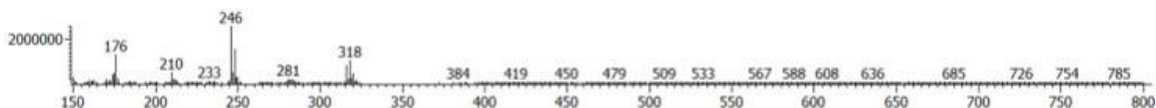


Figure 5-23 The mass spectrum of Unknown 2421 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal is verified to be a match with a high intensity to Unknown 3744.



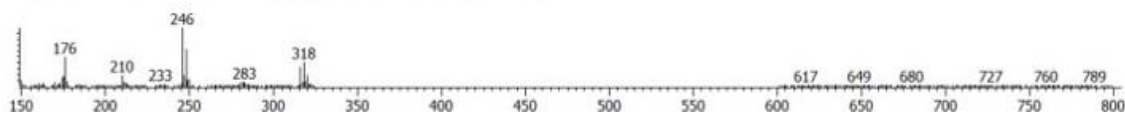


Figure 5-24 The mass spectrum of Unknown 1357 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal is verified to be a match with a high intensity to Unknown 3744.

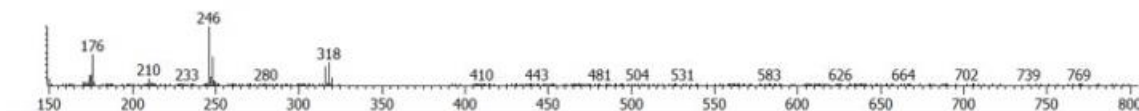


Figure 5-25 The mass spectrum of Unknown 3219 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal is verified to be a match with a high intensity to Unknown 3744.

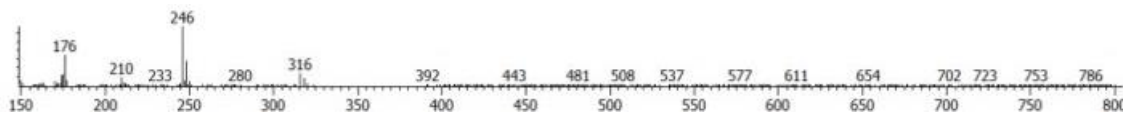


Figure 5-26 The mass spectrum of Unknown 3324 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal is verified to be a match with a high intensity to Unknown 3744.

Each of the mass spectra are clearly the same compound through visual analysis. The fingerprints of the spectra above each contain peaks at 246, 248, 176, 318, and 316. The ratio between the peak areas are also similar to Unknown 3744. Besides the extra masses in the coeluted unknowns, a factor that should be considered is the percent change in the PRRT. In Table 5-1, Unknown 2632, Unknown 1427, Unknown 3500, Unknown 3526, and Unknown 3620 are compared with the matching compound from the same sample. Unknown 2421, Unknown 1357, Unknown 3219, Unknown 3296, and Unknown 3324 are the matching compounds respectively from the same sample. The matching unknowns each had a percent change in the PRRT less than 1% and in the SRRT less than 1.5%. The PRRT percent change in each of the coeluting compounds were all greater than 4%. As a result, the percent change parameter for the PRRT can be reduced.

Table 5-1 The RRTs percent change is compared between the identified coeluted (potentially) compounds and matching compounds from the same sample based on the results of the comparison algorithm. Unknown 3744 where the Unknown 2632, Unknown 1427, Unknown 3500, Unknown 3526, and Unknown 3620 correspond to the coeluting compounds as Unknown 2421, Unknown 1357, Unknown 3219, Unknown 3296, and Unknown 3324 correspond to the matching compounds. Each row includes unknowns that belong to the same file

Coeluting Unknowns	PRRT % Change	SRRT % Change	Matching Unknowns	PRRT % Change	SRRT % Change	% Certainty
2632	4.263%	1.19%	2421	0.679%	1.19%	95.9%
1427	4.263%	0.00%	1357	0.679%	1.19%	97.8%
3500	4.467%	1.22%	3219	0.267%	0.00%	88.4%
3576	4.300%	0.00%	3296	0.253%	0.00%	56.6%
3620	4.474%	0.00%	3324	0.253%	1.23%	50.1%

Out of the remaining seventy-two matches, there were five compounds that were algorithmically identified as a match, but after consultation with the FDA, were not verified as such. Unknown 2911 and Unknown 3495 both had mass spectra matches at low intensities. Notably, the PRRT of each was greater than 4% and another match was found previously in the same sample. Unknown 2698 and Unknown 3230 are the better matching compounds from the same sample respectively. The values of the percent change of the PRRT, SRRT and the certainty measurement are compared in Table 5-2.

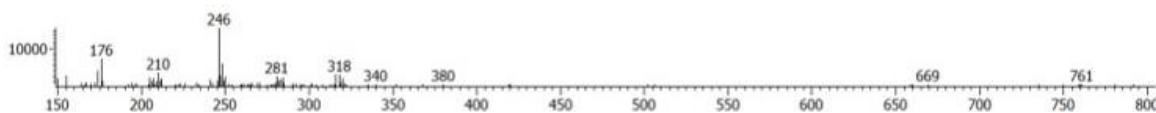


Figure 5-27 The mass spectrum of Unknown 2911 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal visually looks similar to Unknown 3744 at low intensities. The primary relative retention time of the compound is greater than 4% and determined not to be a match.

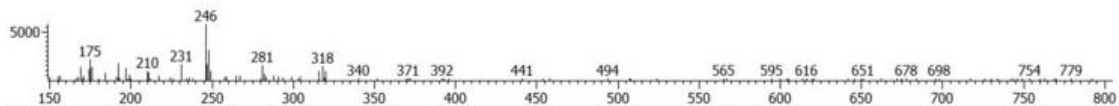


Figure 5-28 The mass spectrum of Unknown 3495 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The peak true signal visually looks similar to Unknown 3744 at low intensities. The primary relative retention time of the compound is greater than 4% and determined not to be a match.

Table 5-2 The percent change of the PRRT and SRRT from the Unknown 3744 are compared between Unknown 2911, Unknown 3495, Unknown 2698, and Unknown 3230 as well as the certainty measurement. Unknown 2911 and Unknown 3495 both had a mass spectra mass of low intensity. Unknown 2698 and Unknown 3230 are from the same samples as the other unknowns respectively. The values that have low percent change in RRT. Another indicator is that the PRRT should be at least less than 4%.

Low Intensity Unknown	PRRT % Change	SRRT % Change	% Certainty	High Intensity Unknown	PRRT % Change	SRRT % Change	% Certainty
2911	4.676%	0.00%	72.7%	2698	0.167%	1.39%	98.7%
3495	4.255%	1.19%	44.2%	3230	0.518%	1.19%	96.2%

From Table 5-2, the certainty percentage calculated by Equation 4-13 are greater than 95% for Unknown 2698 and Unknown 3230. Unknown 2911 and Unknown 3495 both had percentages less than 75% certainty. Notice that the PRRT of Unknown 2911 and 3459 are again greater than 4%. This is another indicator that the PRRT percent change can be reduced from 5%. The SRRT percent change was the same in Unknown 3495 and Unknown 3230. It is possible that the percentage threshold can also be reduced but should at least be 3.5%.

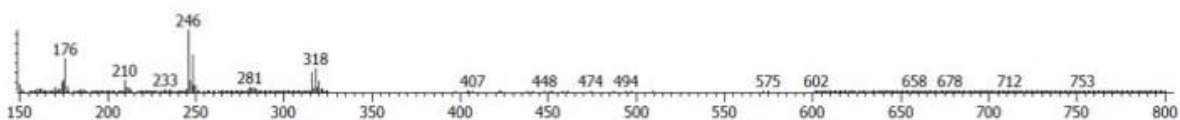


Figure 5-29 The mass spectrum of Unknown 2698 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The compound was confirmed to be a match with a certainty of 98.7%.

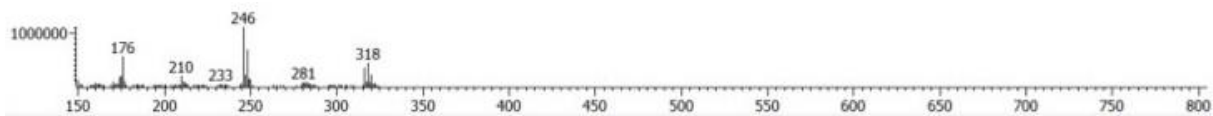


Figure 5-30 The mass spectrum of Unknown 3230 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The compound was confirmed to be a match with a certainty of 96.2%

The remaining three compounds that were algorithmically identified as a match, but after consultation with the FDA, were not verified as such include Unknown 3902, Unknown 3112, and Unknown 3588. Unknown 3902 does not visually have the same ratio of mass 176 as Unknown 3744. Unknown 3902 also had a PRRT and SRRT percent change of 4.446% and 1.20% and was verified not to be a match. Unknown 3112 also was not visually a match in the peak true signal. Though all of the masses in the fingerprint of Unknown 3744 were found, the ratios between each mass was not consistent with Unknown 3744. The PRRT and SRRT percent change of Unknown 3112 was 3.774% and 3.74% respectively. Similarly, Unknown 3588 was not visually equated to be a match because of the ratios in the mass spectrum. The PRRT and SRRT percent change of Unknown 3588 are 4.467% and 0.00%. Unknown 3588 and Unknown 3112 both came from the same sample. In fact, four other compounds were also identified by the algorithm also belonged to the same sample. Analysis of Unknown 3102 (from 5.1.2), Unknown 3112, Unknown 3297, Unknown 3315 (from 5.1.1), Unknown 3330, and Unknown 3588 was completed and summarized in Table 5-3

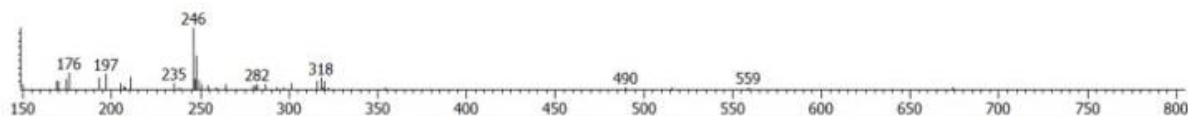


Figure 5-31 The mass spectrum of Unknown 3902 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The compound was not determined to be a match due to the relative retention times.

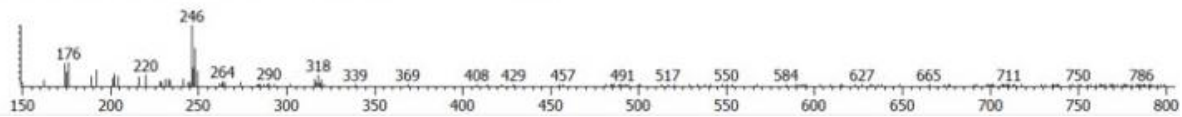


Figure 5-32 The mass spectrum of Unknown 3112 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The compound was not determined to be a match due to the relative retention times.

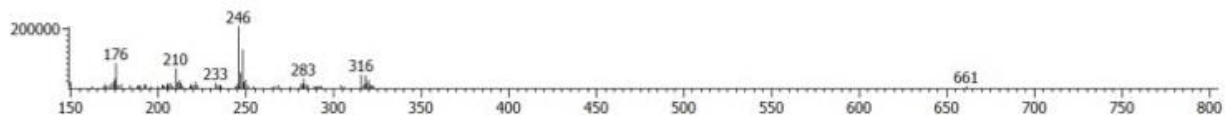


Figure 5-33 The mass spectrum of Unknown 3588 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The compound was not determined to be a match due to the relative retention times.

Unknown 3588 was remarkably similar to the peak true signal of Unknown 3744. Out of the six compounds that were from the same sample, three were identified to have a mass spectrum match visually through consultation with the FDA. This includes Unknown 3297, Unknown 3315, and Unknown 3330. Unknown 3297 sample was determined to be the correct match for the sample as it has the smallest percent change in the PRRT and the highest certainty measurement. The structure of Unknown 3297 was identified to be p-p'-DDE, like Unknown 3744. Unknown 3330 was identified to have an o-p'-DDE structure. As discussed earlier, the p-p'-DDE typically occurs before the o-p'-DDE structure and is illustrated with this example.

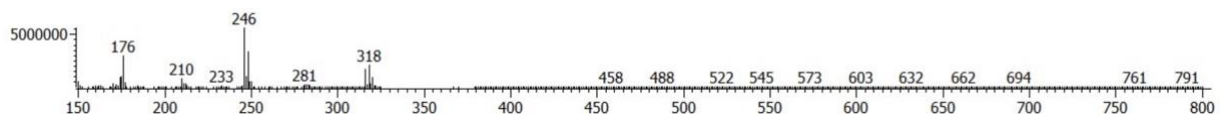


Figure 5-34 The mass spectrum of Unknown 3297 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The compound was the best match in the sample and has the p-p'-DDE structure.

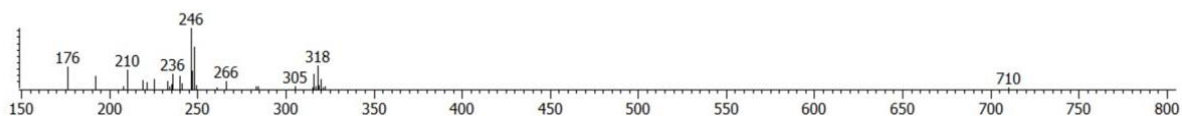


Figure 5-35 The mass spectrum of Unknown 3330 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The structure of the compound was o-p'-DDE and not verified to be a match for the sample.

The others that were not visually a match were Unknown 3102, Unknown 3112, and Unknown 3588. The percent change of the PRRT ranged from 3.74%-4.467%. More evidence that the percent change values of at least the PRRT should be reduced to 3.5% to limit compound comparisons. The SRRT can also be reduced as there were no compounds identified that had a SRRT percent change greater than 2% for all verified matches. Another change to consider algorithmically is to return the best match per sample. Each sample should only have one match. The best match would be selected based on the certainty measurements and the percent change of the RRT values.

Table 5-3 Unknown compounds from the same sample that were identified as either a coelution, match, or base coelution. The PRRT and SRRT percent change are both labeled with their matches that indicate the percent certainty. Those highlighted in green were labeled as a match visually by the chemist expert.

Unknown	Outcome	PRRT % Change	SRRT % Change	% Certainty
3102	Coelution	3.950%	3.74%	-
3112	Match	3.774%	3.74%	48.7%
3297	Match	0.267%	0.00%	98.1%
3315	Base Coelution	0.083%	1.22%	-
3330	Match	0.434%	0.00%	79.2%
3588	Match	4.467%	0.00%	34.9%

The remaining sixty-seven matches were all confirmed visually as a match. Each had a PRRT percent change of less than 1%. The structure of the unknown compounds were identified to be p-p'-DDE like Unknown 3744. Table 5-4 presents the minimum, maximum, and average statistics of the PRRT percent change, SRRT percent change, and the certainty percentage. The average PRRT and SRRT percent change was within 1% of Unknown 3744. Next, the certainty measurements were evaluated. The lowest percent certainty came from Unknown 2819. Recall that the certainty measurement is dependent on the fingerprints' quantification of the mass spectra and the difference in peak areas (Equation 4-13). For Unknown 2819, the fingerprint includes masses 246, 248, and 176. Masses 318 and 316 are not included in the fingerprint of Unknown 2819 as they are outside of the normalized threshold 0.3000. Because of this discrepancy, each of the masses are in both spectra however the value for  $E$  (Equation 4-10) will differ between  $V_{1to2}$  and  $V_{2to1}$ .

Table 5-4 The statistical measurements of the primary relative retention time percent change, secondary relative retention time percent change, and the certainty measurement are listed. The average percent change was less than 1% in the relative retention times and the average certainty was greater than 90%.

	Min	Max	Average
PRRT %	0.00%	0.80%	0.33%
SRRT %	0.00%	3.18%	0.83%
Certainty	19.1%	99.9877%	92.80%

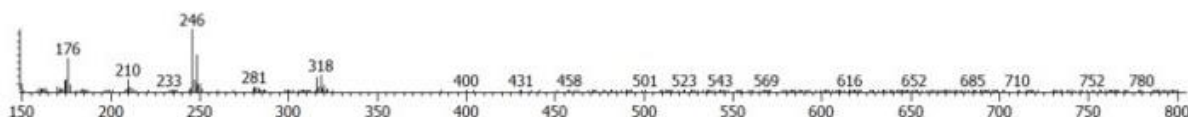


Figure 5-36 The mass spectra of Unknown 2819 where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. The certainty measurement was 19.1% because mass 316 and 318 were not quantified in the fingerprint.

All in all, the algorithm was successful in finding matches throughout a data set. Also compounds that have potential to be coeluting were also identified. Further analysis to confirm coelution would include purchasing a standard from the manufacturer and running the standard in the GCxGC-ToF-MS analyses. Most of the coelutions, base or otherwise, had a different chemical structure and were outside of the PRRT percent change windows. It was concluded that the matching algorithm performed as intended and better performance would occur if the percent change of the PRRT and SRRT are reduced. If the percent change of the PRRT was reduced in the PRRT to 3.5%, then a better performance would occur. Also, if multiple compounds are compared from the same sample and have a result, the algorithm can be modified to return the best match. Now that the algorithm is successful, patterns of the GCxGC image are explored in the next section.



## 5.2. Exploring GCxGC Images to Recognize Patterns Throughout the Samples

A comprehensive two-dimensional gas chromatography image is generated when each sample was processed in the GCxGC-ToF-MS instrument. In each image, there is a corresponding color that indicates the total abundance at the retention time pair. The abundance is equivalent to the summation of the peak areas in corresponding mass spectrum. The GCxGC image has two axes, the primary retention time (PRT) and the secondary retention time (SRT). The PRT ranges between 450 to 2300 seconds. The SRT ranges from 0 to 3 seconds. The purpose of this section is to explore patterns that (potentially) emerge in the GCxGC image. Instead of investigating the entirety of the image, each image is separated into *windows*, based on the target POPs. Similar to section 5.1, a case study for window five was investigated.

Window five is defined by the target compounds PCB-101 and PCB-123. The location of the target compounds has previously been identified during the target analysis of each sample. The location of PCB-123 was used to calculate the relative retention times (RRT), as expressed in 4.1.1, of each of the 63,501 analytes in the GCxGC image. Each RRT pairs were used to analyze the GCxGC image of window five. A data matrix was created for each unique RRT pair. If a sample has an area associated in the GCxGC image at the RRT location, the area was recorded. Otherwise, zero is stored in the data matrix. Afterwards, each of the areas were normalized based on the abundance of PCB-123 as discussed in section 4.2. The location of each RRT pair over the seventy-two samples is illustrated in Figure 5-37 in the discrete GCxGC image.

There are 51,143 unique RRT pair locations found in all seventy-two samples. The PCA was performed on the 72 by 51,143 data matrix and resulted in a 72 by 71 data matrix. The  $p$  variables of 51,143 was reduces the  $m$  variables of 71. Each of the  $m$  variables correlated to the principal components (PCs). The scree plot describes how the seventy-one PCs describe the

variance in Figure 5-87. The scree plot has an inflection point at PC 7. The PC 1-7 describe 55.4936% of the variance in the data set as PC 8-71 describes 44.4064%.

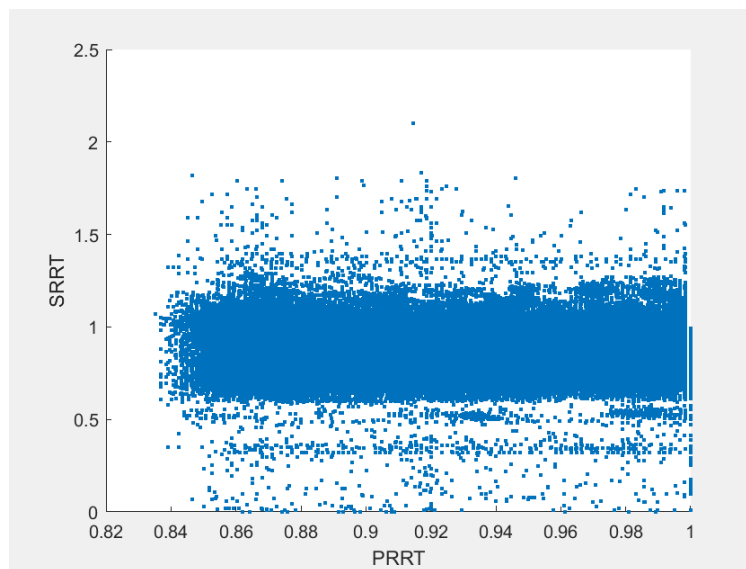


Figure 5-37 The Discrete GCxGC image of every unique relative retention time location (PRRT, SRRT) described over all the samples. There seem to be overlapping points that can be identified in the principal component analysis. Recall that there are 63,501 analytes stored identified in the GCxGC image of window five. There are 51,143 unique RRT locations found in all 72 samples.

The PCA was performed on the 72 by 51,143 data matrix and resulted in a seventy-two by seventy-one data matrix. A combination of the 51,143 locations was reduced to seventy-one different PCs. The scree plot describes how the seventy-one PCs describe the variance in Figure 5-37. The scree plot has an inflection point at PC 7. The PC 1-7 describe 55.4936% of the variance in the data set as PC 8-71 describes 44.4064%.

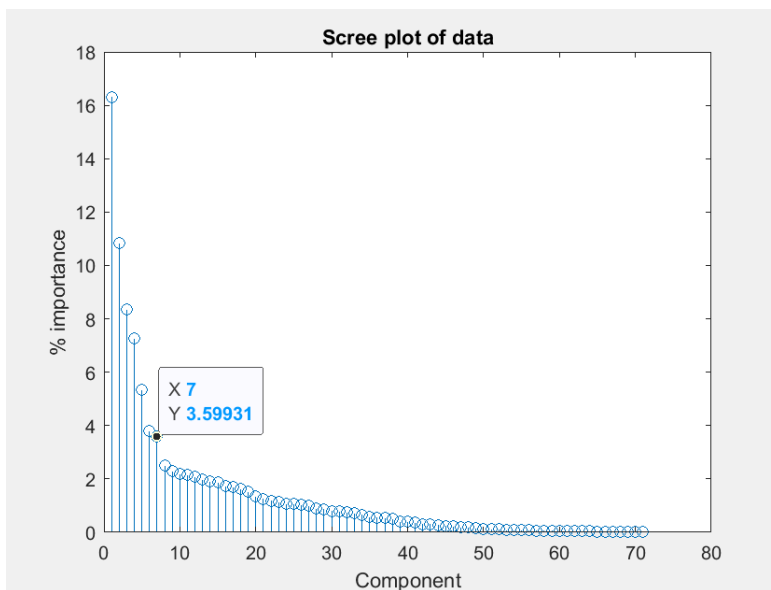


Figure 5-38 Scree Plot of PCA performed over all GCxGC relative locations. Each component explains a percentage of the total variance within the original data matrix. The data matrix stored the normalized areas, based on PCB-123, at each RRT in Figure 5-37

Based on the shape of the scree plot, the first seven components identify files that have large abundances of a compound that are not found in the other samples. The other components do not have a significant amount of weight as they account for 2.5% or less of the total variance of the data matrix. A reason that this may have occurred is that if a file does not have an area at a certain RRT location, a zero was inserted into the data matrix. More samples would be similar if they did not have a significant amount of analytes. In the future, it would be worth investigating the average area of a section of the GCxGC image. Recall that for the mass spectra, similarities were found typically within 1% of the PRRT and SRRT. Taking this into account, instead of using distinct unique locations the analysis could be performed over areas that are within a percent change of the unique RRT pairs.

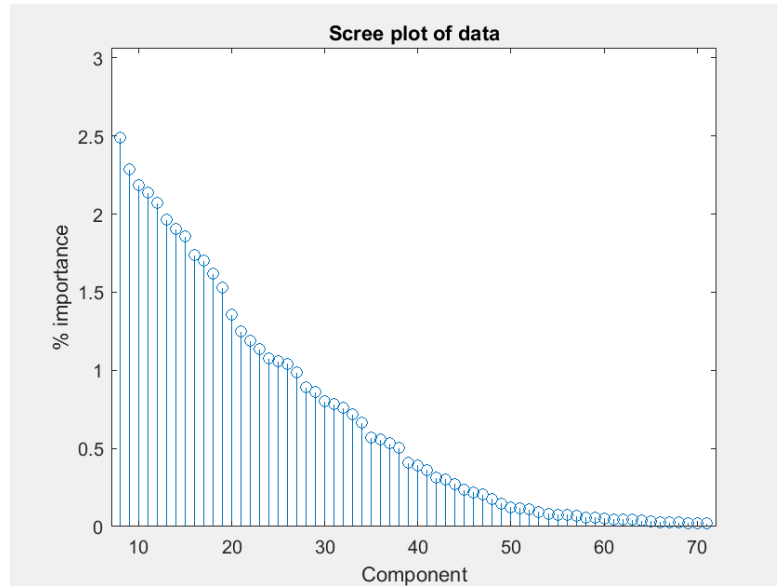


Figure 5-39 The scree plot is zoomed in on principal components 8-71. Components 8-28 represent at least 1% of the total variance in the data matrix. Components 29-38 explain 0.5% to 1% of the total variance. The remaining compounds 39 to 71 account for less than 0.5%, with principal components 50 to 71 accounting for barely a percentage. When components explain a low percentage of the total data matrix, a near linear relationship is described based on the components.

### 5.2.1. Performing the Clustering of the Scores from the Principal Component Analysis

The total sum of the  $L_2$  normalization of all seventy-one components is graphed in Figure 5-41. The samples that have the largest scores, from highest to lowest, are 49, 48, 30, 28, 47, 59 and 31. PCs 1 to 7 are the most significant in defining the  $L_2$  normalization score for these samples. There are a handful of samples that have an  $L_2$  normalization score that is less than one sixth of sample 49. The contribution of PCs 8 to 71 of the  $L_2$  normalization score is graphed in Figure 5-42. The lowest scores in the figure are the same files indicated earlier; sample 28, 30, 31, 47, 48, 49, and 59. This illustrates that the components 8 to 71 have low effects on those samples.

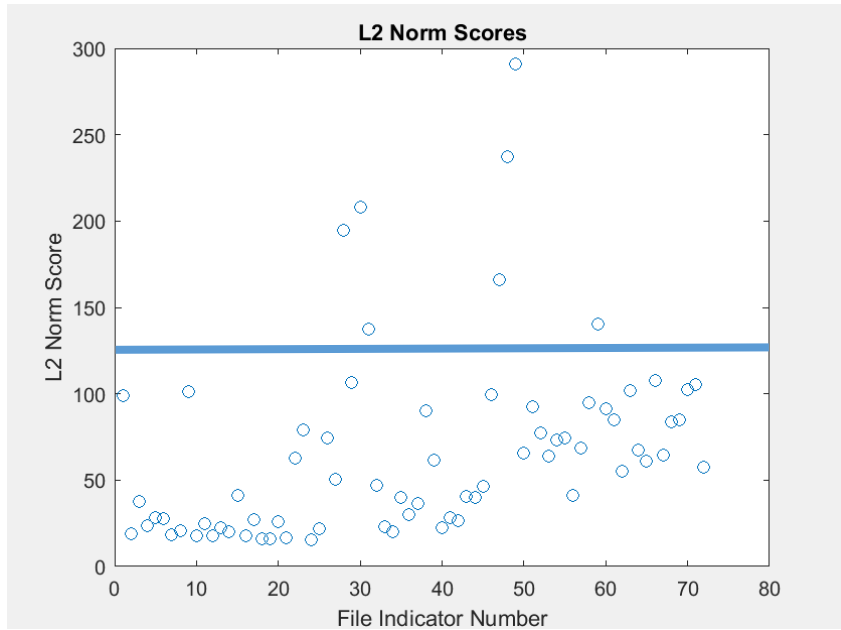


Figure 5-40 The  $L_2$  normalization scores of all 72 samples based on all seventy-one principal components. The blue line indicates the seven samples that have high correlation to the first seven components.

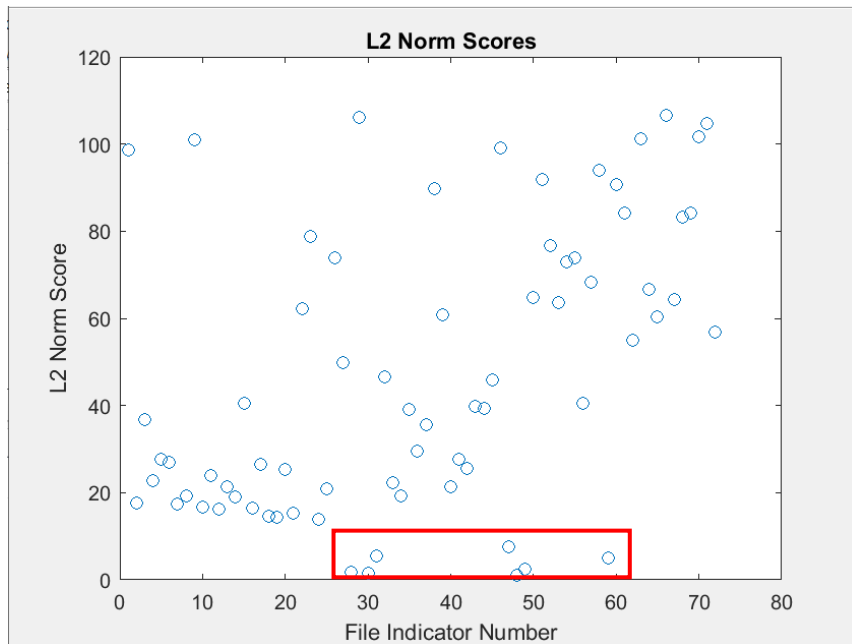


Figure 5-41 The  $L_2$  normalization scores for the 72 samples of principal components 8-71. The red box indicates files 28, 30, 31, 47, 48, 49, and 59 that have low associations to principal components 8 to 71.

The scores that have similar magnitudes seem to have locations in the GCxGC image that play a role in the PC. The k-means clustering was applied to organize the samples into different clusters. To determine the best representation of the data, the cophenetic correlation coefficient was calculated through MATLAB software as described in section 4.2.2. The squared Euclidean distance measurement is the default for the k-means clustering algorithm. The dendrogram using the squared Euclidean distance is in Figure 5-42. The cophenetic correlation coefficient for the squared Euclidean distance was 0.7360. The dendrogram of the sum of absolute distance was produced in Figure 5-44.

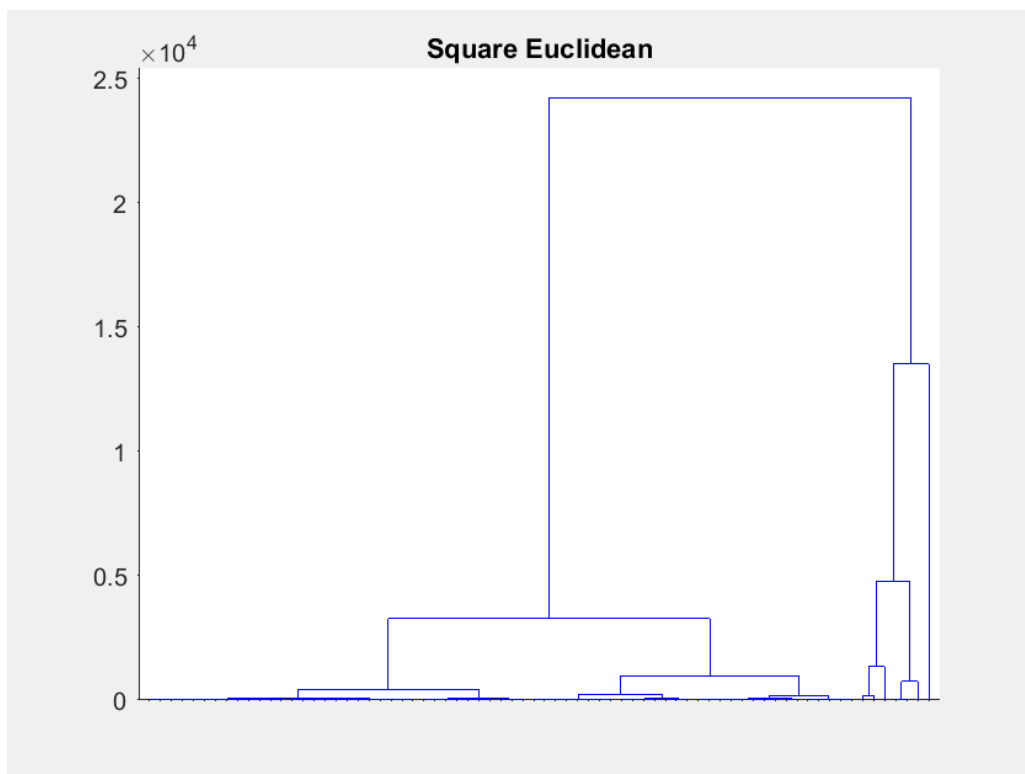


Figure 5-42 The dendrogram of the Square Euclidean Distance. The cophenetic correlation for the hierarchical tree is 0.7360. The square Euclidean distance therefore is not the best representation of the data when using k-means

The cophenetic correlation coefficient of the sum of absolute distance is 0.8513. Between figure 5-42 and 5-43, it is clear that the sum of absolute distances measurement, referred to as Cityblock in MATLAB, represents the data better than the squared Euclidean distance. There

seem to be three distinct clusters. The evaluation tool using the silhouette method confirmed this hypothesis and determined that the optimal number of clusters is three.

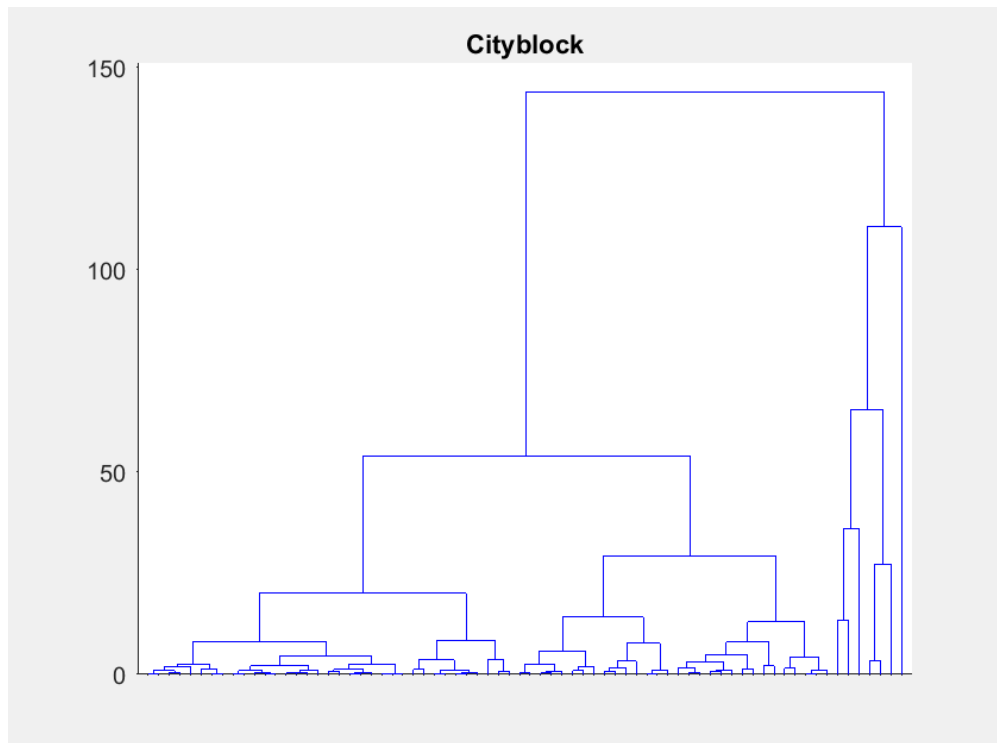


Figure 5-43 The dendrogram of the sum of absolute distances referred to as Cityblock in MATLAB. The right most tree defines the first seven samples into one cluster. The remaining samples seem to evenly belong to two separate clusters. The add cophenetic correlation coefficient is 0.8513 and is a better representation of the data.

The clustering was performed on the graph on Figure 5-41 where the results are illustrated in Figure 5-45. Based on the clusters shown in Figure 5-42 and Figure 5-43, it was surprising to see that samples 31 and 59 were not included in the yellow cluster, as even in the Cityblock dendrogram, seven samples are in the right most grouping. Regardless, the samples 28, 30, 47, 48, and 49 are grouped into one cluster. The clustering was repeated 1,000 times internally to produce the best result.

Once the clusters were determined, the cluster with the least amount of data points was removed from the  $L_2$  normalization graph. Another k-means clustering was performed still

utilizing the sum of absolute distance measurement between data points. A summary of samples within respective clusters is listed below in Table 5-5.

Table 5-5 The colors of each box indicated the sample numbers that are within the same color cluster of Figure 5-45. Samples 31 and 59 were expected to be the outliers of the  $L_2$  normalization score.

2 4 5 6 7 8 10 11 12 13 14 16 17 18 19 20 21 24 25 33 34 36 40 41 42	3 15 27 32 35 37 43 44 45 56	22 26 39 50 53 54 55 57 62 64 65 67 72
23 38 51 52 60 61 68 69	1 9 29 46 58 63 66 70 71	31 59

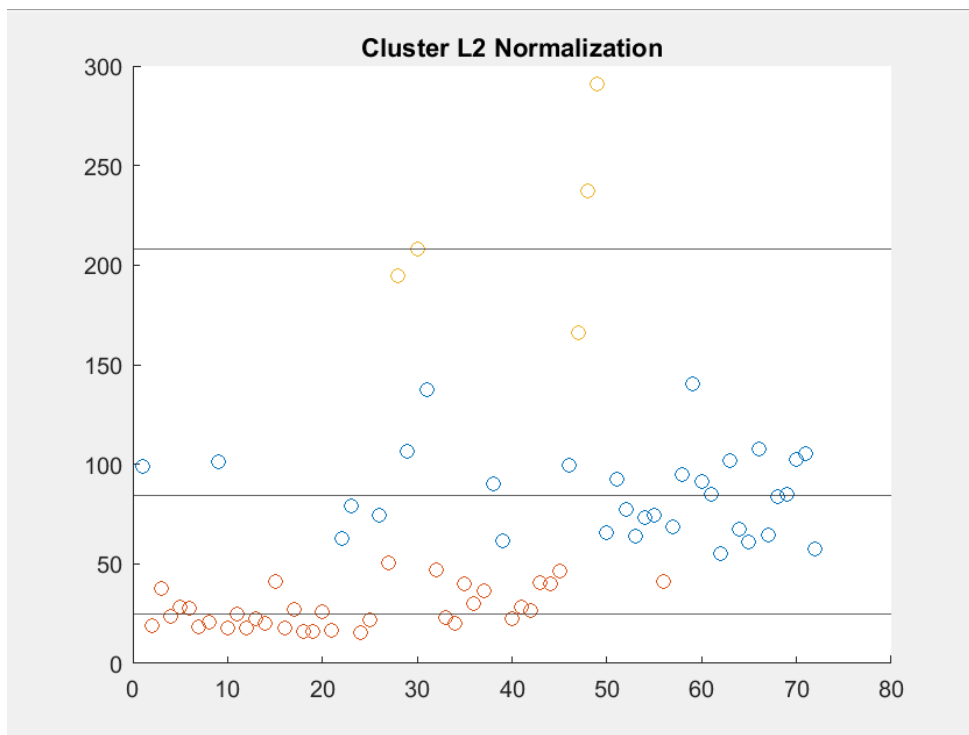


Figure 5-44 The first clustering of all seventy-two samples based on the total  $L_2$  normalization score is shown. There are three clusters. The five samples in the yellow cluster identify the most variant samples. The blue and red clusters have samples that seem to be very close to one another. Clusters blue and red are re-clustered and shown in Figure 5-46. The x-axis is the sample numbers as the y-axis is the  $L_2$  normalization score



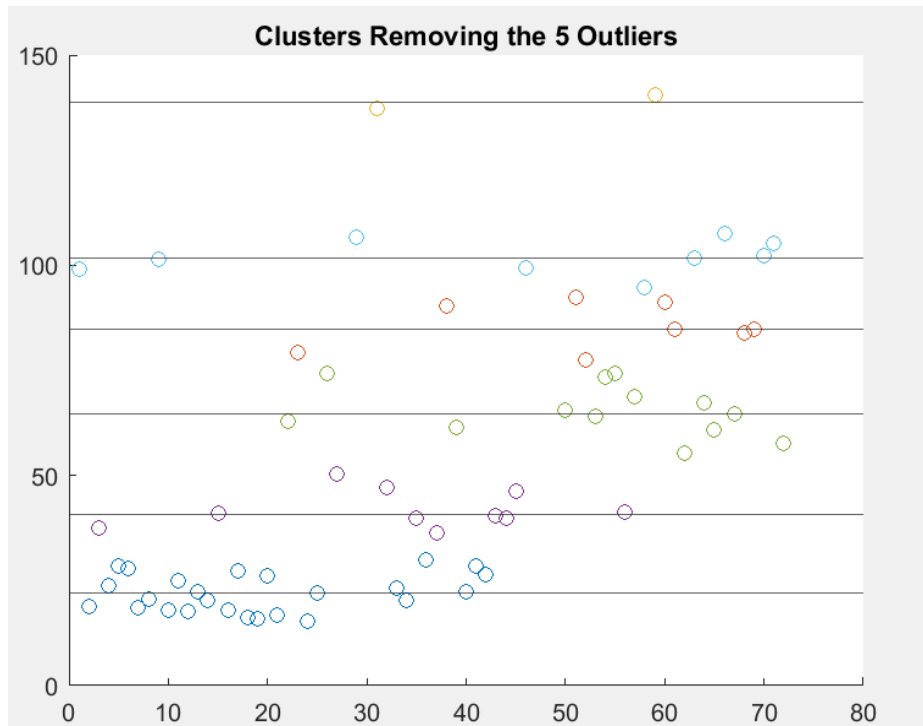


Figure 5-45 The second clustering of the  $L_2$  normalization scores not including sample 28, 30, 47, 48, and 49. The clustering representation has a cophenetic correlation coefficient of 0.9629 when clustering the data points above using the sum of absolute distance measurement. The x-axis is the sample numbers as the y-axis is the  $L_2$  normalization score

Unsurprisingly sample 31 and 59 are clustered together. These samples are outliers to the red and blue clusters in Figure 5-44, which is why they are isolated from the remaining data points. The lines indicate the centroid location of each cluster in the figures. Once the clusters were established, data on where the sample was taken and if the sample was organic or not was looked at.

There is no clear indication that location or whether the milk is organic is a factor in the clustering of the  $L_2$  normalization scores. The six samples that were organic showed up in four separate clusters. In terms of the geographical location that the samples were taken from, multiple clusters had the same states within them. There again is no clear indication that location played a factor in the compounds identified in each sample. It was hypothesized that the

geographical location could matter because there may be a greater concentration of certain POPs based on local dairy cow feed and fertilizer used on the farms.

### 5.2.2. Investigating Scores of Principal Components and their Meaning

Instead of looking at the total  $L_2$  normalization score, only the first seven scores of each sample were calculated as well as the  $L_2$  normalization score based on PC eight to seventy-one. A score is associated to each of the seventy-one PCs for each sample. A graph of the  $L_2$  normalization, as described in Equation 4-17, of PC 1 to 7 vs PC 8 to 71 is shown in Figure 5-40. Samples 28, 30, 31, 47, 48, 49, and 59 each have a high score association with PC 1 to 7 and account for 55.4936% of the total variance. The other samples do not have an  $L_2$  score greater than 120 and are mostly described by PC 8 to 71.

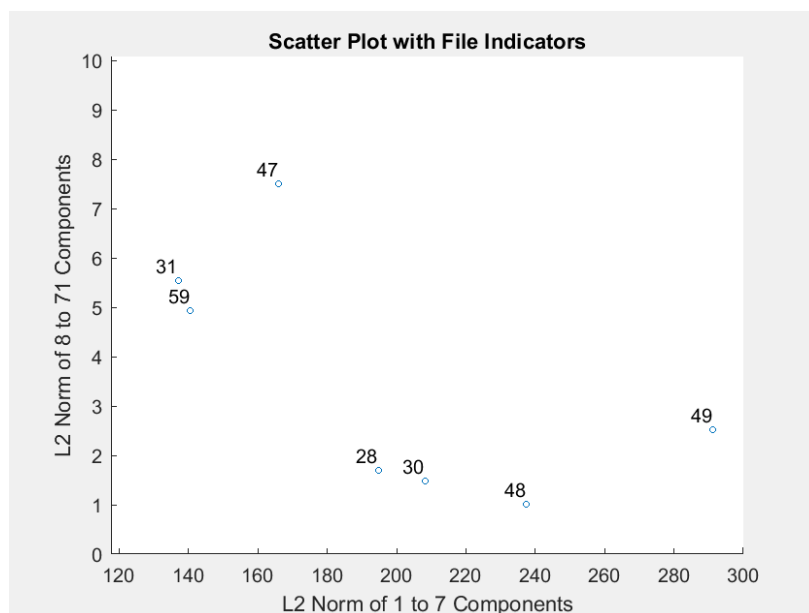


Figure 5-46 Scatter Plot of Files of the  $L_2$  normalization scores of PCs 1 to 7 vs PCs 8 to 71. The sample numbers that are shown have high associations with the 55.4936% of the total variance

The intensity of the GCxGC image at a RRT location dictates the variance of the data set in the PCA. Discrete GCxGC images of each sample were created to illustrate each compound found at the RRT location and the intensity of each normalized area with respect to PCB-123.

The discrete GCxGC image and the actual GCxGC image of each sample are located in Appendix B organized by each cluster that have the highest  $L_2$  normalization scores to the lowest  $L_2$  normalization scores. The red data points indicate areas with high intensities, followed by the green which indicates PCB-123, followed by yellow that indicates intensities from 0.5 to 1 and finally blue which are low intensity areas in the discrete GCxGC image.

The seven samples 28, 30, 31, 47, 48, 49, and 59 all have high intensity locations in the discrete GCxGC image. The PCA prioritized the high intensity locations when defining the first seven components. PCs eight to seventy-one indicate more of the linear relationships between the samples based on the low percentage associated in the scree plots. It was hypothesized that each cluster would have GCxGC images that look similar, however that wasn't necessarily the case. Since the  $L_2$  normalization combines the scores of multiple components, samples could have similar values based on separate PCs.

To investigate the scores further, the  $L_2$  normalization score of each PC is graphed and located in Appendix C. For example, the  $L_2$  normalization score of PC one is graphed in Figure 5-46. By observation, the PC 1 identifies sample 49 as an outlier as PC one correlates to 16% of the total variance between each sample. Samples 28, 30, 47, and 48 have at least one location in common with sample 49 as they have a very small  $L_2$  normalization score. Conclusions of which samples have a commonality can be assessed based on the score graphs.

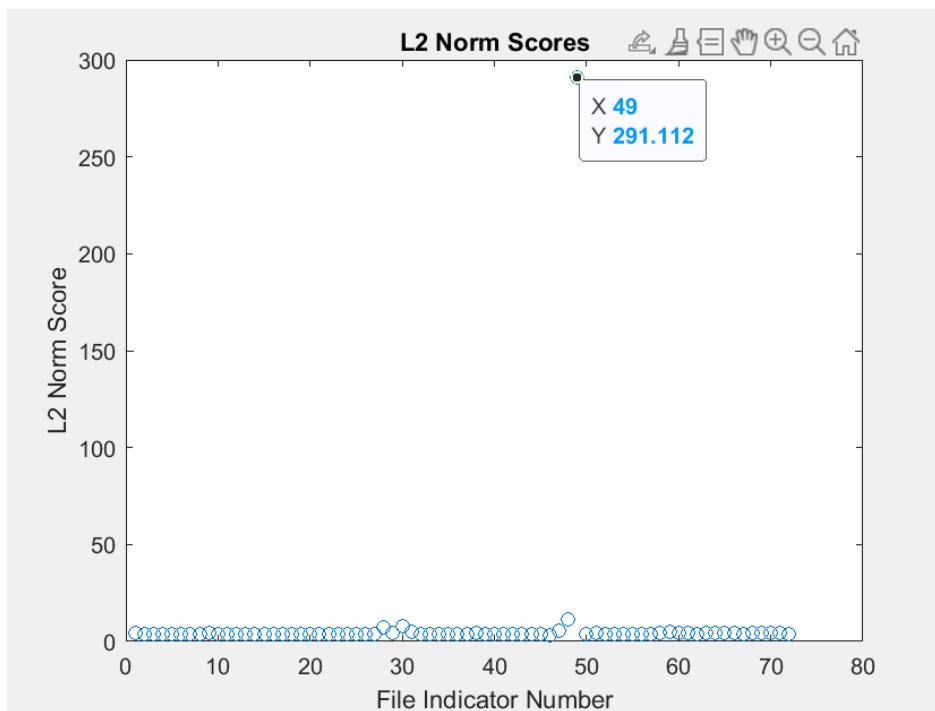


Figure 5-47 The  $L_2$  normalization score of principal component 1 where sample 49 is highly correlated with. Sample 28, 30, 47, and 48 each have at least one location that is in common with sample 49 since their value is not zero

The individual scores of the rest of the components are in Appendix C. Each illustrate which samples have similarities based on location of the GCxGC image. Similarities between samples can then be evaluated where other factors, such as whether the sample is organic or the location where the sample was acquired, can also be considered.

In conclusion, the scores of individual principal components relate samples to one another. Samples that have the areas in the same relative retention time location of the discrete GCxGC image will both have a value greater than zero in the principal component score graphs. To identify the locations in the GCxGC image, the coefficients of each principal component need to be investigated.

### 5.2.3. Investigating Coefficients of Principal Components and their Meaning

Each PC is a weighted linear combination of every 51,143 relative retention time locations in the discrete GCxGC image as seen in Figure 5-37. The weights indicate the importance of each location in the discrete GCxGC image. Locations with zero weight do not contribute to the PC. Location with low weights describe locations that have an area that are not distinctive between samples. Finally, locations with high weights indicate there are compounds that have high intensities that separate samples from one another. To understand the weights of each location, the coefficients of a PC were normalized with respect to the maximum weight within the same PC.

Insight of each PC can be made by observing the normalized weights of the coefficients. The graphs for PC 2 to 71 are located in Appendix D. The magnitude of the weights indicate how much the area at the unique RRT location correlates with the total variance of the data set. A positive weight of the coefficient indicates that an increase in the normalized area would also increase the variance of the principal component, as the score for a sample will increase. Similarly, a negative weight demonstrates that a decrease in the normalized area would increase the variance of the principal component, as the score for a sample will again increase.

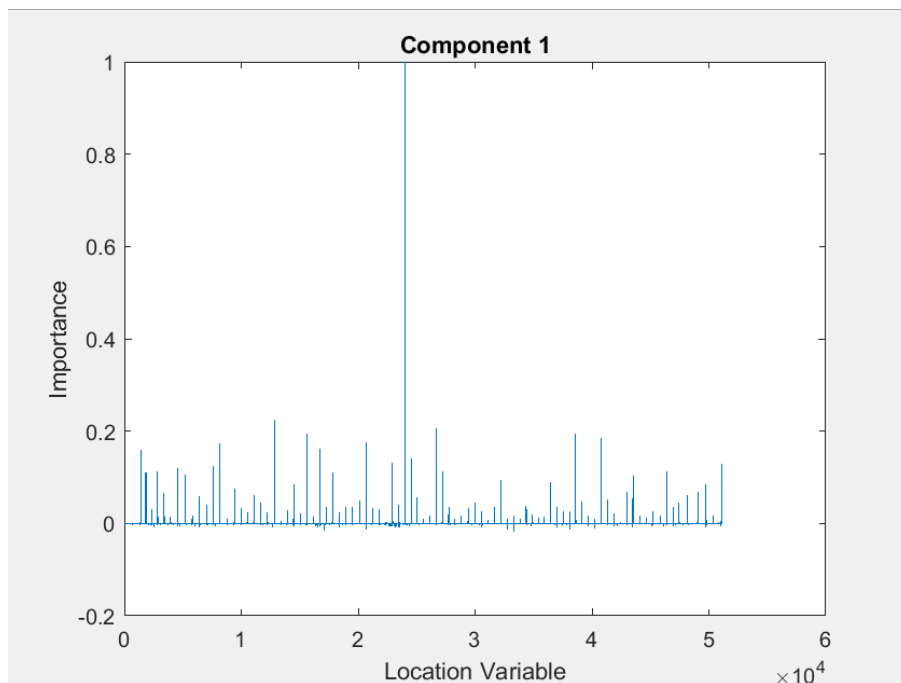


Figure 5-48 The normalized coefficients of principal component 1 indicates how relevant the locations of the discrete GCxGC image are for the principal component score.

Insight of the first PC can be gathered by observing Figure 5-48, which illustrates the normalized weights. There seems to be one clear location that dominates PC 1. A MATLAB function was developed where the RRT locations that are greater than or less than a certain threshold of the normalized weights can be identified. The function was used to pinpoint the location with the most importance i.e., has a value of one. The most dominant location corresponds to an area collected in sample 49. The RRT location of (0.9188, 1) has a total abundance of 159,242,849 and is the largest area concentration out of all seventy-two samples. The absolute retention time location is (1102.08, 1.459). The location shows a very high intensity in the GCxGC image found in Appendix B.

Overall, the coefficients of the principal components specify how each location in the discrete GCxGC image contribute to the individual score of the component. When observations of the scores and coefficients are combined, locations that are shared between multiple samples

can be identified. The compounds designated can then be used in the mass spectra algorithm to determine if the compounds are matches or (potential) coelutions. Recall that PCs with a low percentage in the scree plot tend to show more of a linear or near-constant relationship while higher percentages identify locations that are unique to samples.

### **5.3. Further Discussion of the Principal Component Analysis and its Limitations**

The coefficients and individual scores generated from the PCA can identify similarities between samples. Samples that have similar scores in an individual PC indicate that there are locations, based on the coefficients that generate the score, which are within both samples. Comparisons between mass spectra can be done based on the coefficient locations that are greater than or less than a threshold of the normalized masses.

Recall that the mass spectra algorithm considers compounds currently that are less than 5% change in the PRRT and SRRT. The threshold value can be reduced to 3.5%. The PCA associates the variance to distinct RRT pairs, not RRTs within a certain percent change of one another. The average matching compounds found were within 1% change in the PRRT and SRRT. Instead of having 51,143 unique locations, combining areas that are within 1% change of one another may be a better representation of the data. It was rare to see compounds that were matches with 0.00% change in the PRRT and SRRT. If the PCA was performed with the variables representing each unique area that are within 1% change of one another, a better variance between each area can be completed.

Another limitation in the PCA is that the  $L_2$  normalization scores remove the sign of the individual principal component scores. Though the graphs indicate the relevance of the samples, information is lost when only the magnitudes are considered. There may be files that have a large negative score for a PC. If that's the case, then the negative coefficients are of more interest to

the sample. Also, samples that differ in sign allude to the fact that the samples do not correlate with one another. Including the sign could indicate files that are vastly different.

The PCA is still significant in identifying locations of interest. Though the PCA does not take into consideration the percent change in the RRT pair locations, the mass spectra algorithm will compare all the compounds that are within that percent change. All in all, the mass spectra algorithm performed exceptionally well and the PCA was able to identify compounds for further investigation and which samples have similar areas based on the individual scores.



## CHAPTER 6: CONCLUDING REMARKS

The goal of this research was to create an algorithm that can identify matches and (potential) coelutions across all samples, identify compounds that are unique to a sample or within multiple samples, and find patterns in the overall data matrix.

The data was collected from a GCxGC-ToF-MS instrument. The mass spectrometry of each sample was evaluated with each other to determine if there were any matching compounds or coelutions that occurred. Overall, the algorithm designed worked well in identifying matching compounds. Each sample had a match of the DDE identified, as 77 matches were returned when analyzing Unknown 3744. There were five samples that returned multiple matches which is not possible. Only the best match should be returned by the algorithm if multiple matches were found in a sample. Another fix would be to reduce the percent change threshold of the PRRT to 3.5-4%. The coelutions could not be confirmed as a manufactured standard of the compound would have to be processed on the instrument to determine the presence of the compound. The algorithm for 'base coelution' or 'coelution' did perform as expected and identified compound that can be further investigated. All in all, the mass spectra comparisons performed exceptionally well when identifying compounds that are matches.

The data from the GCxGC image was also explored. PCA was performed over all the areas at a RRT location, normalized with respect to PCB-123 for window 5. Each PC affects the score of the samples where the coefficients of the PC identifies locations that impact the variance of the original dataset. No distinct pattern was found within the samples, however, insight to locations and which samples are similar to each other can be gather based on the observation of the individual score graphs.

There are numerous ideas that can be applied in future work to identify patterns across samples. First, analysis can be done using the scores of the individual components. This would be done by identifying the samples that have a value greater than zero in the individual score graphs. A node graph could be created where the weight between samples is increased if a principal component's individual score correlates with both samples. The graph will indicate how many samples are corresponded together, and after quantification, clustering can be done for similar files.

Using the node graph, multiple windows should be evaluated. Each window can have a representative graph that can be expanded to include the information for all the windows. Insight on how windows are similar or different may drive more research.

A node graph could also be used to relate the amount of compound matches from the mass spectra algorithm there are between samples. Large amounts of matches between samples can address the importance of the geographical location. It can also indicate if organic samples are more similar.

Another aspect to consider is to identify compounds that are within all samples by counting the number of matches. If there are compounds that only appear in a handful of samples, then further investigation if other unique compounds also belong to the same subsection occur throughout multiple windows.

Building off section 5.3, the information collected from the PCA can identify abundant compounds. It would be of interest to analyze matches throughout the entire sample set to see if the abundant compound appears in other samples, even at low intensities.

Also, again continuing the discussion from 5.3, it would be relevant to try to use the PCA over a data matrix that focuses on areas that are within a 1% change of the distinct retention

time. The areas should still be normalized as the PCA is dependent on the magnitudes of the variables. Identifying areas that are within a small region of the GCxGC image identifies locations in the GCxGC image that are most likely to be a match. Investigating based on the distinct relative retention time pair means that if there was an area that differed by 0.03% in the primary relative retention time, they were viewed to be similar. Based on the mass spectra comparisons, the low percent change has a likelihood of producing a match and should be further researched.

Finally, another approach for future work would be to identify the locations of significance in each of the principal components. Distinguish which relative locations were used throughout each component to classify each locations importance.

Overall, the framework of interpretation of the principal component analysis is given. There are multiple routes that the future work can take utilizing this information. The key factors from the research were how target compounds can identify non-target compounds using relative retention times as well as how to classify non-target compounds within the data matrix.

### **6.1. Key Contributions**

The most valuable information attained from the presented research is that relative retention times based on the target compounds can account for the drifts in the absolute retention time. Since each sample has the same target compounds, where each have all been previously identified over the past decade, the relative retention time measurements can be applied to every sample. On average, matching analytes across samples are typically within 1% of another analyte's primary relative retention.

Discussion of how to interpret the principal component analysis was also provided. Locations of interest based on the relative retention times were indicated from the scores and coefficients of each principal component.

As a result of this research, a MATLAB resource was produced. Chemists can then do their own analysis and further investigate other windows or refine results in window 5. The algorithms and methodologies discussed can be applied over any large data repository of chemometric measurements. Other lipophilic samples, such as eggs and fish, could also undergo the same analysis.

APPENDIX A: MASS SPECTRA

Figure A-1 The mass spectrum of Unknown 3381 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All masses that are within the fingerprint of Unknown 3744 are present however there are other masses that are included in the fingerprint of Unknown 3381. .... 74

Figure A-2 The mass spectrum of Unknown 2723 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 2723..... 74

Figure A-3 The mass spectrum of Unknown 2741 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 2741..... 74

Figure A-4 The mass spectrum of Unknown 2760 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 2760..... 74

Figure A-5 The mass spectrum of Unknown 3855 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 3855..... 75

Figure A-6 The mass spectrum of Unknown 3476 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 3476..... 75

Figure A-7 The mass spectrum of Unknown 2466 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 2723..... 75

Figure A-8 The mass spectrum of Unknown 3914 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 2723..... 75

Figure A-9 The mass spectrum of Unknown 3102 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 3102..... 76

Figure A-10 The mass spectrum of Unknown 3199 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 3199..... 76

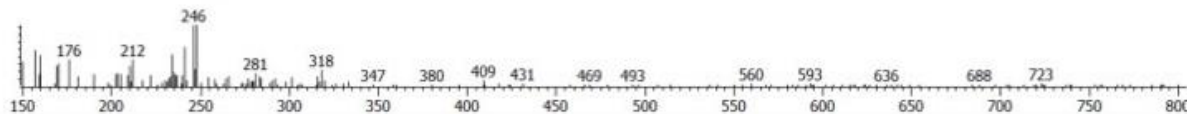


Figure A-1 The mass spectrum of Unknown 3381 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All masses that are within the fingerprint of Unknown 3744 are present however there are other masses that are included in the fingerprint of Unknown 3381.

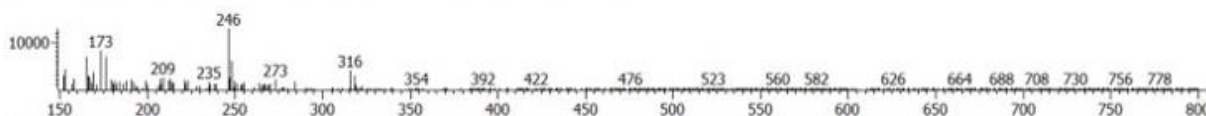


Figure A-2 The mass spectrum of Unknown 2723 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 2723.

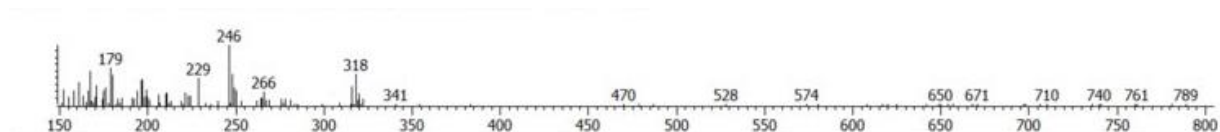


Figure A-3 The mass spectrum of Unknown 2741 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 2741.

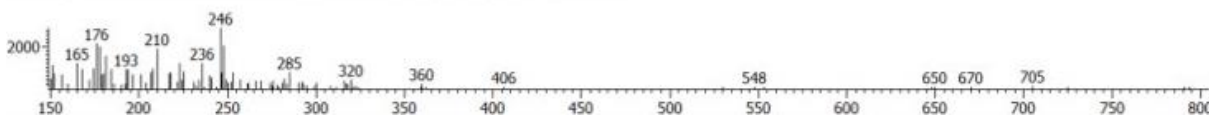


Figure A-4 The mass spectrum of Unknown 2760 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 2760

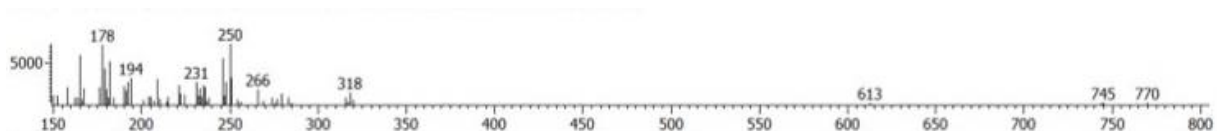


Figure A-5 The mass spectrum of Unknown 3855 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 3855.

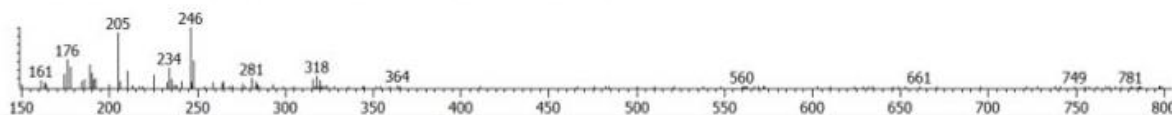


Figure A-6 The mass spectrum of Unknown 3476 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 3476.

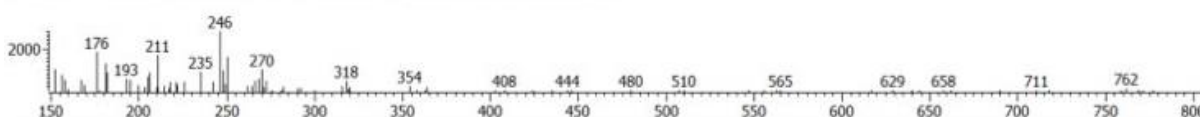


Figure A-7 The mass spectrum of Unknown 2466 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 2723.

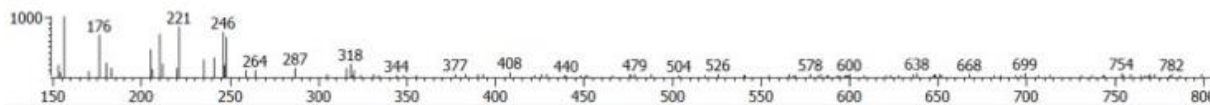


Figure A-8 The mass spectrum of Unknown 3914 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 2723.



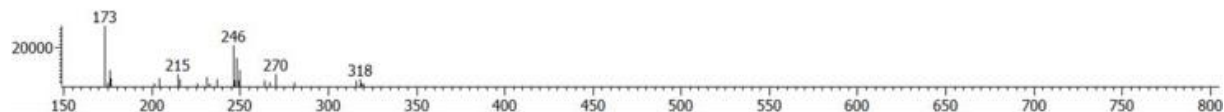


Figure A-9 The mass spectrum of Unknown 3102 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 3102.

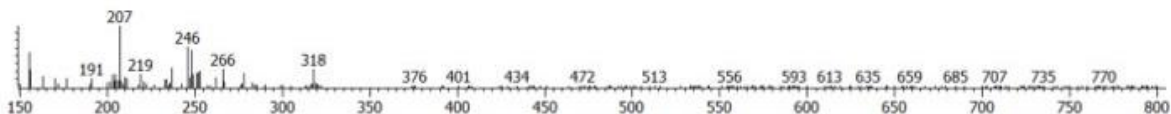


Figure A-10 The mass spectrum of Unknown 3199 labeled as Coelution where the x-axis represents the mass of each ion and the y-axis represents the abundance of the mass based on their peak area. All the masses from Unknown 3744 are included however there are more masses within the fingerprint of Unknown 3199.

## APPENDIX B: GCxGC IMAGES

Figure B-1 The discrete GCxGC image of Sample 28 that belongs to cluster 1 .....	84
Figure B-2 The real GCxGC image of Sample 28 that belongs to cluster 1 .....	84
Figure B-3 The discrete GCxGC image of sample 30 that belongs to cluster 1 .....	85
Figure B-4 The real GCxGC image of Sample 30 that belongs to cluster 1 .....	85
Figure B-5 The discrete GCxGC image of Sample 47 that belongs to cluster 1 .....	86
Figure B-6 The real GCxGC image of Sample 47 that belongs to cluster 1 .....	86
Figure B-7 The discrete GCxGC image of Sample 48 that belongs to cluster 1 .....	87
Figure B-8 The real GCxGC image of Sample 48 that belongs to cluster 1 .....	87
Figure B-9 The discrete GCxGC image of Sample 49 that belongs to cluster 1 .....	88
Figure B-10 The real GCxGC image of Sample 49 that belongs to cluster 1 .....	88
Figure B-11 The discrete GCxGC image of Sample 31 that belongs to cluster 2 .....	89
Figure B-12 The real GCxGC image of Sample 31 that belongs to cluster 2 .....	89
Figure B-13 The discrete GCxGC image of Sample 59 that belongs to cluster 2 .....	90
Figure B-14 The real GCxGC image of Sample 59 that belongs to cluster 2 .....	90
Figure B-15 The discrete GCxGC image of Sample 1 that belongs to cluster 3 .....	91
Figure B-16 The real GCxGC image of Sample 1 that belongs to cluster 3 .....	91
Figure B-17 The discrete GCxGC image of Sample 9 that belongs to cluster 3 .....	92
Figure B-18 The real GCxGC image of Sample 9 that belongs to cluster 3 .....	92
Figure B-19 The discrete GCxGC image of Sample 29 that belongs to cluster 3 .....	93
Figure B-20 The real GCxGC image of Sample 29 that belongs to cluster 3 .....	93
Figure B-21 The discrete GCxGC image of Sample 46 that belongs to cluster 3 .....	94
Figure B-22 The real GCxGC image of Sample 46 that belongs to cluster 3 .....	94
Figure B-23 The discrete GCxGC image of Sample 58 that belongs to cluster 3 .....	95
Figure B-24 The real GCxGC image of Sample 58 that belongs to cluster 3 .....	95

Figure B-25 The discrete GCxGC image of Sample 63 that belongs to cluster 3 .....	96
Figure B-26 The real GCxGC image of Sample 63 that belongs to cluster 3 .....	96
Figure B-27 The discrete GCxGC image of Sample 66 that belongs to cluster 3 .....	97
Figure B-28 The real GCxGC image of Sample 66 that belongs to cluster 3 .....	97
Figure B-29 The discrete GCxGC image of Sample 70 that belongs to cluster 3 .....	98
Figure B-30 The real GCxGC image of Sample 70 that belongs to cluster 3 .....	98
Figure B-31 The discrete GCxGC image of Sample 71 that belongs to cluster 3 .....	99
Figure B-32 The real GCxGC image of Sample 71 that belongs to cluster 3 .....	99
Figure B-33 The discrete GCxGC image of Sample 23 that belongs to cluster 4 .....	100
Figure B-34 The real GCxGC image of Sample 23 that belongs to cluster 4 .....	100
Figure B-35 The discrete GCxGC image of Sample 38 that belongs to cluster 4 .....	101
Figure B-36 The real GCxGC image of Sample 38 that belongs to cluster 4 .....	101
Figure B-37 The discrete GCxGC image of Sample 51 that belongs to cluster 4 .....	102
Figure B-38 The real GCxGC image of Sample 51 that belongs to cluster 4 .....	102
Figure B-39 The discrete GCxGC image of Sample 52 that belongs to cluster 4 .....	103
Figure B-40 The real GCxGC image of Sample 52 that belongs to cluster 4 .....	103
Figure B-41 The discrete GCxGC image of Sample 60 that belongs to cluster 4 .....	104
Figure B-42 The real GCxGC image of Sample 60 that belongs to cluster 4 .....	104
Figure B-43 The discrete GCxGC image of Sample 61 that belongs to cluster 4 .....	105
Figure B-44 The real GCxGC image of Sample 61 that belongs to cluster 4 .....	105
Figure B-45 The discrete GCxGC image of Sample 68 that belongs to cluster 4 .....	106
Figure B-46 The real GCxGC image of Sample 68 that belongs to cluster 4 .....	106
Figure B-47 The discrete GCxGC image of Sample 69 that belongs to cluster 4 .....	107
Figure B-48 The real GCxGC image of Sample 69 that belongs to cluster 4 .....	107
Figure B-49 The discrete GCxGC image of Sample 22 that belongs to cluster 5 .....	108

Figure B-50 The real GCxGC image of Sample 22 that belongs to cluster 5 .....	108
Figure B-51 The discrete GCxGC image of Sample 26 that belongs to cluster 5 .....	109
Figure B-52 The real GCxGC image of Sample 26 that belongs to cluster 5 .....	109
Figure B-53 The discrete GCxGC image of Sample 39 that belongs to cluster 5 .....	110
Figure B-54 The real GCxGC image of Sample 39 that belongs to cluster 5 .....	110
Figure B-55 The discrete GCxGC image of Sample 50 that belongs to cluster 5 .....	111
Figure B-56 The real GCxGC image of Sample 50 that belongs to cluster 5 .....	111
Figure B-57 The discrete GCxGC image of Sample 53 that belongs to cluster 5 .....	112
Figure B-58 The real GCxGC image of Sample 53 that belongs to cluster 5 .....	112
Figure B-59 The discrete GCxGC image of Sample 54 that belongs to cluster 5 .....	113
Figure B-60 The real GCxGC image of Sample 54 that belongs to cluster 5 .....	113
Figure B-61 The discrete GCxGC image of Sample 55 that belongs to cluster 5 .....	114
Figure B-62 The real GCxGC image of Sample 55 that belongs to cluster 5 .....	114
Figure B-63 The discrete GCxGC image of Sample 57 that belongs to cluster 5 .....	115
Figure B-64 The real GCxGC image of Sample 57 that belongs to cluster 5 .....	115
Figure B-65 The discrete GCxGC image of Sample 62 that belongs to cluster 5 .....	116
Figure B-66 The real GCxGC image of Sample 62 that belongs to cluster 5 .....	116
Figure B-67 The discrete GCxGC image of Sample 64 that belongs to cluster 5 .....	117
Figure B-68 The real GCxGC image of Sample 64 that belongs to cluster 5 .....	117
Figure B-69 The discrete GCxGC image of Sample 65 that belongs to cluster 5 .....	118
Figure B-70 The real GCxGC image of Sample 65 that belongs to cluster 5 .....	118
Figure B-71 The discrete GCxGC image of Sample 67 that belongs to cluster 5 .....	119
Figure B-72 The real GCxGC image of Sample 67 that belongs to cluster 5 .....	119
Figure B-73 The discrete GCxGC image of Sample 72 that belongs to cluster 5 .....	120
Figure B-74 The real GCxGC image of Sample 72 that belongs to cluster 5 .....	120

Figure B-75 The discrete GCxGC image of Sample 3 that belongs to cluster 6.....	121
Figure B-76 The real GCxGC image of Sample 3 that belongs to cluster 6 .....	121
Figure B-77 The discrete GCxGC image of Sample 15 that belongs to cluster 6.....	122
Figure B-78 The real GCxGC image of Sample 15 that belongs to cluster 6 .....	122
Figure B-79 The discrete GCxGC image of Sample 27 that belongs to cluster 6.....	123
Figure B-80 The real GCxGC image of Sample 27 that belongs to cluster 6 .....	123
Figure B-81 The discrete GCxGC image of Sample 32 that belongs to cluster 6.....	124
Figure B-82 The real GCxGC image of Sample 32 that belongs to cluster 6 .....	124
Figure B-83 The discrete GCxGC image of Sample 37 that belongs to cluster 6.....	125
Figure B-84 The real GCxGC image of Sample 37 that belongs to cluster 6 .....	125
Figure B-85 The discrete GCxGC image of Sample 43 that belongs to cluster 6.....	126
Figure B-86 The real GCxGC image of Sample 43 that belongs to cluster 6 .....	126
Figure B-87 The discrete GCxGC image of Sample 44 that belongs to cluster 6.....	127
Figure B-88 The real GCxGC image of Sample 44 that belongs to cluster 6 .....	127
Figure B-89 The discrete GCxGC image of Sample 45 that belongs to cluster 6.....	128
Figure B-90 The real GCxGC image of Sample 45 that belongs to cluster 6 .....	128
Figure B-91 The discrete GCxGC image of Sample 46 that belongs to cluster 6.....	129
Figure B-92 The real GCxGC image of Sample 46 that belongs to cluster 6 .....	129
Figure B-93 The discrete GCxGC image of Sample 2 that belongs to cluster 7.....	130
Figure B-94 The real GCxGC image of Sample 2 that belongs to cluster 7 .....	130
Figure B-95 The discrete GCxGC image of Sample 4 that belongs to cluster 7.....	131
Figure B-96 The real GCxGC image of Sample 4 that belongs to cluster 7 .....	131
Figure B-97 The discrete GCxGC image of Sample 5 that belongs to cluster 7.....	132
Figure B-98 The real GCxGC image of Sample 5 that belongs to cluster 7 .....	132
Figure B-99 The discrete GCxGC image of Sample 6 that belongs to cluster 7.....	133

Figure B-100	The real GCxGC image of Sample 6 that belongs to cluster 7 .....	133
Figure B-101	The discrete GCxGC image of Sample 7 that belongs to cluster 7 .....	134
Figure B-102	The real GCxGC image of Sample 7 that belongs to cluster 7 .....	134
Figure B-103	The discrete GCxGC image of Sample 8 that belongs to cluster 7 .....	135
Figure B-104	The real GCxGC image of Sample 8 that belongs to cluster 7 .....	135
Figure B-105	The discrete GCxGC image of Sample 10 that belongs to cluster 7 .....	136
Figure B-106	The real GCxGC image of Sample 10 that belongs to cluster 7 .....	136
Figure B-107	The discrete GCxGC image of Sample 11 that belongs to cluster 7 .....	137
Figure B-108	The real GCxGC image of Sample 11 that belongs to cluster 7 .....	137
Figure B-109	The discrete GCxGC image of Sample 12 that belongs to cluster 7 .....	138
Figure B-110	The real GCxGC image of Sample 12 that belongs to cluster 7 .....	138
Figure B-111	The discrete GCxGC image of Sample 13 that belongs to cluster 7 .....	139
Figure B-112	The real GCxGC image of Sample 13 that belongs to cluster 7 .....	139
Figure B-113	The discrete GCxGC image of Sample 14 that belongs to cluster 7 .....	140
Figure B-114	The real GCxGC image of Sample 14 that belongs to cluster 7 .....	140
Figure B-115	The discrete GCxGC image of Sample 16 that belongs to cluster 7 .....	141
Figure B-116	The real GCxGC image of Sample 16 that belongs to cluster 7 .....	141
Figure B-117	The discrete GCxGC image of Sample 17 that belongs to cluster 7 .....	142
Figure B-118	The real GCxGC image of Sample 17 that belongs to cluster 7 .....	142
Figure B-119	The discrete GCxGC image of Sample 18 that belongs to cluster 7 .....	143
Figure B-120	The real GCxGC image of Sample 18 that belongs to cluster 7 .....	143
Figure B-121	The discrete GCxGC image of Sample 19 that belongs to cluster 7 .....	144
Figure B-122	The real GCxGC image of Sample 19 that belongs to cluster 7 .....	144
Figure B-123	The discrete GCxGC image of Sample 20 that belongs to cluster 7 .....	145
Figure B-124	The real GCxGC image of Sample 20 that belongs to cluster 7 .....	145

Figure B-125 The discrete GCxGC image of Sample 21 that belongs to cluster 7 .....	146
Figure B-126 The real GCxGC image of Sample 21 that belongs to cluster 7 .....	146
Figure B-127 The discrete GCxGC image of Sample 24 that belongs to cluster 7 .....	147
Figure B-128 The real GCxGC image of Sample 24 that belongs to cluster 7 .....	147
Figure B-129 The discrete GCxGC image of Sample 25 that belongs to cluster 7 .....	148
Figure B-130 The real GCxGC image of Sample 42 that belongs to cluster 7 .....	148
Figure B-131 The discrete GCxGC image of Sample 33 that belongs to cluster 7 .....	149
Figure B-132 The real GCxGC image of Sample 33 that belongs to cluster 7 .....	149
Figure B-133 The discrete GCxGC image of Sample 34 that belongs to cluster 7 .....	150
Figure B-134 The real GCxGC image of Sample 34 that belongs to cluster 7 .....	150
Figure B-135 The discrete GCxGC image of Sample 36 that belongs to cluster 7 .....	151
Figure B-136 The real GCxGC image of Sample 36 that belongs to cluster 7 .....	151
Figure B-137 The discrete GCxGC image of Sample 40 that belongs to cluster 7 .....	152
Figure B-138 The real GCxGC image of Sample 40 that belongs to cluster 7 .....	152
Figure B-139 The discrete GCxGC image of Sample 41 that belongs to cluster 7 .....	153
Figure B-140 The real GCxGC image of Sample 41 that belongs to cluster 7 .....	153
Figure B-141 The discrete GCxGC image of Sample 42 that belongs to cluster 7 .....	154
Figure B-142 The real GCxGC image of Sample 42 that belongs to cluster 7 .....	154

Each discrete GCxGC image of window 5 has two axes. The x-axis is the primary relative retention time and the y-axis is the secondary relative retention times. The red data points indicate normalized areas based on PCB-123 that are greater than 1, followed by the green which indicates the target compound PCB-123, followed by yellow that indicates the normalized areas in the range of 0.5 to 1 and finally blue which are low intensity areas in the discrete GCxGC image.

Each of the real GCxGC images are between PCB-101 and PCB-123 to indicate window 5. The absolute primary and secondary retention time represent the two axes. The color of the chromatogram represents the abundance of a compound at a retention time pair. A graph of the intensities is below the chromatogram. The primary and secondary retention times represent the x-axis where the y-axis is the intensity.



Cluster 1: Samples 28, 30, 47, 48, and 49 that have the highest  $L_2$  normalization scores.

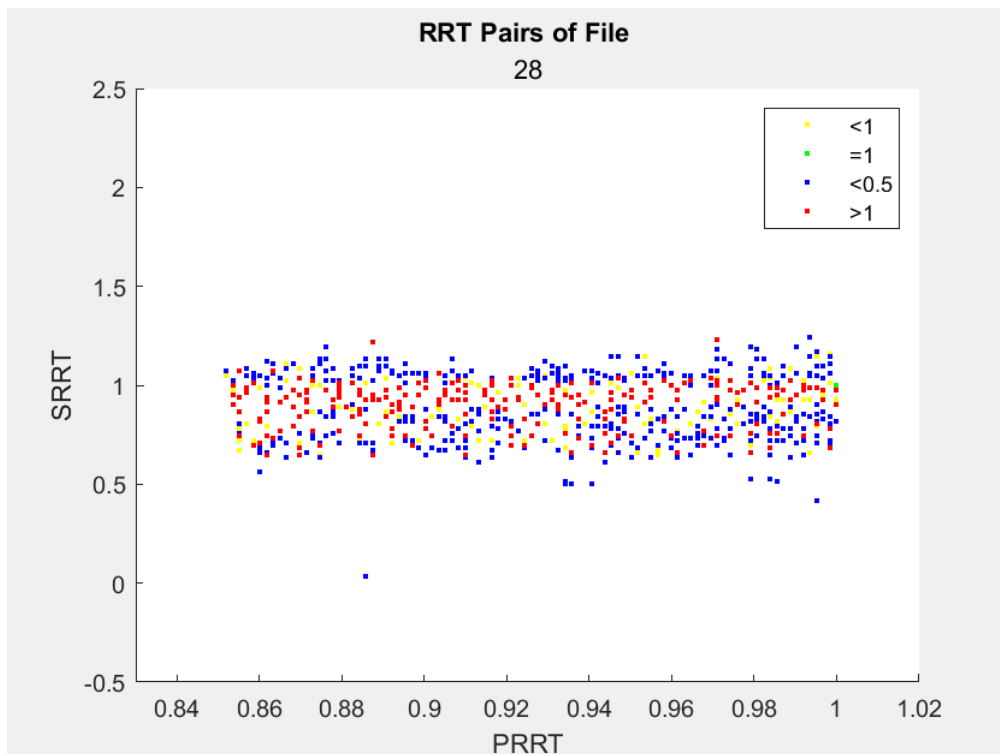


Figure B-1 The discrete GCxGC image of Sample 28 that belongs to cluster 1

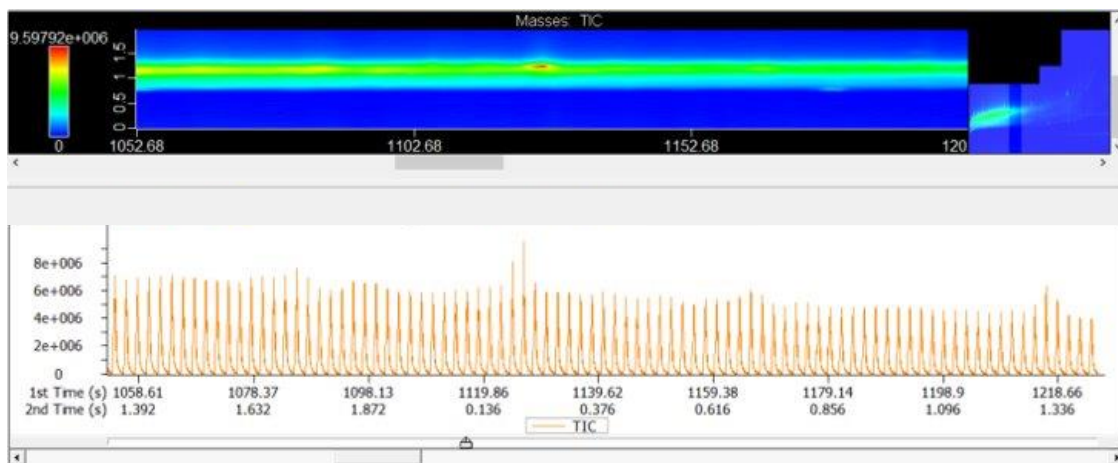


Figure B-2 The real GCxGC image of Sample 28 that belongs to cluster 1

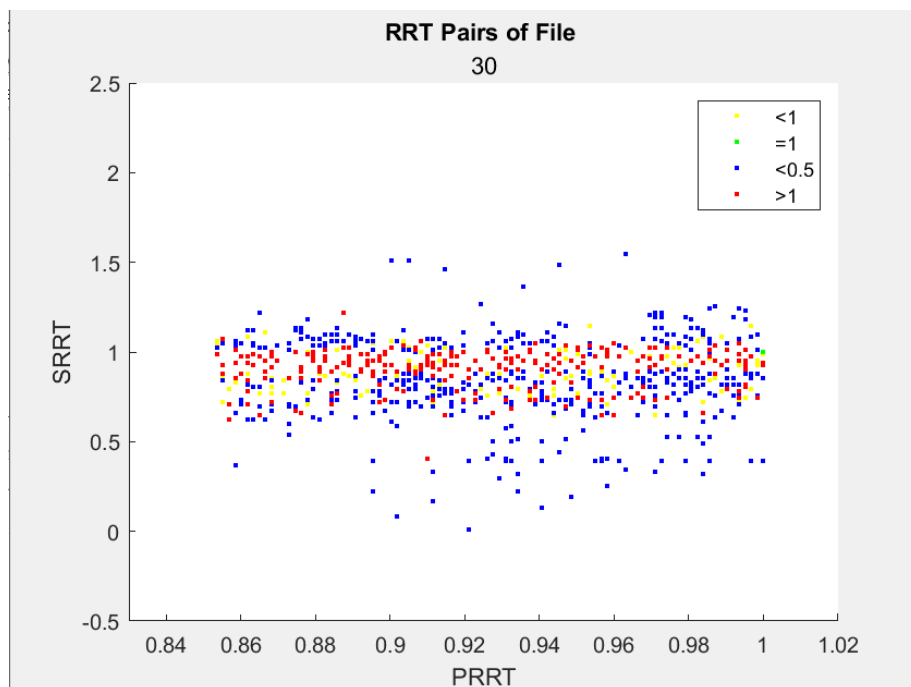


Figure B-3 The discrete GCxGC image of sample 30 that belongs to cluster 1

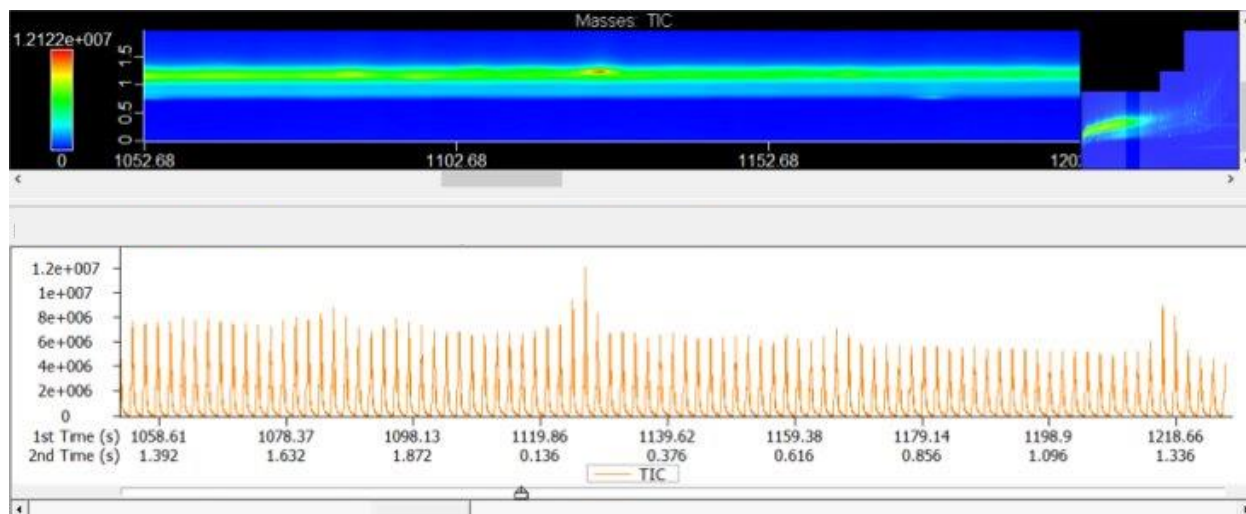


Figure B-4 The real GCxGC image of Sample 30 that belongs to cluster 1

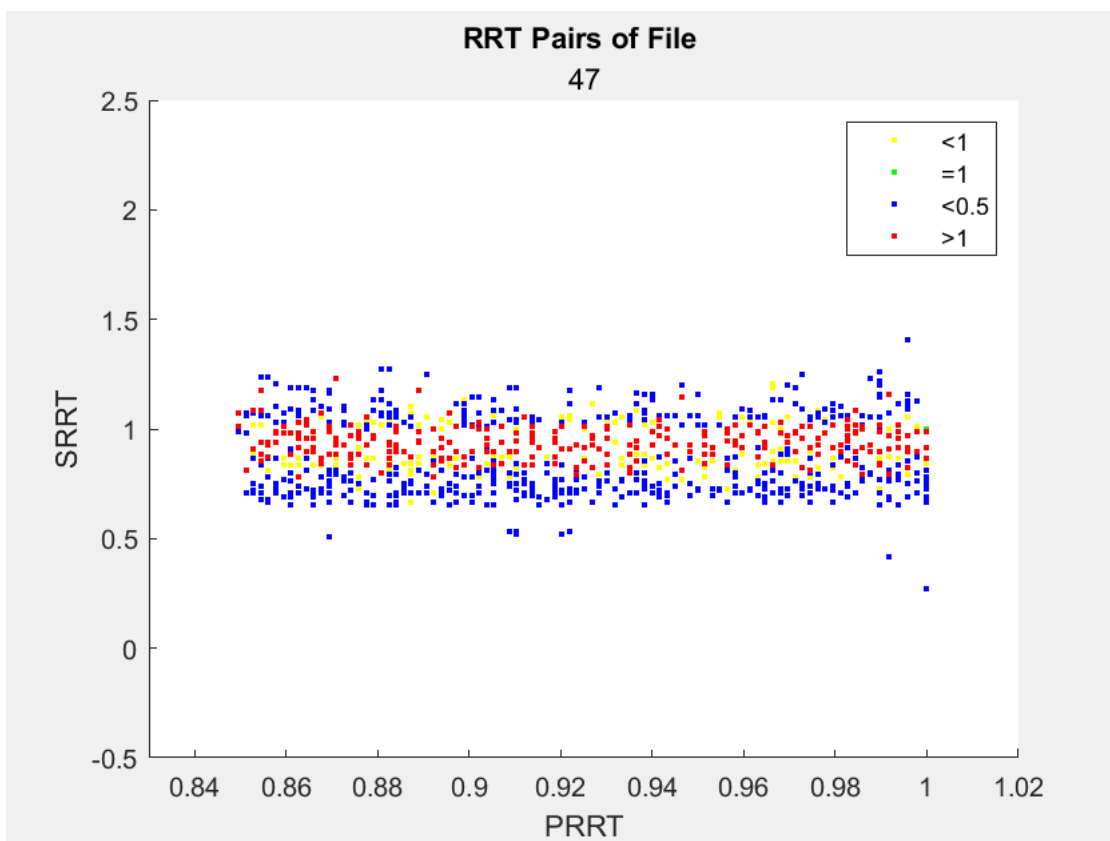


Figure B-5 The discrete GCxGC image of Sample 47 that belongs to cluster 1

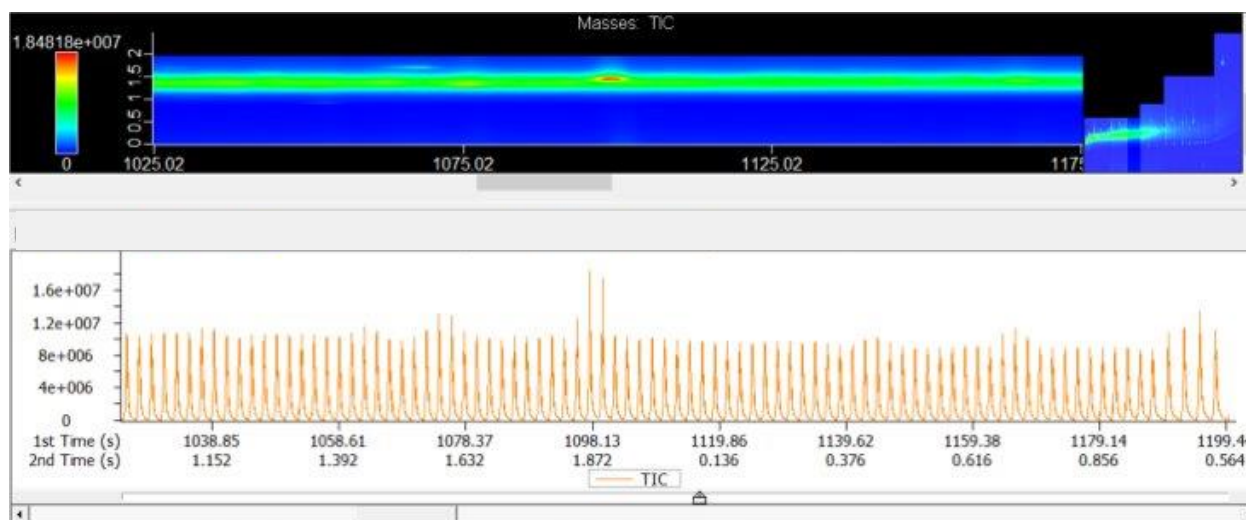


Figure B-6 The real GCxGC image of Sample 47 that belongs to cluster 1

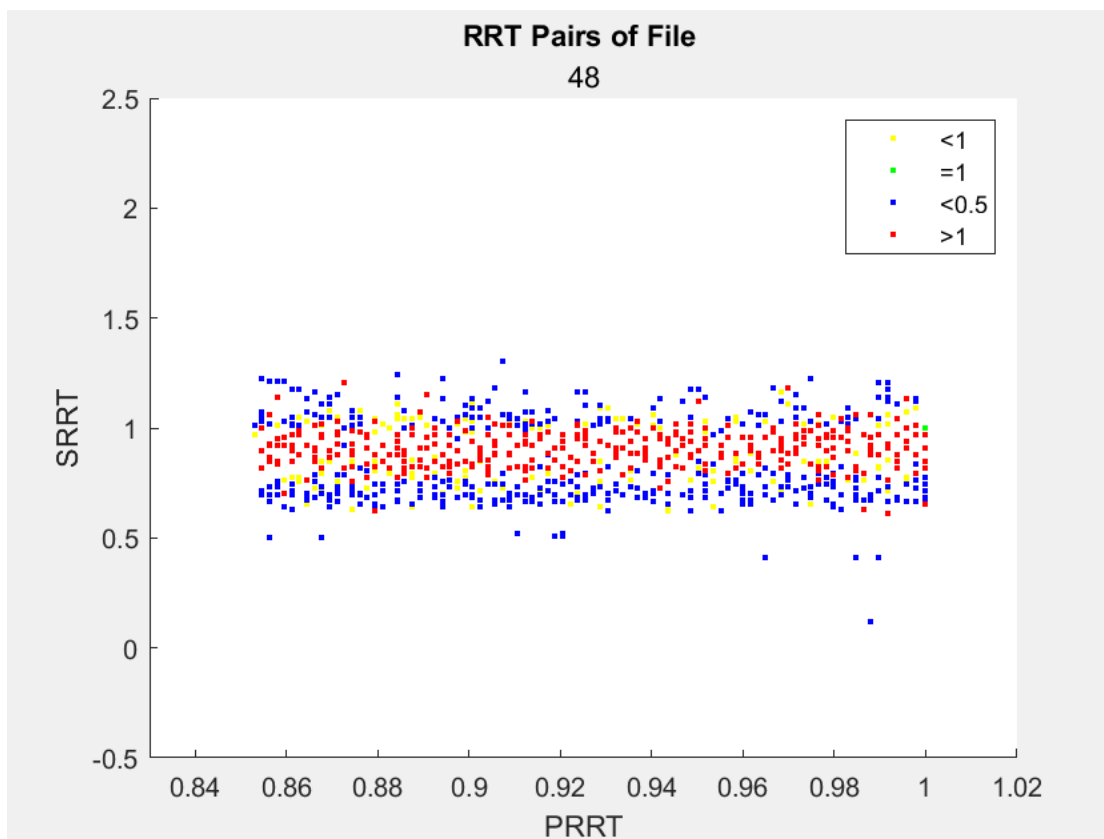


Figure B-7 The discrete GCxGC image of Sample 48 that belongs to cluster 1

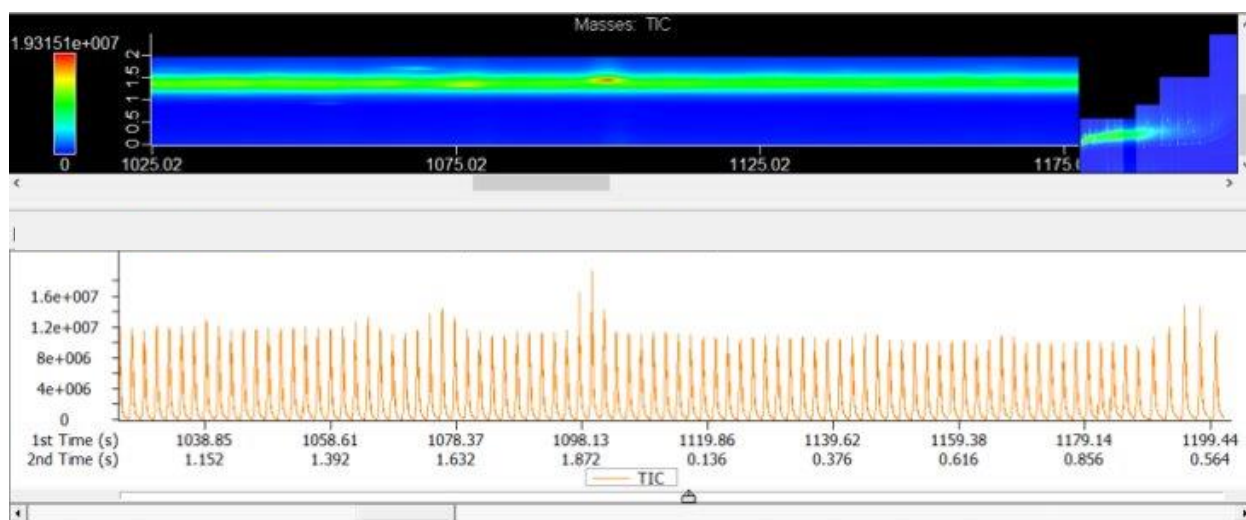


Figure B-8 The real GCxGC image of Sample 48 that belongs to cluster 1

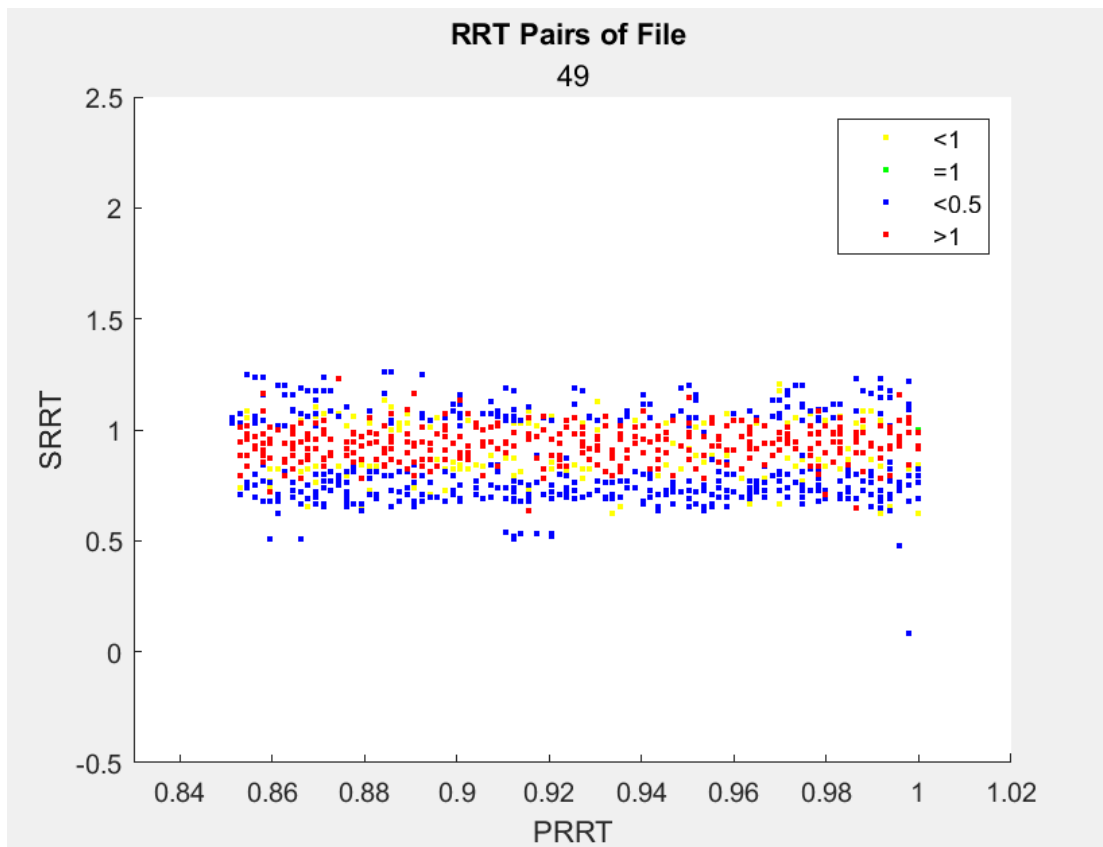


Figure B-9 The discrete GCxGC image of Sample 49 that belongs to cluster 1

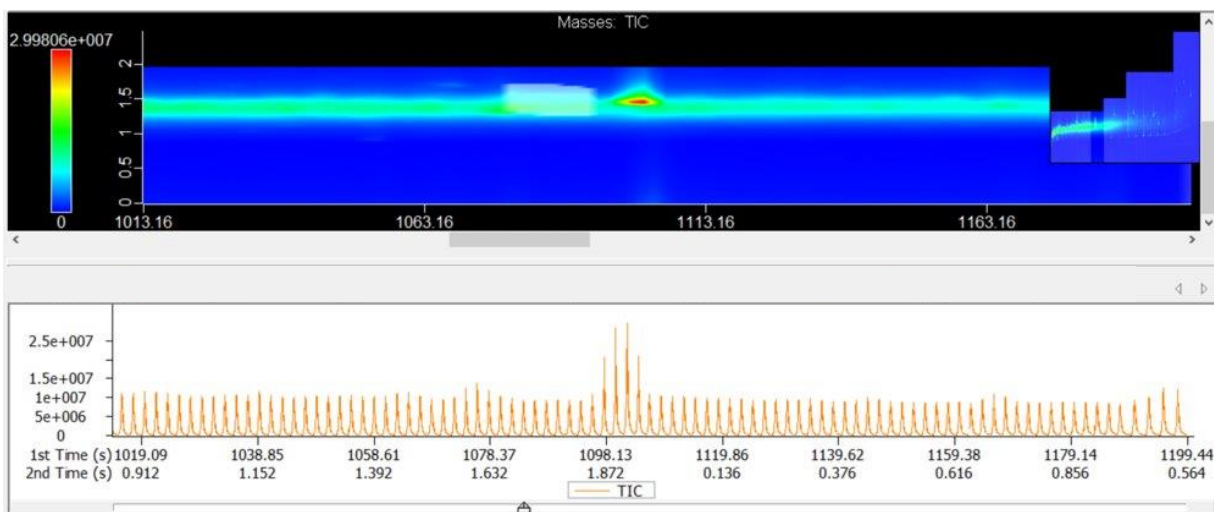


Figure B-10 The real GCxGC image of Sample 49 that belongs to cluster 1

Cluster 2: Samples 31, and 59

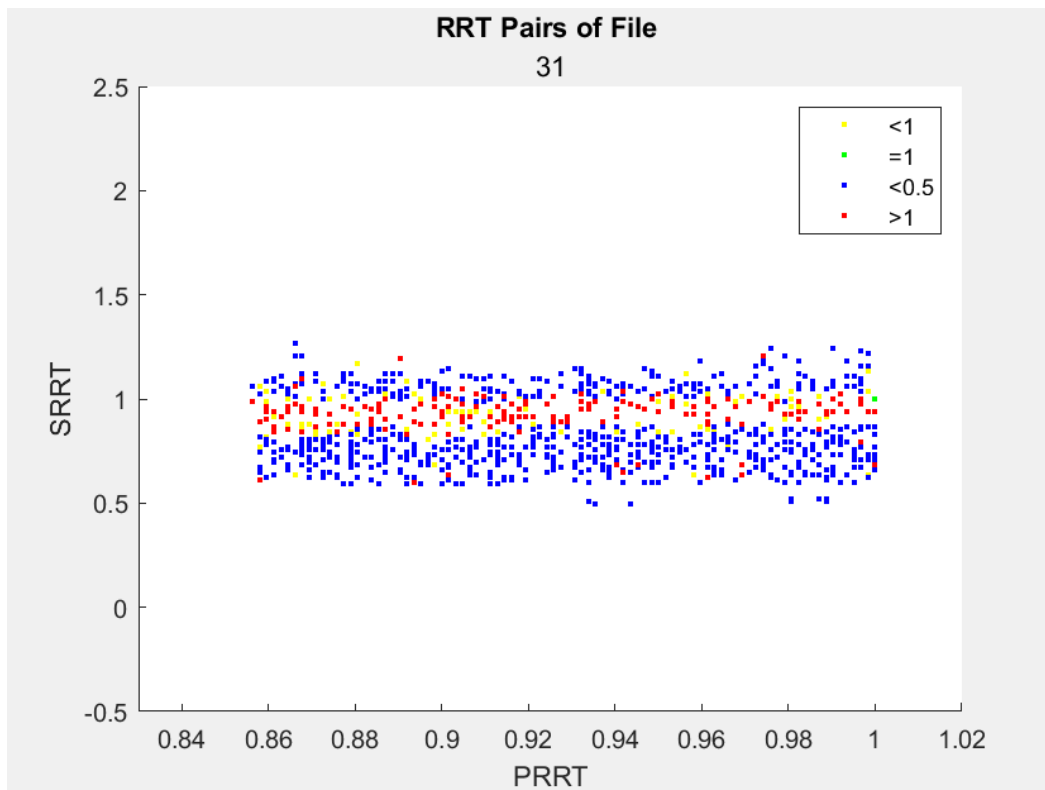


Figure B-11 The discrete GCxGC image of Sample 31 that belongs to cluster 2

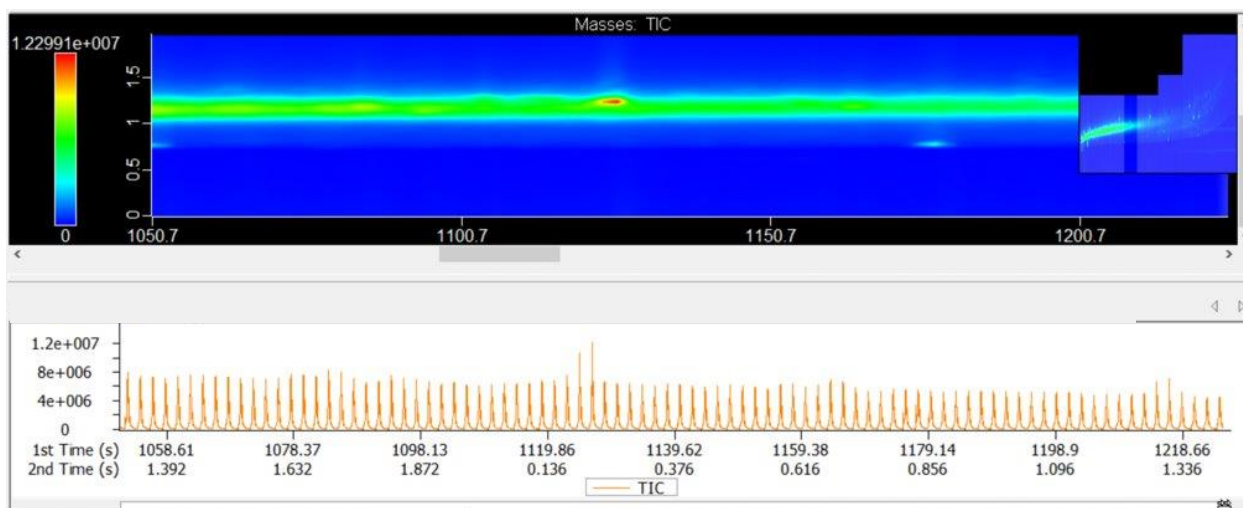


Figure B-12 The real GCxGC image of Sample 31 that belongs to cluster 2

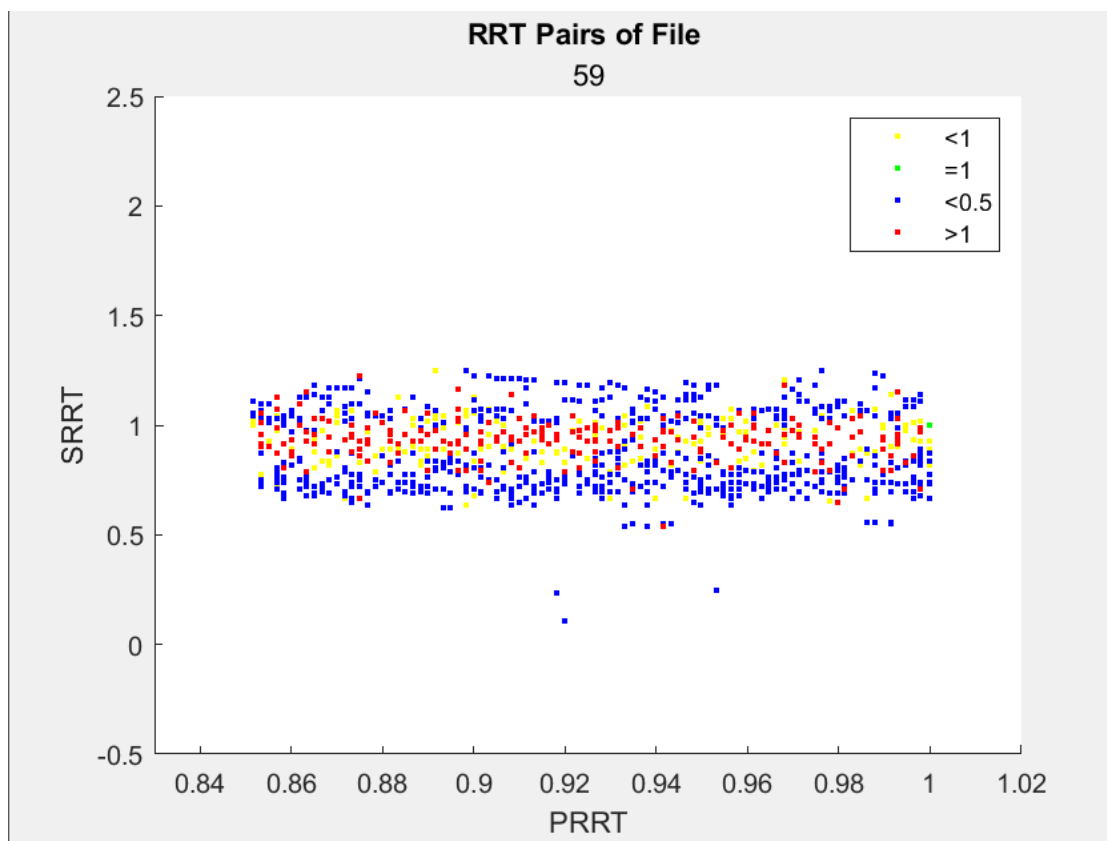


Figure B-13 The discrete GCxGC image of Sample 59 that belongs to cluster 2

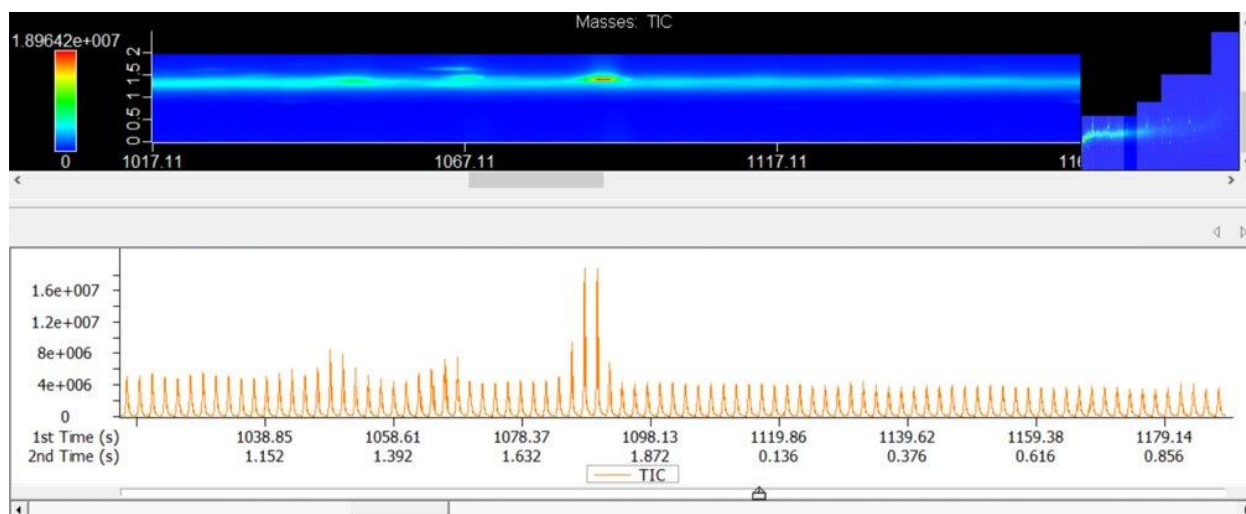


Figure B-14 The real GCxGC image of Sample 59 that belongs to cluster 2

Cluster 3: Samples 1, 9, 29, 46, 58, 63, 66, 70, and 71

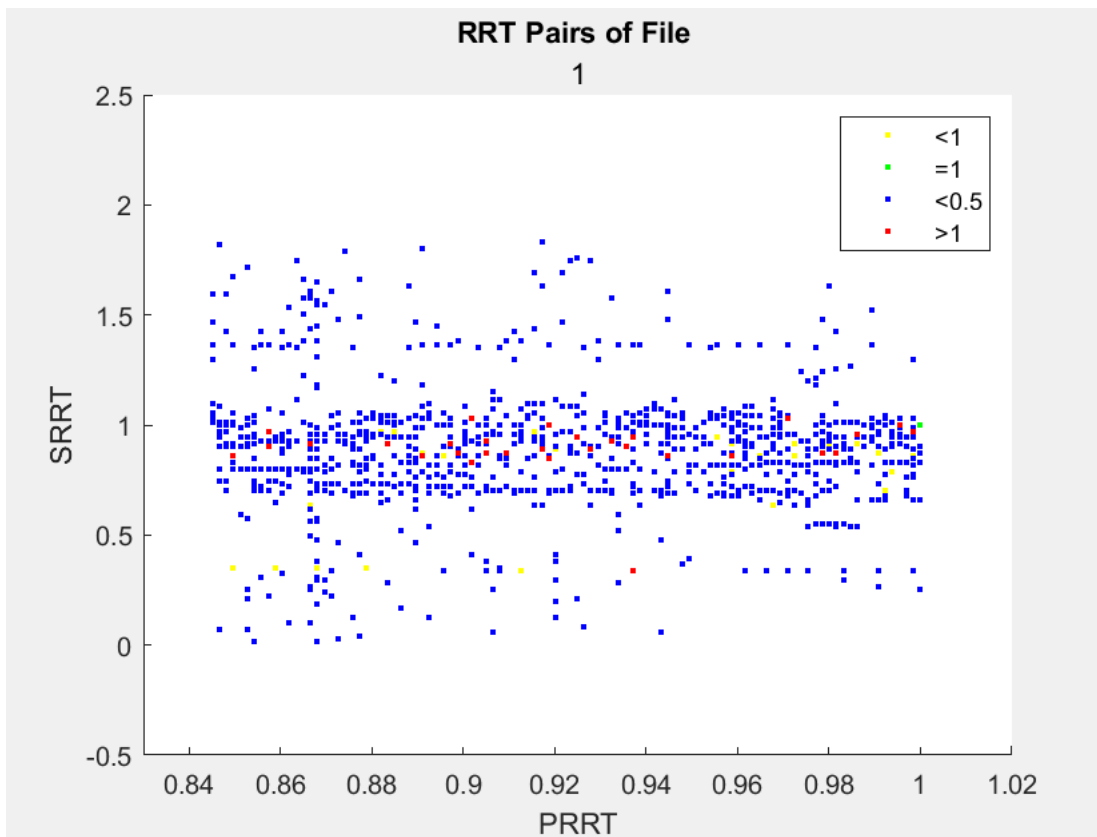


Figure B-15 The discrete GCxGC image of Sample 1 that belongs to cluster 3

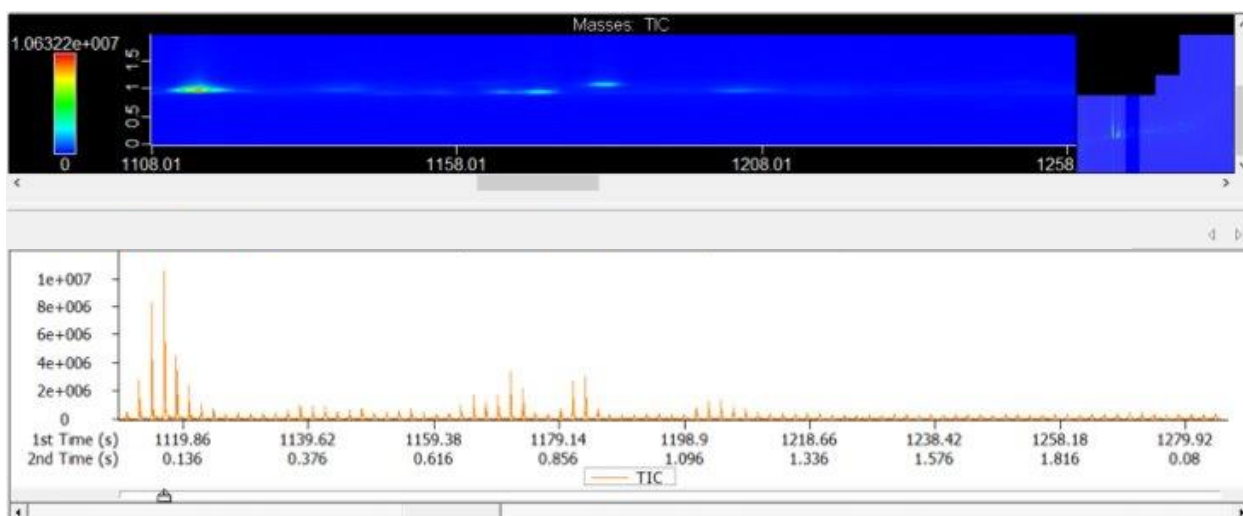


Figure B-16 The real GCxGC image of Sample 1 that belongs to cluster 3



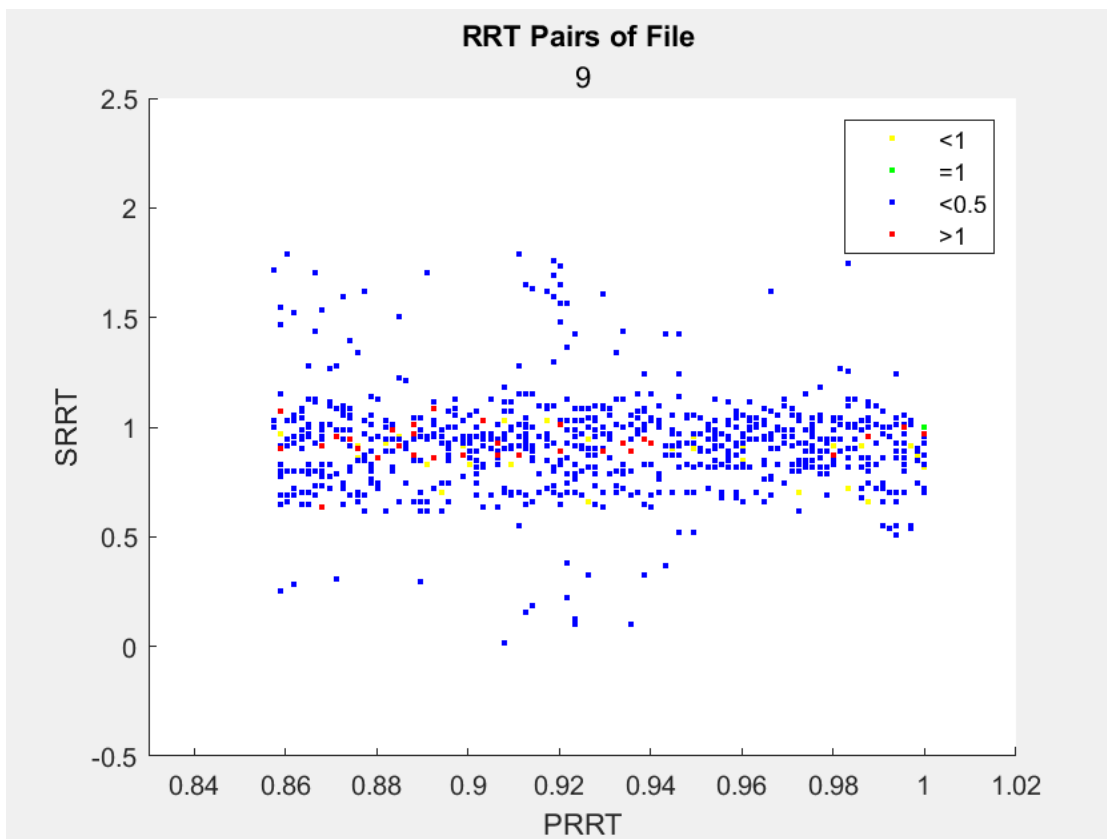


Figure B-17 The discrete GCxGC image of Sample 9 that belongs to cluster 3

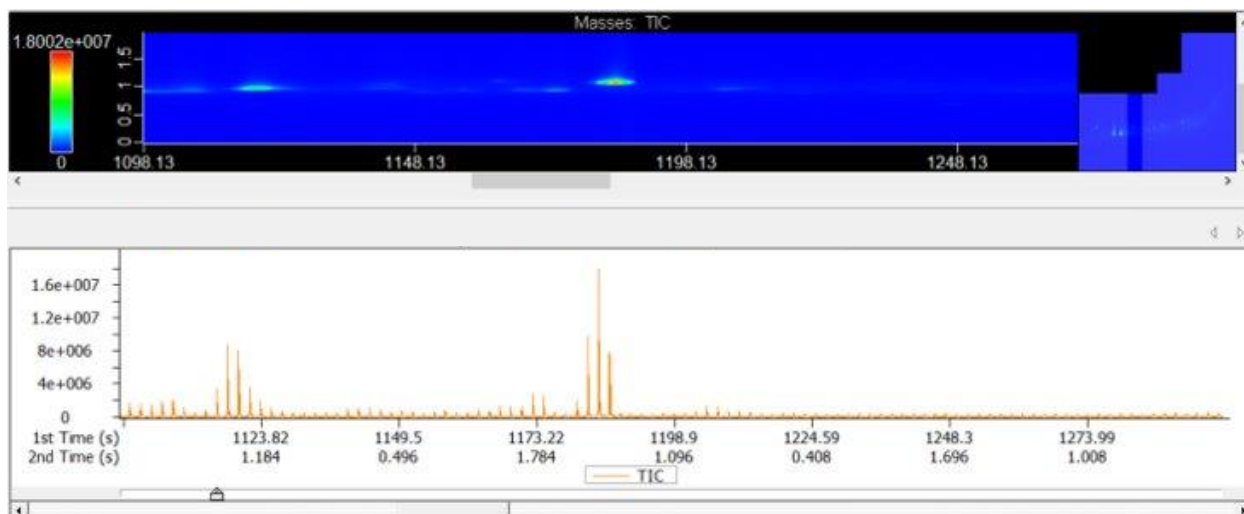


Figure B-18 The real GCxGC image of Sample 9 that belongs to cluster 3

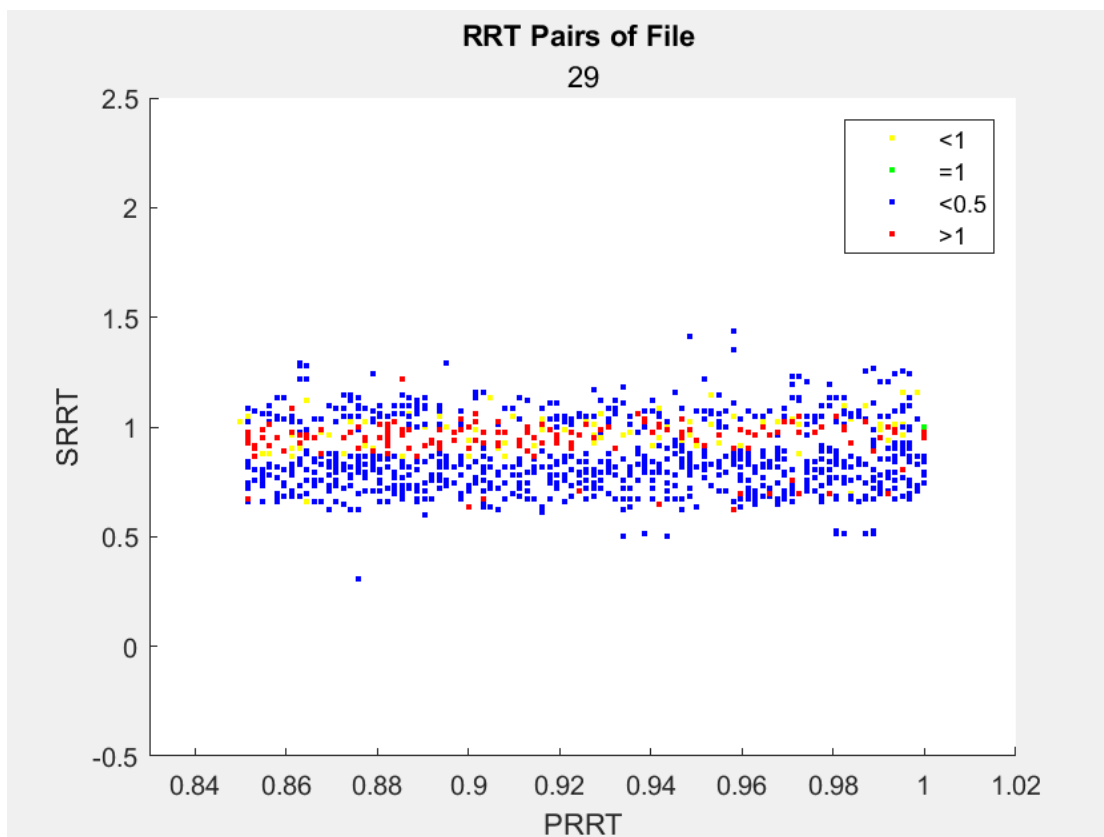


Figure B-19 The discrete GCxGC image of Sample 29 that belongs to cluster 3

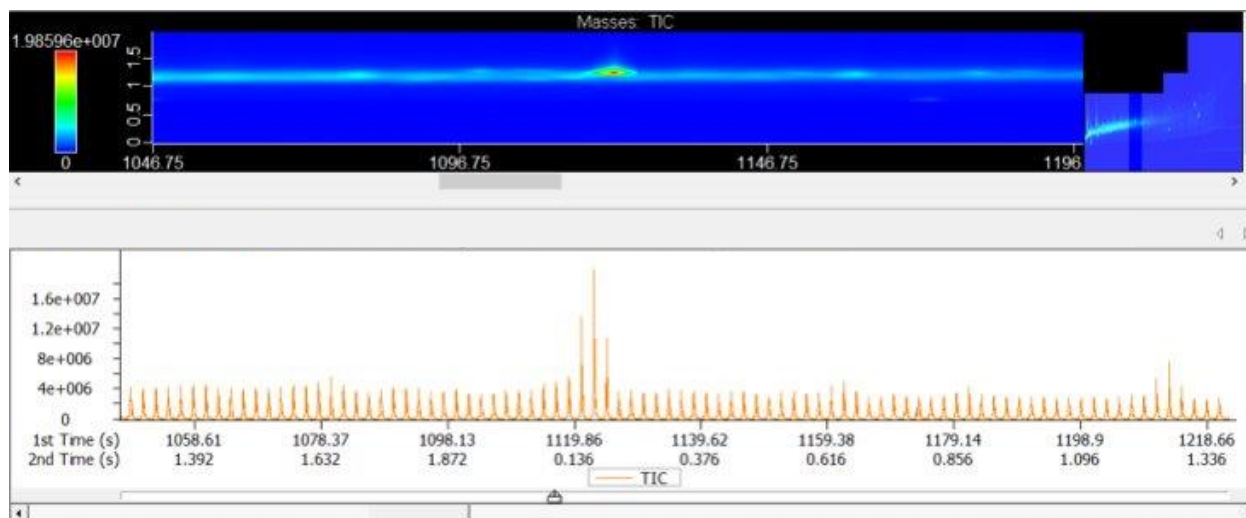


Figure B-20 The real GCxGC image of Sample 29 that belongs to cluster 3

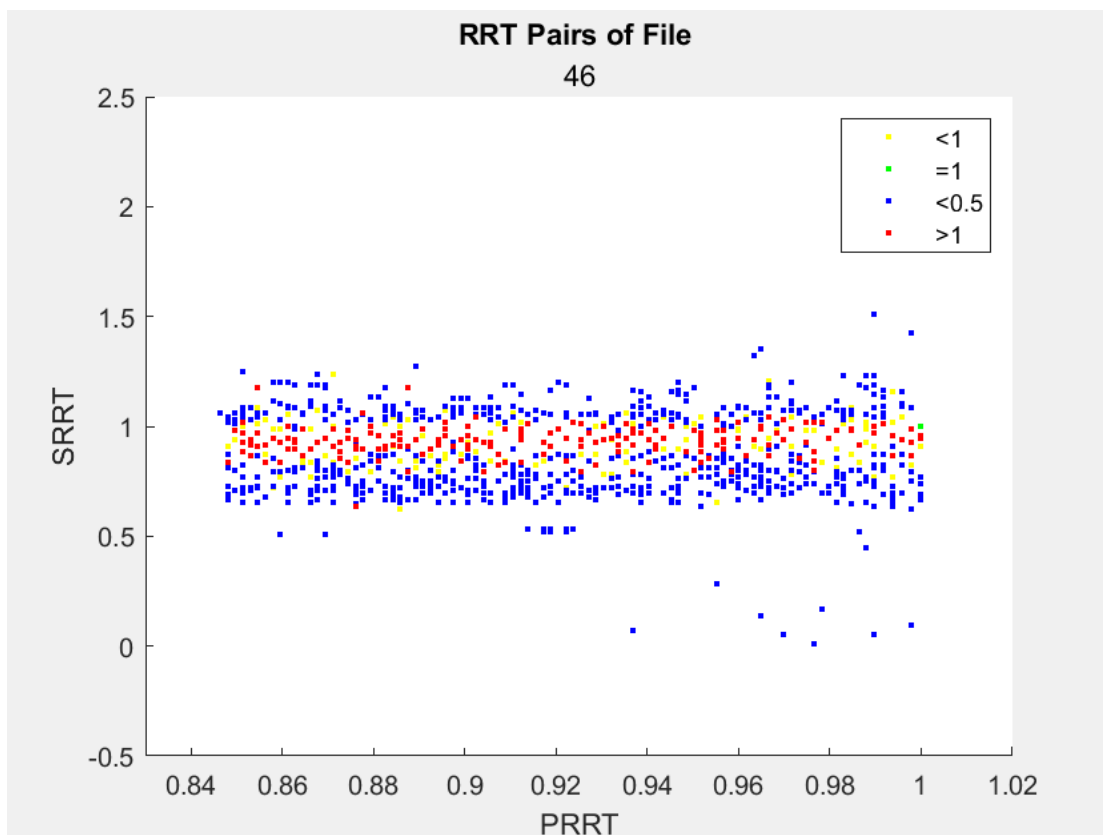


Figure B-21 The discrete GCxGC image of Sample 46 that belongs to cluster 3

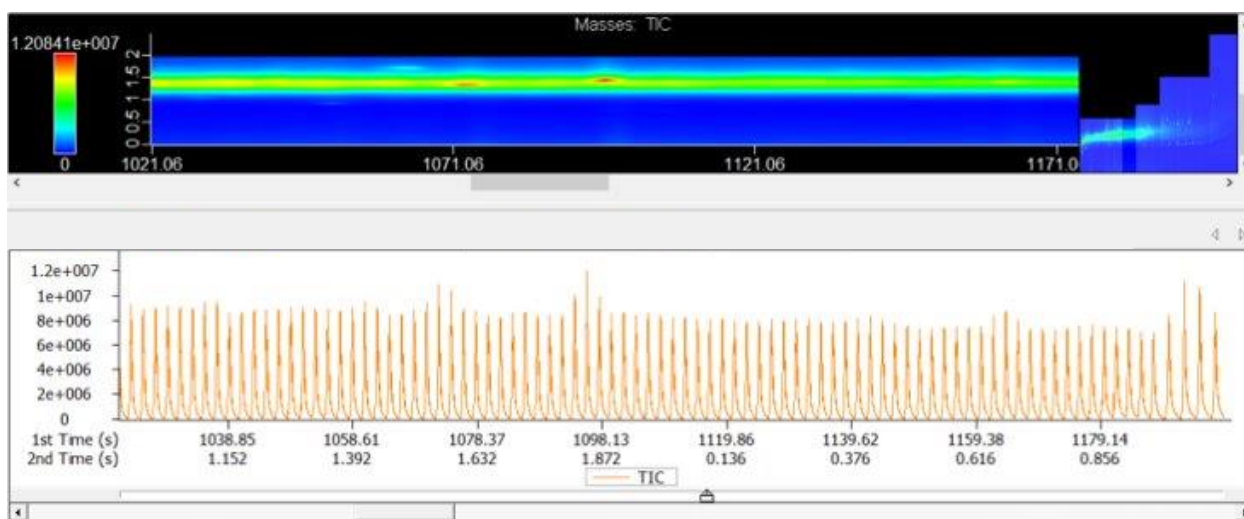


Figure B-22 The real GCxGC image of Sample 46 that belongs to cluster 3

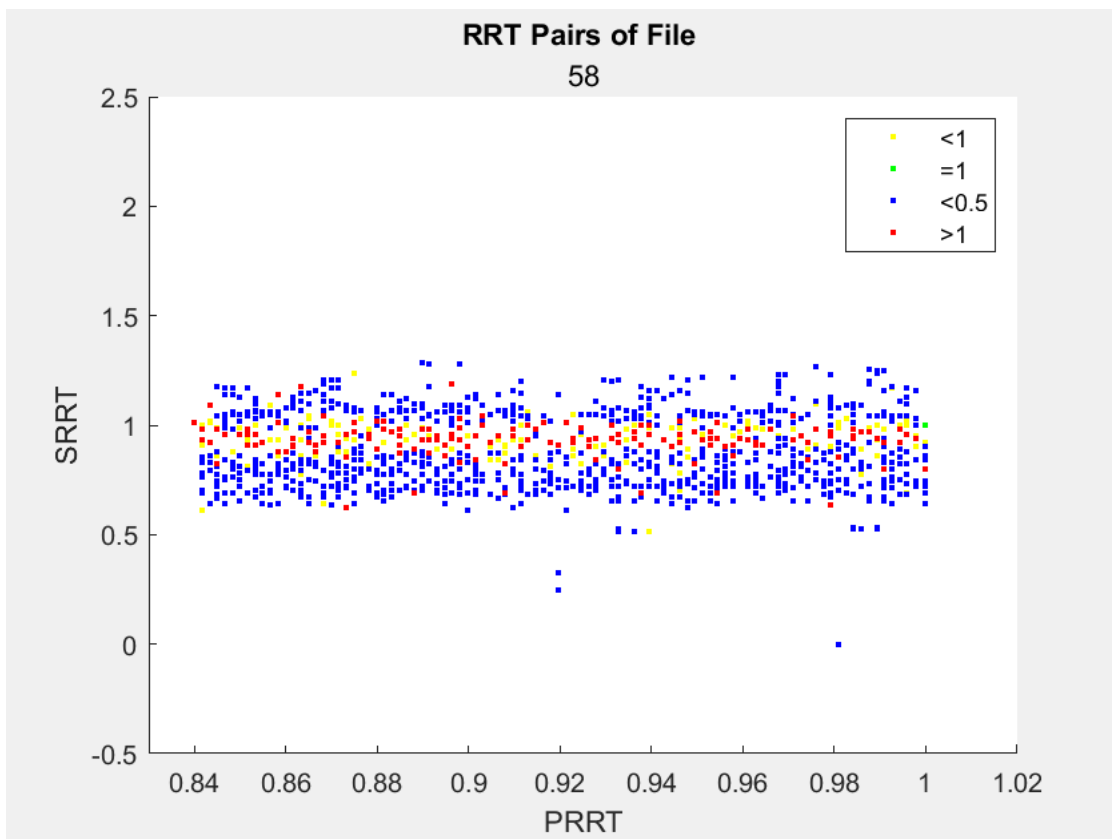


Figure B-23 The discrete GCxGC image of Sample 58 that belongs to cluster 3

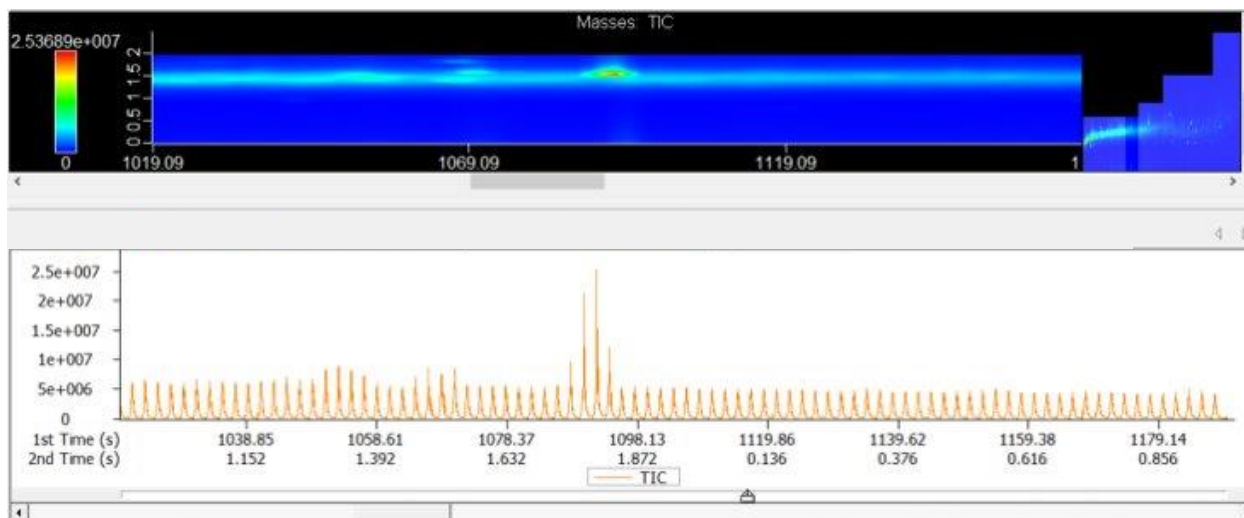


Figure B-24 The real GCxGC image of Sample 58 that belongs to cluster 3

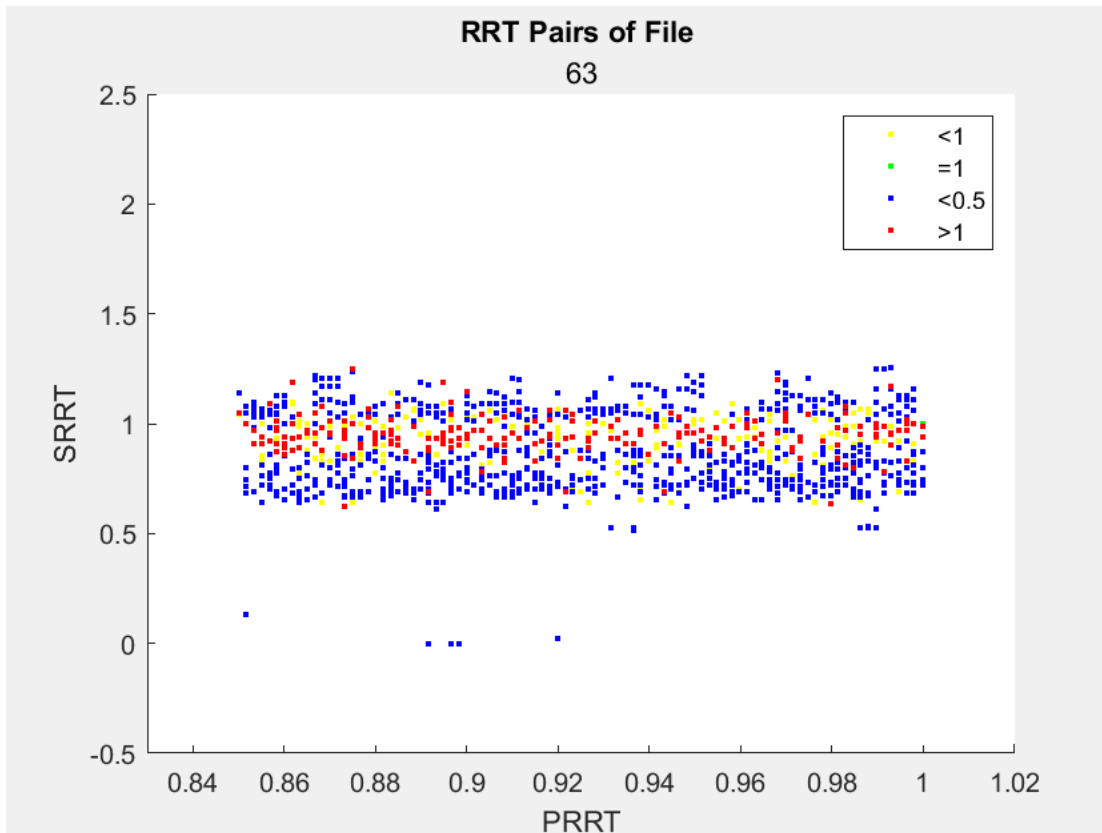


Figure B-25 The discrete GCxGC image of Sample 63 that belongs to cluster 3

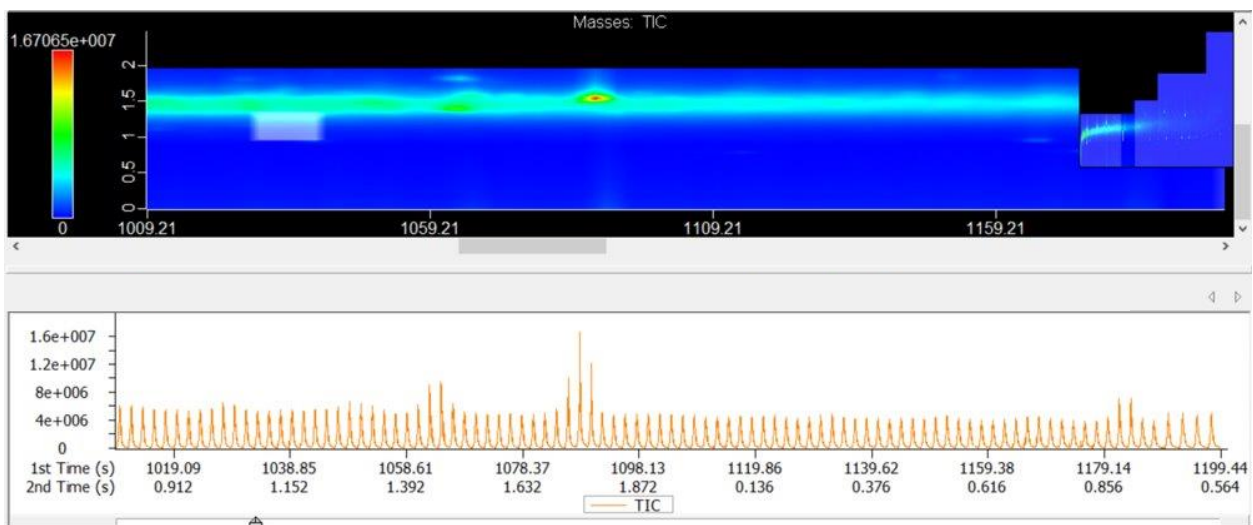


Figure B-26 The real GCxGC image of Sample 63 that belongs to cluster 3

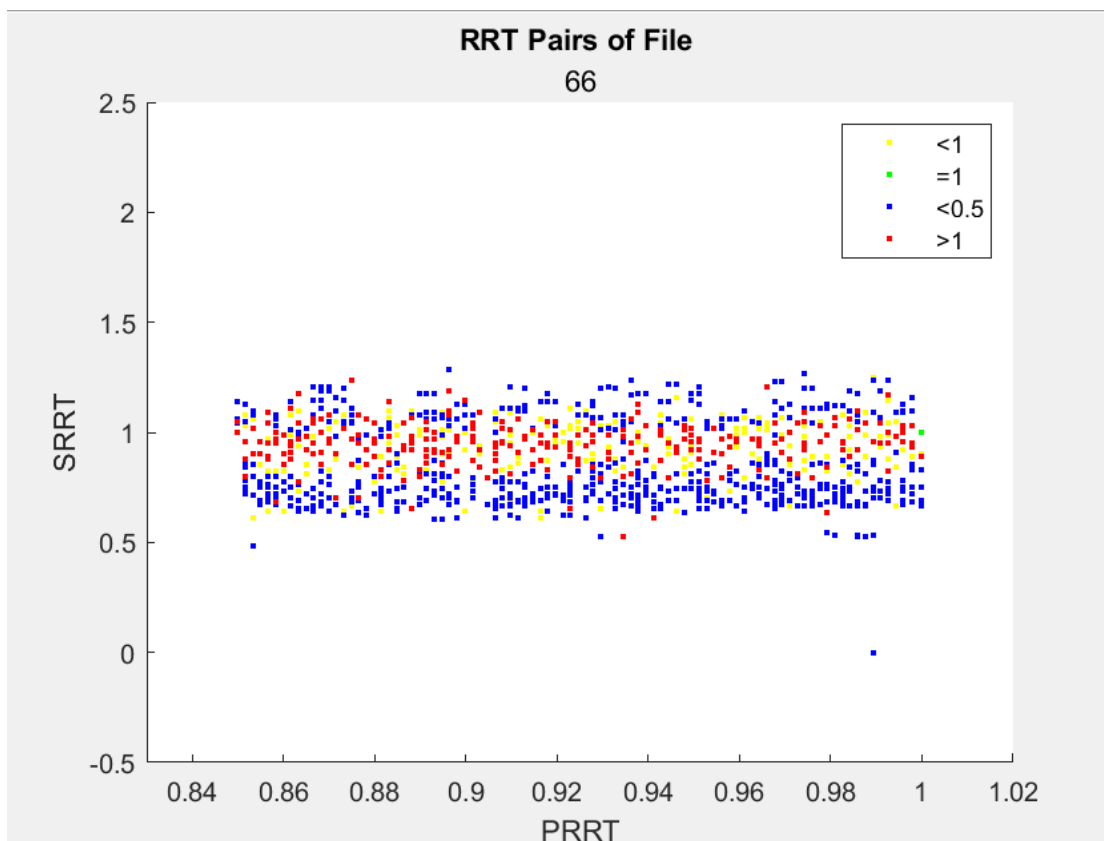


Figure B-27 The discrete GCxGC image of Sample 66 that belongs to cluster 3

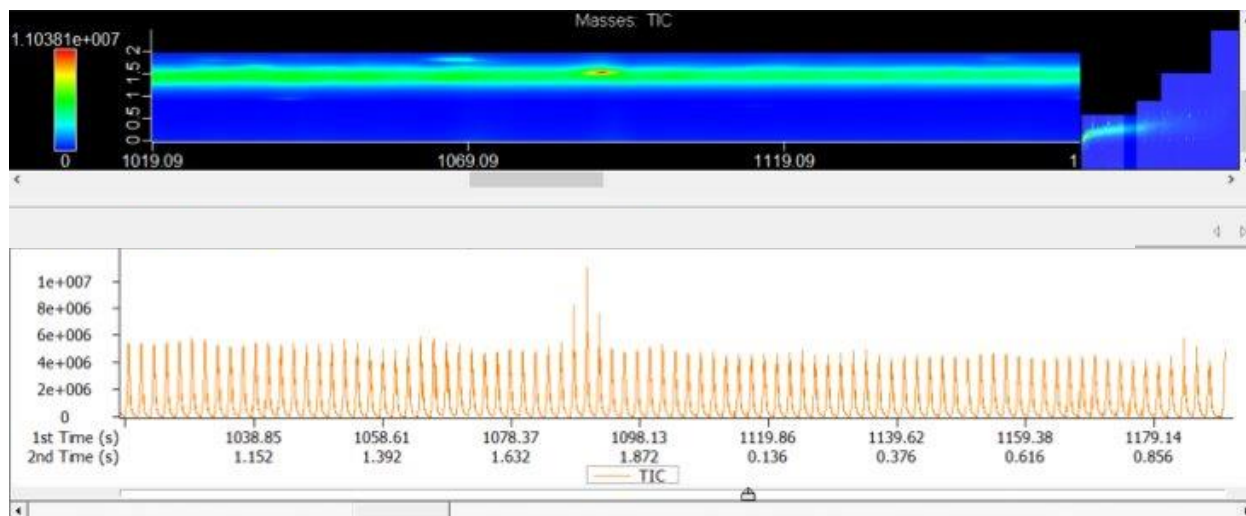


Figure B-28 The real GCxGC image of Sample 66 that belongs to cluster 3

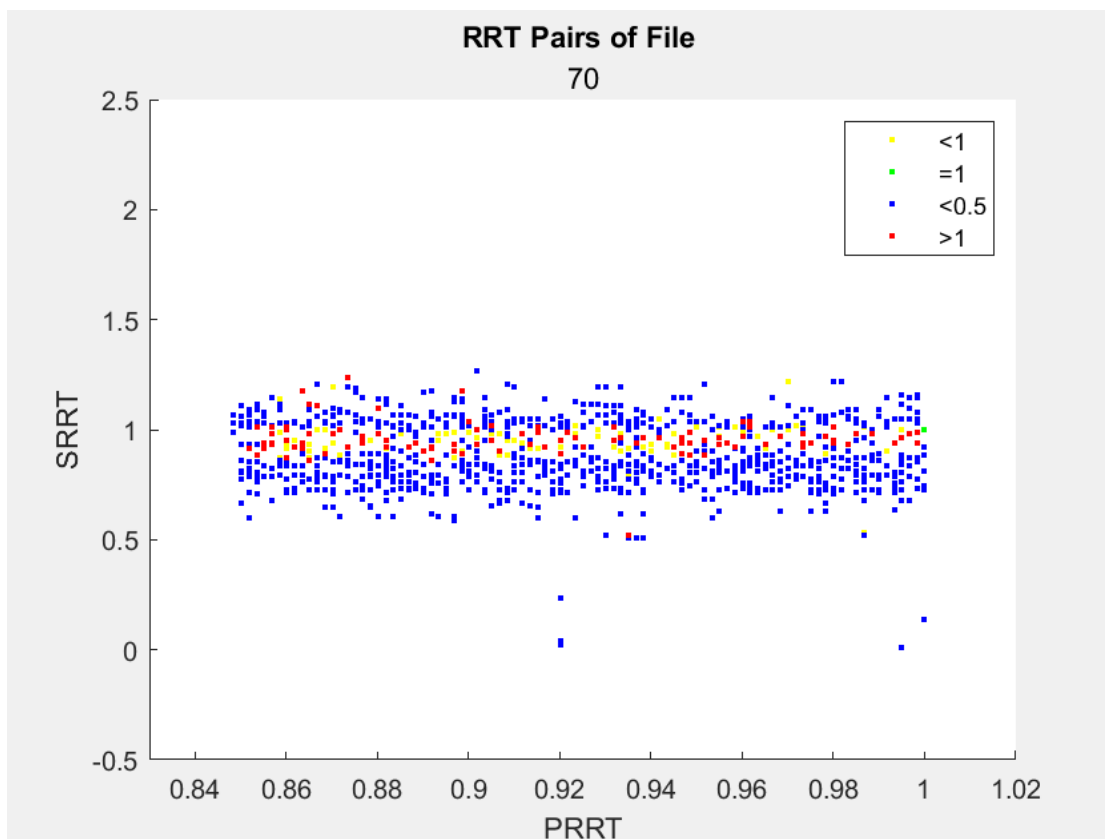


Figure B-29 The discrete GCxGC image of Sample 70 that belongs to cluster 3

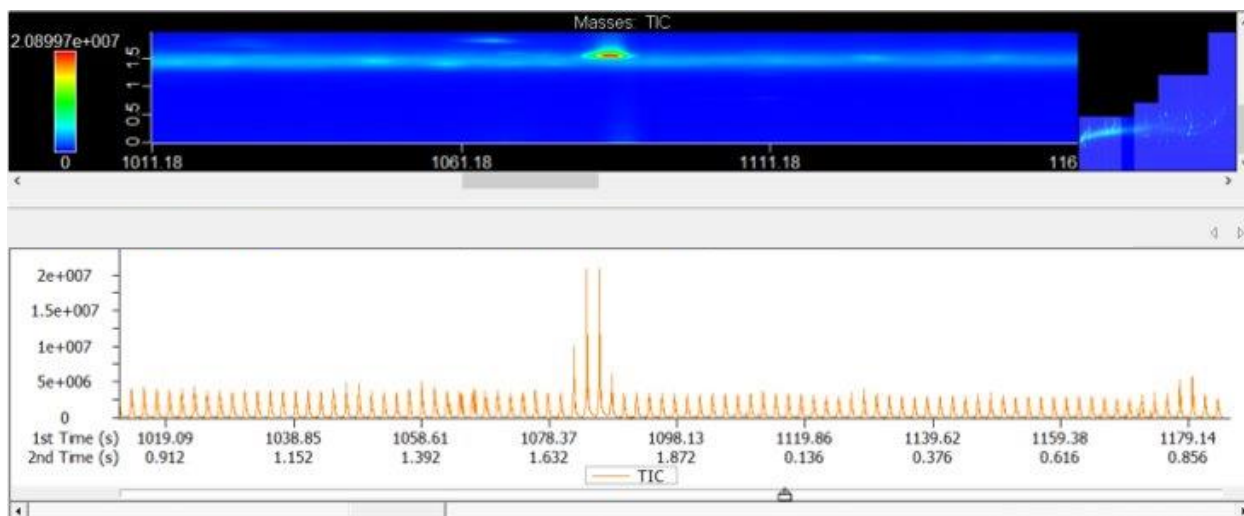


Figure B-30 The real GCxGC image of Sample 70 that belongs to cluster 3

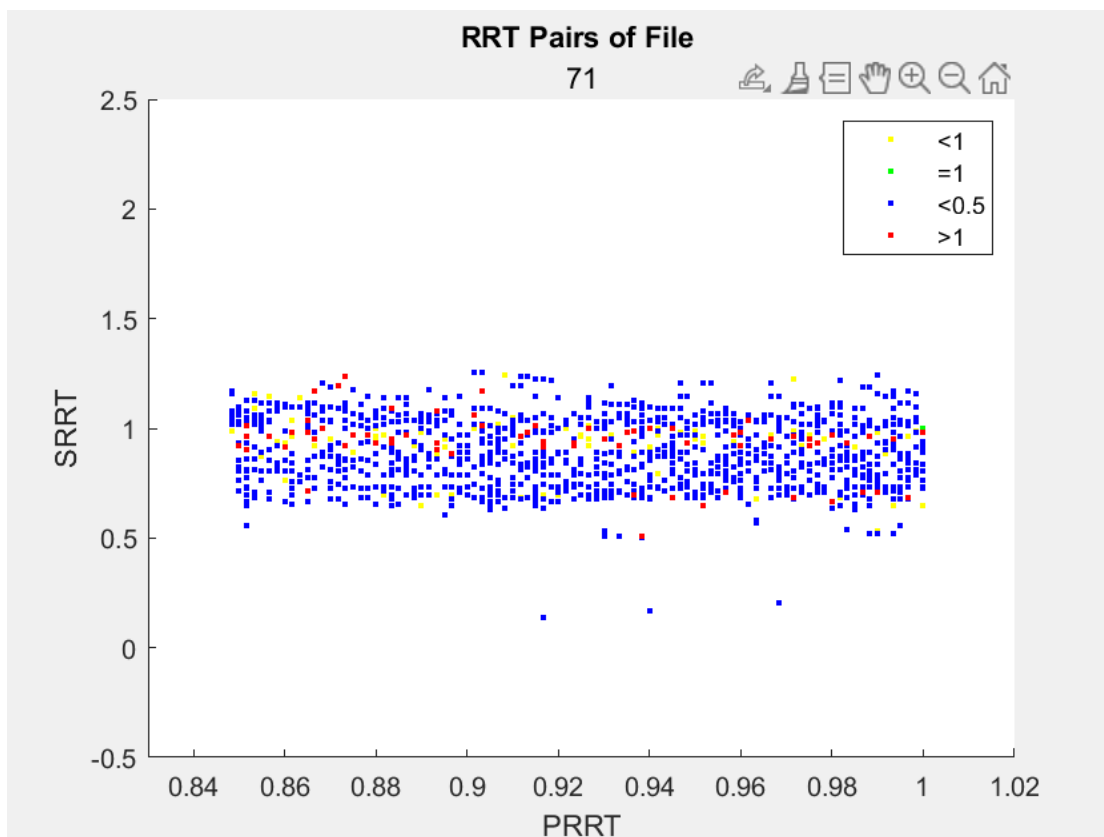


Figure B-31 The discrete GCxGC image of Sample 71 that belongs to cluster 3

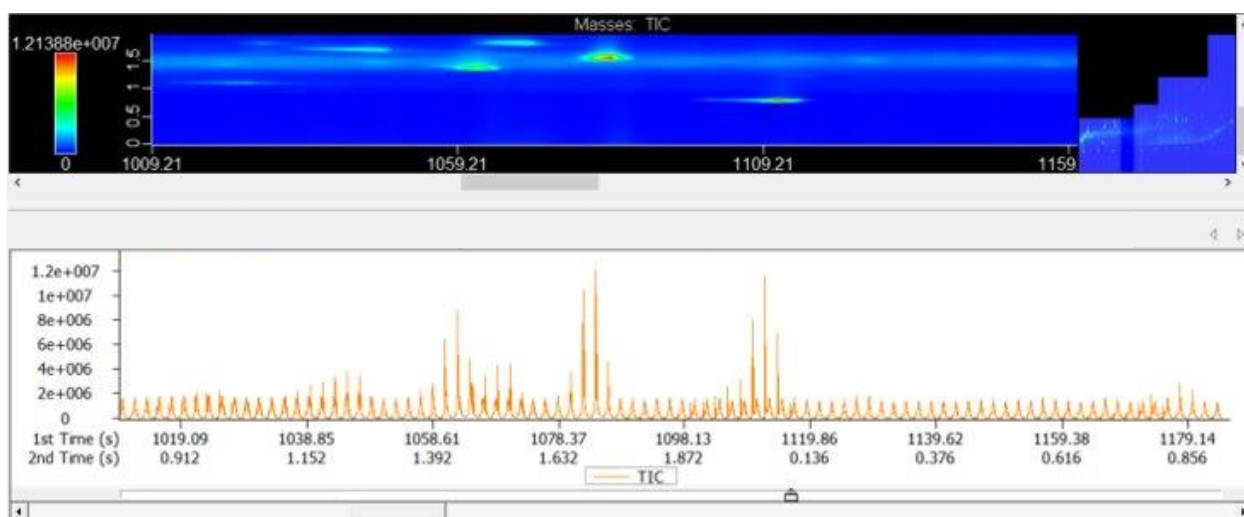


Figure B-32 The real GCxGC image of Sample 71 that belongs to cluster 3



Cluster 4: Samples 23, 38, 51, 52, 60, 61, 68, and 69

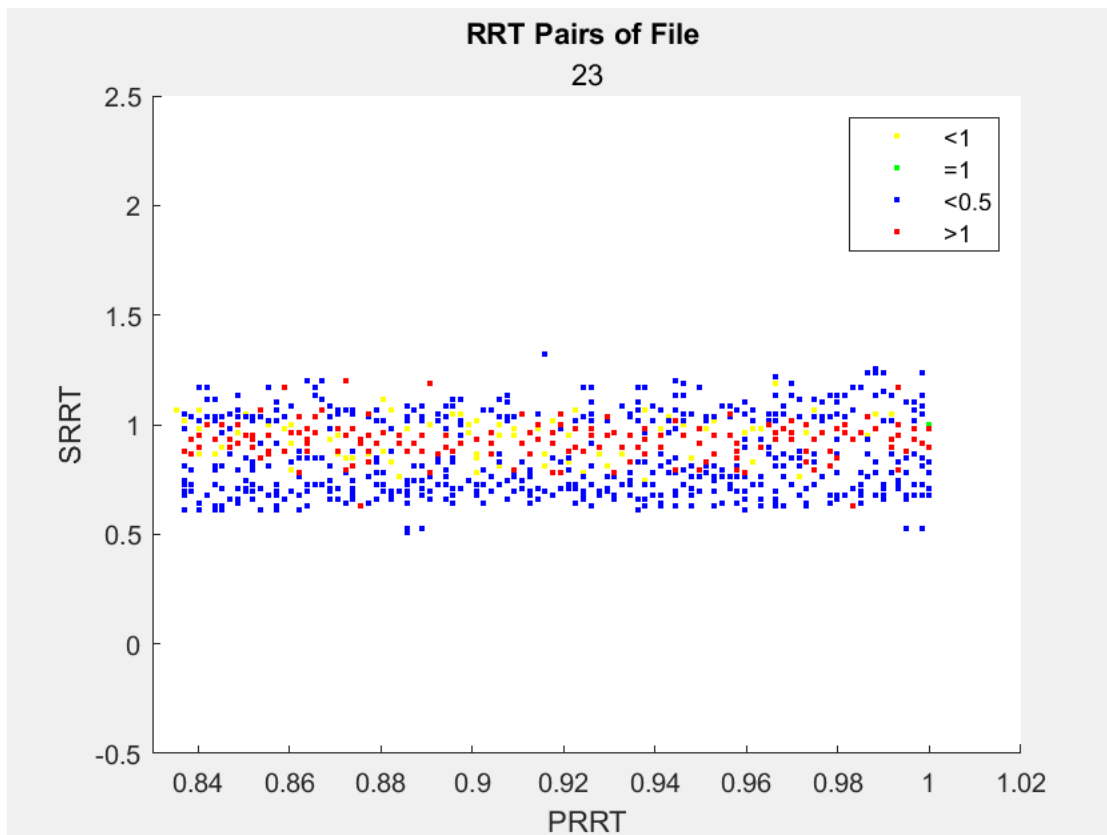


Figure B-33 The discrete GCxGC image of Sample 23 that belongs to cluster 4

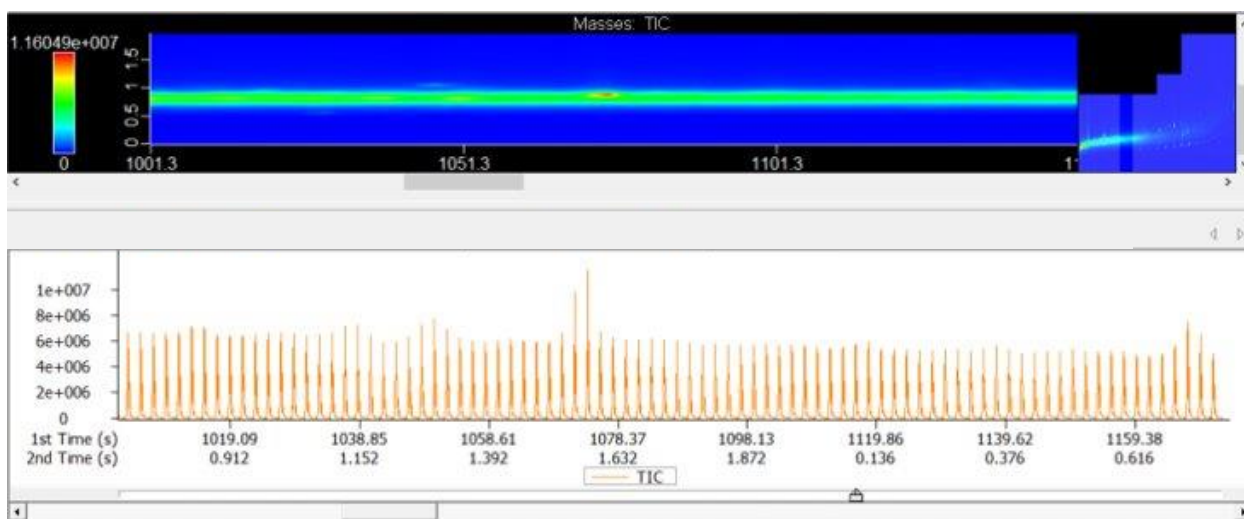


Figure B-34 The real GCxGC image of Sample 23 that belongs to cluster 4

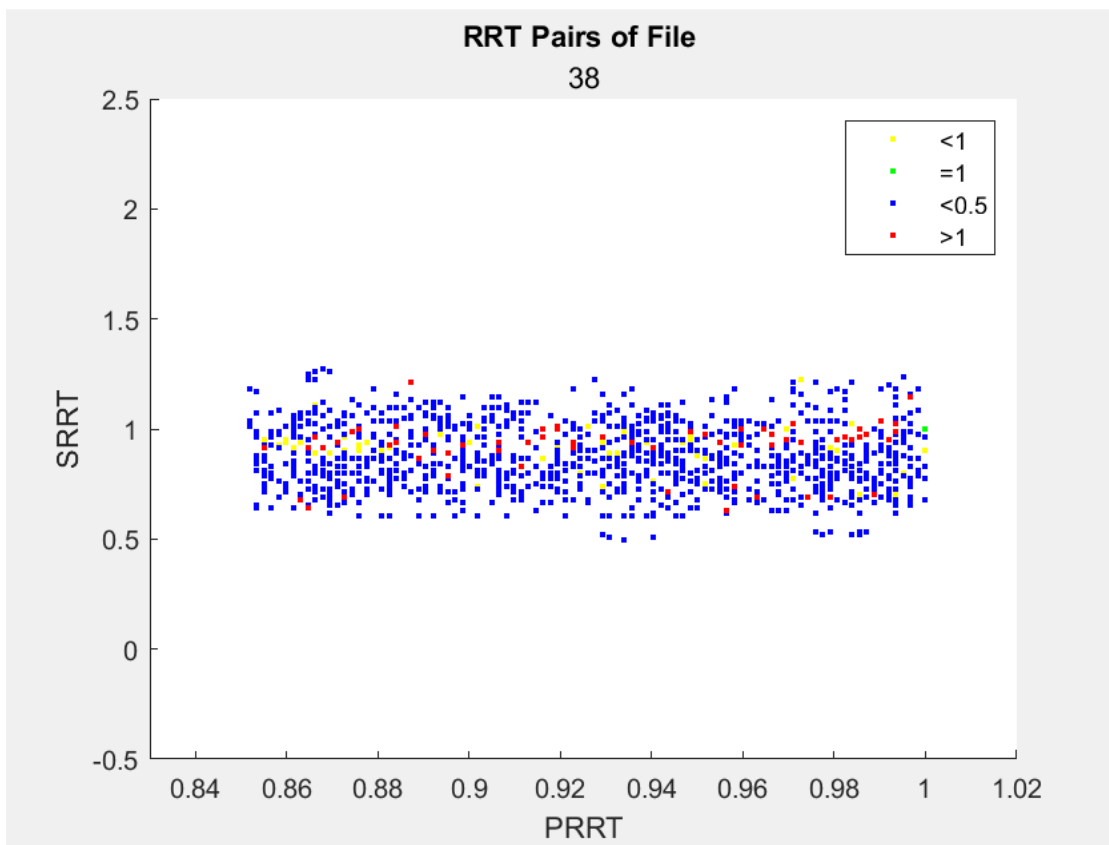


Figure B-35 The discrete GCxGC image of Sample 38 that belongs to cluster 4

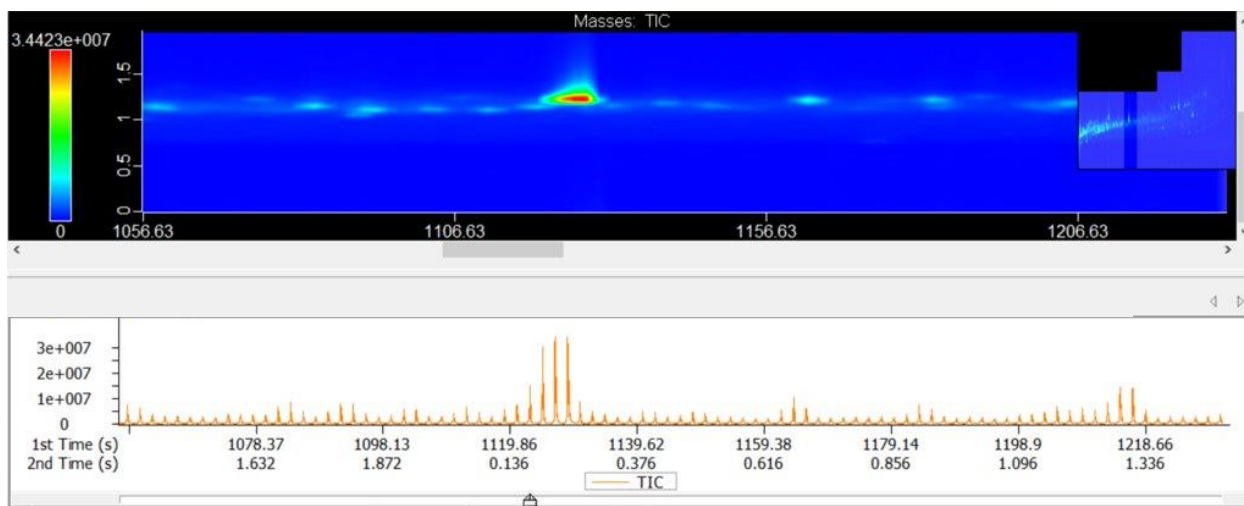


Figure B-36 The real GCxGC image of Sample 38 that belongs to cluster 4

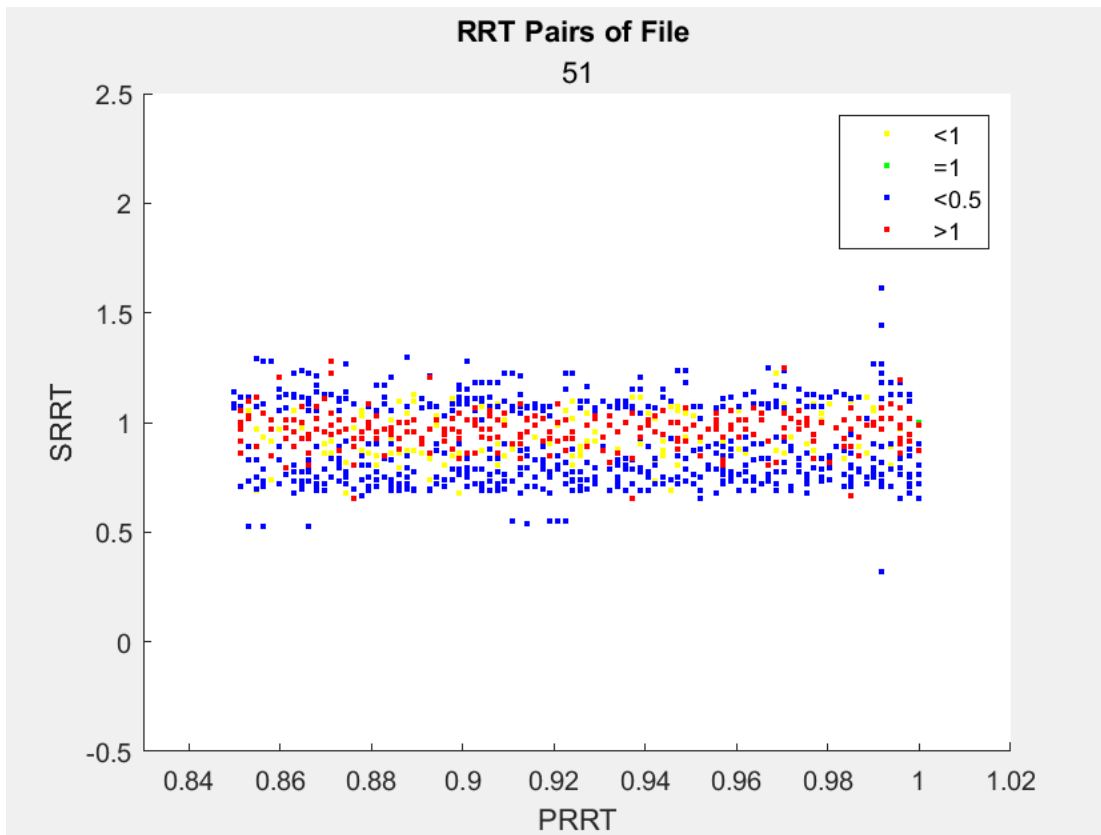


Figure B-37 The discrete GCxGC image of Sample 51 that belongs to cluster 4

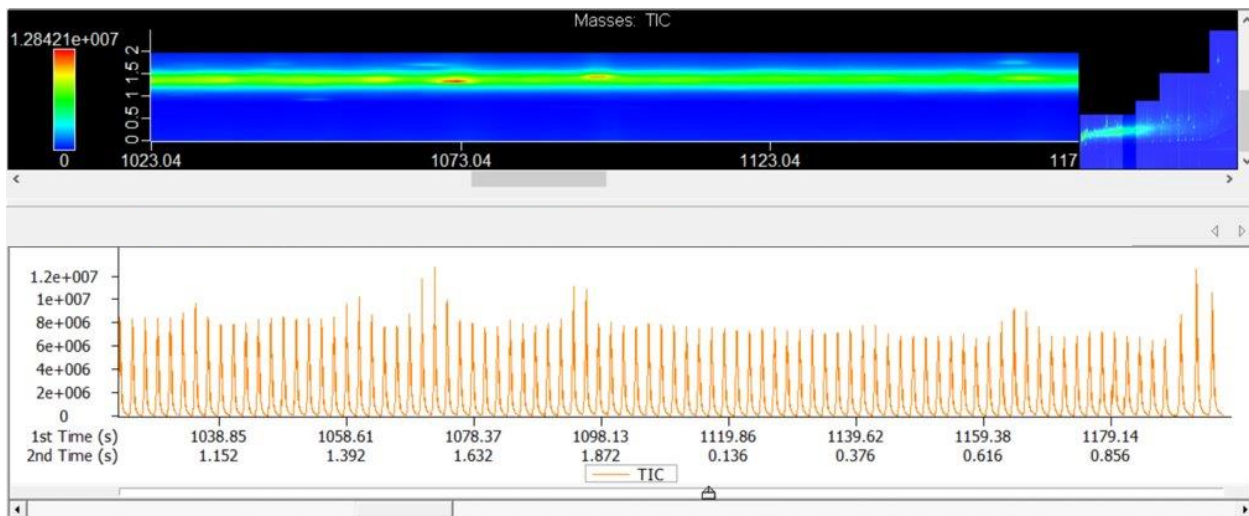


Figure B-38 The real GCxGC image of Sample 51 that belongs to cluster 4

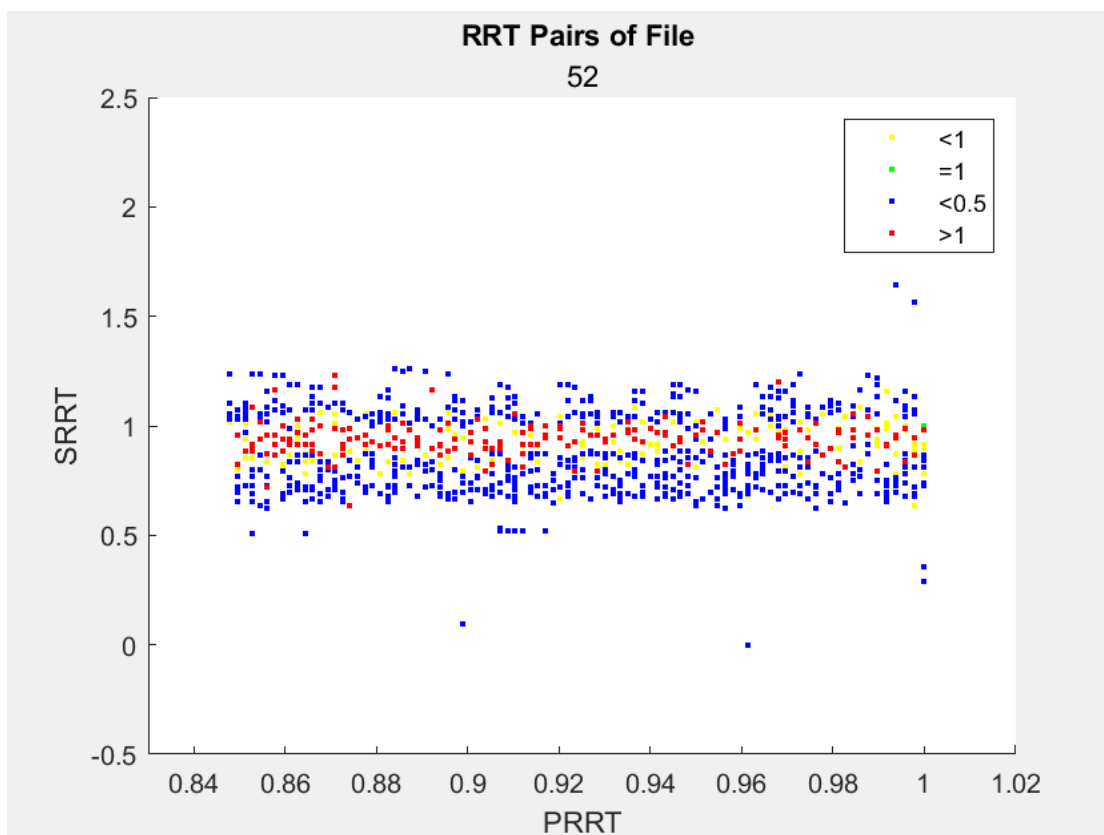


Figure B-39 The discrete GCxGC image of Sample 52 that belongs to cluster 4

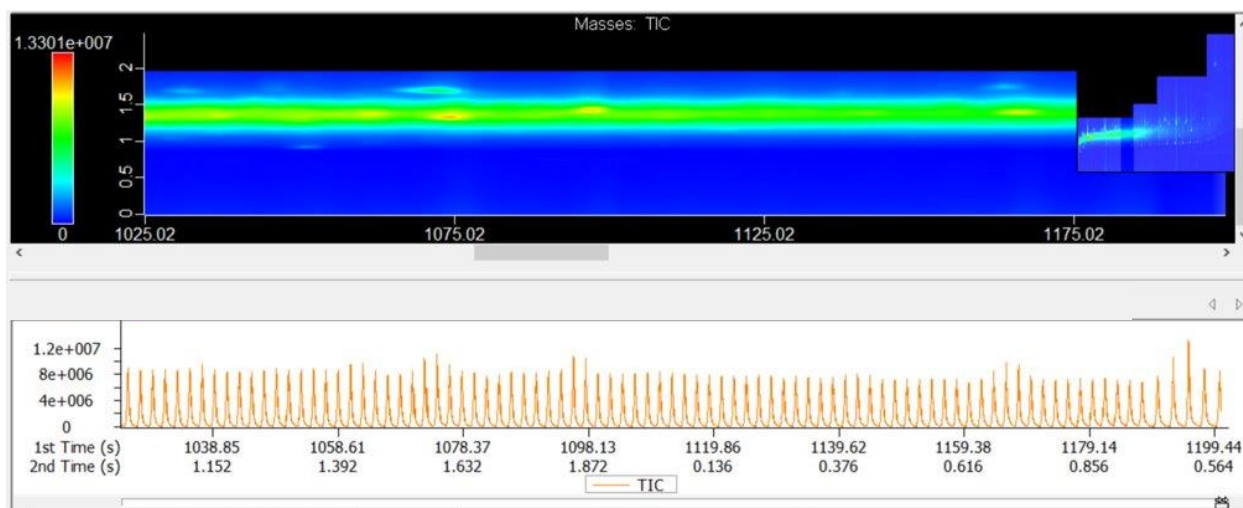


Figure B-40 The real GCxGC image of Sample 52 that belongs to cluster 4

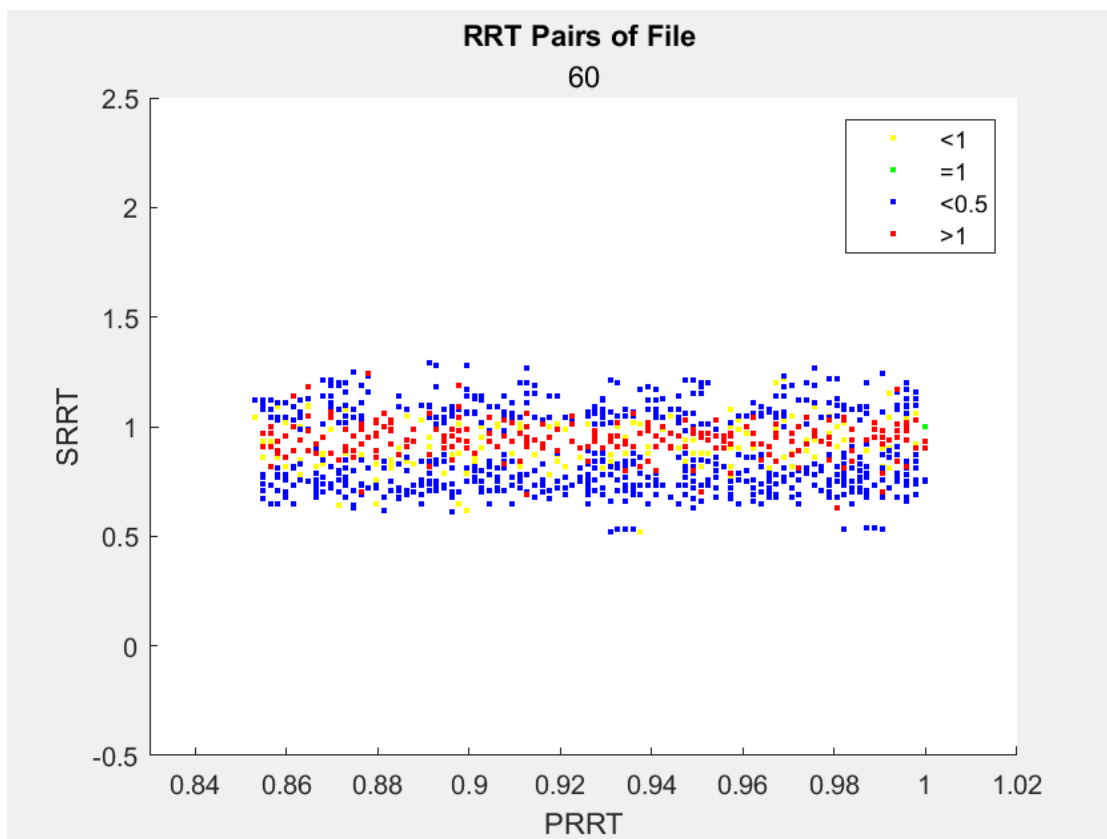


Figure B-41 The discrete GCxGC image of Sample 60 that belongs to cluster 4

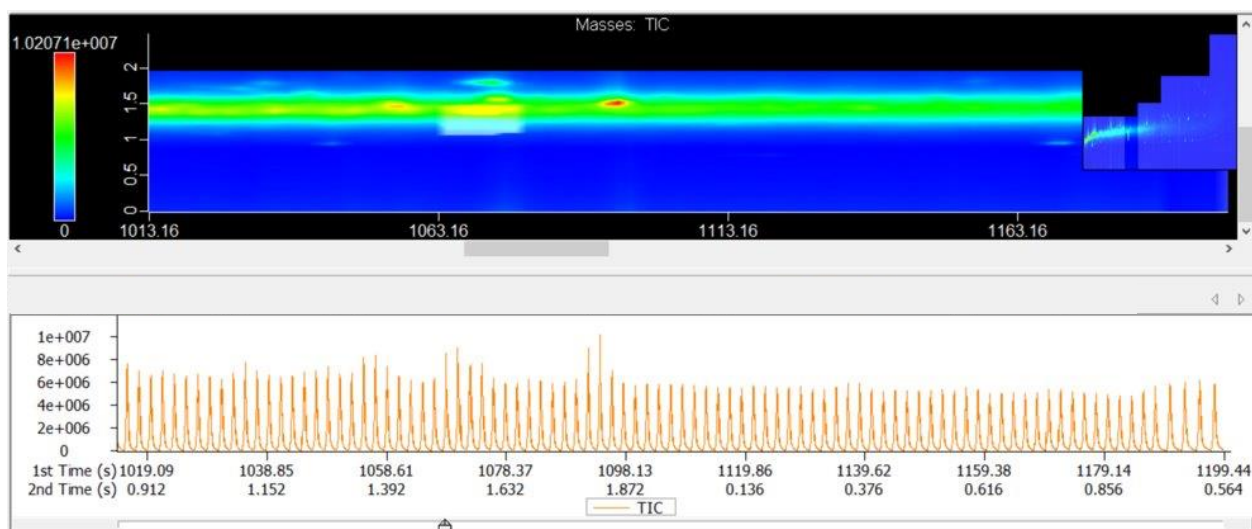


Figure B-42 The real GCxGC image of Sample 60 that belongs to cluster 4

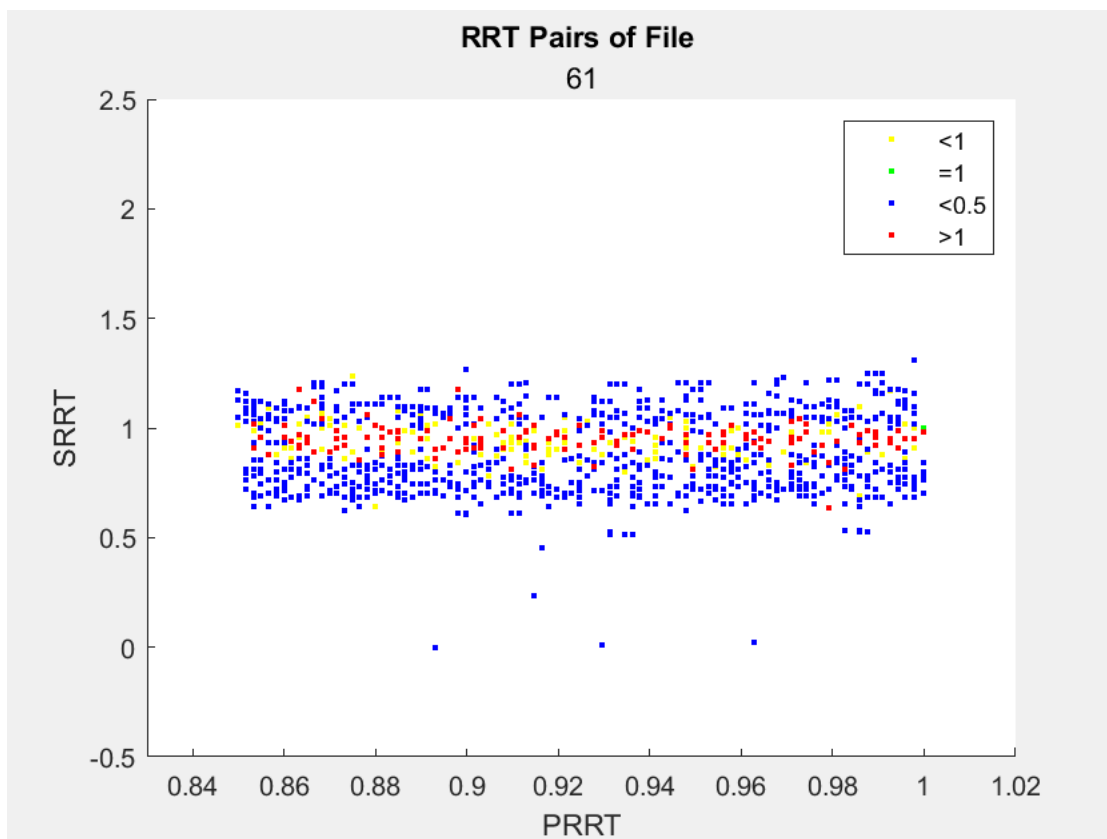


Figure B-43 The discrete GCxGC image of Sample 61 that belongs to cluster 4

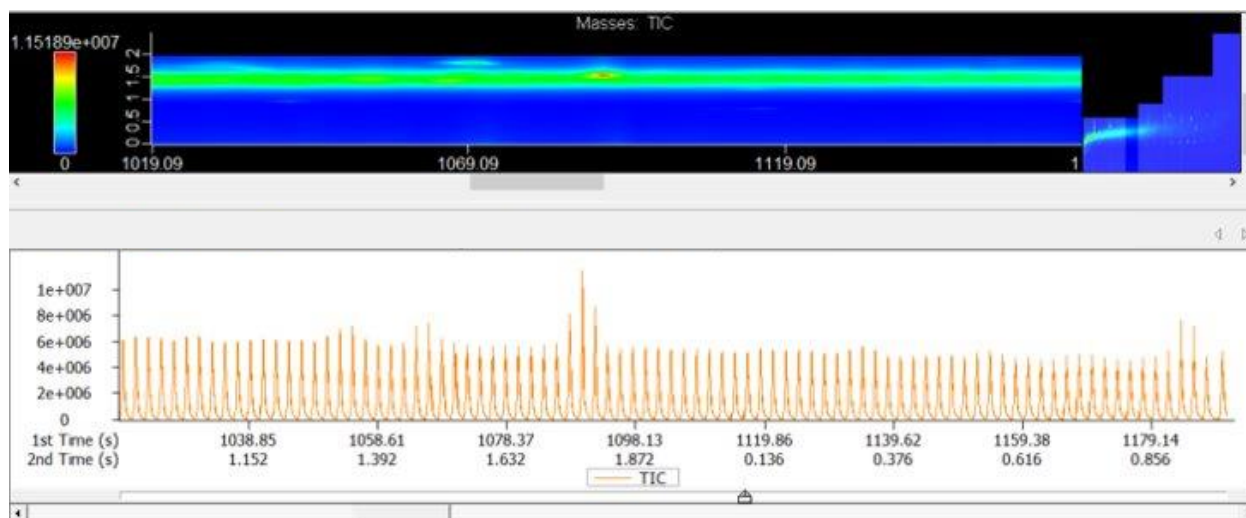


Figure B-44 The real GCxGC image of Sample 61 that belongs to cluster 4

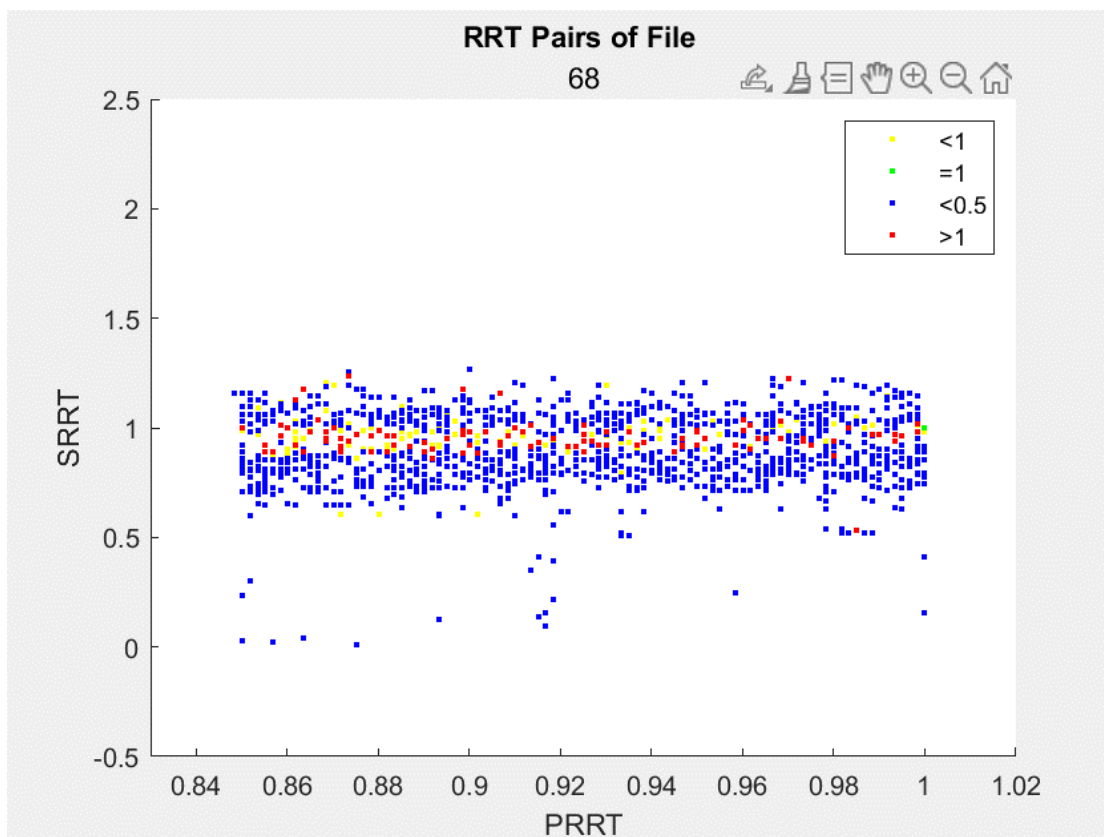


Figure B-45 The discrete GCxGC image of Sample 68 that belongs to cluster 4

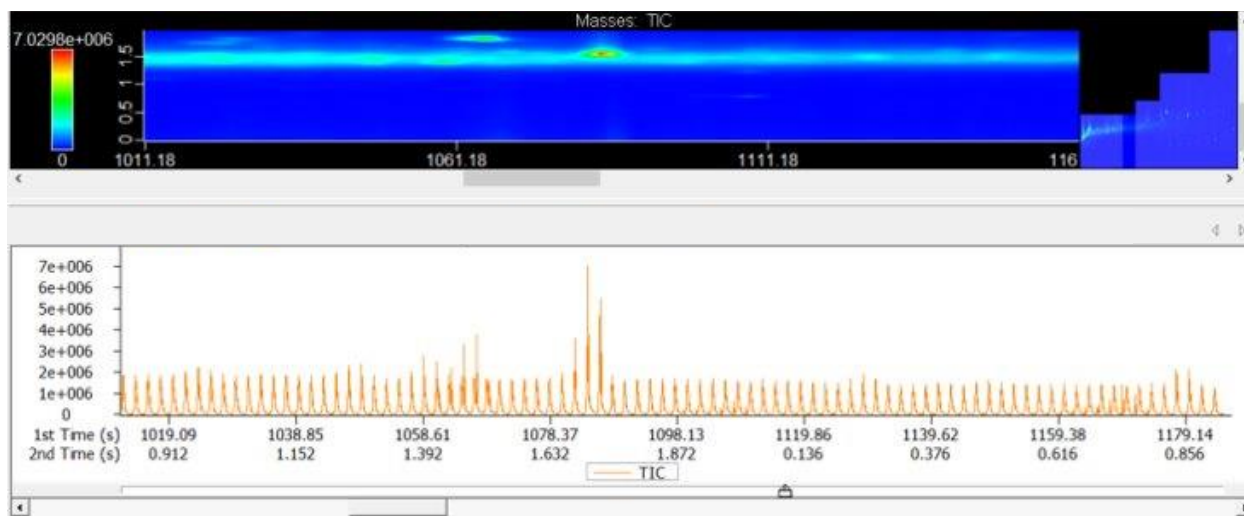


Figure B-46 The real GCxGC image of Sample 68 that belongs to cluster 4

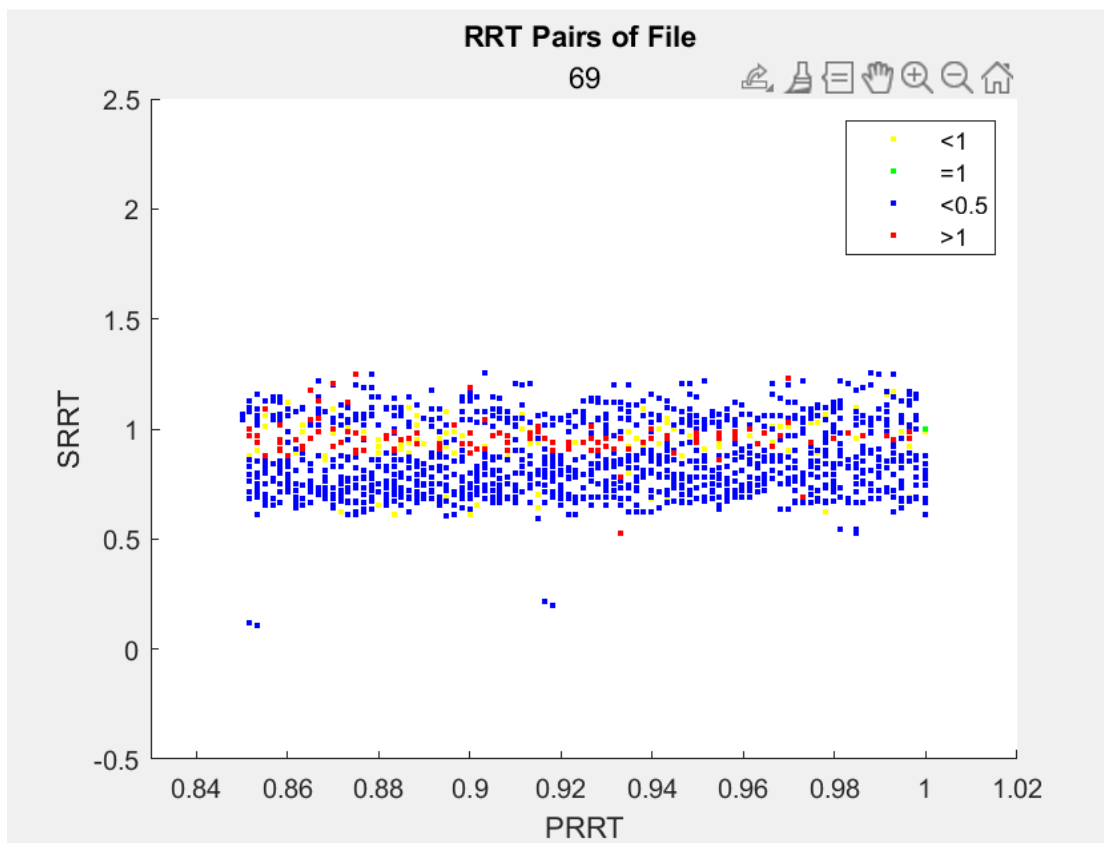


Figure B-47 The discrete GCxGC image of Sample 69 that belongs to cluster 4

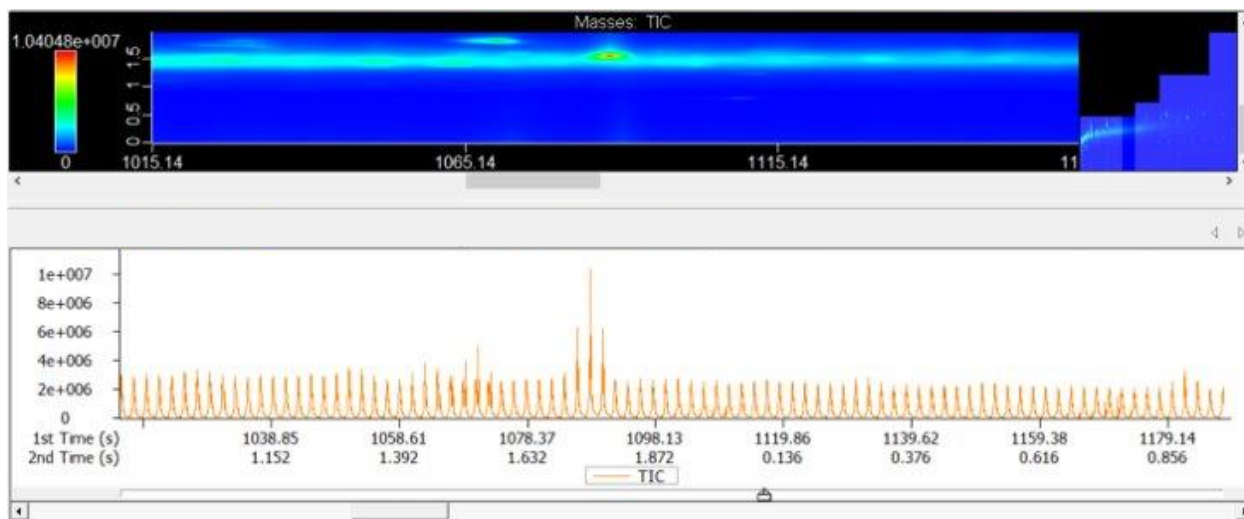


Figure B-48 The real GCxGC image of Sample 69 that belongs to cluster 4



Cluster 5: Samples 22, 26, 39, 50, 53, 54, 55, 57, 62, 64, 65, 67, and 72

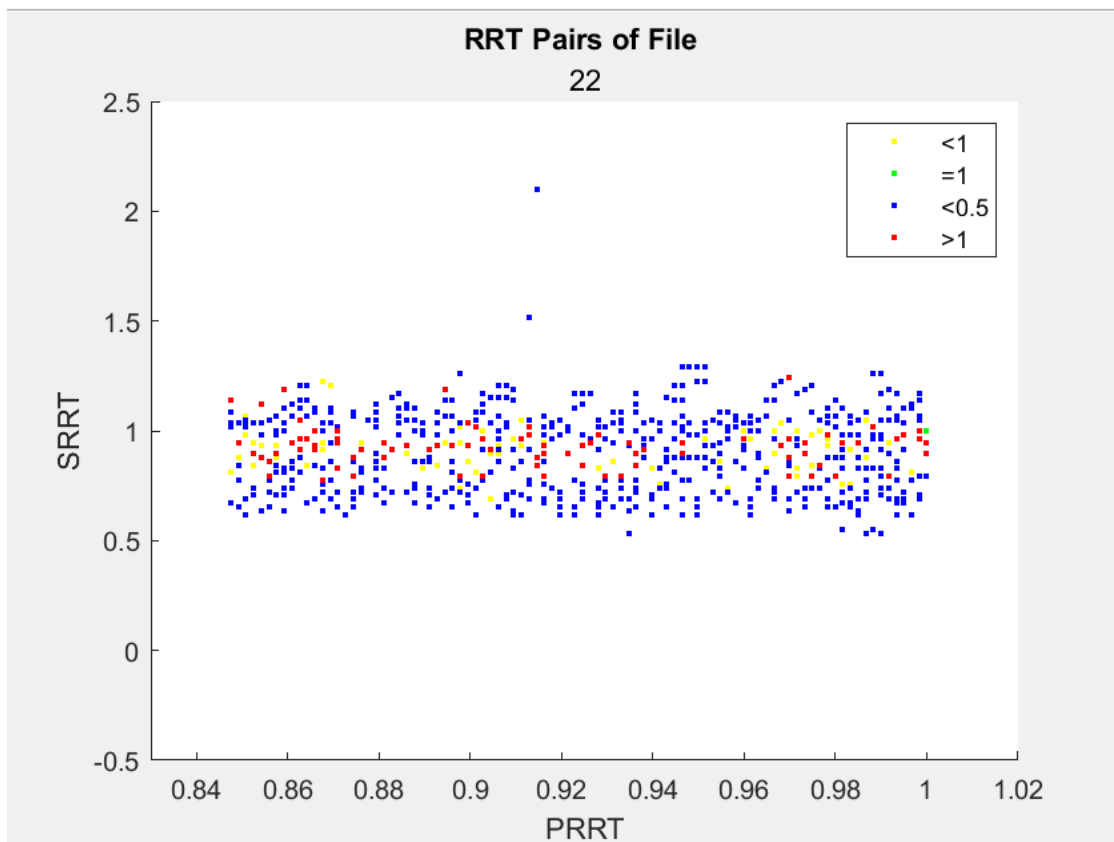


Figure B-49 The discrete GCxGC image of Sample 22 that belongs to cluster 5

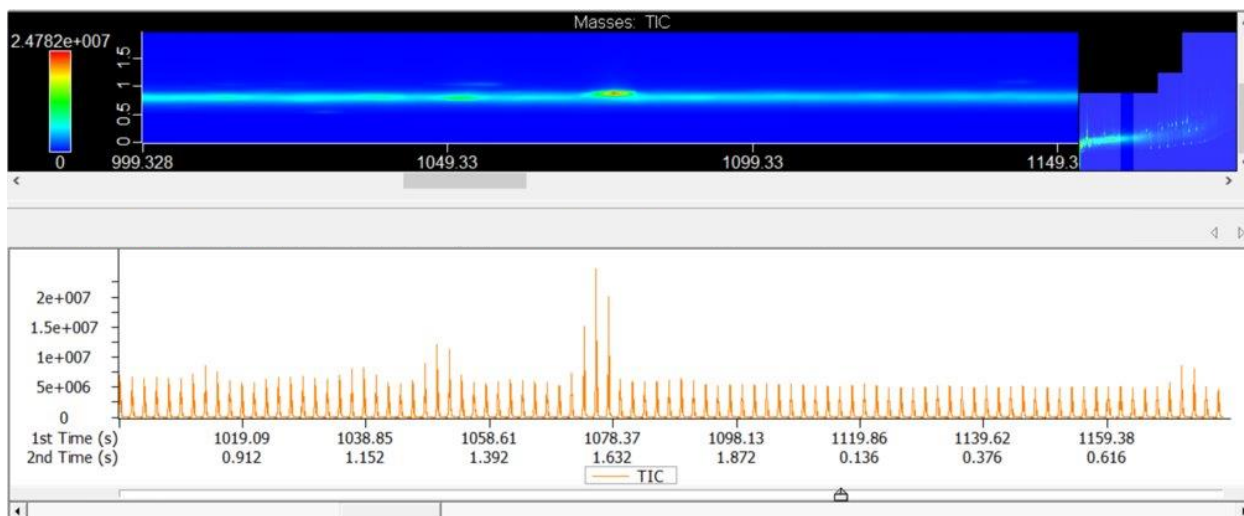


Figure B-50 The real GCxGC image of Sample 22 that belongs to cluster 5

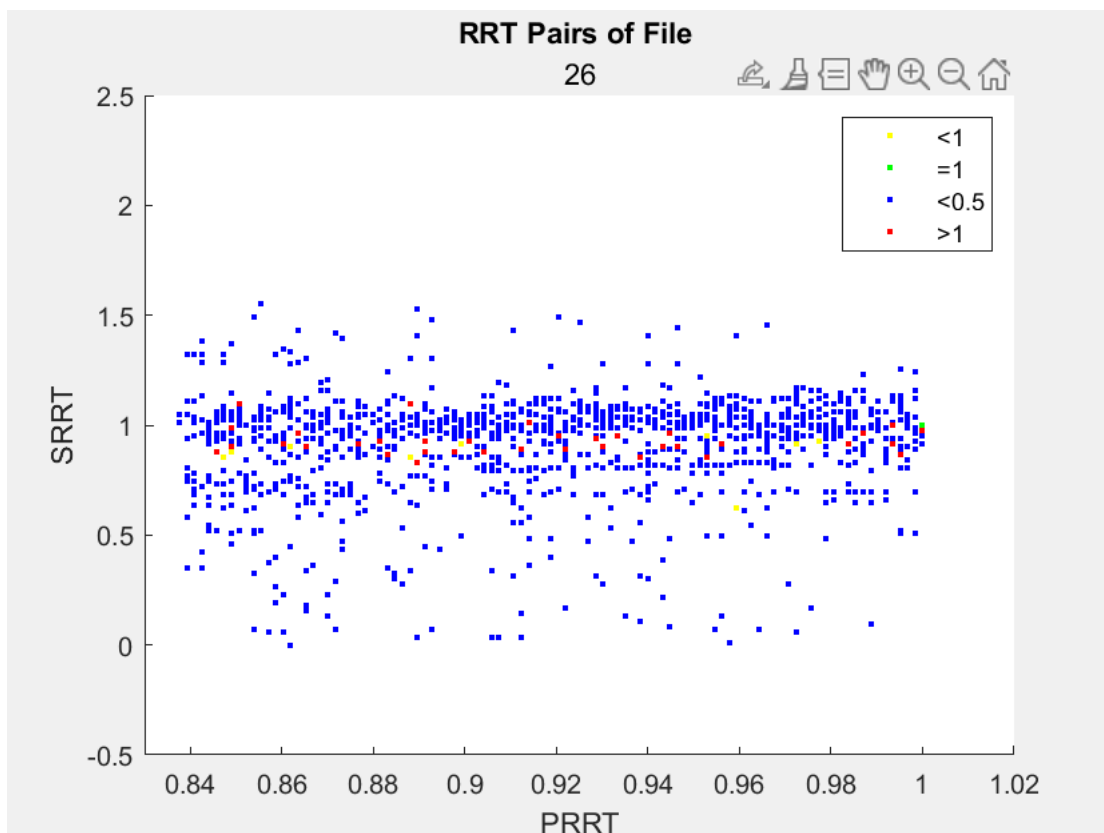


Figure B-51 The discrete GCxGC image of Sample 26 that belongs to cluster 5

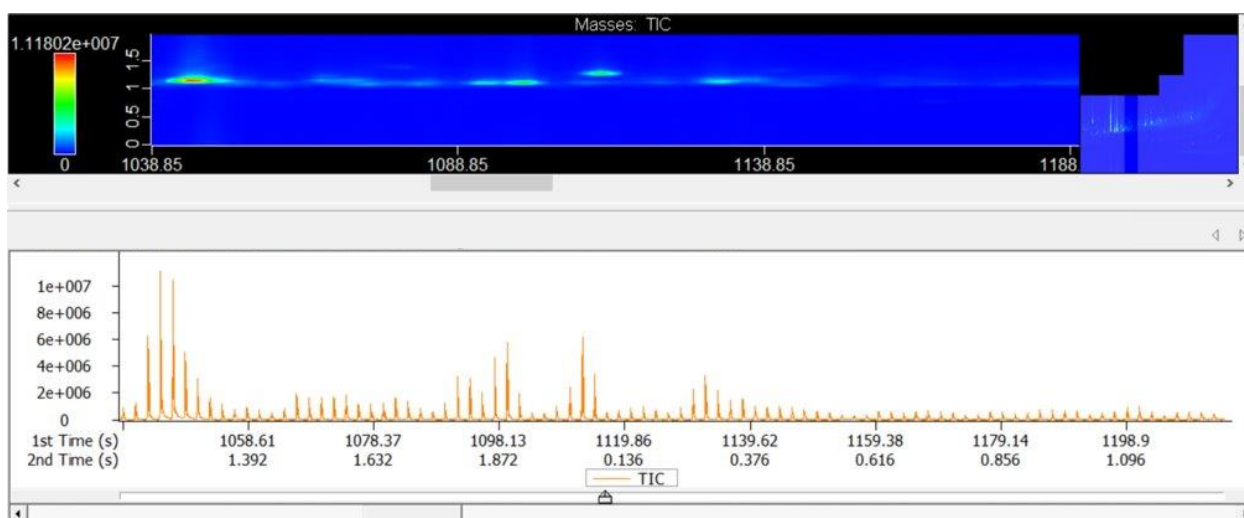


Figure B-52 The real GCxGC image of Sample 26 that belongs to cluster 5

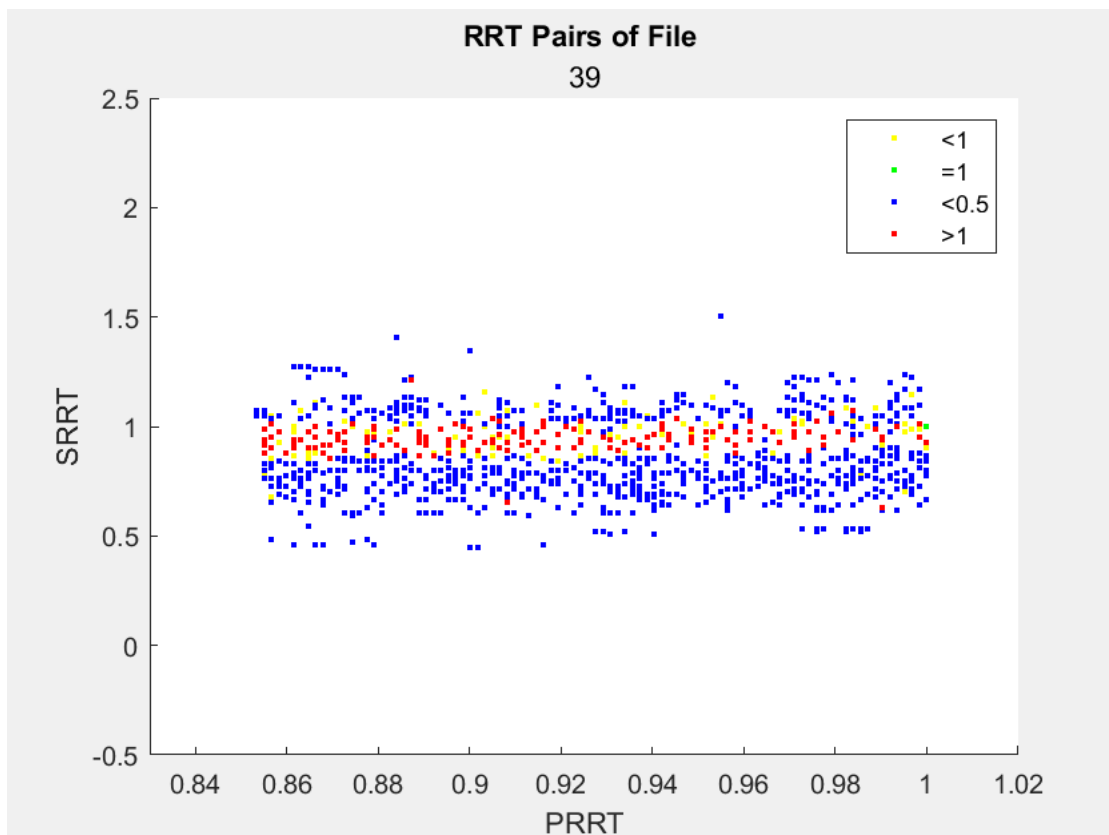


Figure B-53 The discrete GCxGC image of Sample 39 that belongs to cluster 5

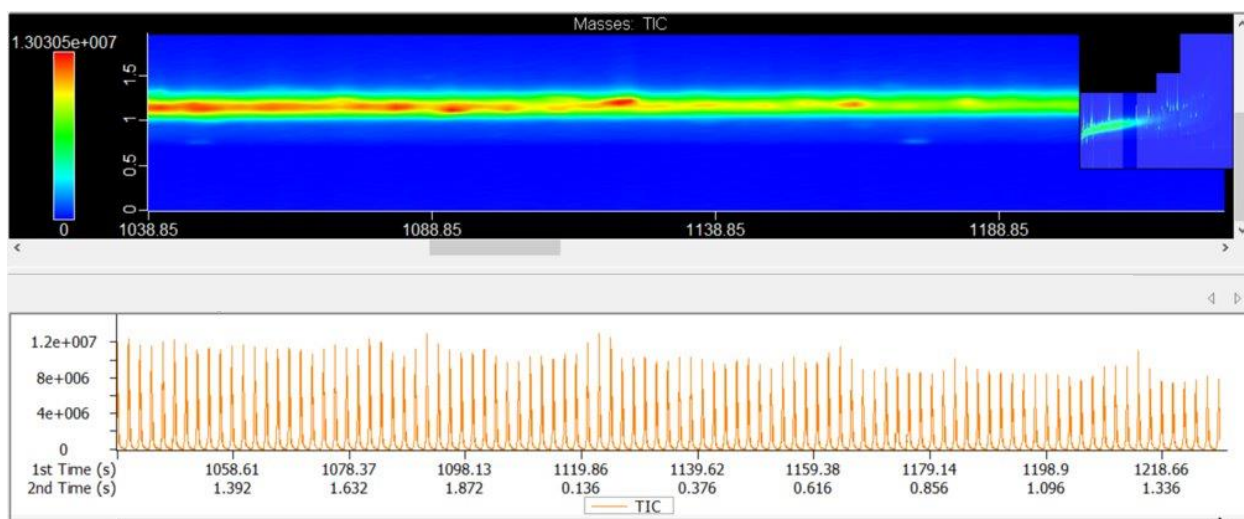


Figure B-54 The real GCxGC image of Sample 39 that belongs to cluster 5

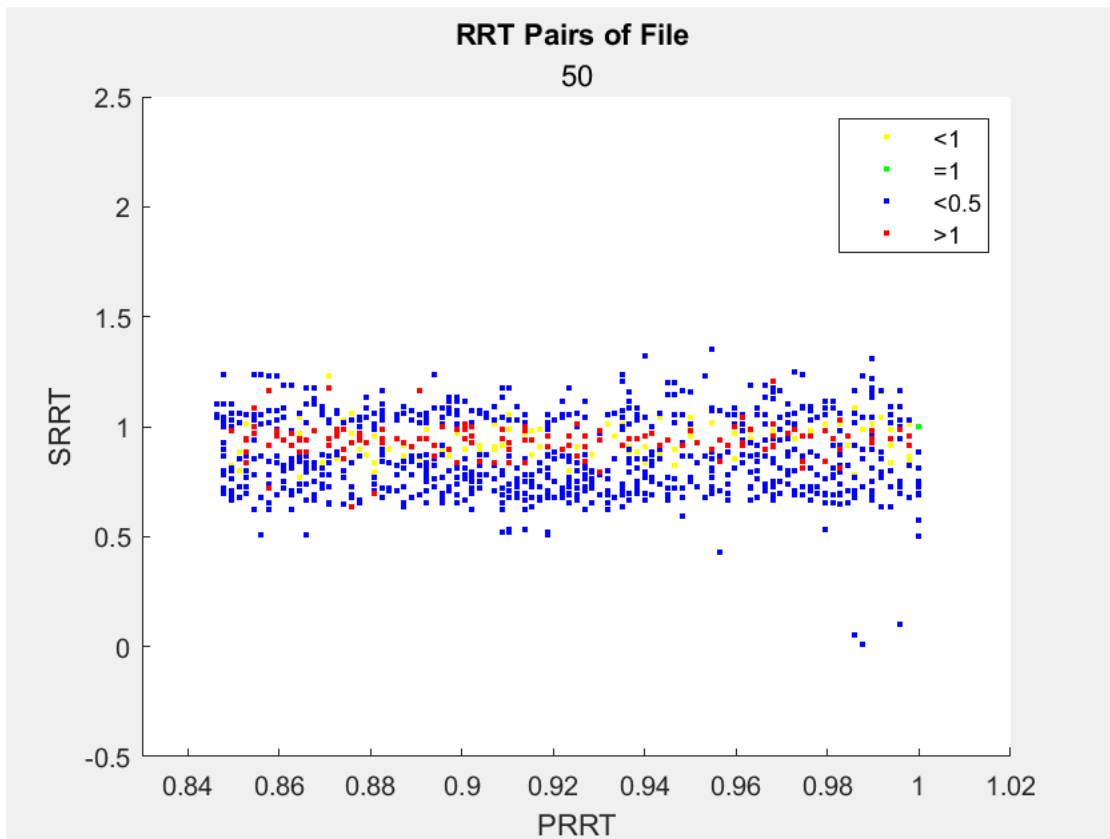


Figure B-55 The discrete GCxGC image of Sample 50 that belongs to cluster 5

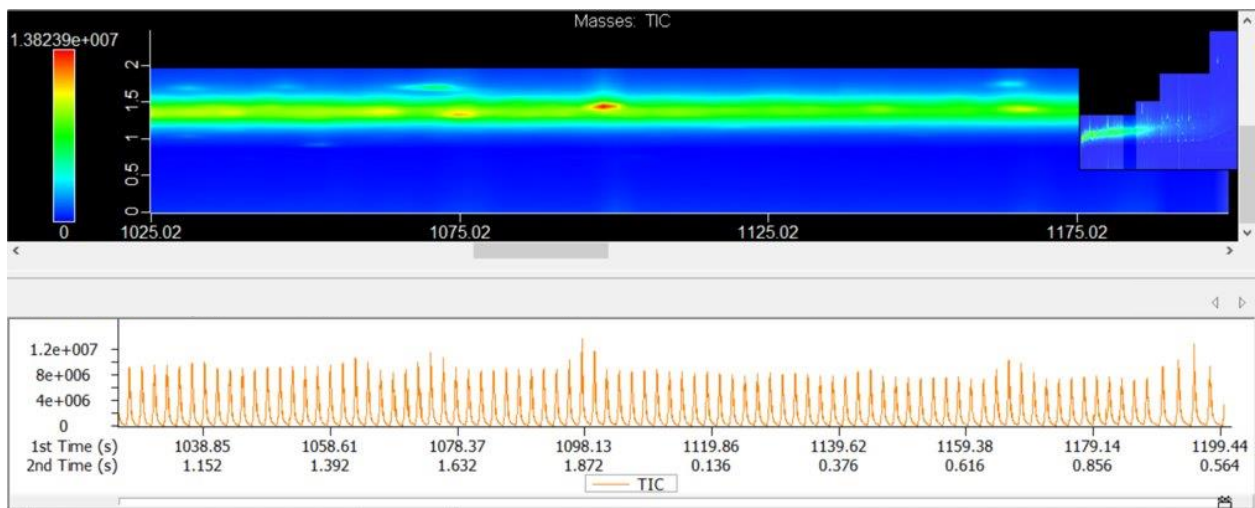


Figure B-56 The real GCxGC image of Sample 50 that belongs to cluster 5

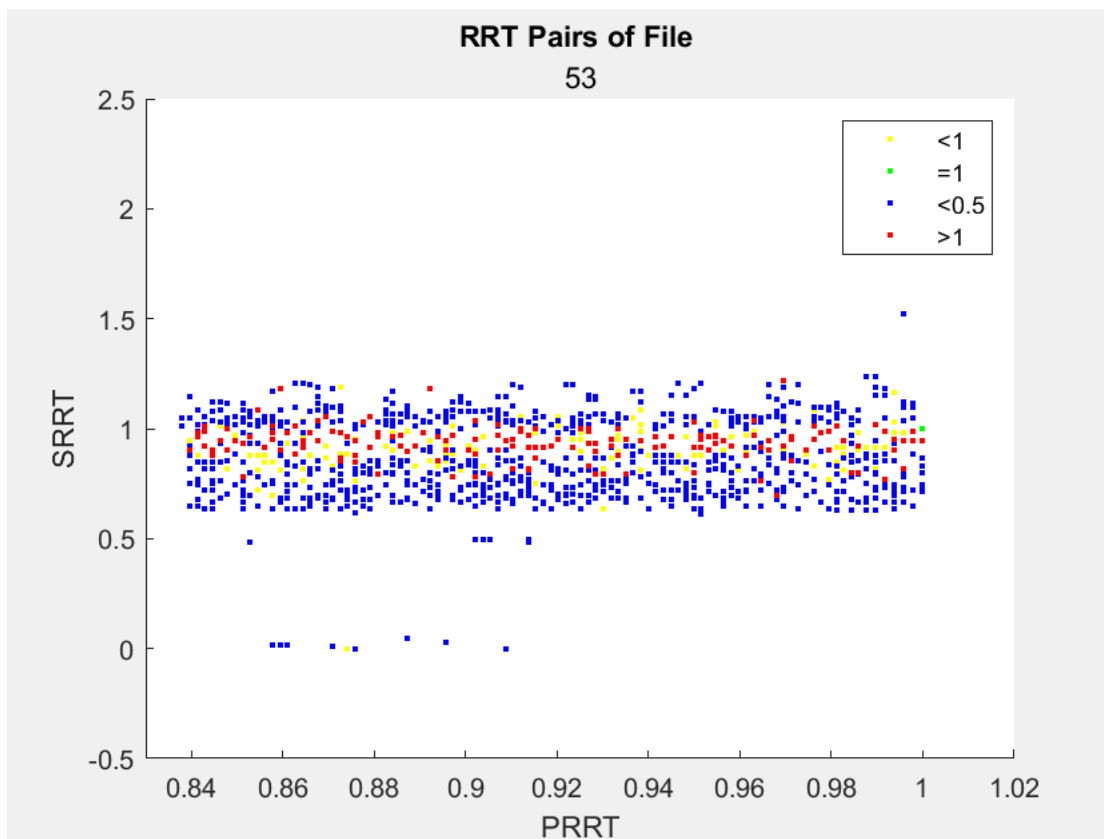


Figure B-57 The discrete GCxGC image of Sample 53 that belongs to cluster 5

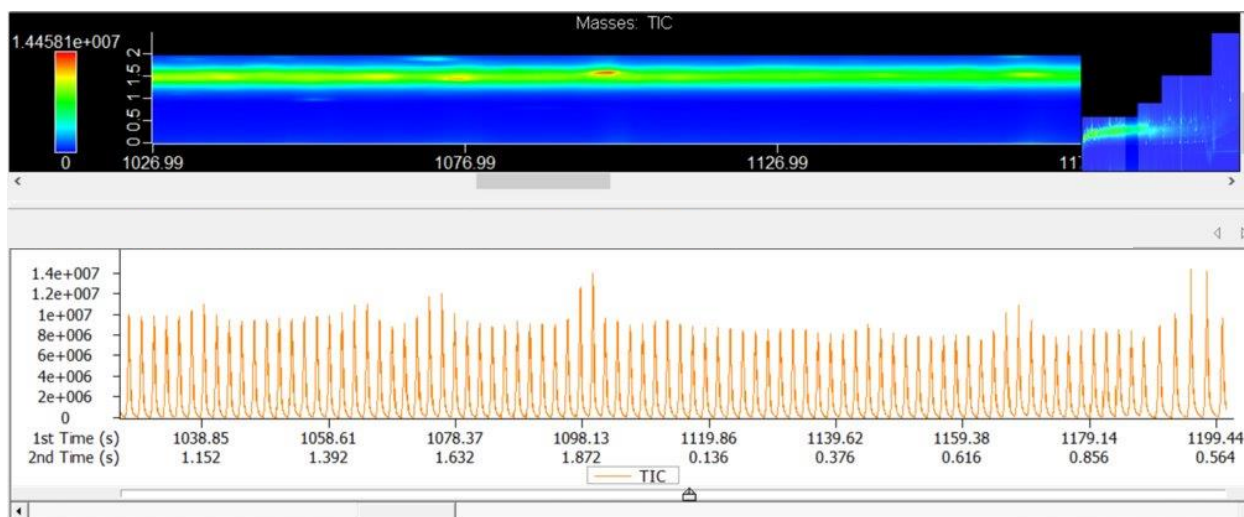


Figure B-58 The real GCxGC image of Sample 53 that belongs to cluster 5

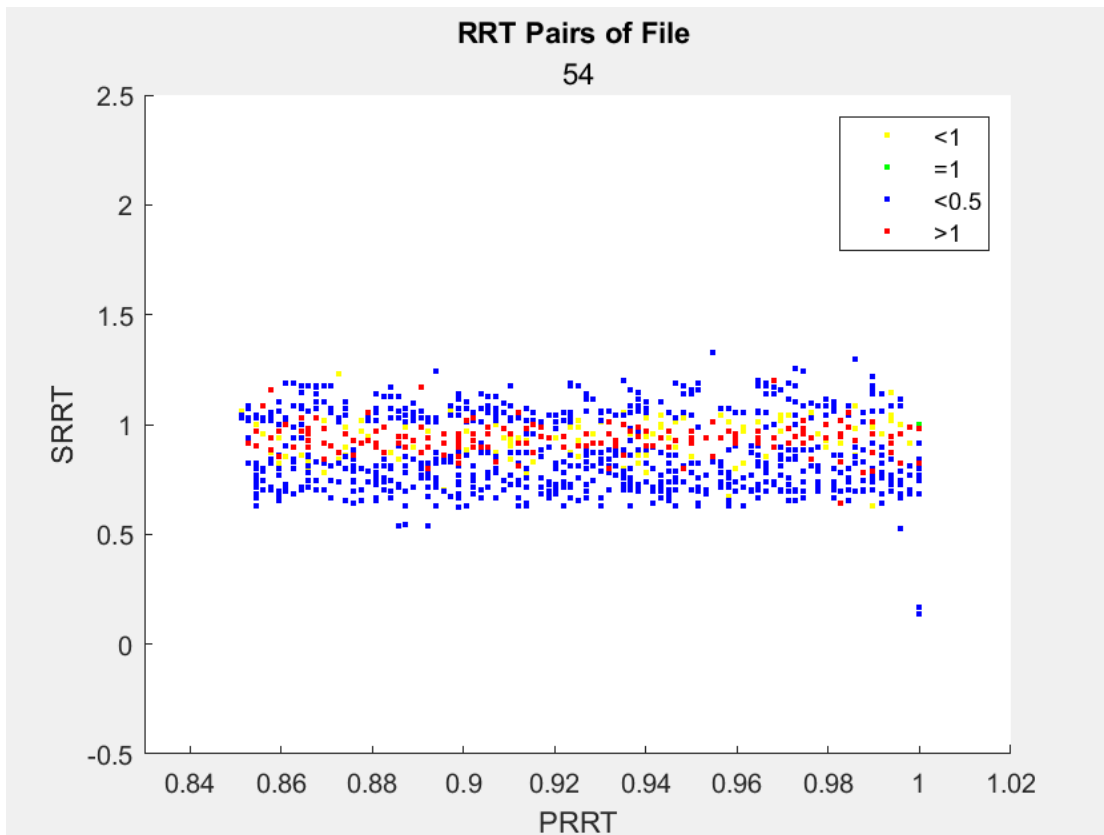


Figure B-59 The discrete GCxGC image of Sample 54 that belongs to cluster 5

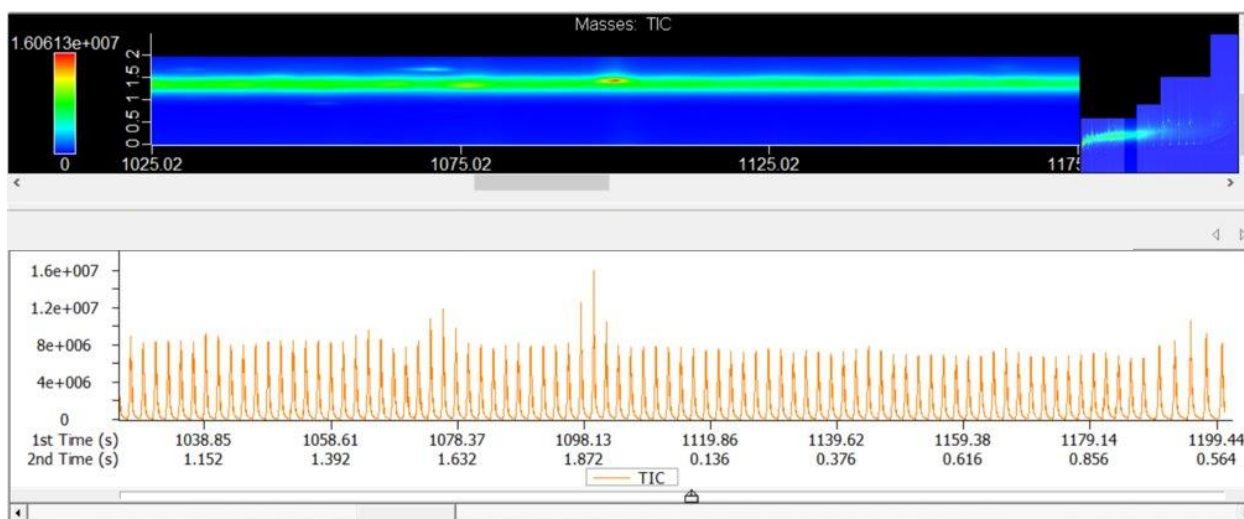


Figure B-60 The real GCxGC image of Sample 54 that belongs to cluster 5

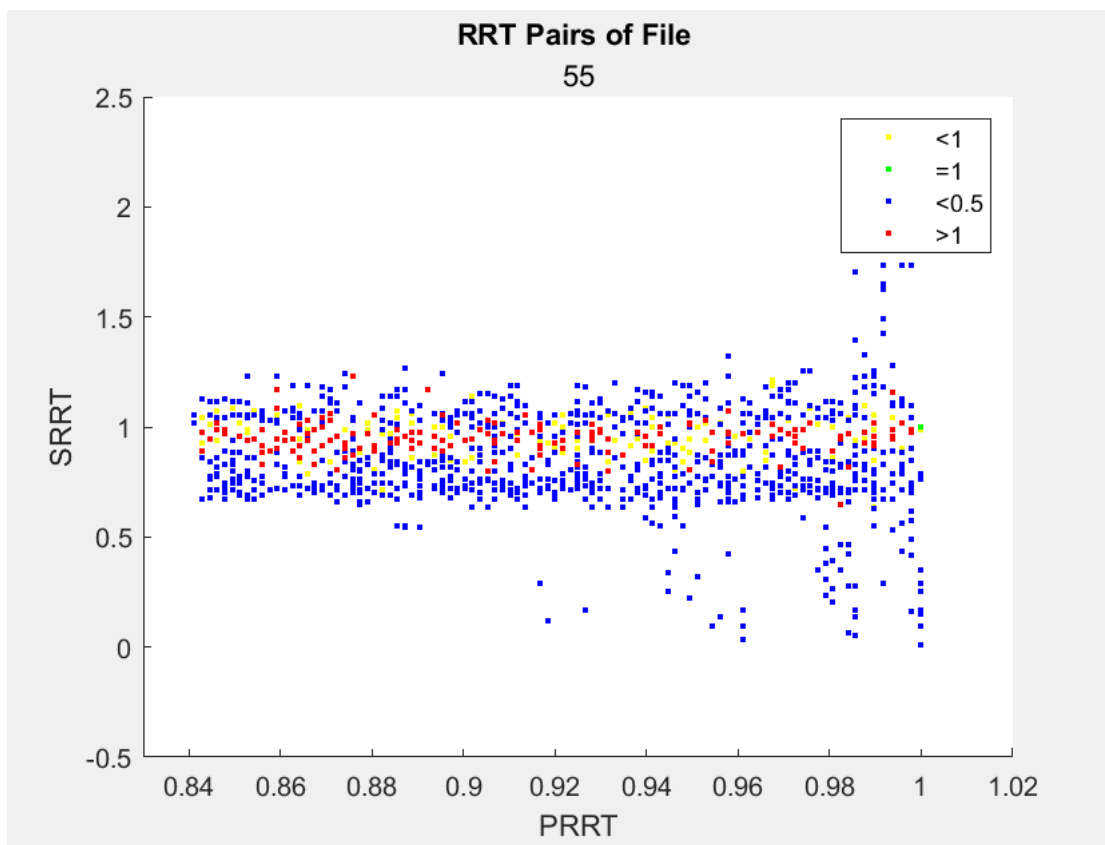


Figure B-61 The discrete GCxGC image of Sample 55 that belongs to cluster 5

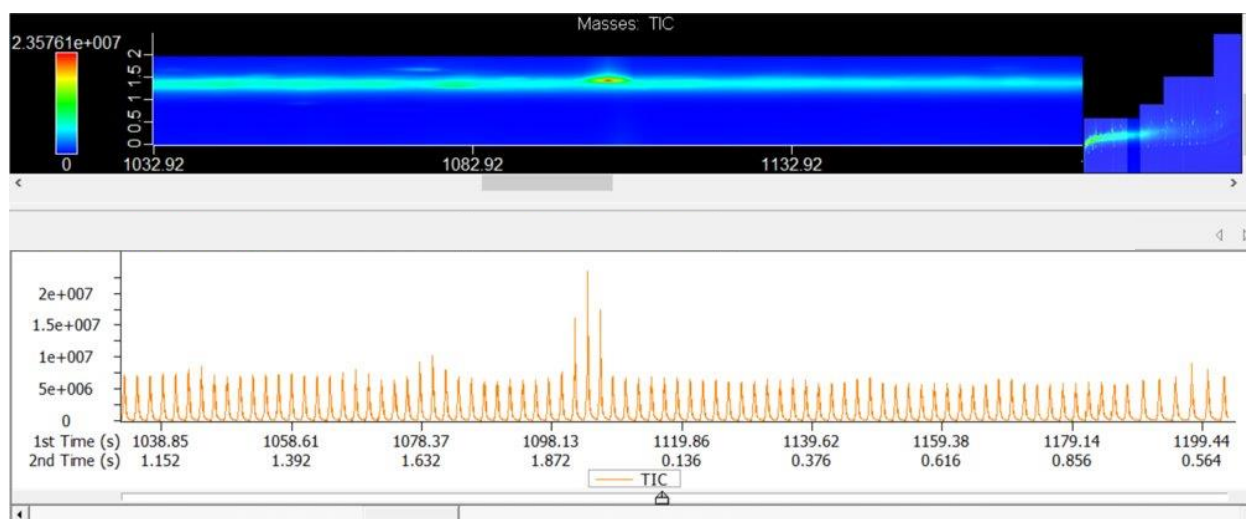


Figure B-62 The real GCxGC image of Sample 55 that belongs to cluster 5

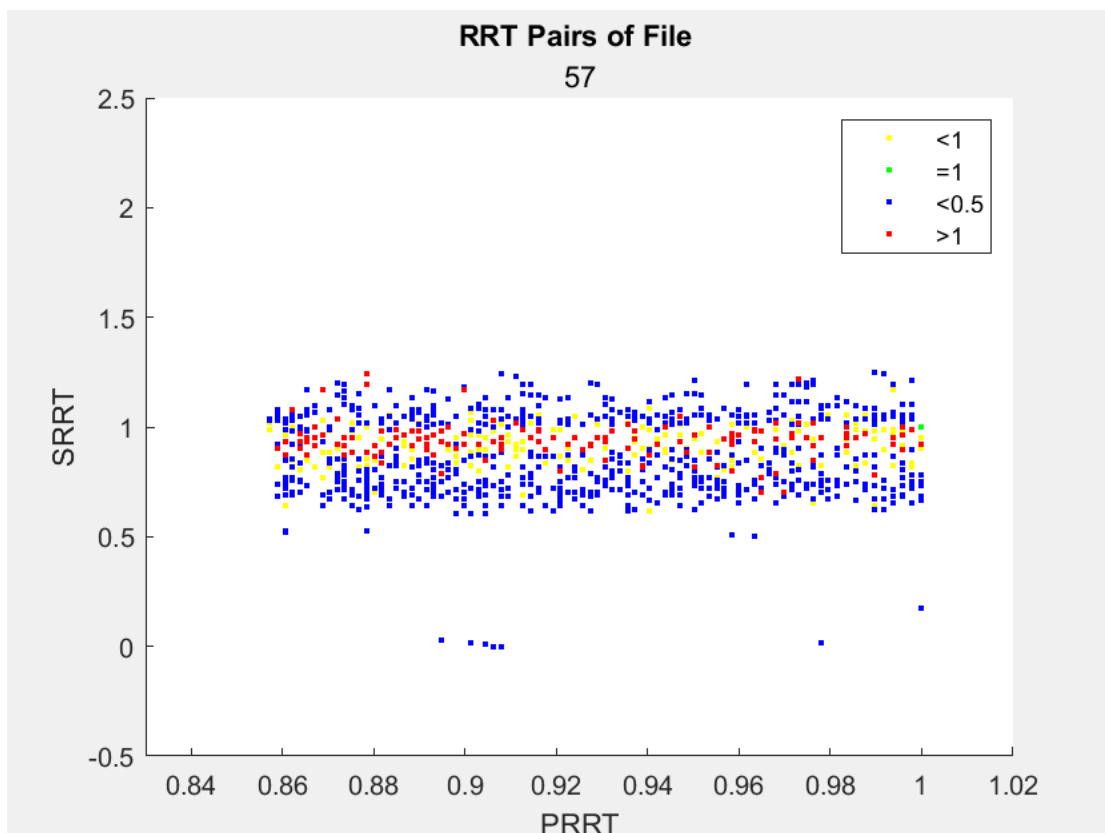


Figure B-63 The discrete GCxGC image of Sample 57 that belongs to cluster 5

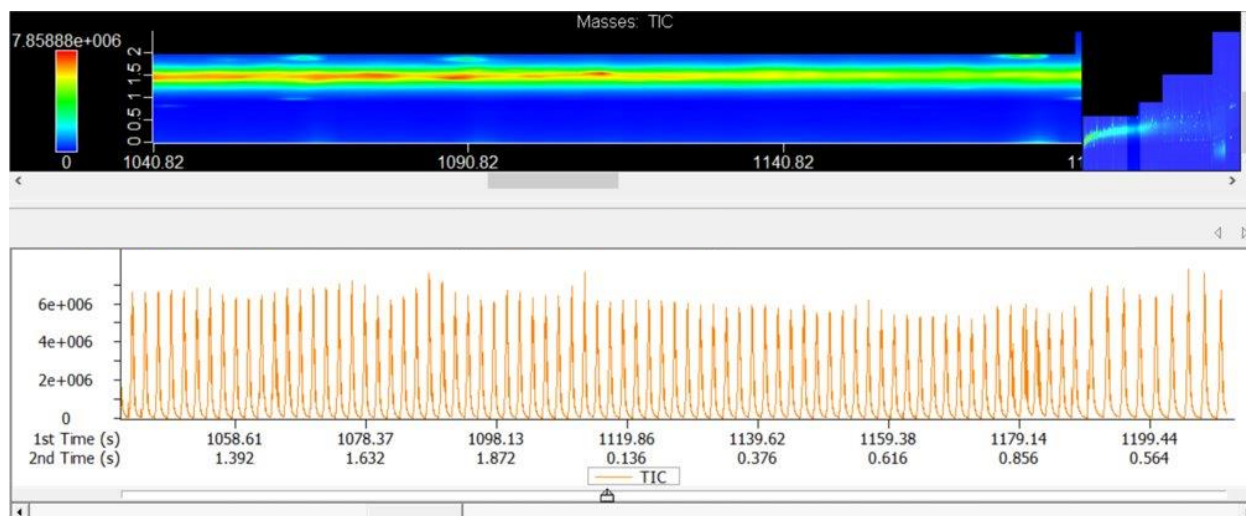


Figure B-64 The real GCxGC image of Sample 57 that belongs to cluster 5



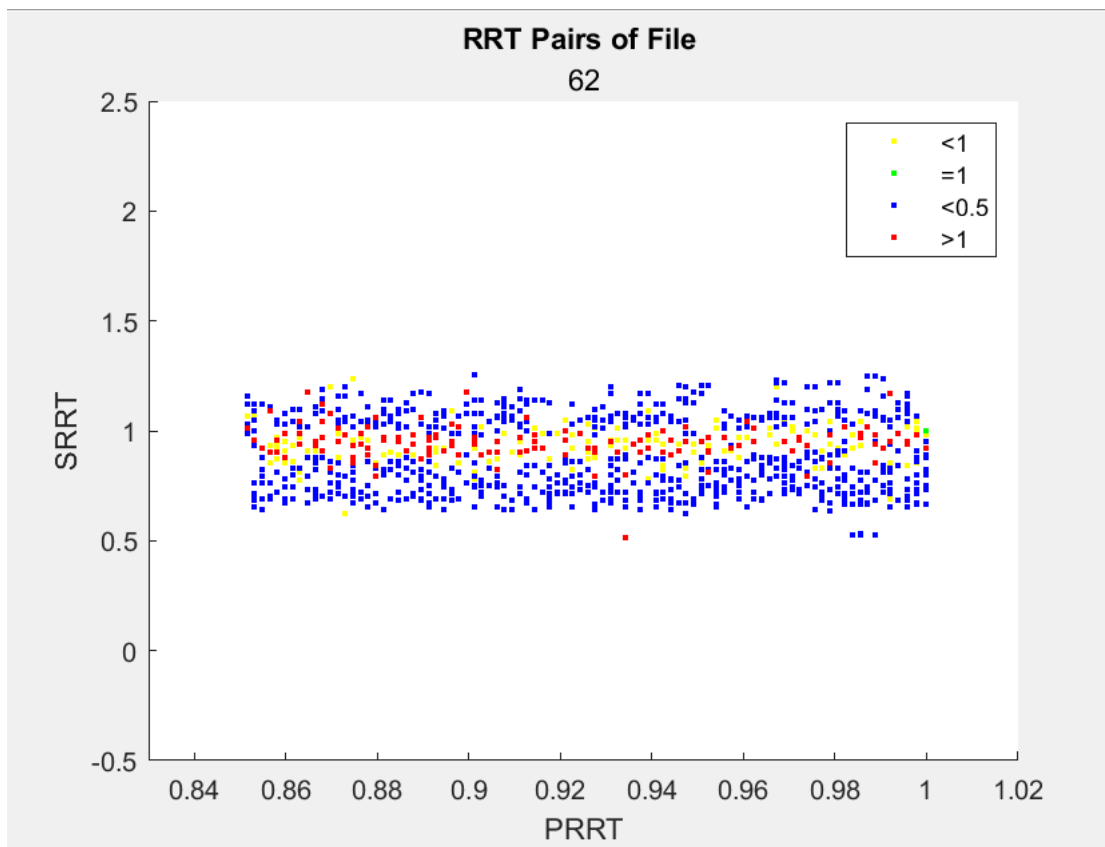


Figure B-65 The discrete GCxGC image of Sample 62 that belongs to cluster 5

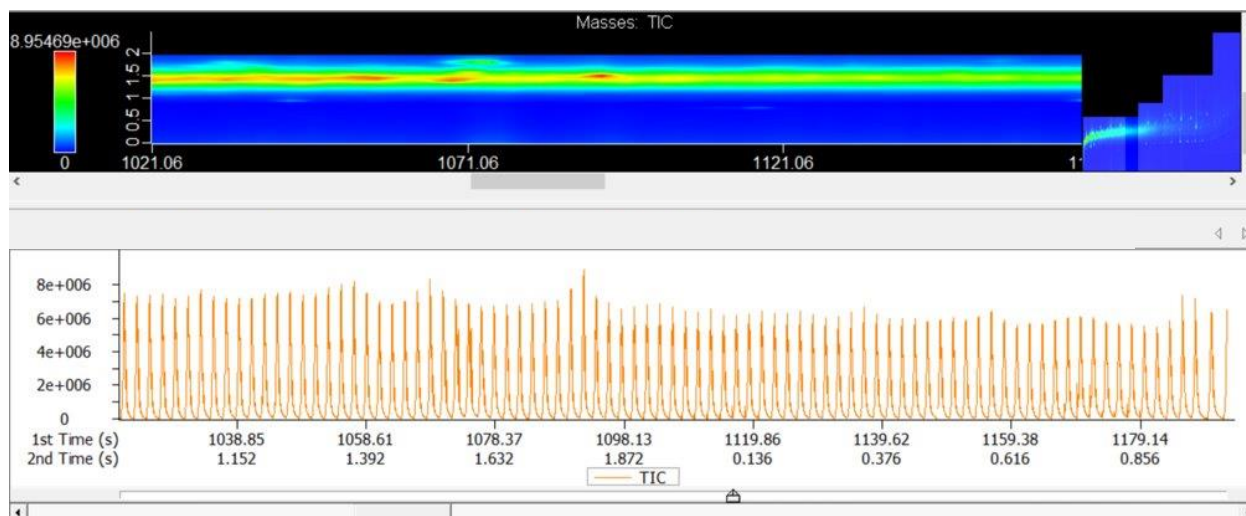


Figure B-66 The real GCxGC image of Sample 62 that belongs to cluster 5

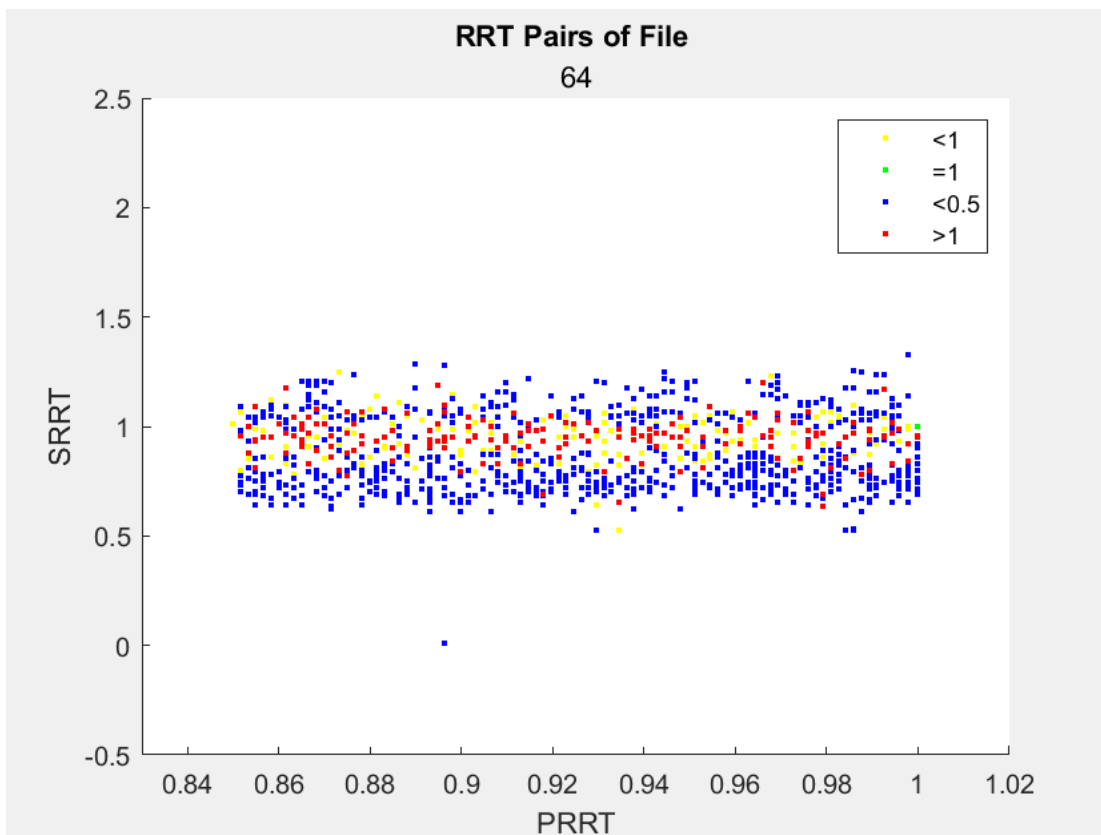


Figure B-67 The discrete GCxGC image of Sample 64 that belongs to cluster 5

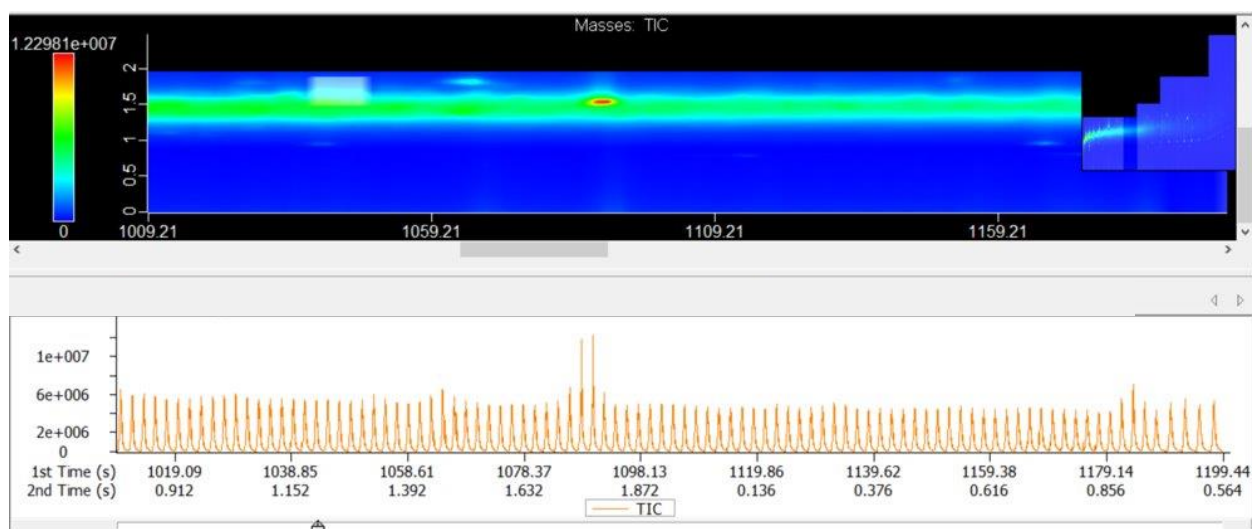


Figure B-68 The real GCxGC image of Sample 64 that belongs to cluster 5

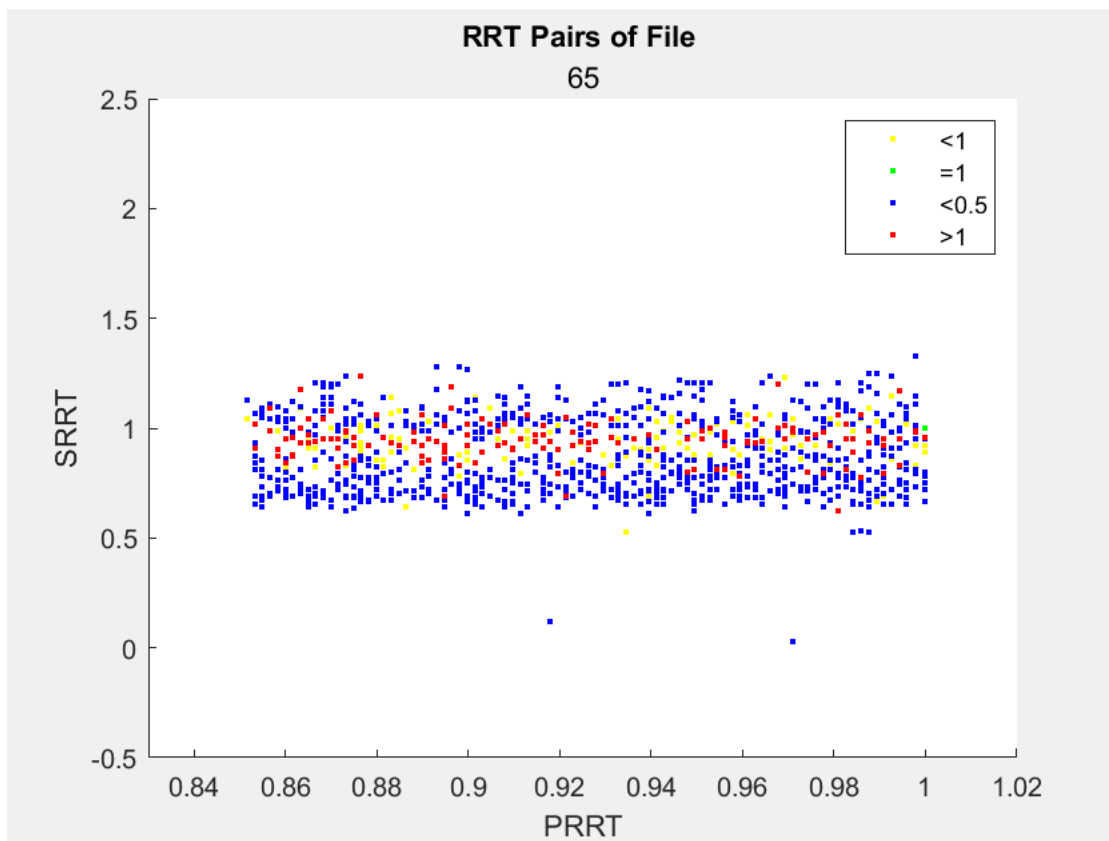


Figure B-69 The discrete GCxGC image of Sample 65 that belongs to cluster 5

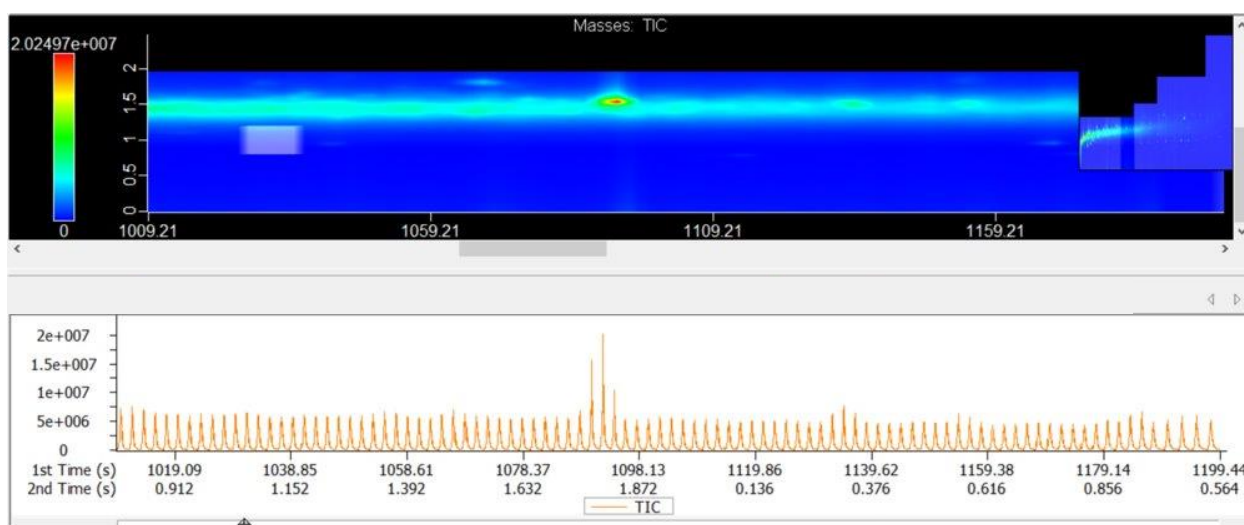


Figure B-70 The real GCxGC image of Sample 65 that belongs to cluster 5

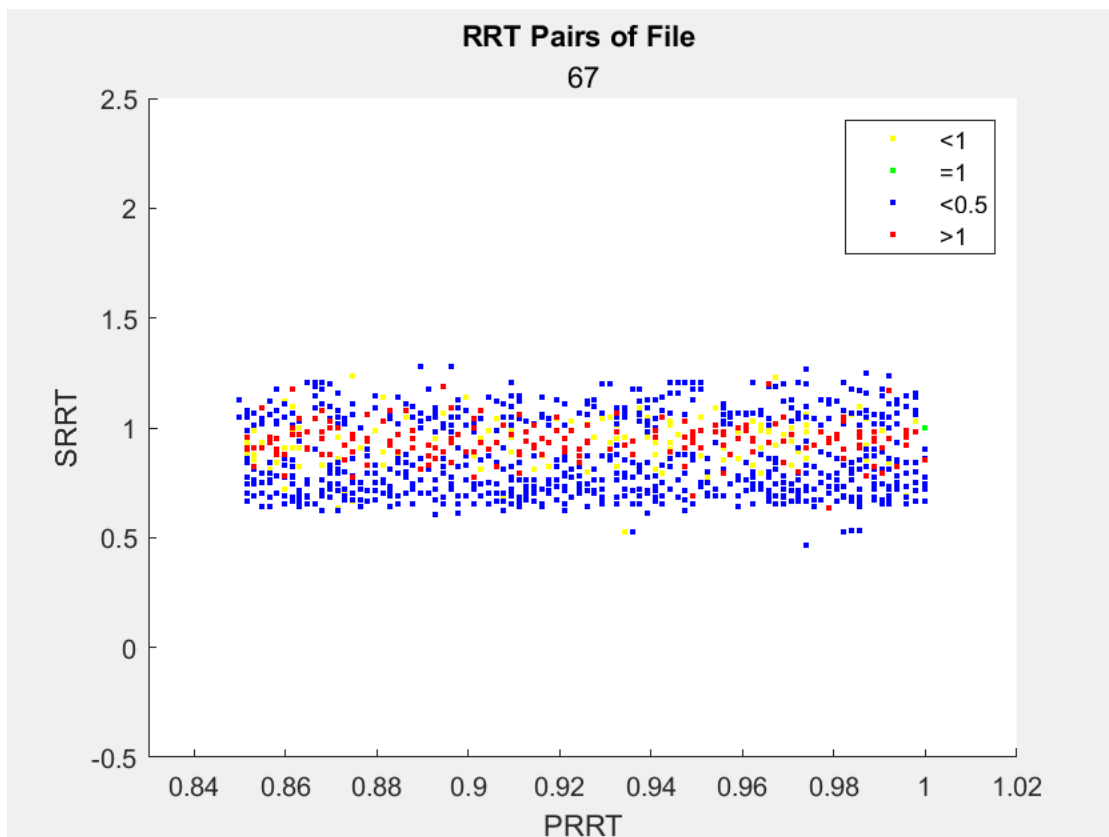


Figure B-71 The discrete GCxGC image of Sample 67 that belongs to cluster 5

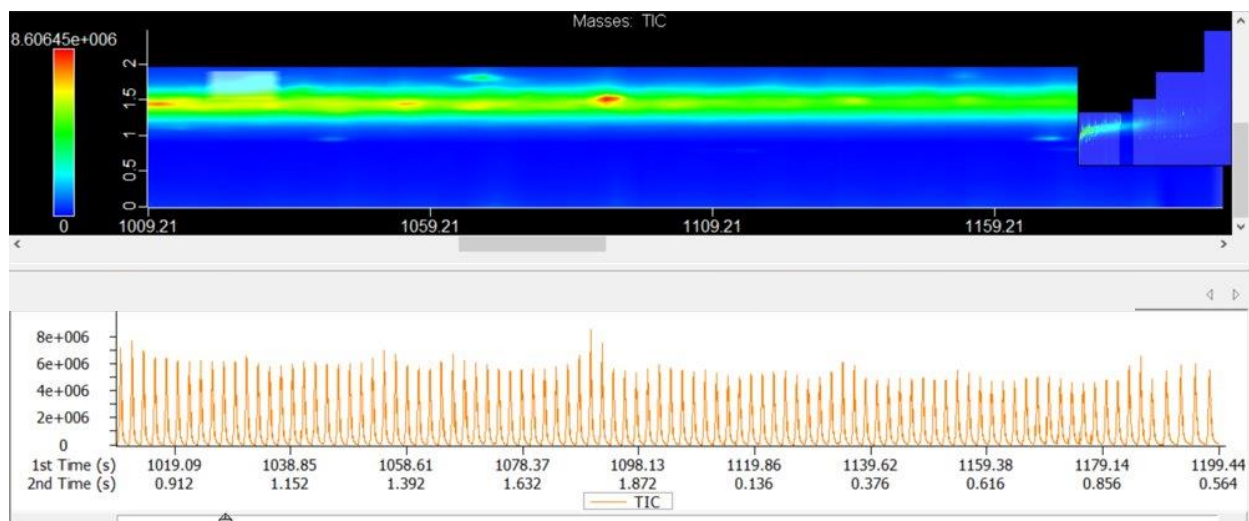


Figure B-72 The real GCxGC image of Sample 67 that belongs to cluster 5

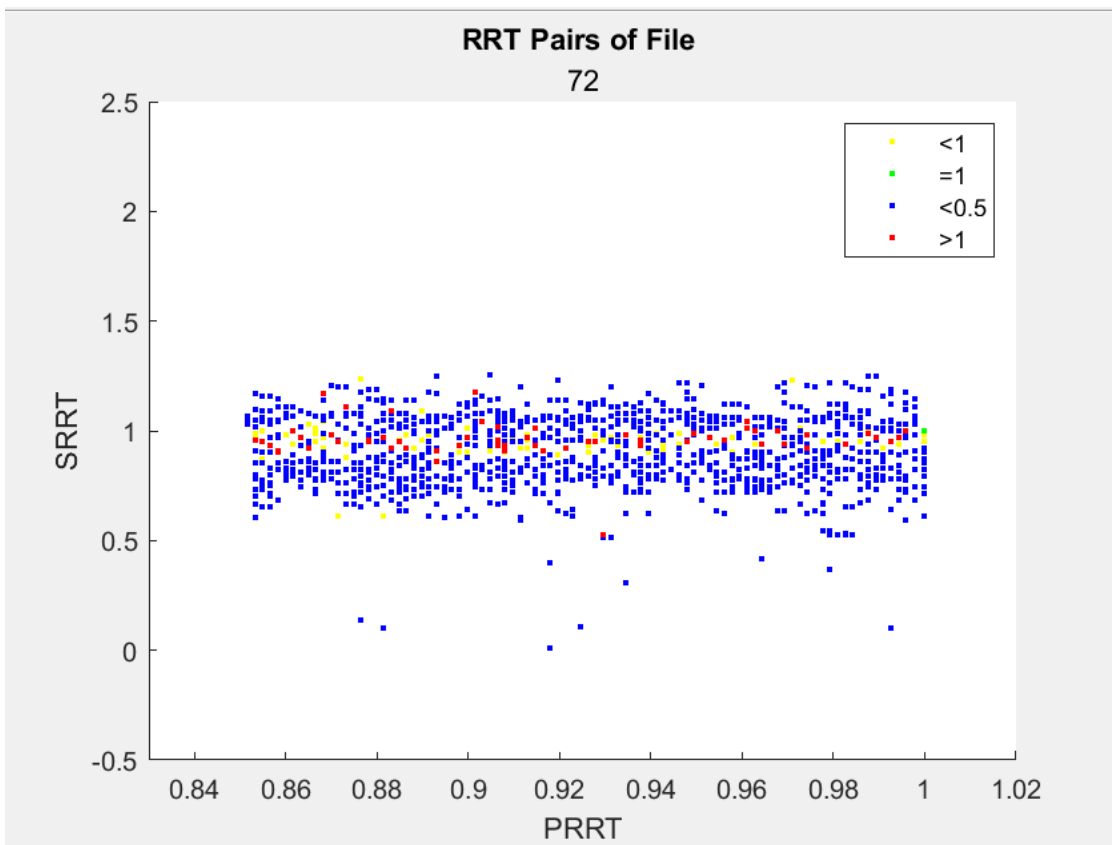


Figure B-73 The discrete GCxGC image of Sample 72 that belongs to cluster 5

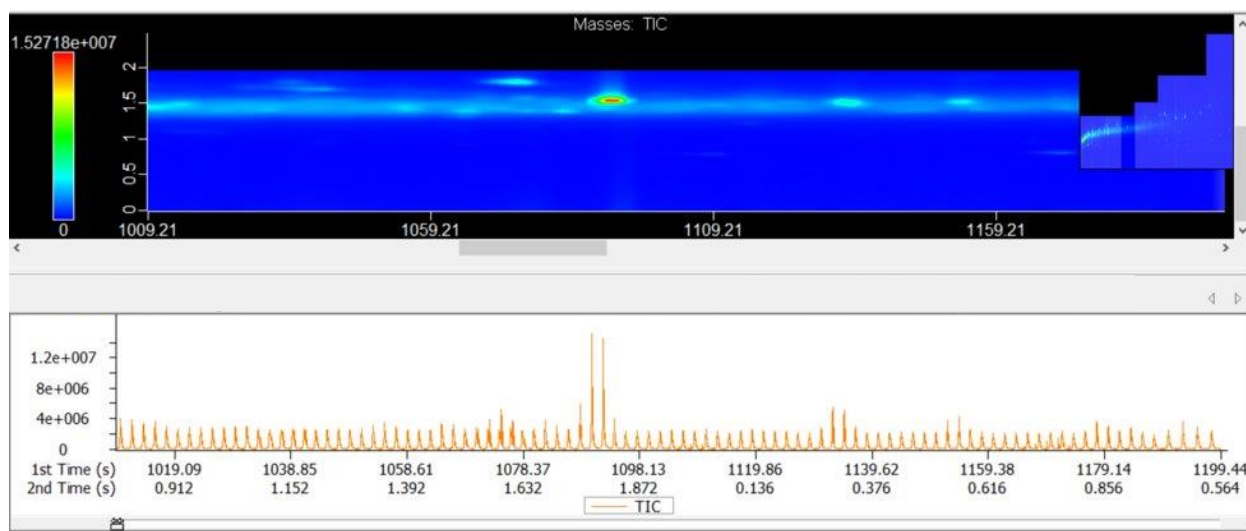


Figure B-74 The real GCxGC image of Sample 72 that belongs to cluster 5

Cluster 6: Sample 3, 15, 27, 32, 37, 43, 44, 45, 46

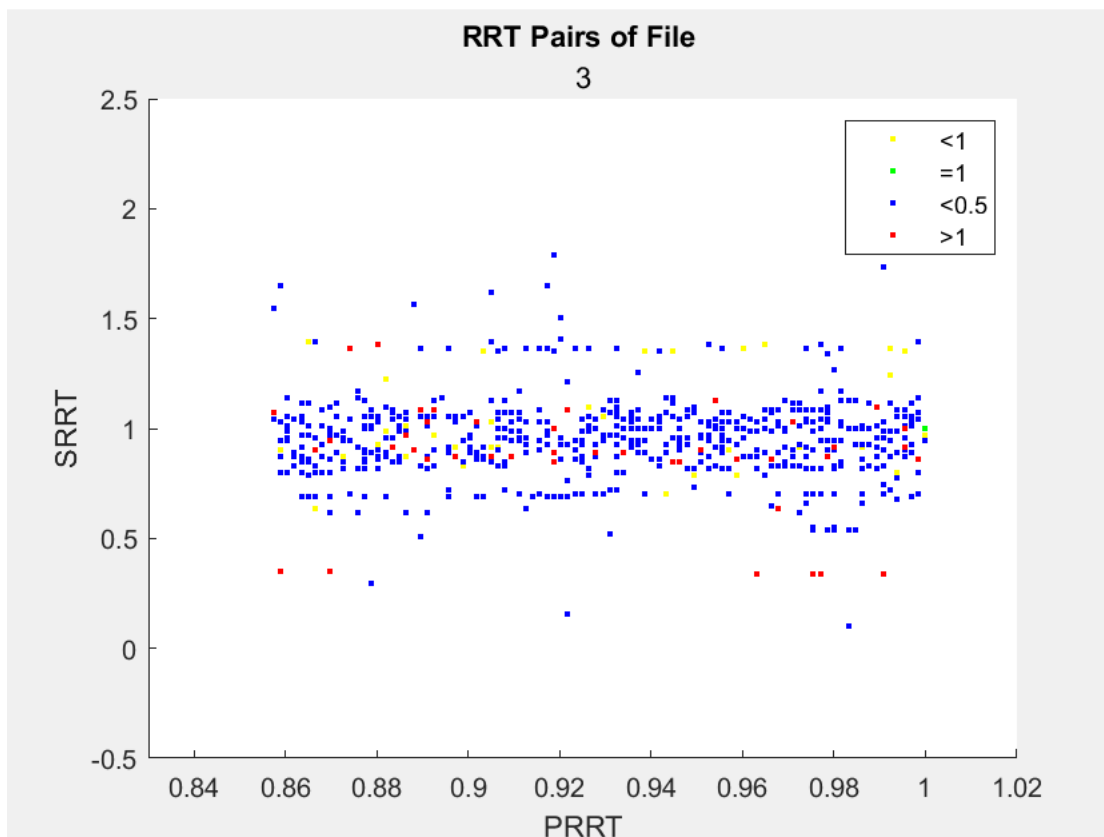


Figure B-75 The discrete GCxGC image of Sample 3 that belongs to cluster 6

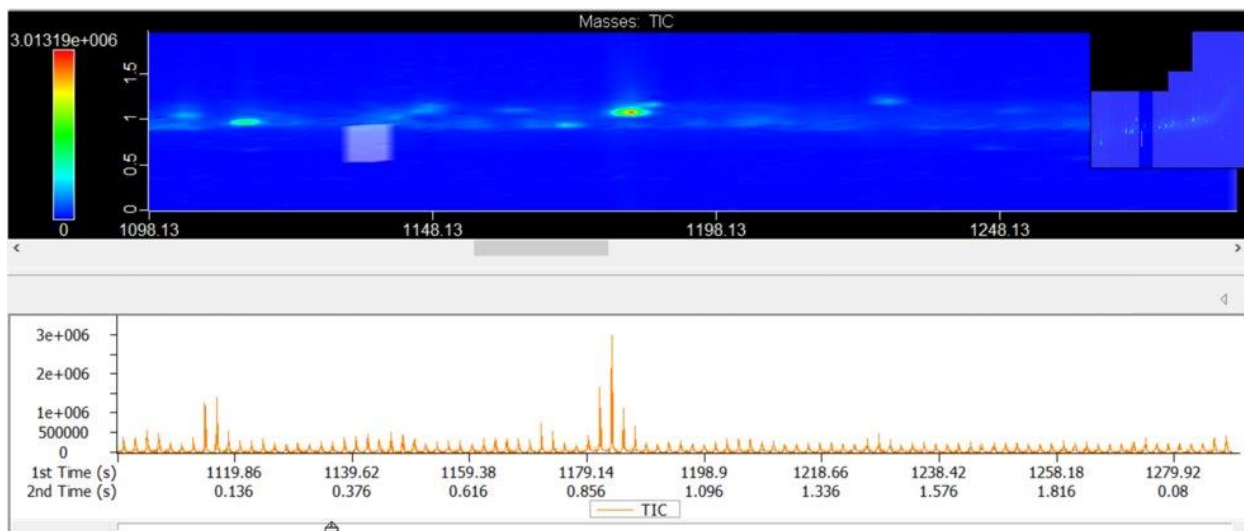


Figure B-76 The real GCxGC image of Sample 3 that belongs to cluster 6

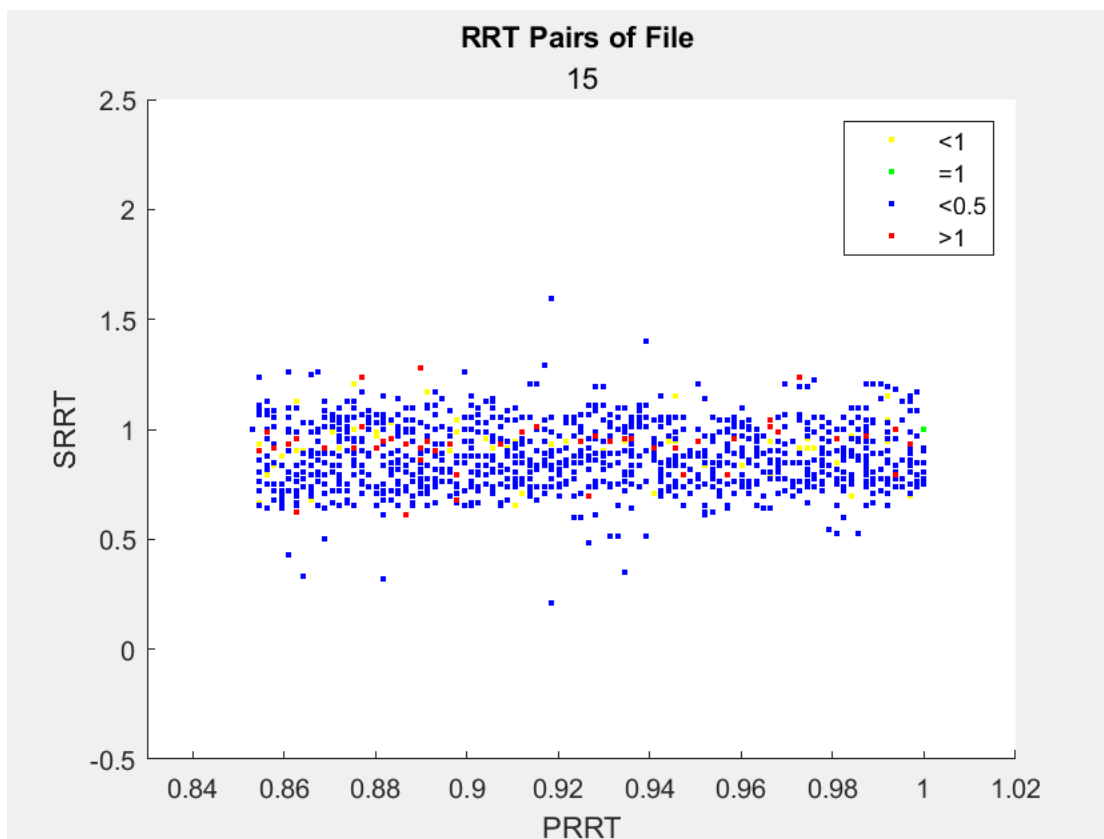


Figure B-77 The discrete GCxGC image of Sample 15 that belongs to cluster 6

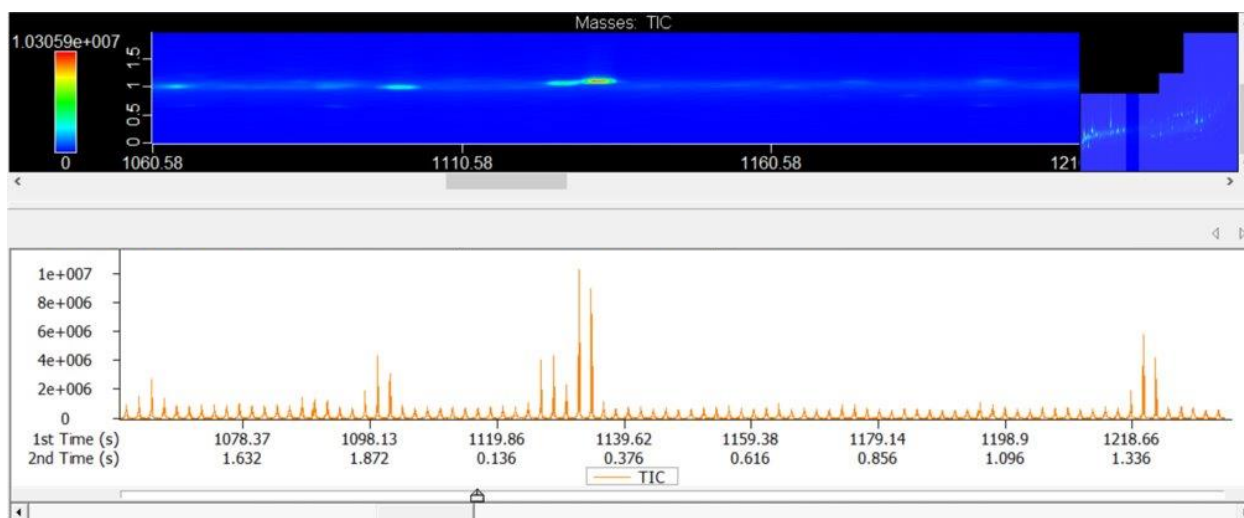


Figure B-78 The real GCxGC image of Sample 15 that belongs to cluster 6

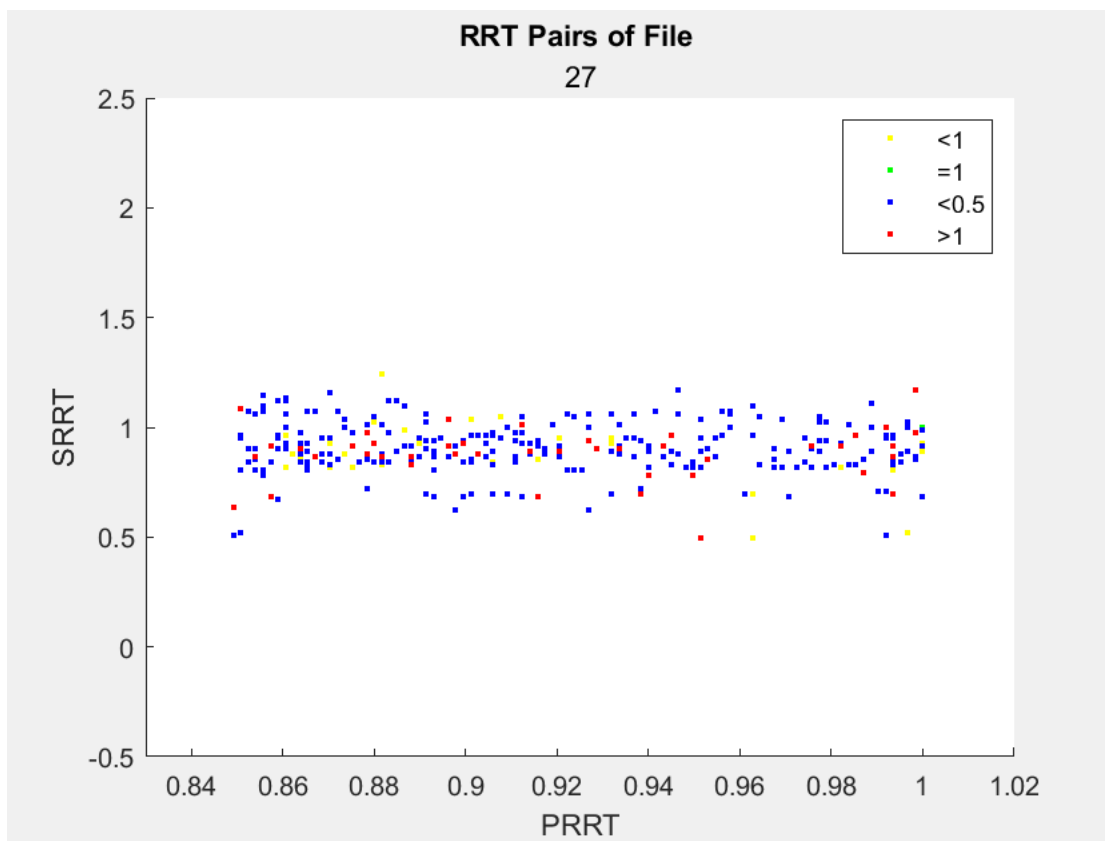


Figure B-79 The discrete GCxGC image of Sample 27 that belongs to cluster 6

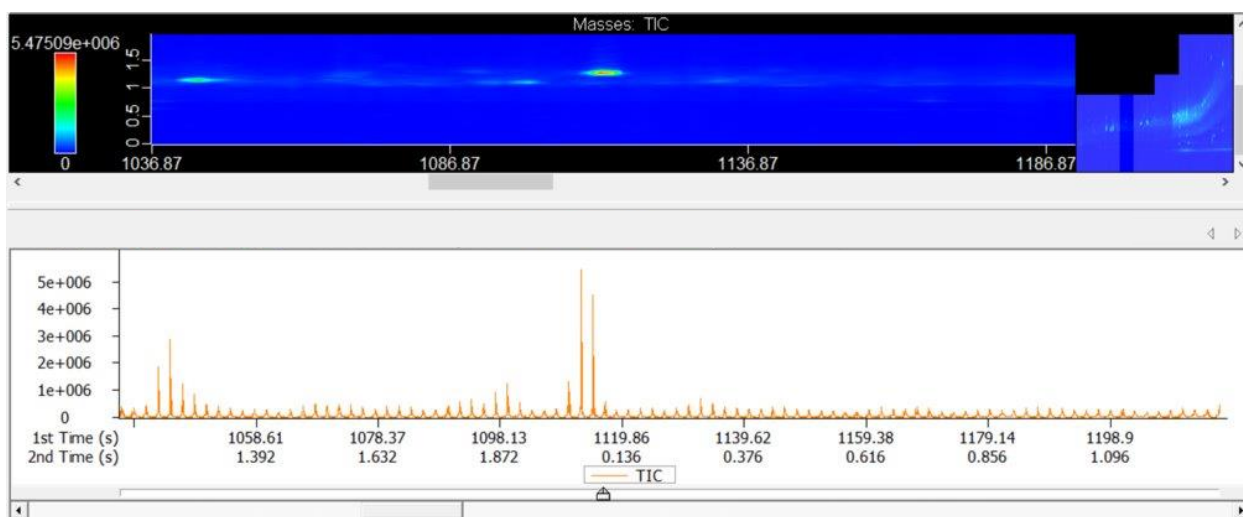


Figure B-80 The real GCxGC image of Sample 27 that belongs to cluster 6



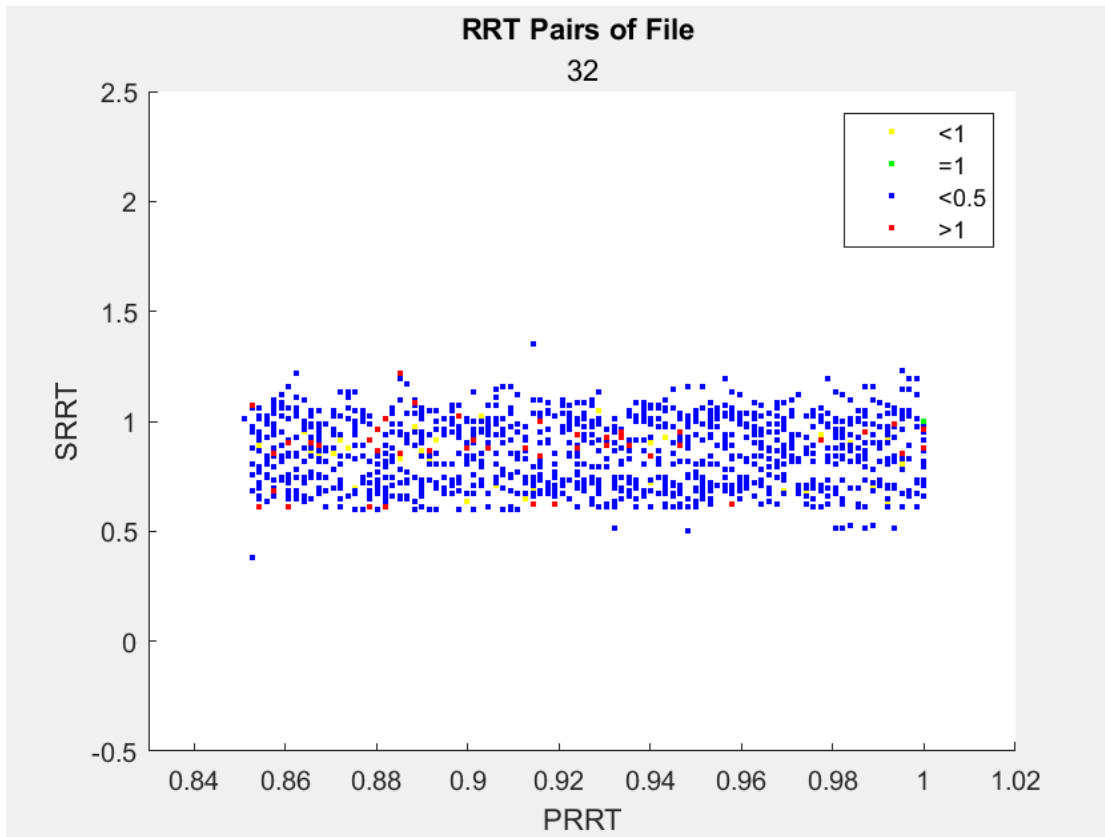


Figure B-81 The discrete GCxGC image of Sample 32 that belongs to cluster 6

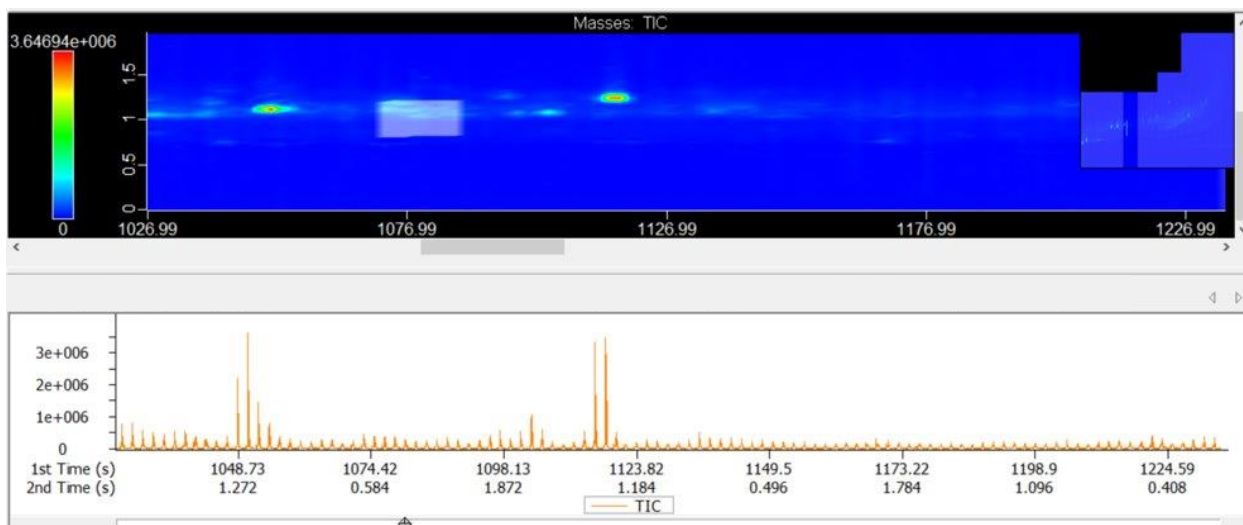


Figure B-82 The real GCxGC image of Sample 32 that belongs to cluster 6

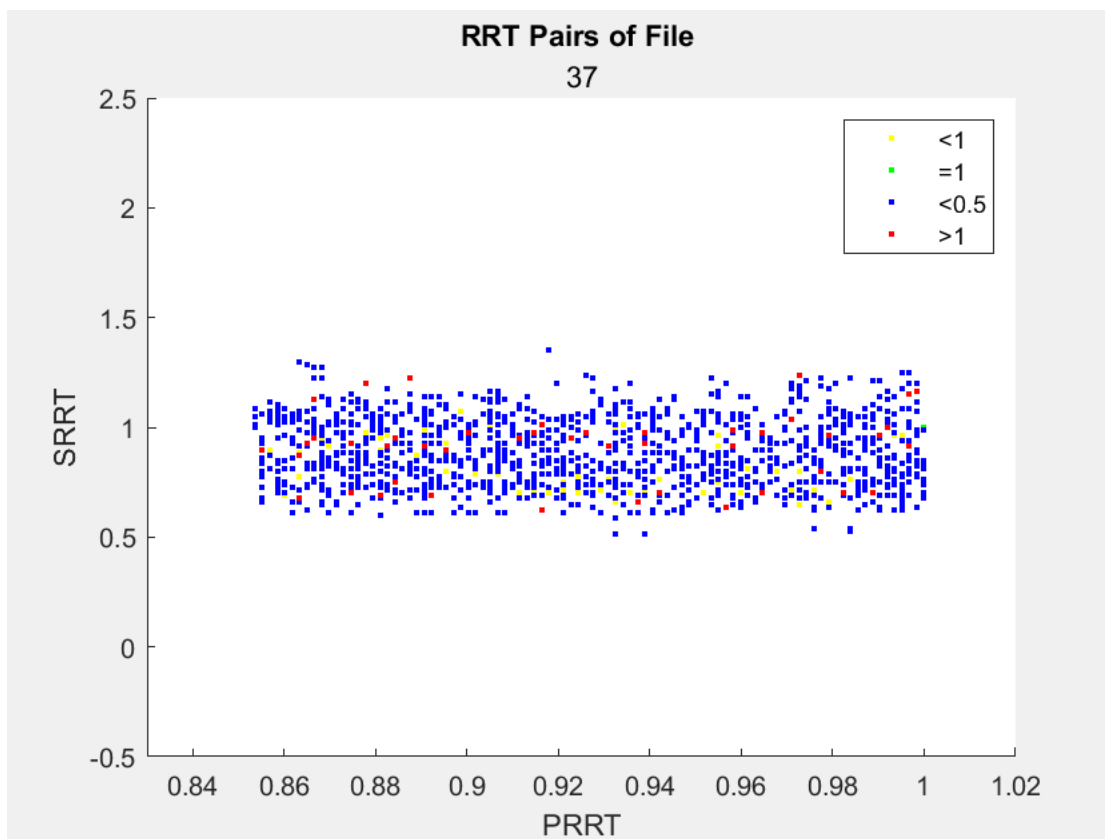


Figure B-83 The discrete GCxGC image of Sample 37 that belongs to cluster 6

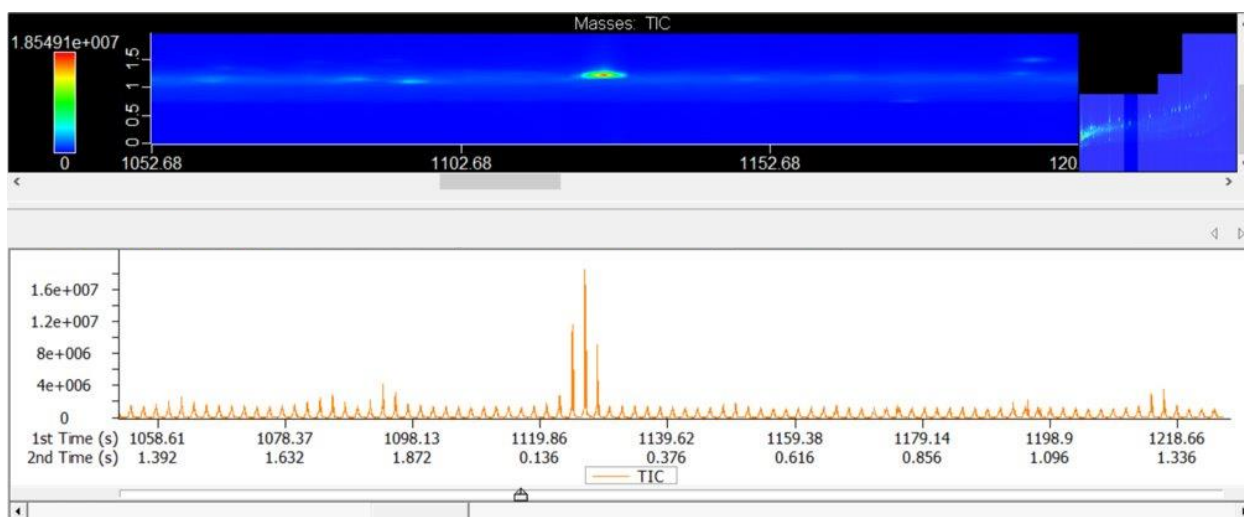


Figure B-84 The real GCxGC image of Sample 37 that belongs to cluster 6

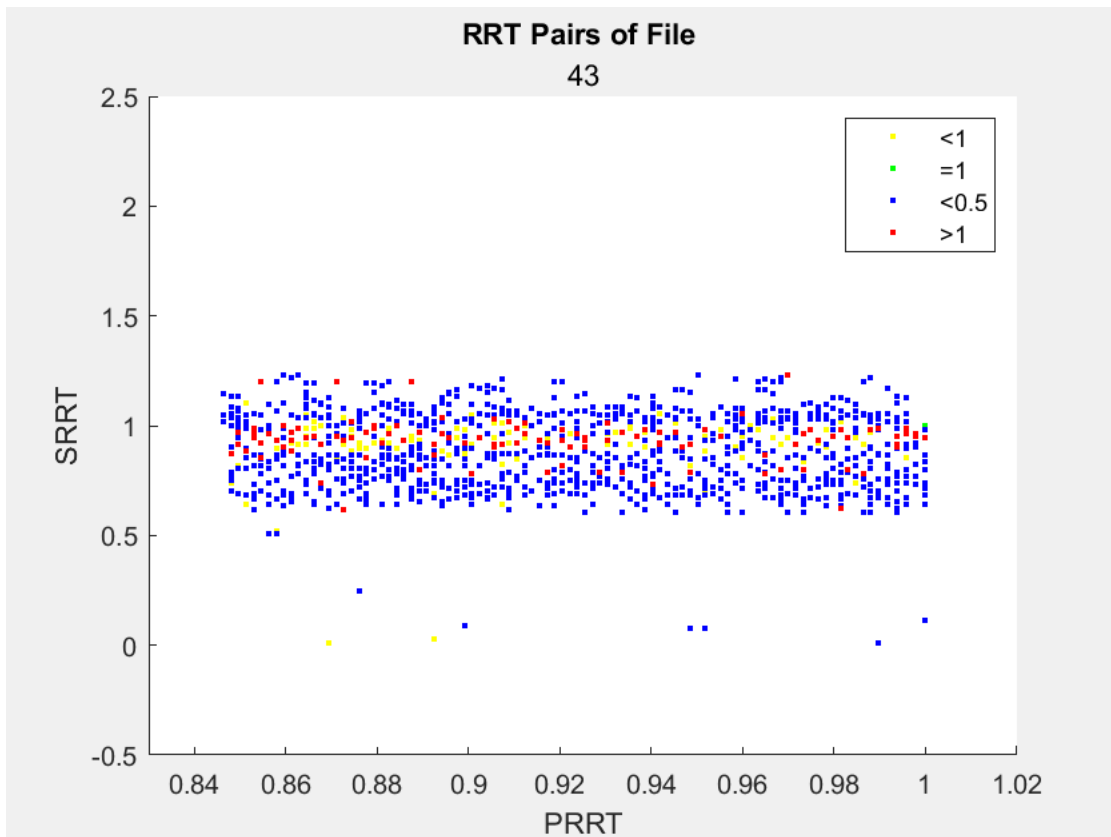


Figure B-85 The discrete GCxGC image of Sample 43 that belongs to cluster 6

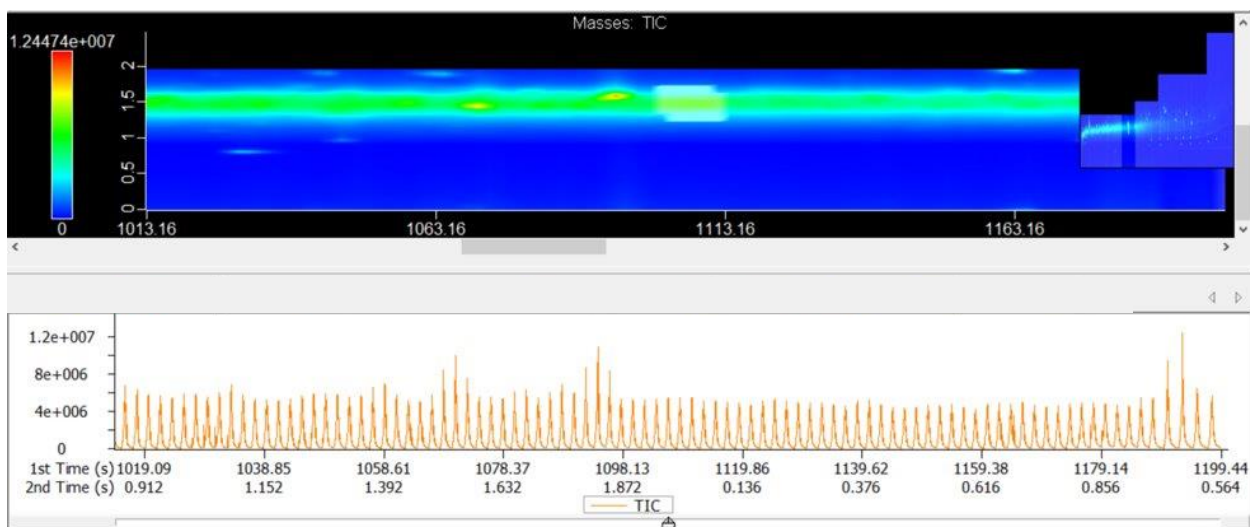


Figure B-86 The real GCxGC image of Sample 43 that belongs to cluster 6

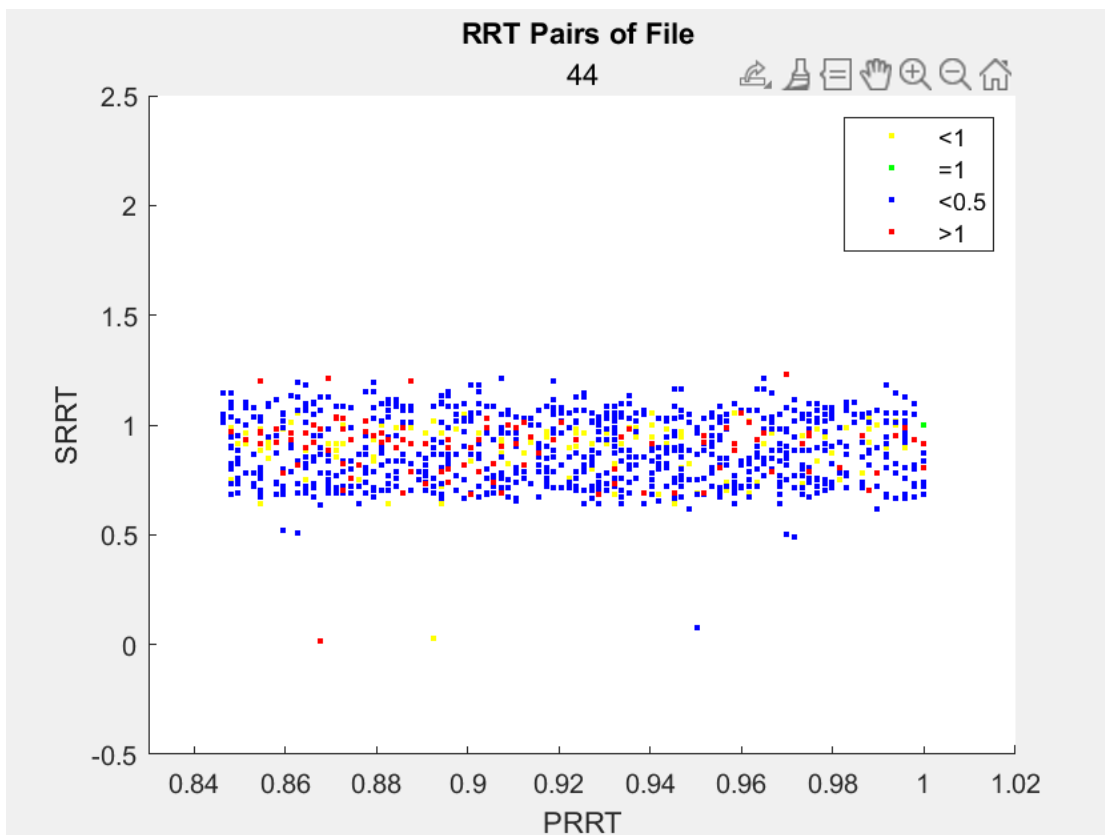


Figure B-87 The discrete GCxGC image of Sample 44 that belongs to cluster 6

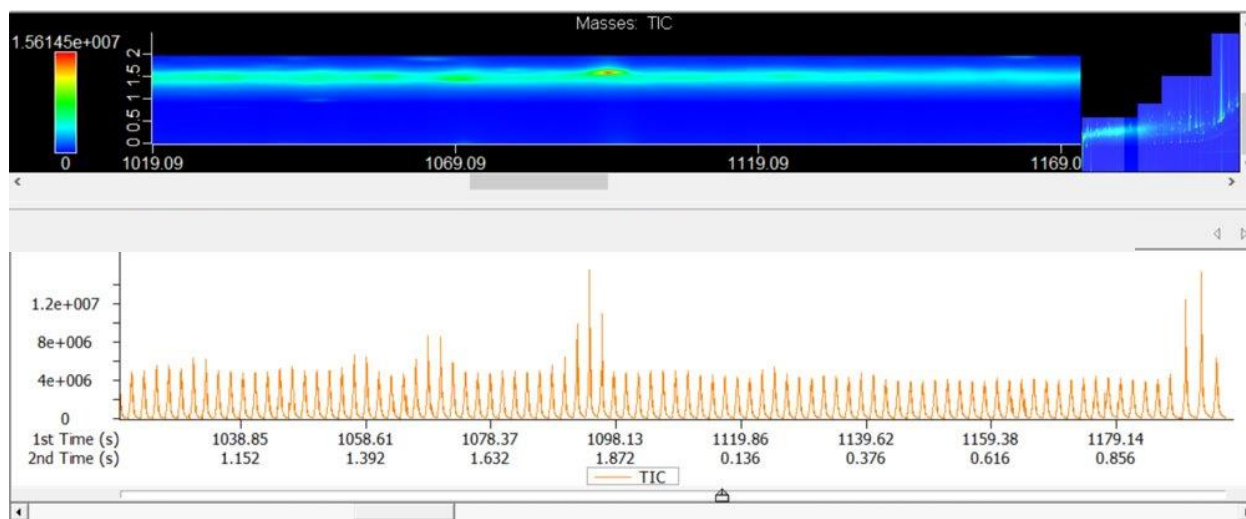


Figure B-88 The real GCxGC image of Sample 44 that belongs to cluster 6

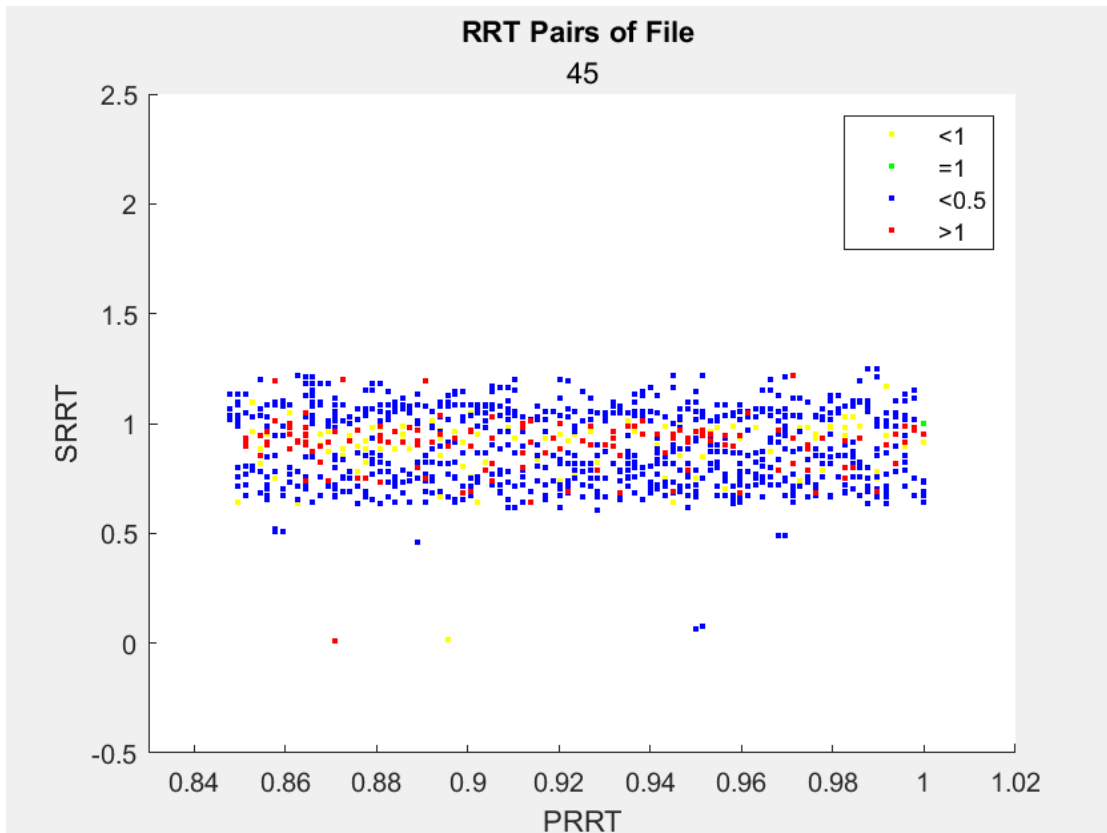


Figure B-89 The discrete GCxGC image of Sample 45 that belongs to cluster 6

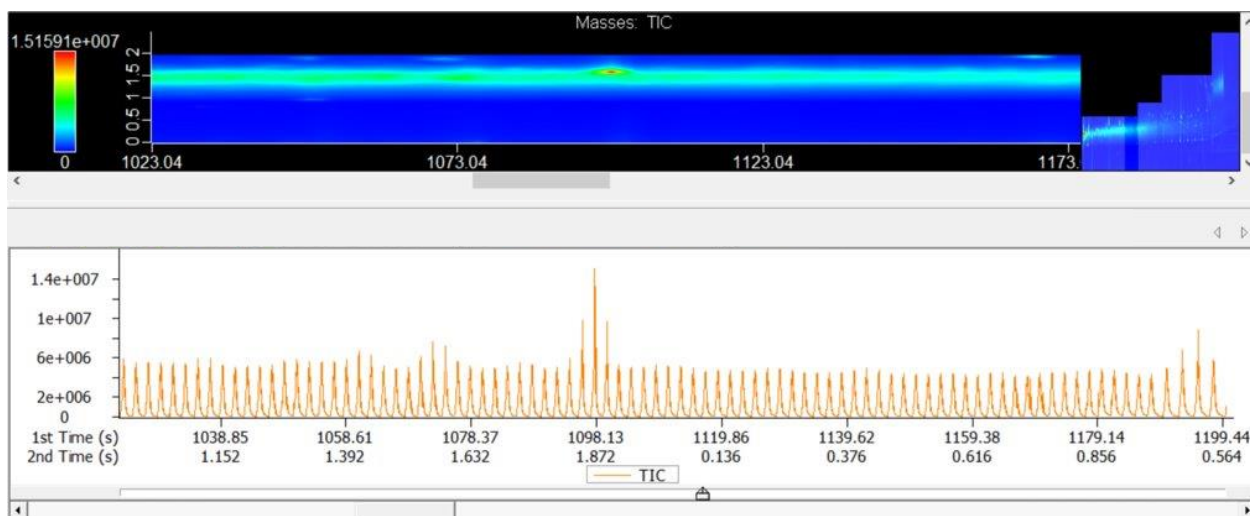


Figure B-90 The real GCxGC image of Sample 45 that belongs to cluster 6

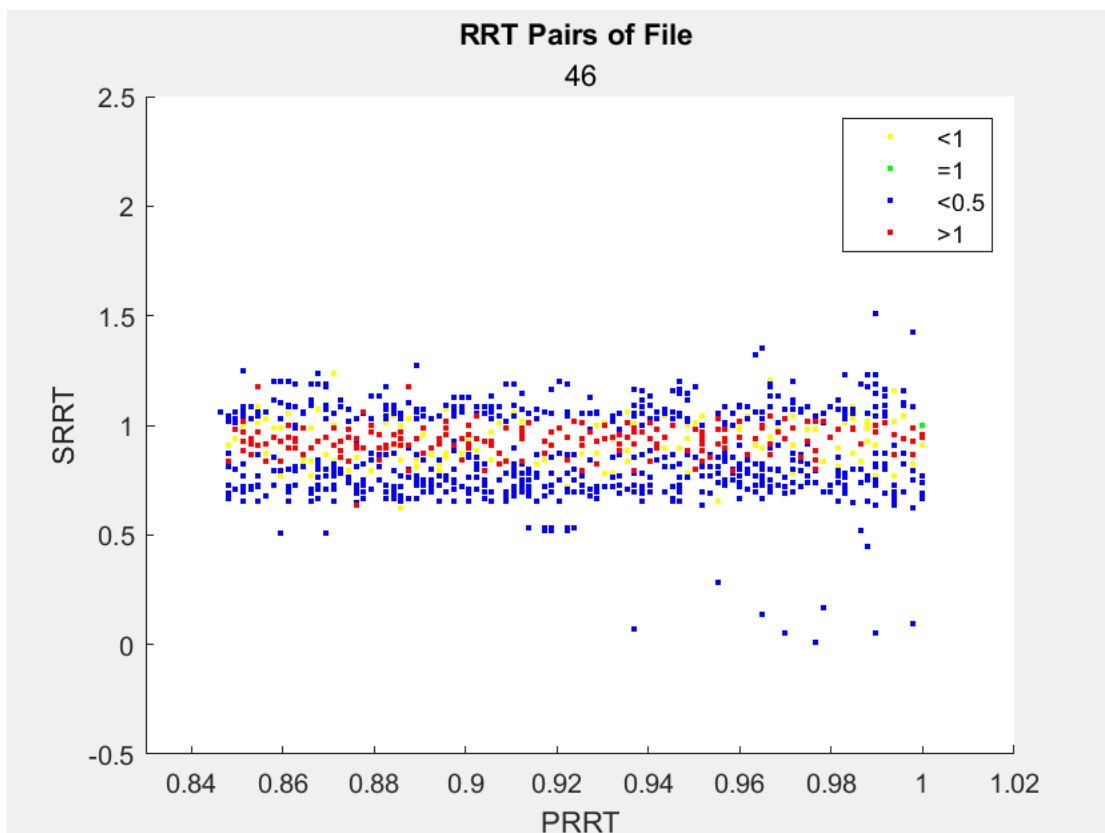


Figure B-91 The discrete GCxGC image of Sample 46 that belongs to cluster 6

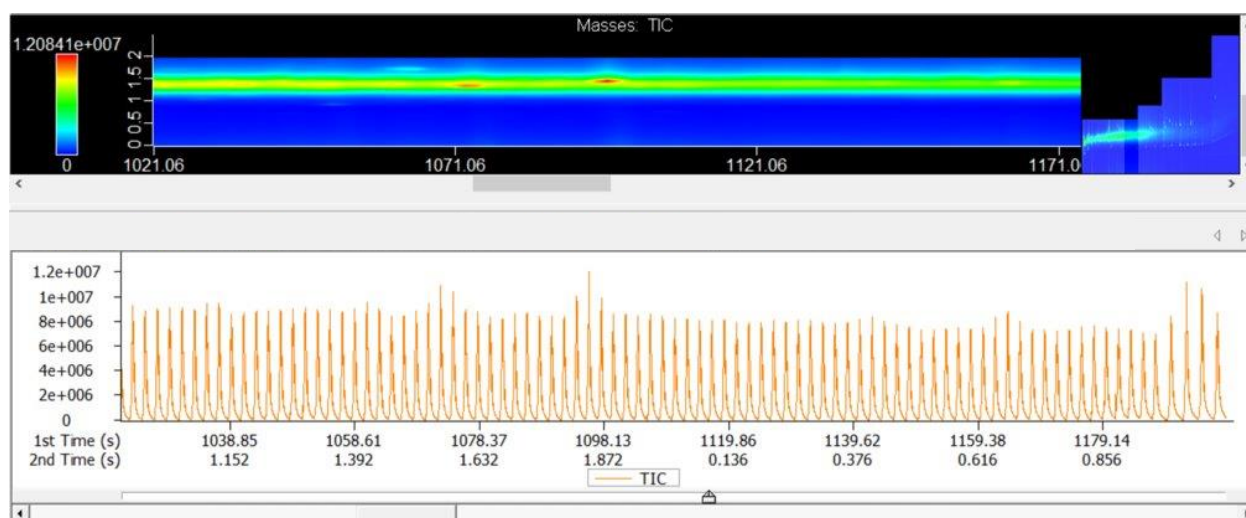


Figure B-92 The real GCxGC image of Sample 46 that belongs to cluster 6

Cluster 7: 2, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 24, 25, 33, 34, 36, 40, 41, and 42

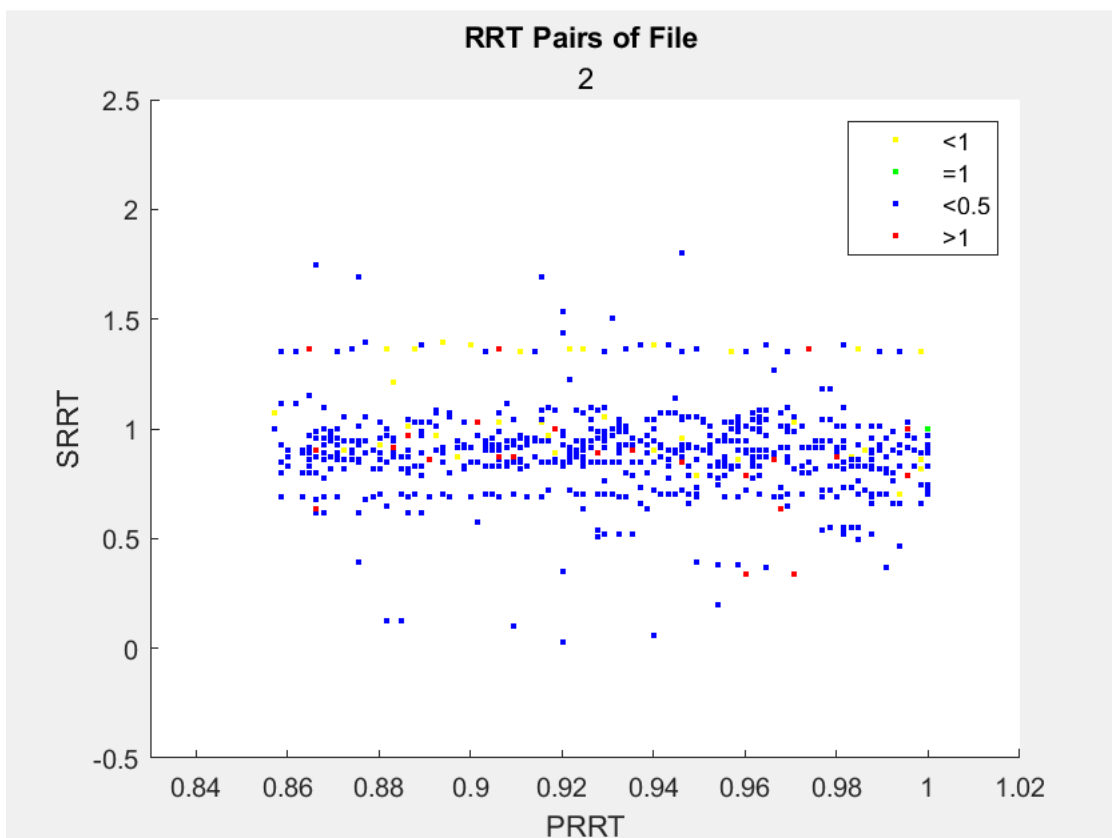


Figure B-93 The discrete GCxGC image of Sample 2 that belongs to cluster 7

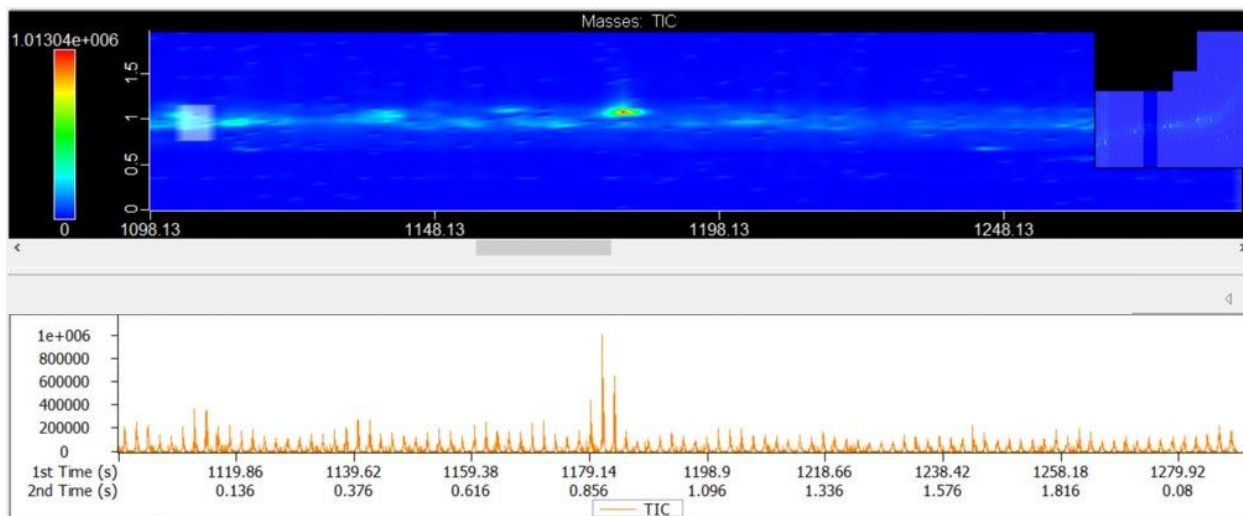


Figure B-94 The real GCxGC image of Sample 2 that belongs to cluster 7

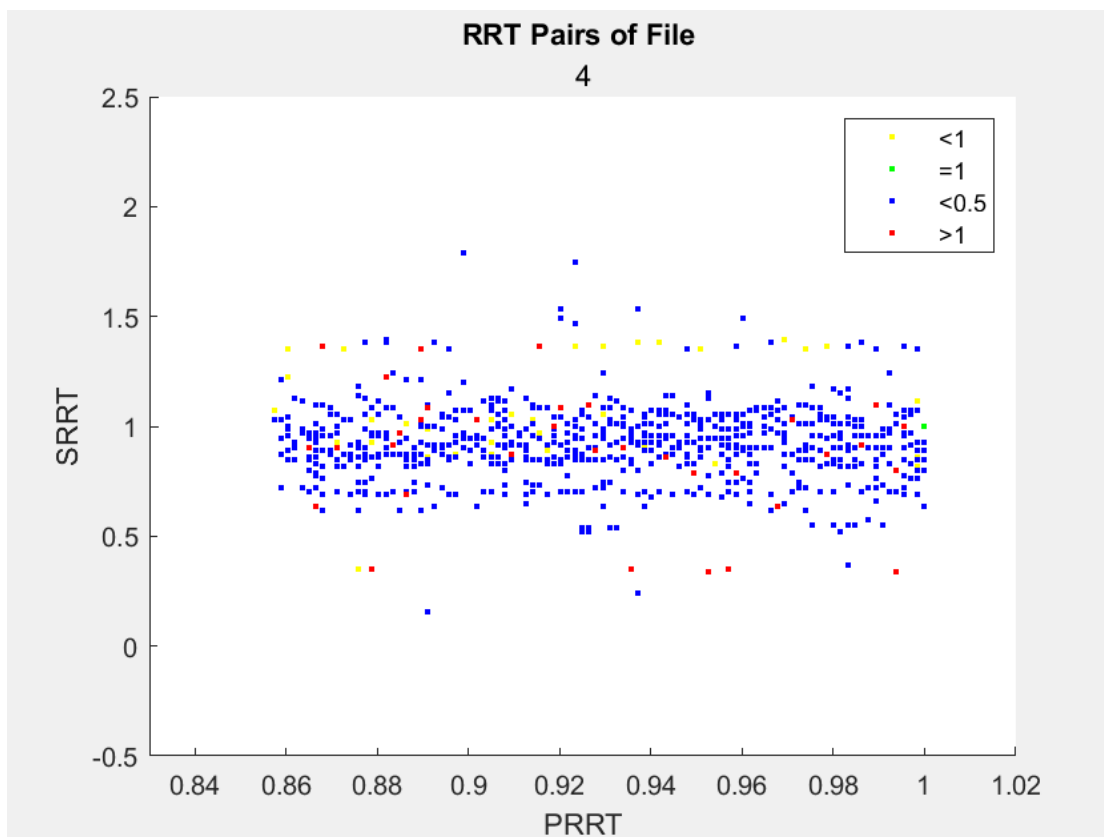


Figure B-95 The discrete GCxGC image of Sample 4 that belongs to cluster 7

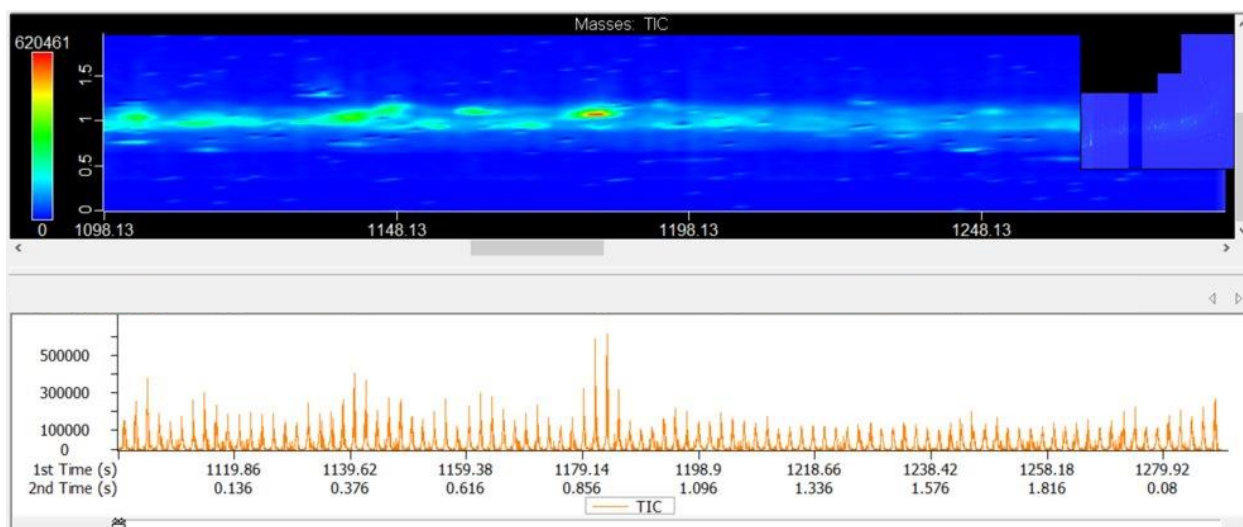


Figure B-96 The real GCxGC image of Sample 4 that belongs to cluster 7



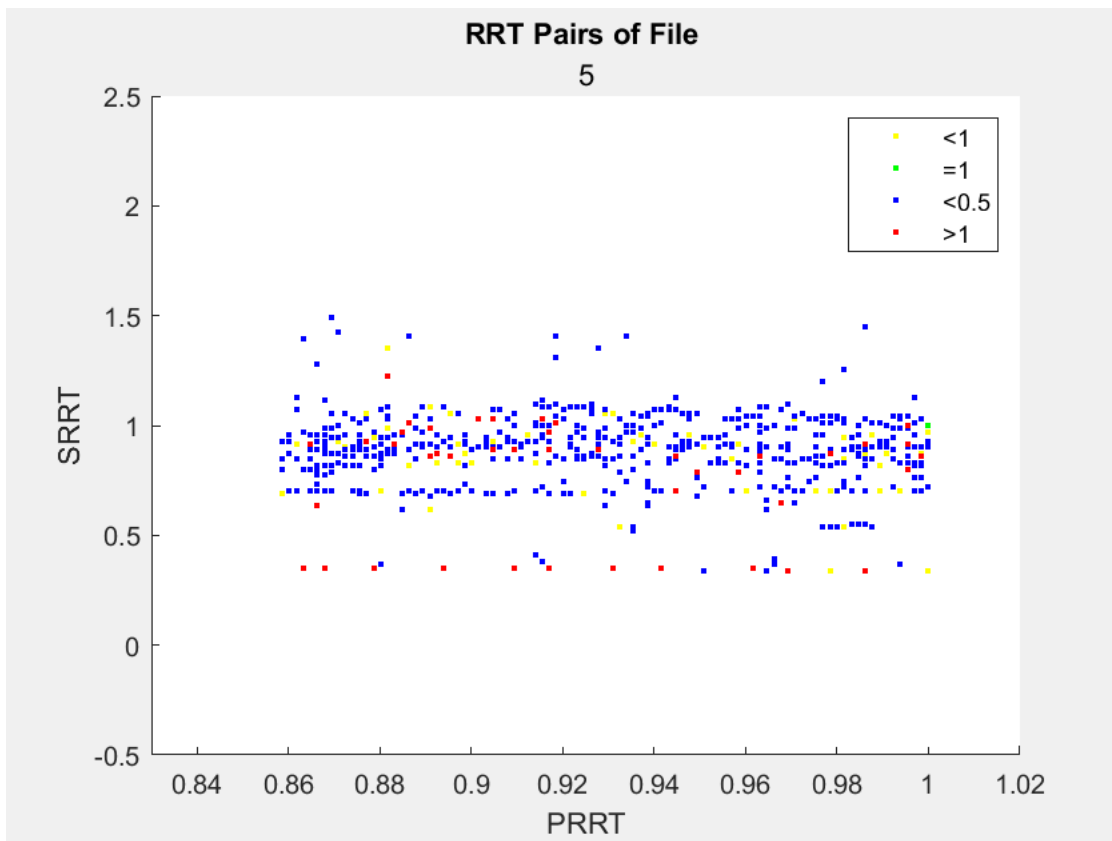


Figure B-97 The discrete GCxGC image of Sample 5 that belongs to cluster 7

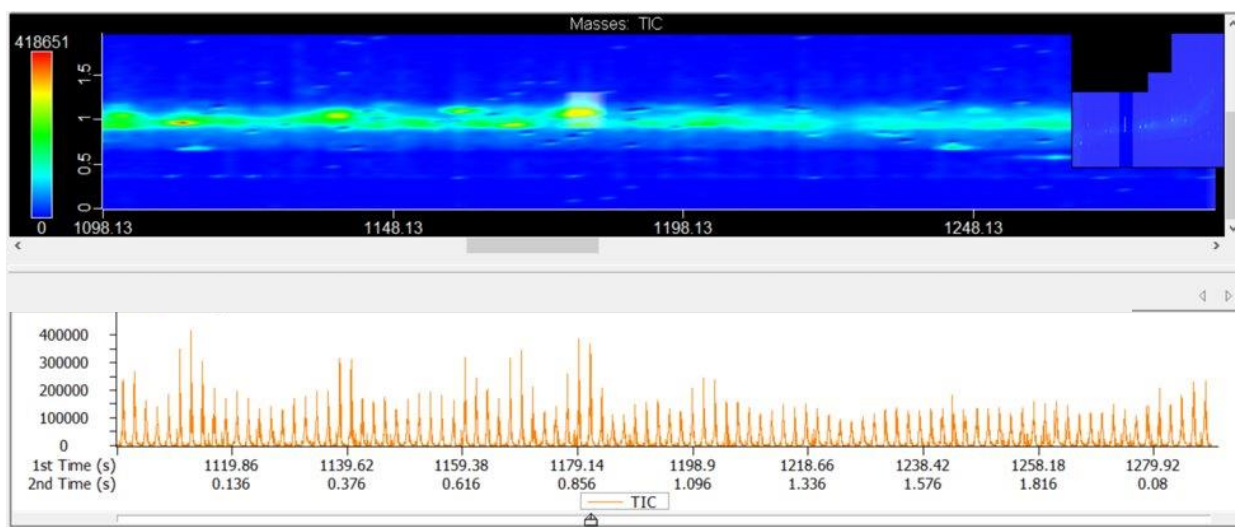


Figure B-98 The real GCxGC image of Sample 5 that belongs to cluster 7

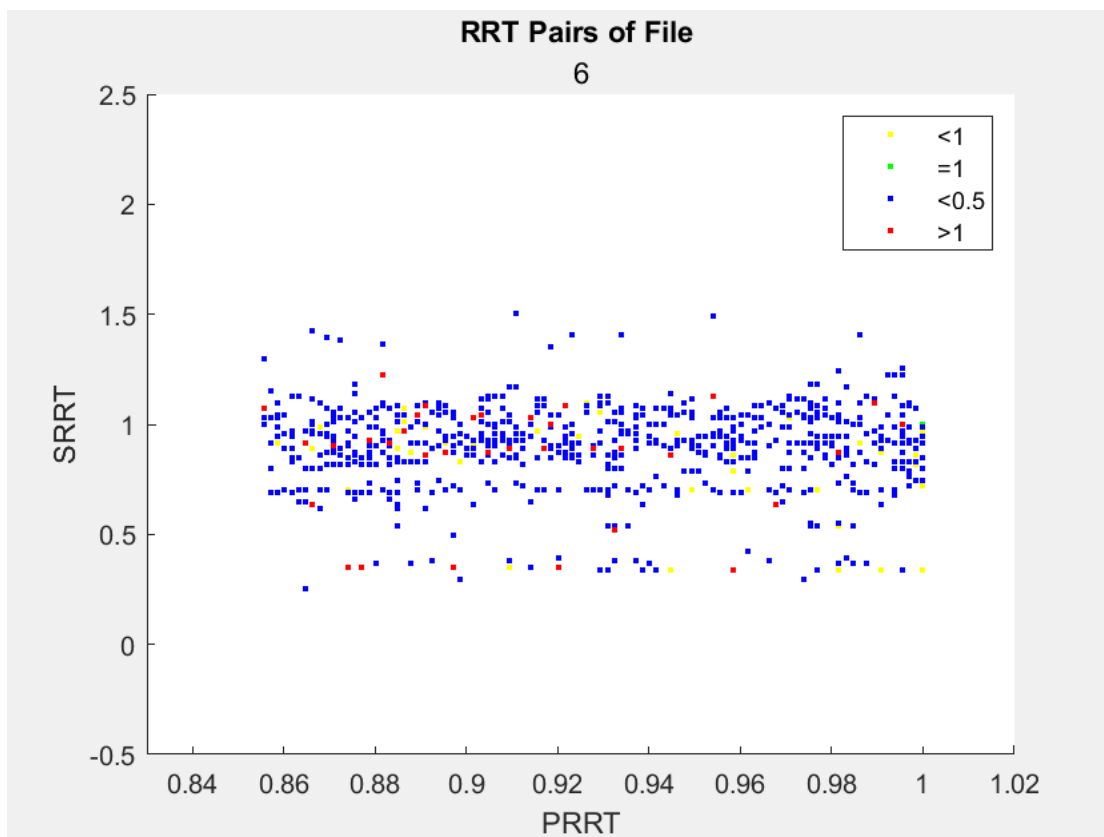


Figure B-99 The discrete GCxGC image of Sample 6 that belongs to cluster 7

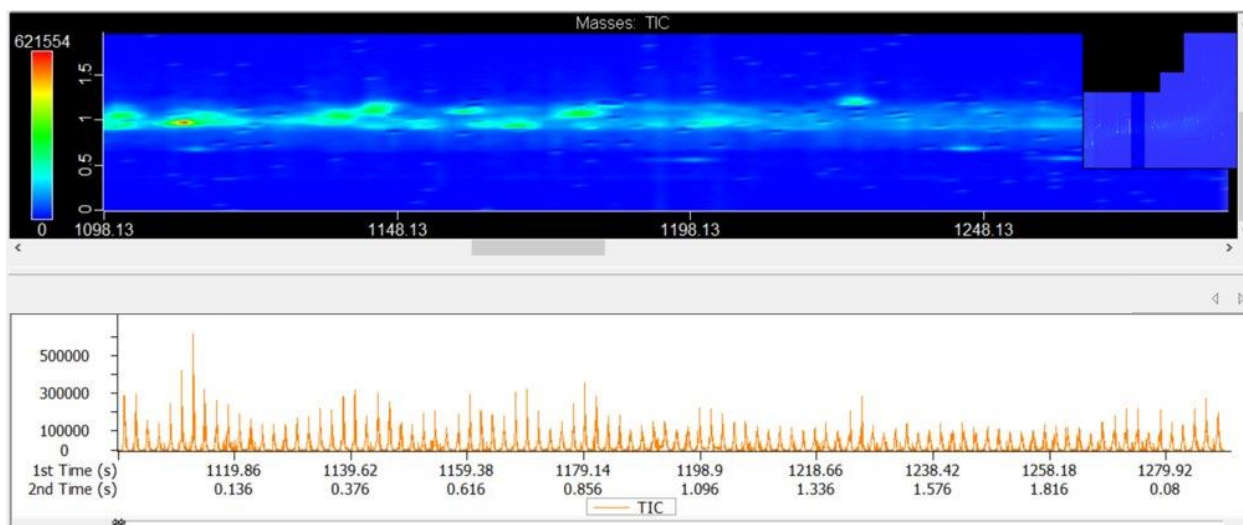


Figure B-100 The real GCxGC image of Sample 6 that belongs to cluster 7

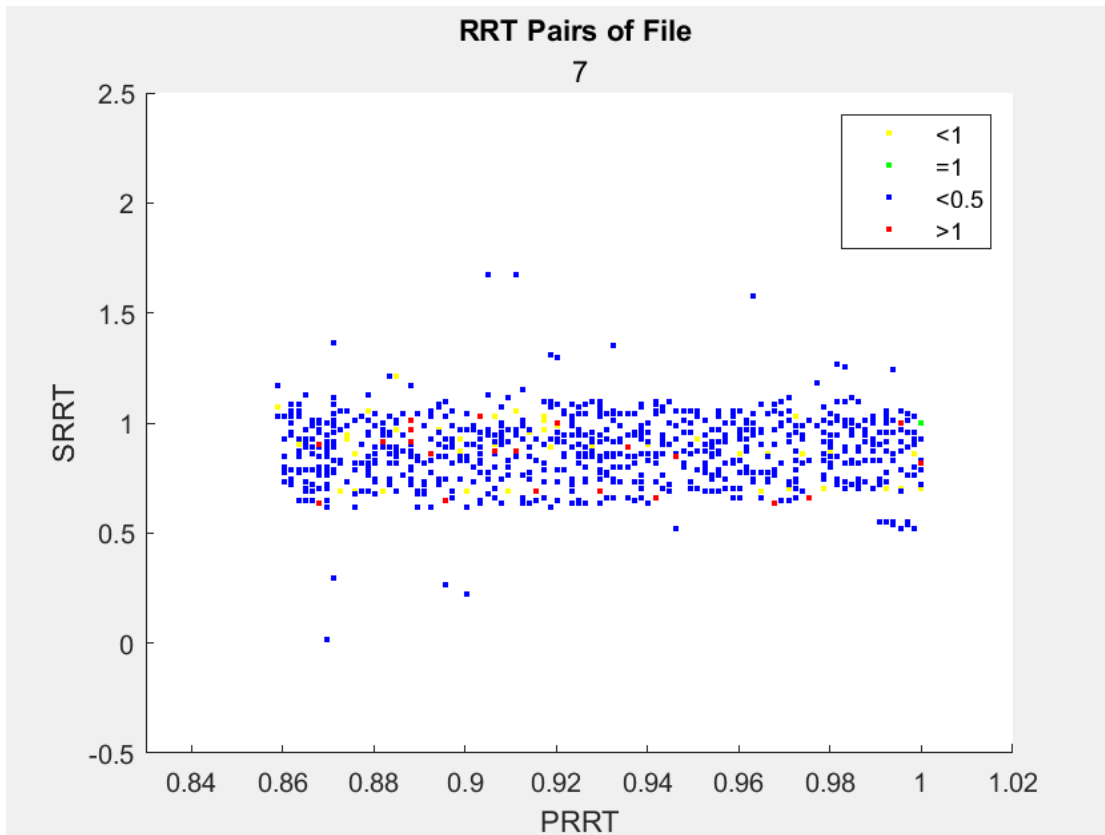


Figure B-101 The discrete GCxGC image of Sample 7 that belongs to cluster 7

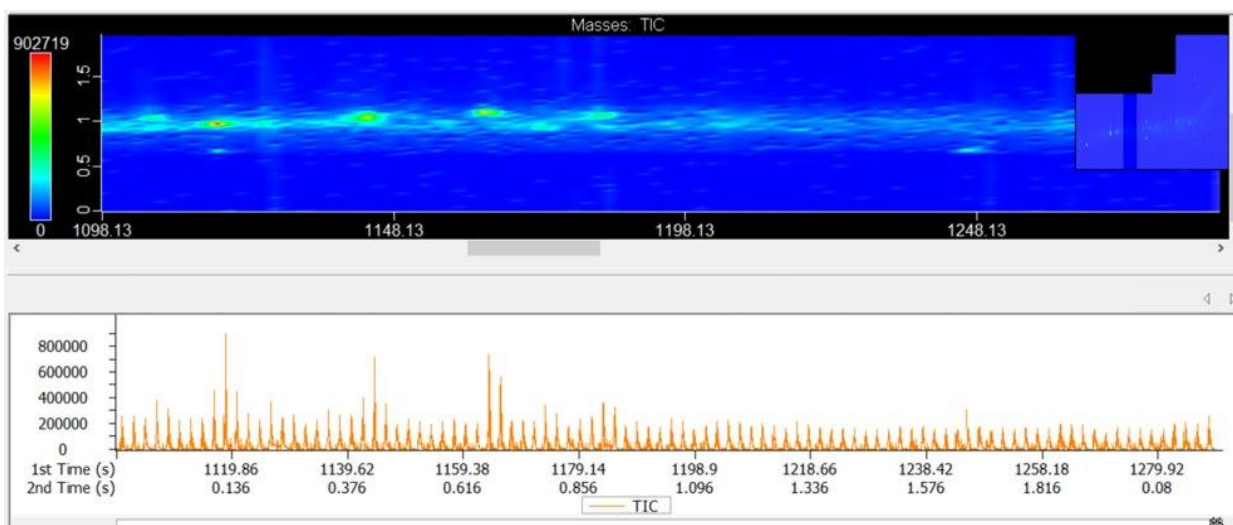


Figure B-102 The real GCxGC image of Sample 7 that belongs to cluster 7

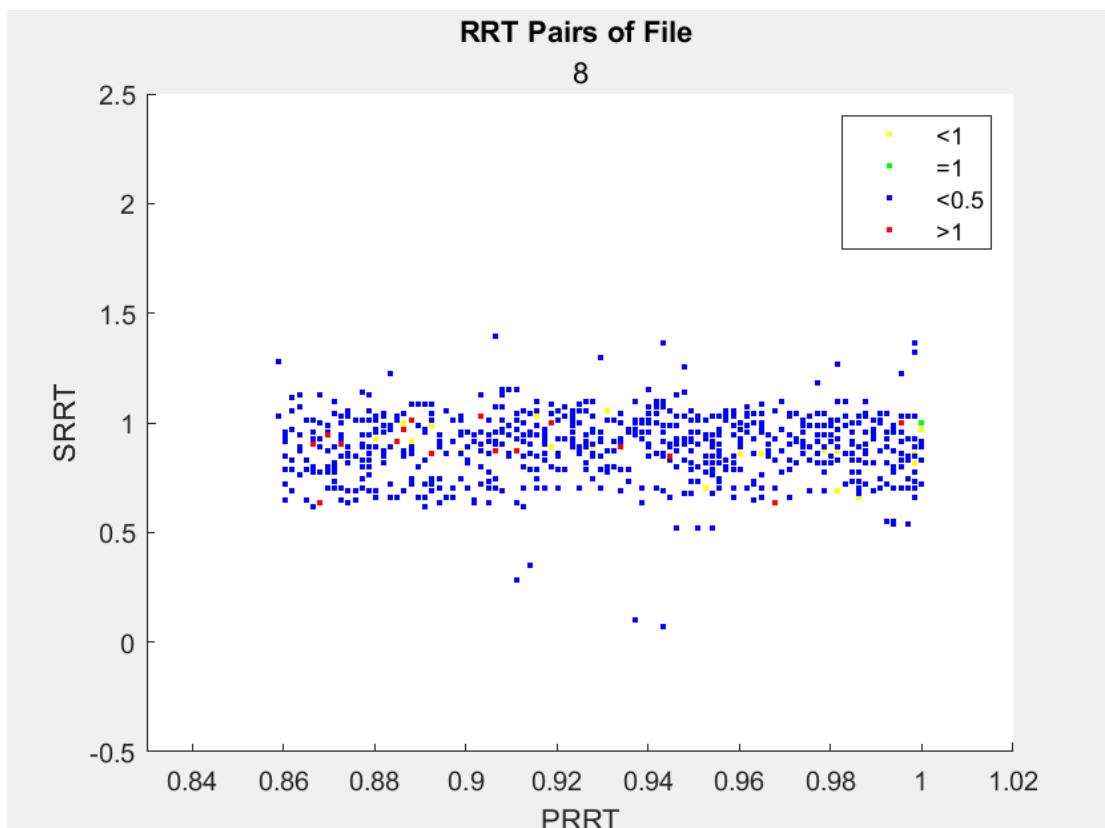


Figure B-103 The discrete GCxGC image of Sample 8 that belongs to cluster 7

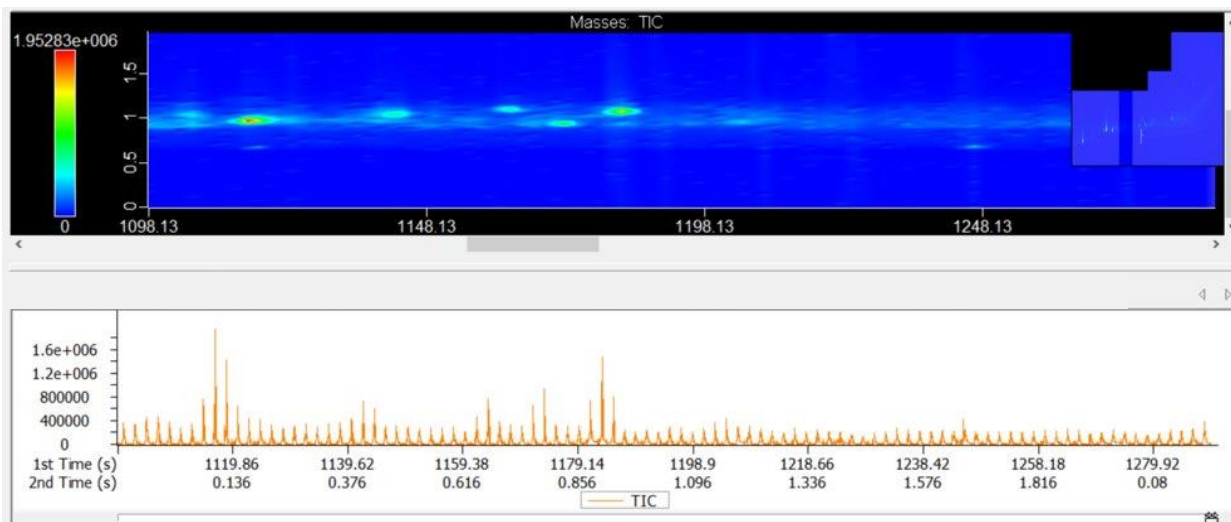


Figure B-104 The real GCxGC image of Sample 8 that belongs to cluster 7

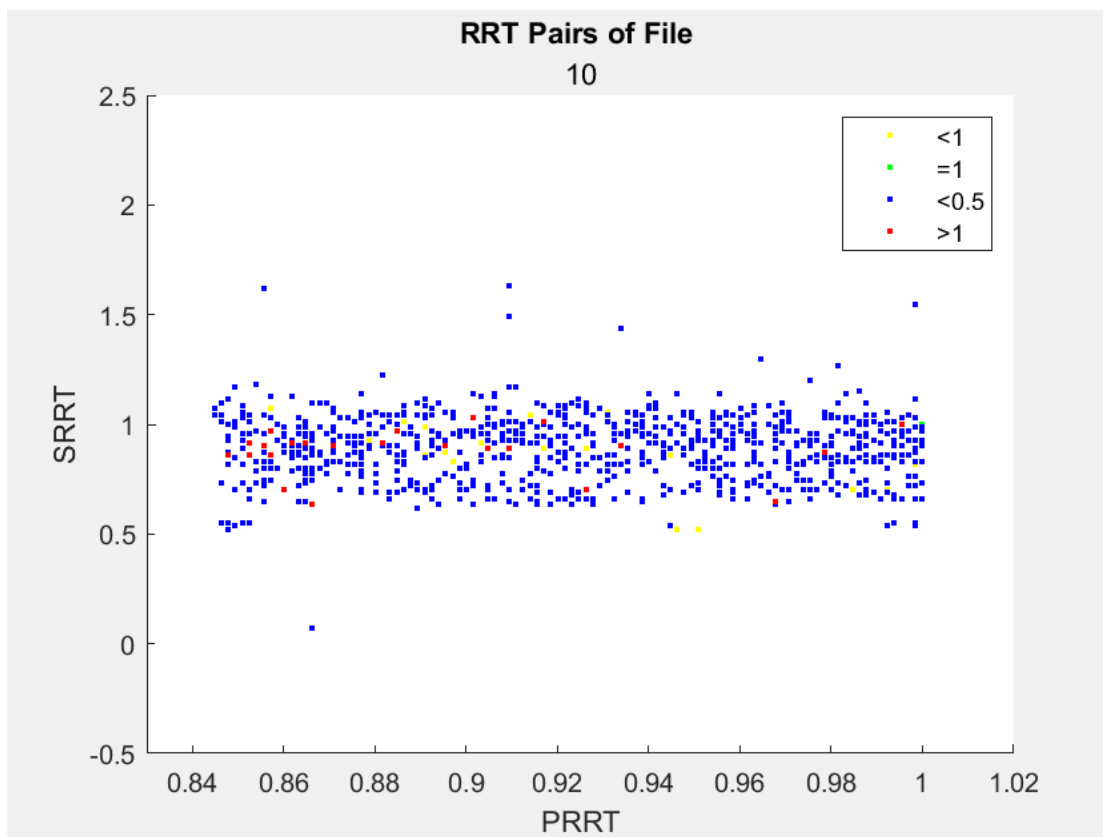


Figure B-105 The discrete GCxGC image of Sample 10 that belongs to cluster 7

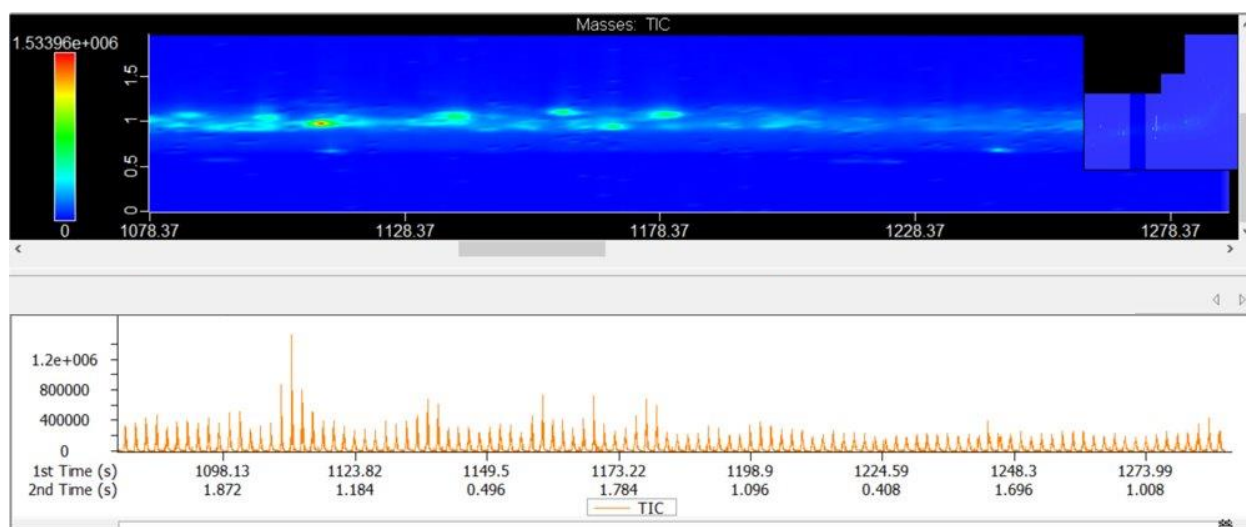


Figure B-106 The real GCxGC image of Sample 10 that belongs to cluster 7

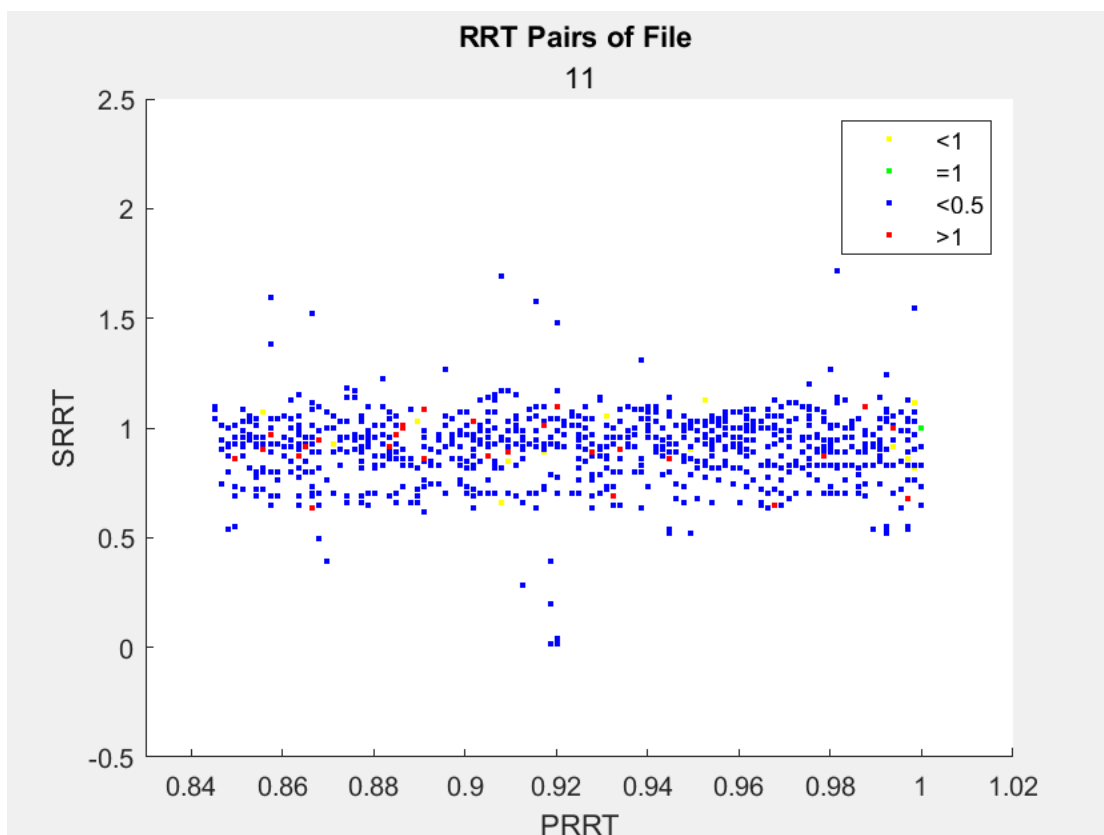


Figure B-107 The discrete GCxGC image of Sample 11 that belongs to cluster 7

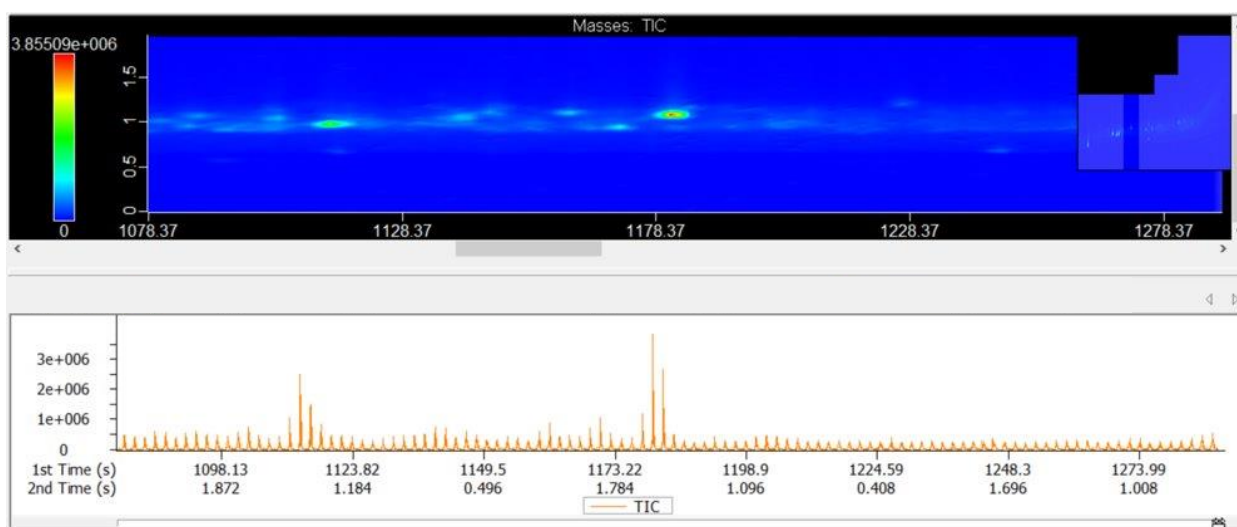


Figure B-108 The real GCxGC image of Sample 11 that belongs to cluster 7

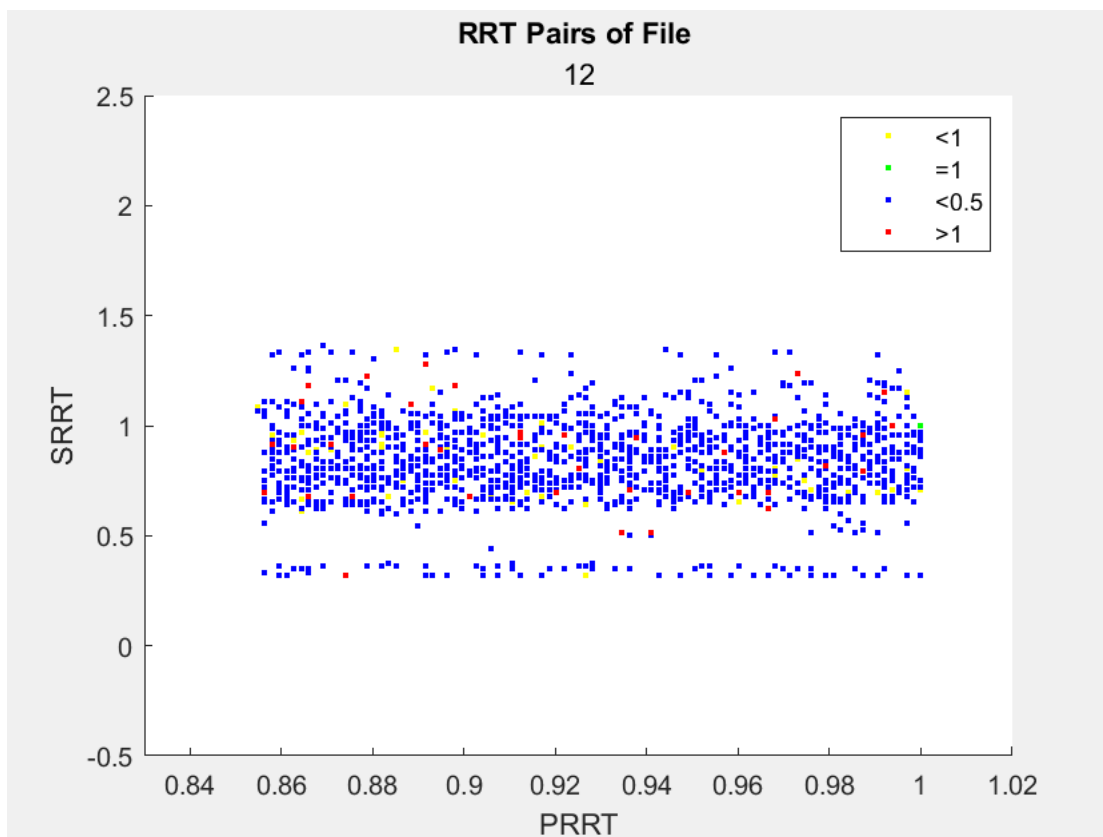


Figure B-109 The discrete GCxGC image of Sample 12 that belongs to cluster 7

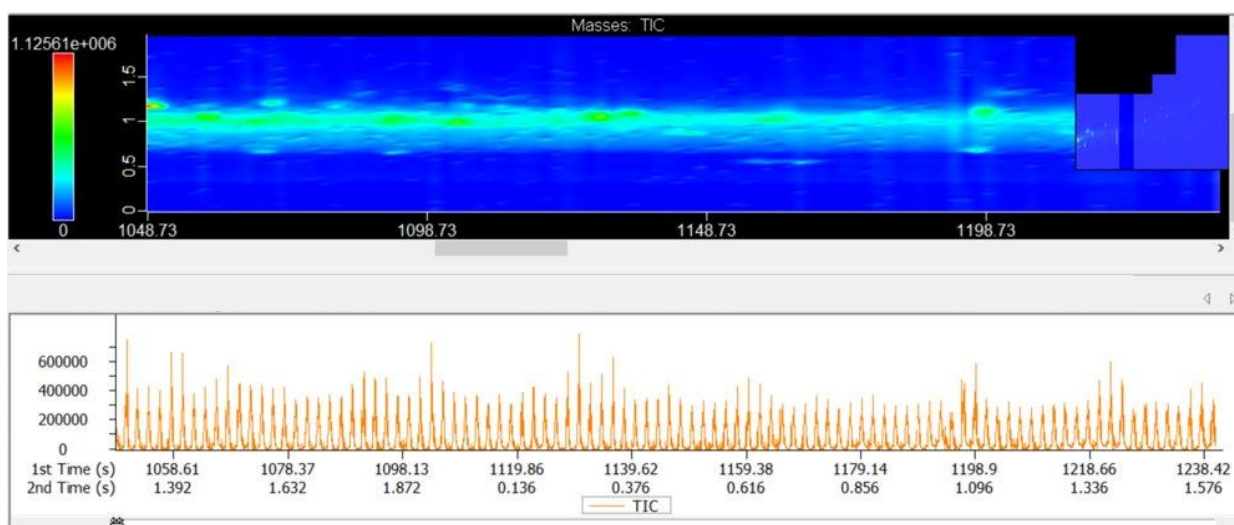


Figure B-110 The real GCxGC image of Sample 12 that belongs to cluster 7

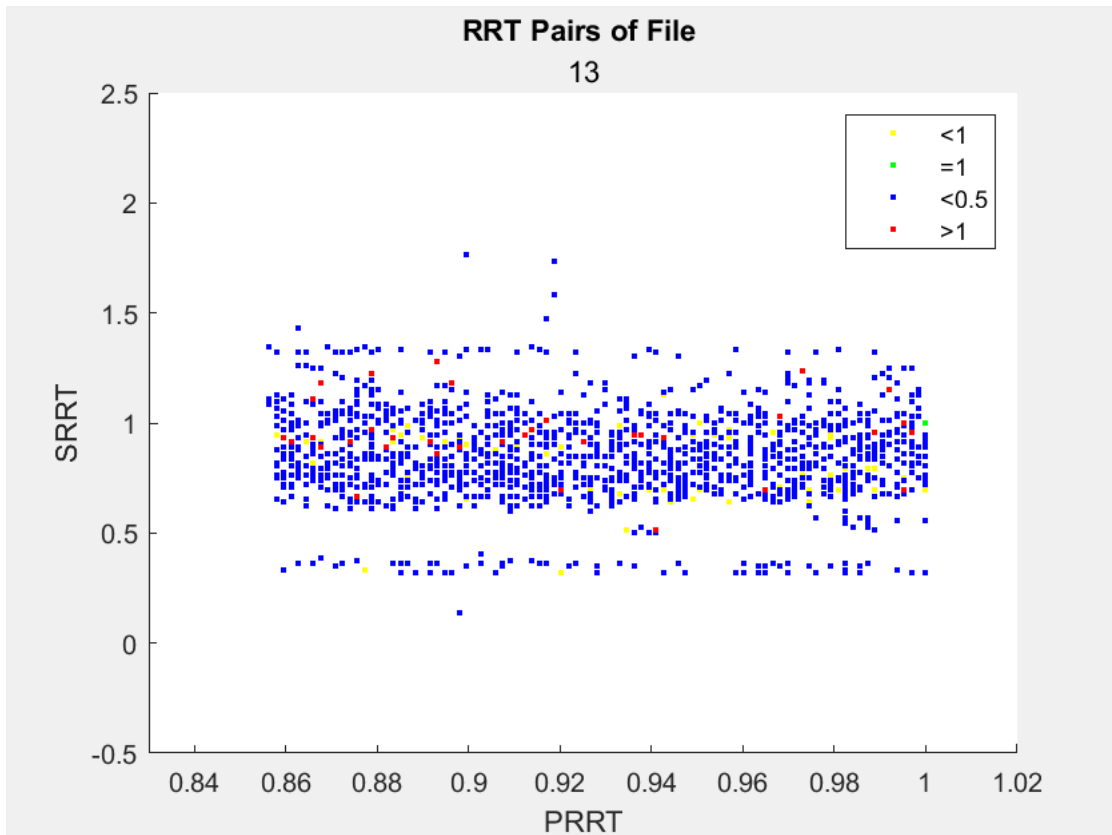


Figure B-111 The discrete GCxGC image of Sample 13 that belongs to cluster 7

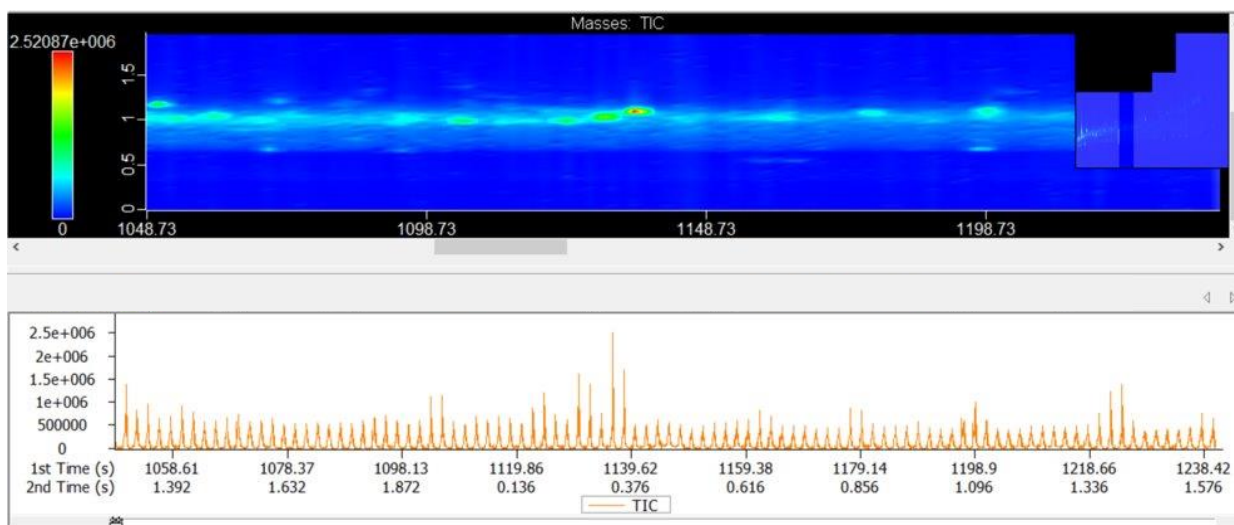


Figure B-112 The real GCxGC image of Sample 13 that belongs to cluster 7



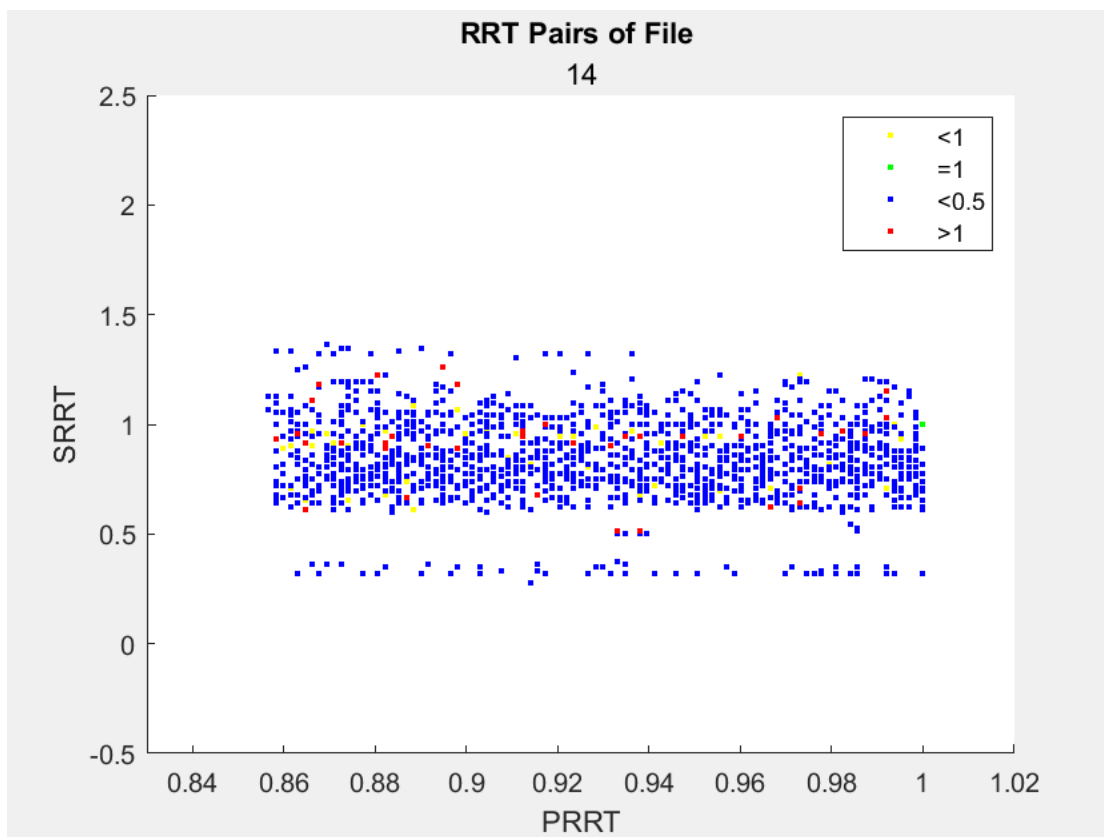


Figure B-113 The discrete GCxGC image of Sample 14 that belongs to cluster 7

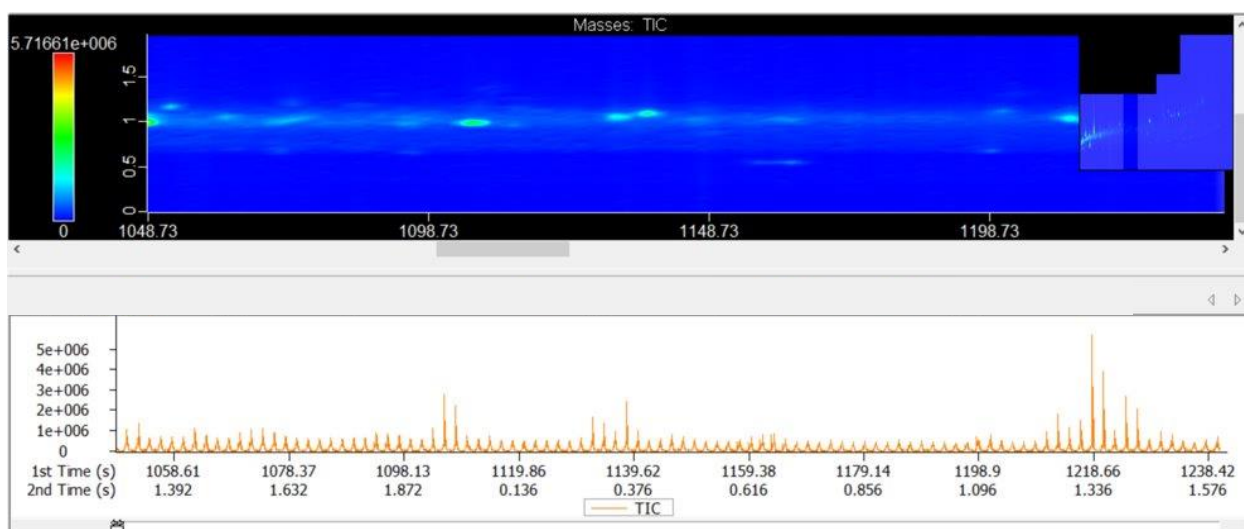


Figure B-114 The real GCxGC image of Sample 14 that belongs to cluster 7

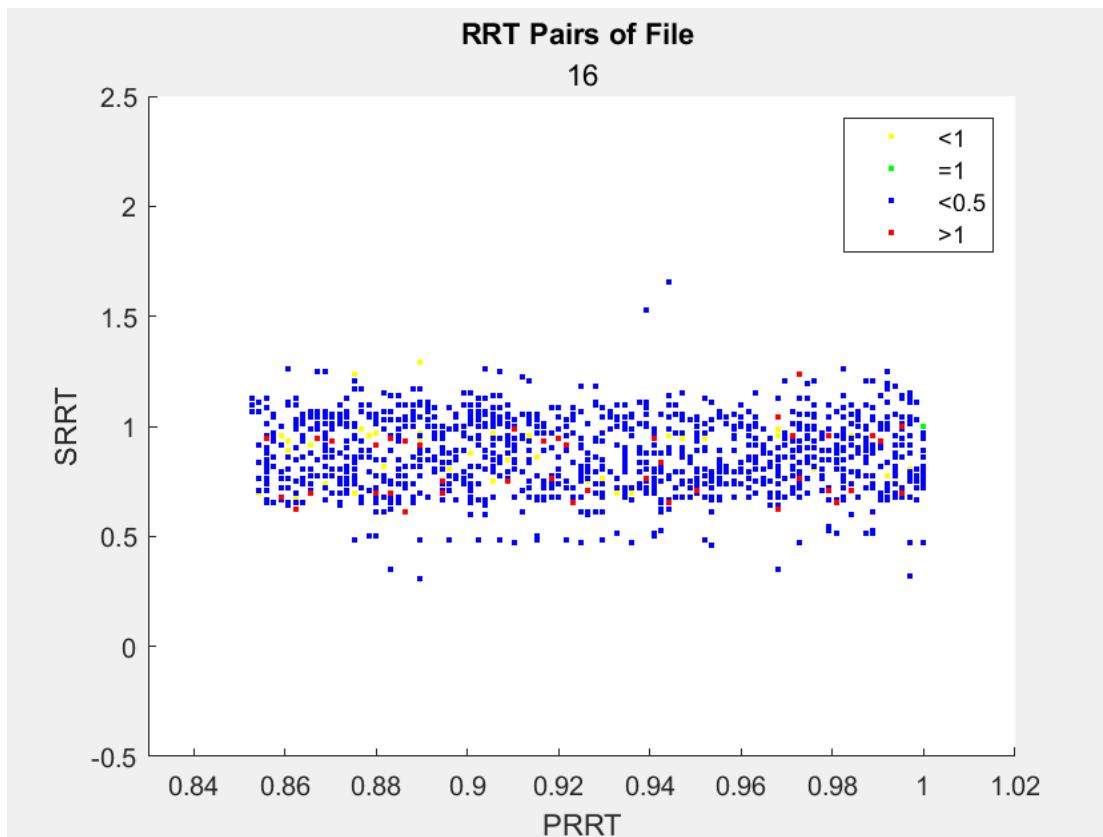


Figure B-115 The discrete GCxGC image of Sample 16 that belongs to cluster 7

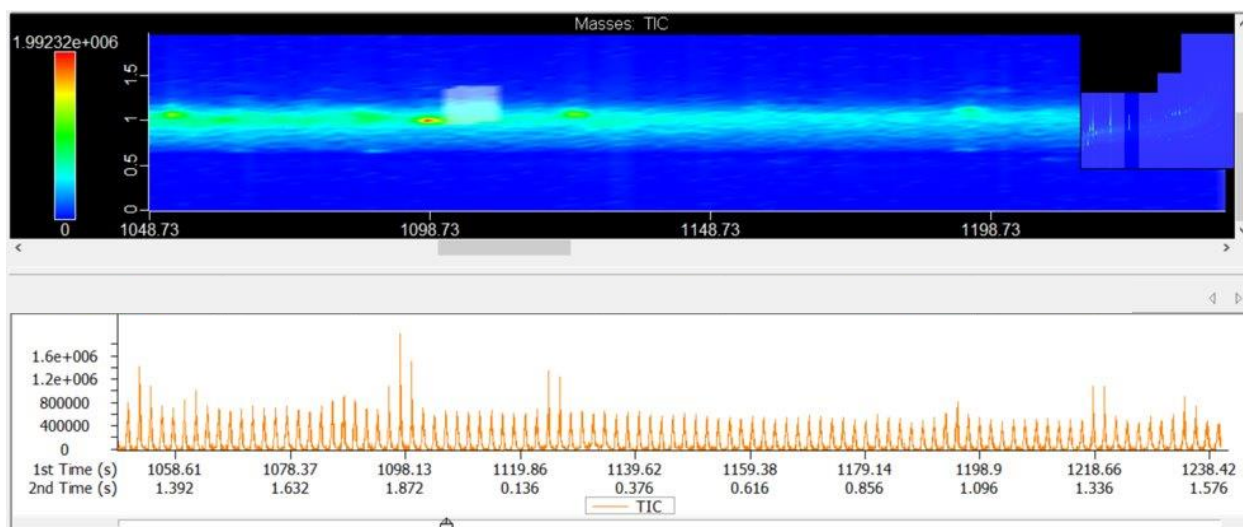


Figure B-116 The real GCxGC image of Sample 16 that belongs to cluster 7

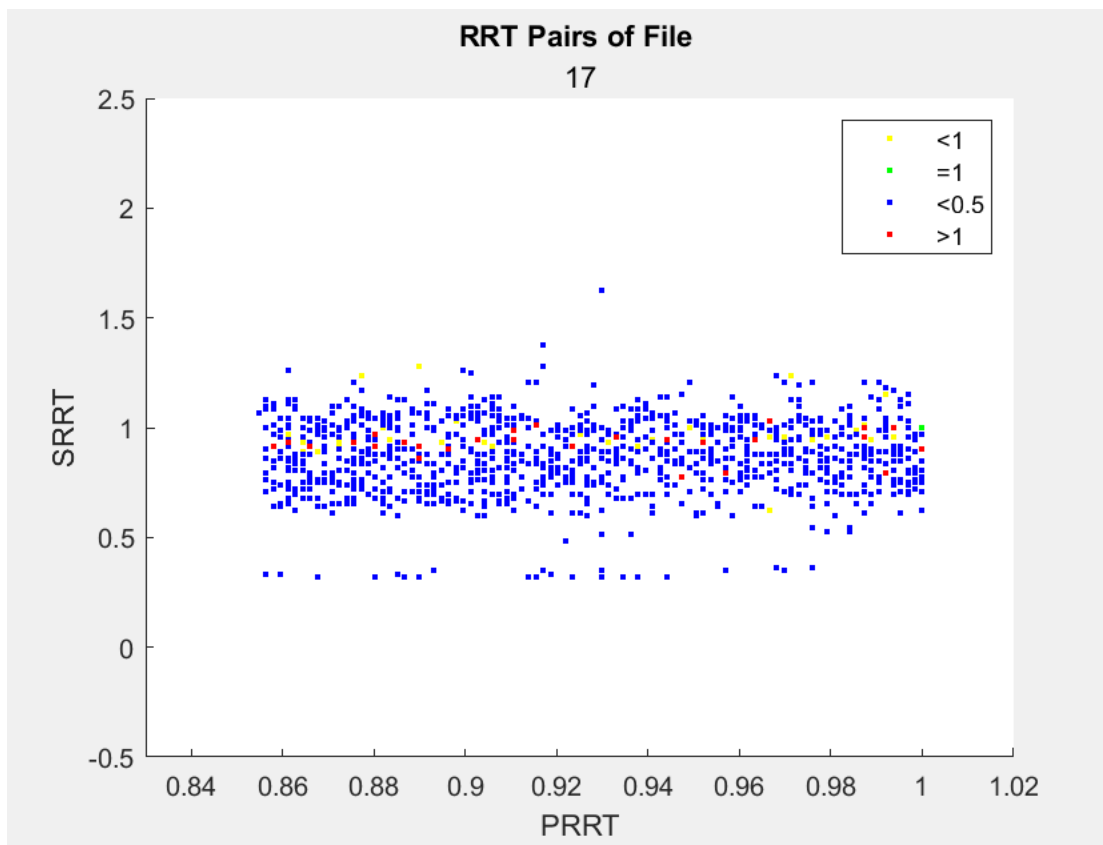


Figure B-117 The discrete GCxGC image of Sample 17 that belongs to cluster 7

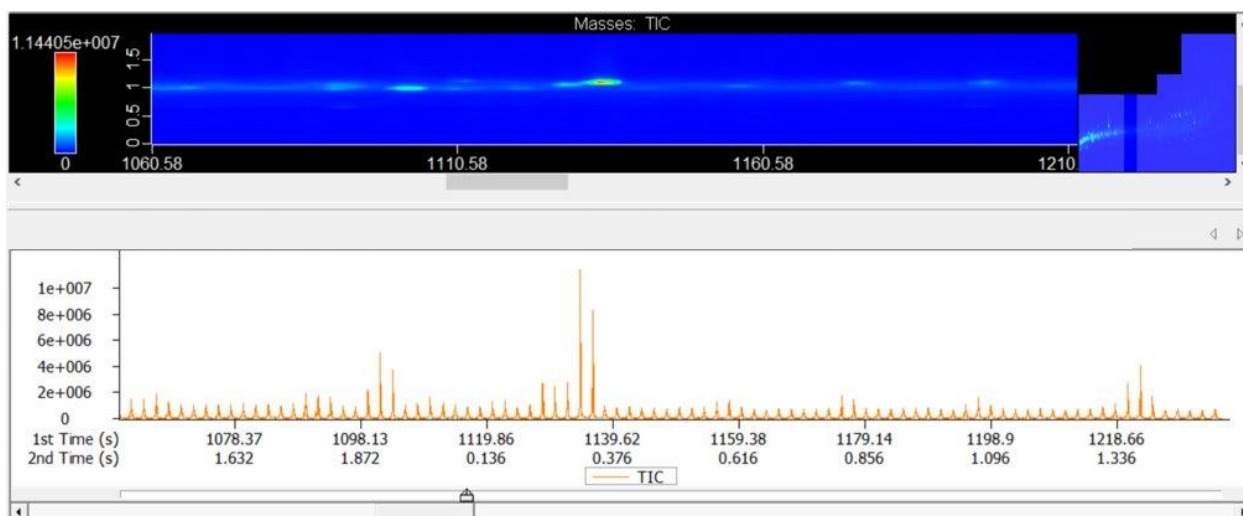


Figure B-118 The real GCxGC image of Sample 17 that belongs to cluster 7

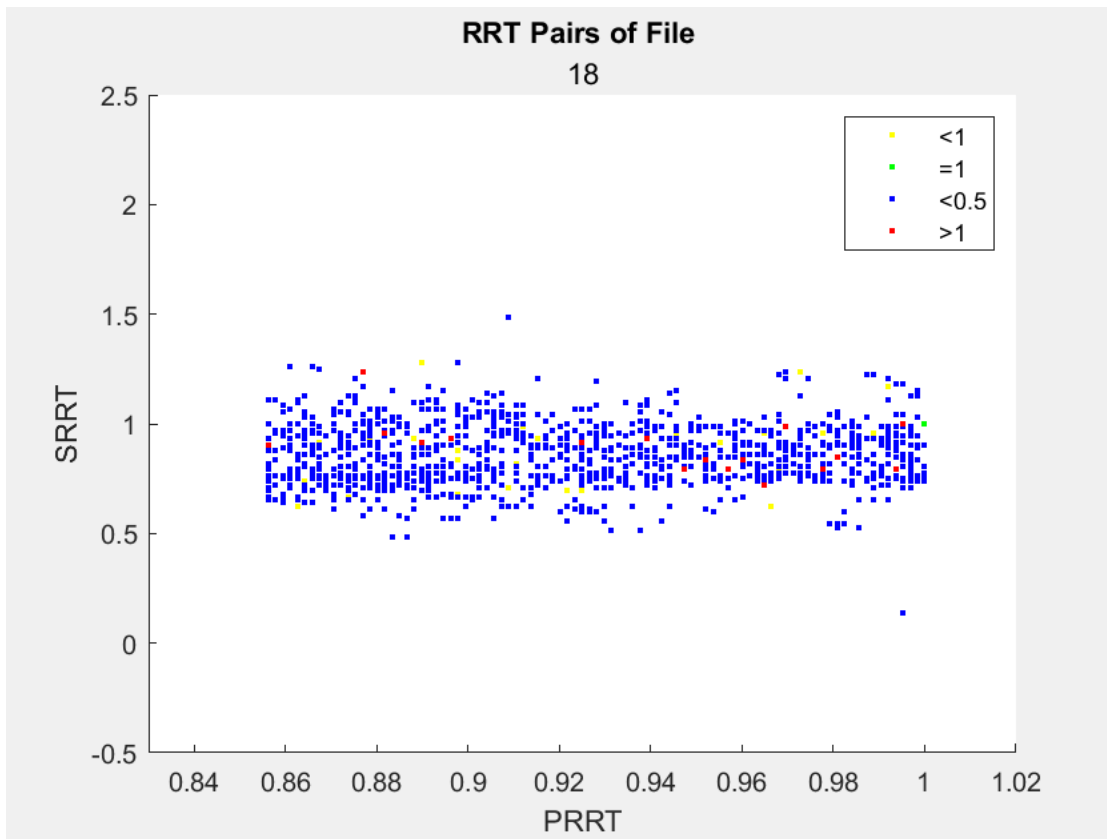


Figure B-119 The discrete GCxGC image of Sample 18 that belongs to cluster 7

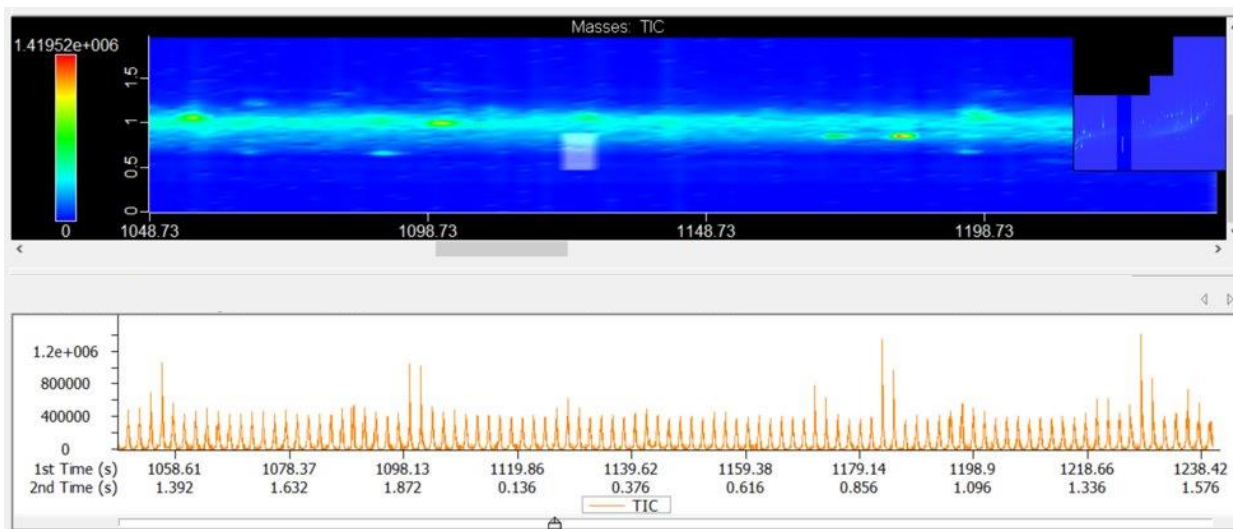


Figure B-120 The real GCxGC image of Sample 18 that belongs to cluster 7

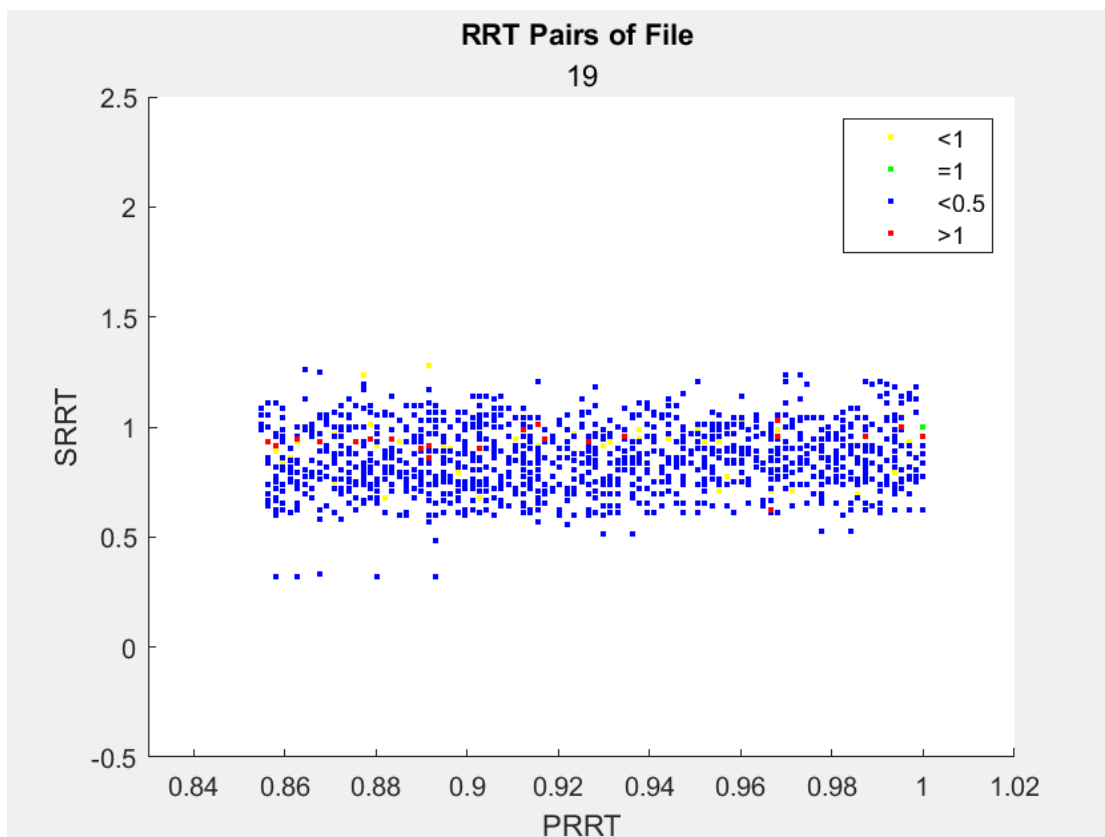


Figure B-121 The discrete GCxGC image of Sample 19 that belongs to cluster 7

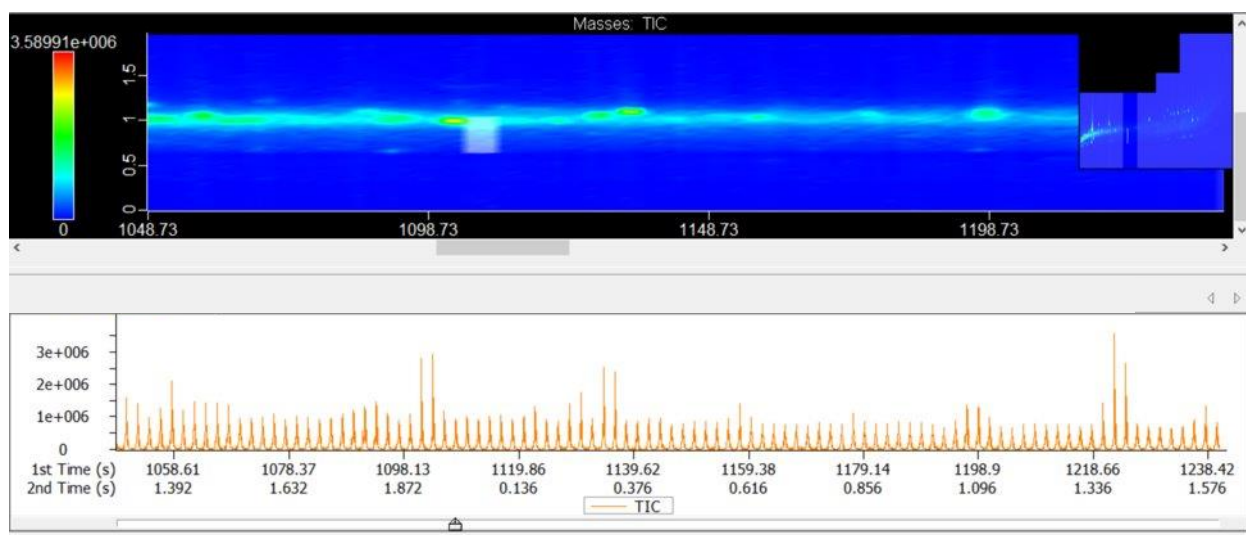


Figure B-122 The real GCxGC image of Sample 19 that belongs to cluster 7

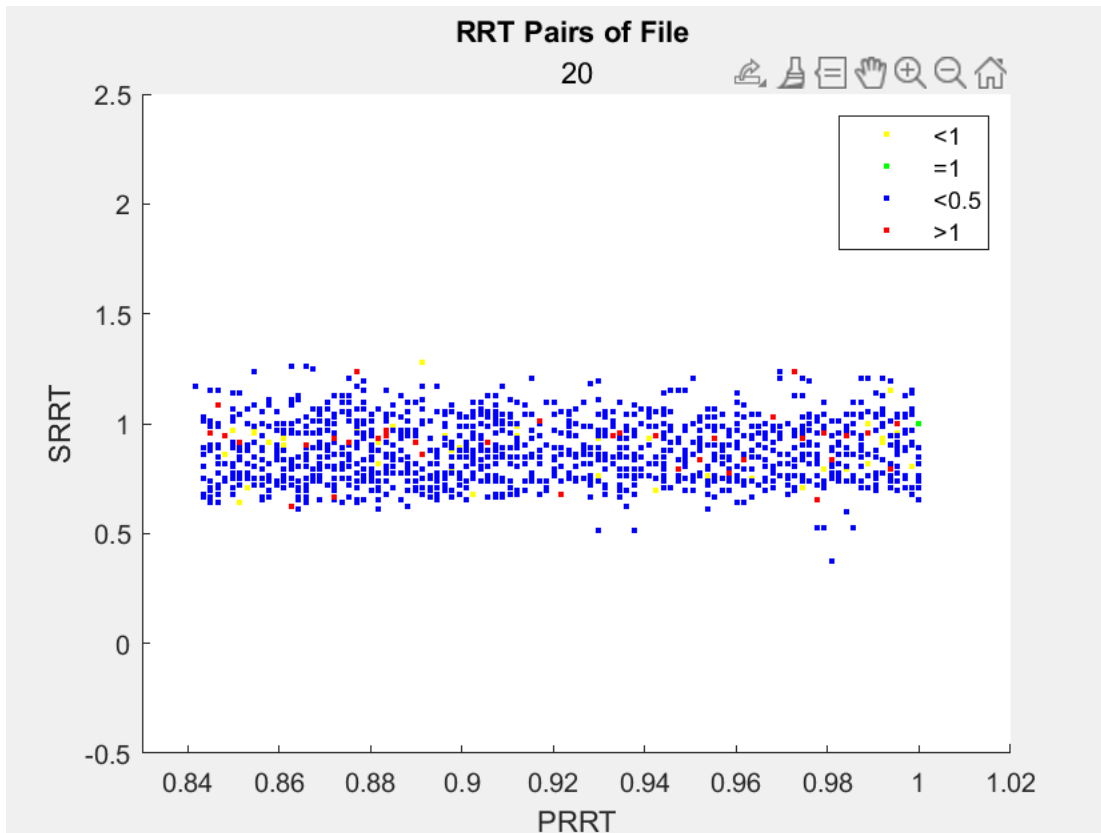


Figure B-123 The discrete GCxGC image of Sample 20 that belongs to cluster 7

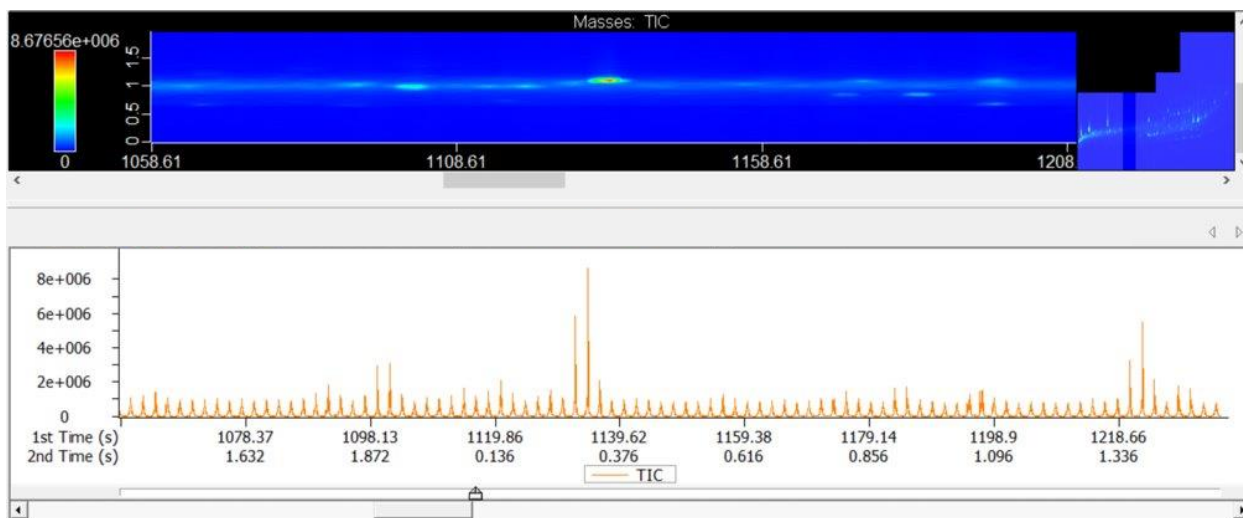


Figure B-124 The real GCxGC image of Sample 20 that belongs to cluster 7

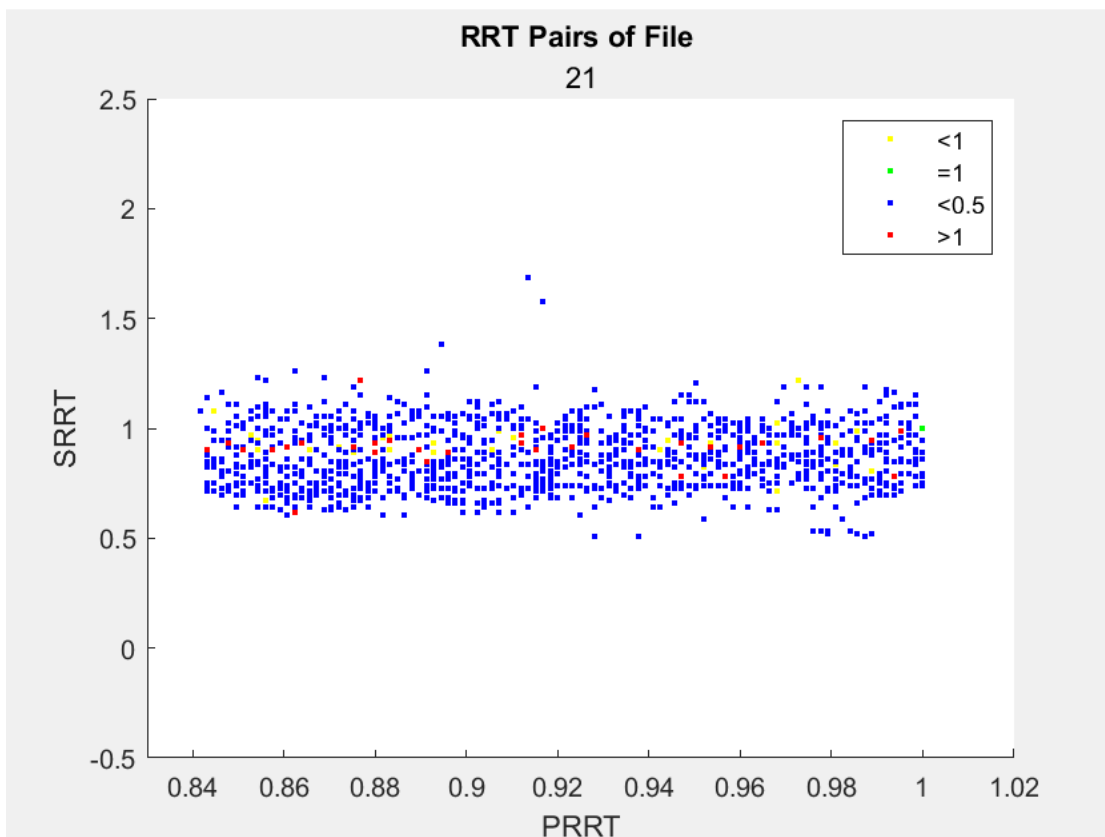


Figure B-125 The discrete GCxGC image of Sample 21 that belongs to cluster 7

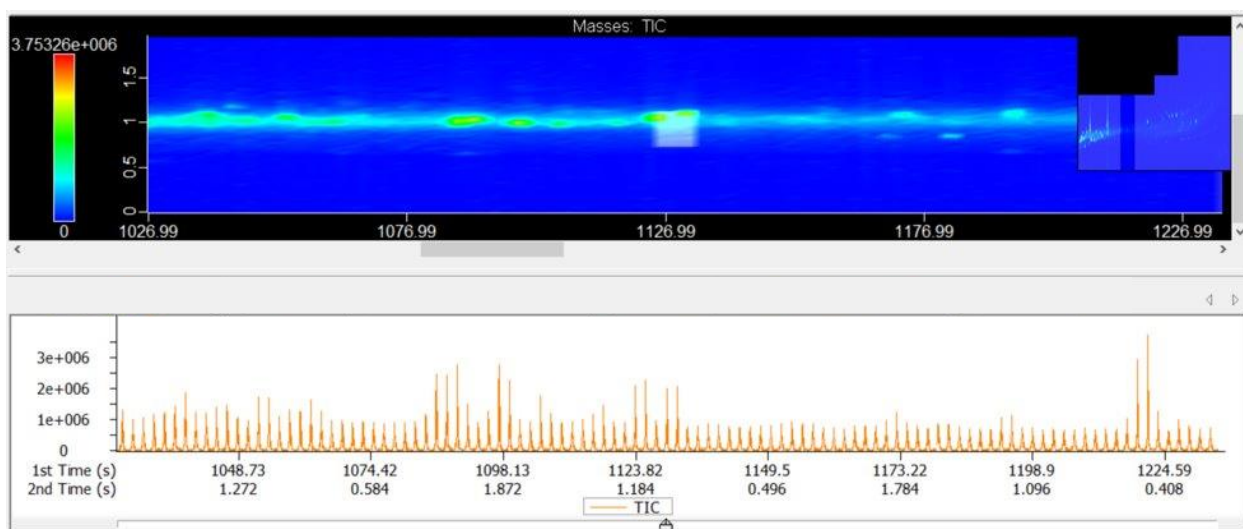


Figure B-126 The real GCxGC image of Sample 21 that belongs to cluster 7

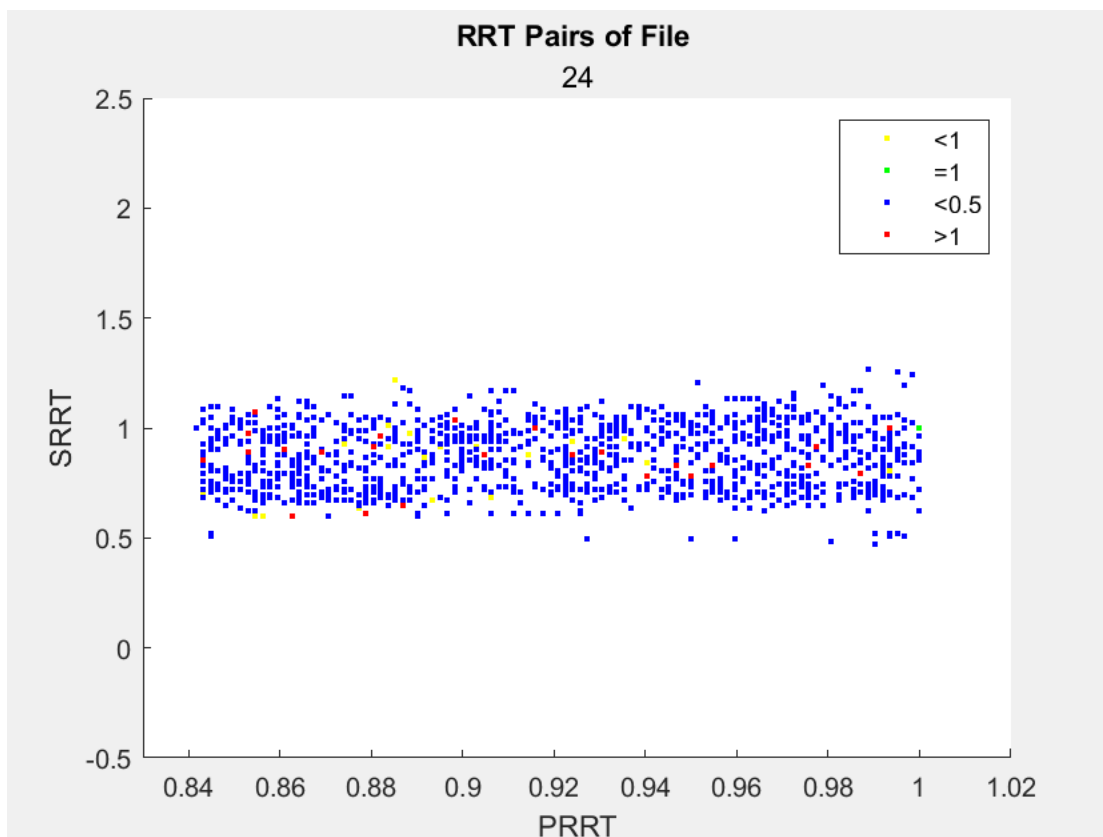


Figure B-127 The discrete GCxGC image of Sample 24 that belongs to cluster 7

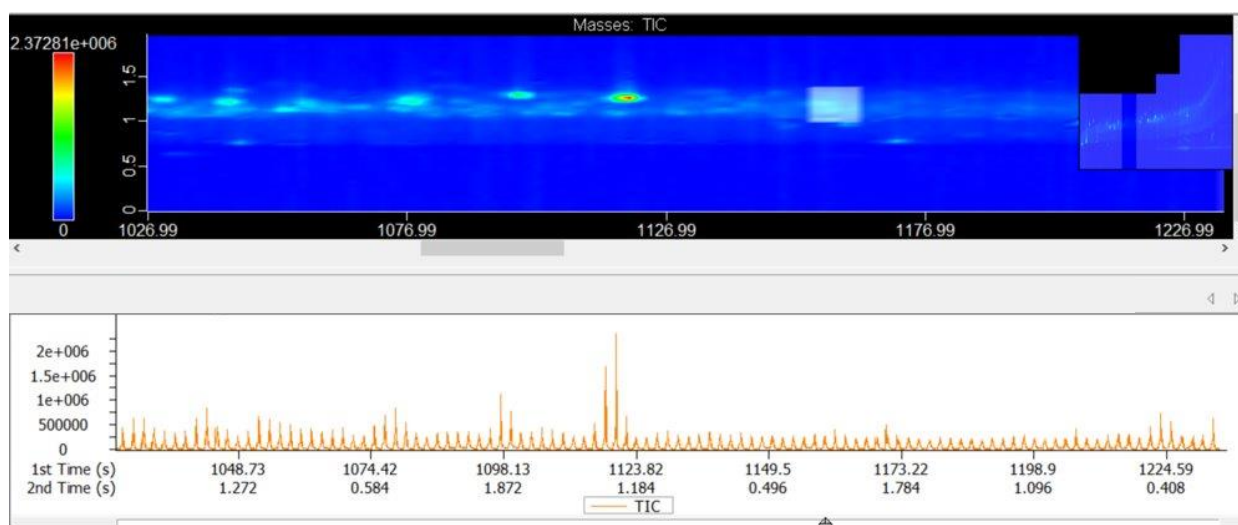


Figure B-128 The real GCxGC image of Sample 24 that belongs to cluster 7



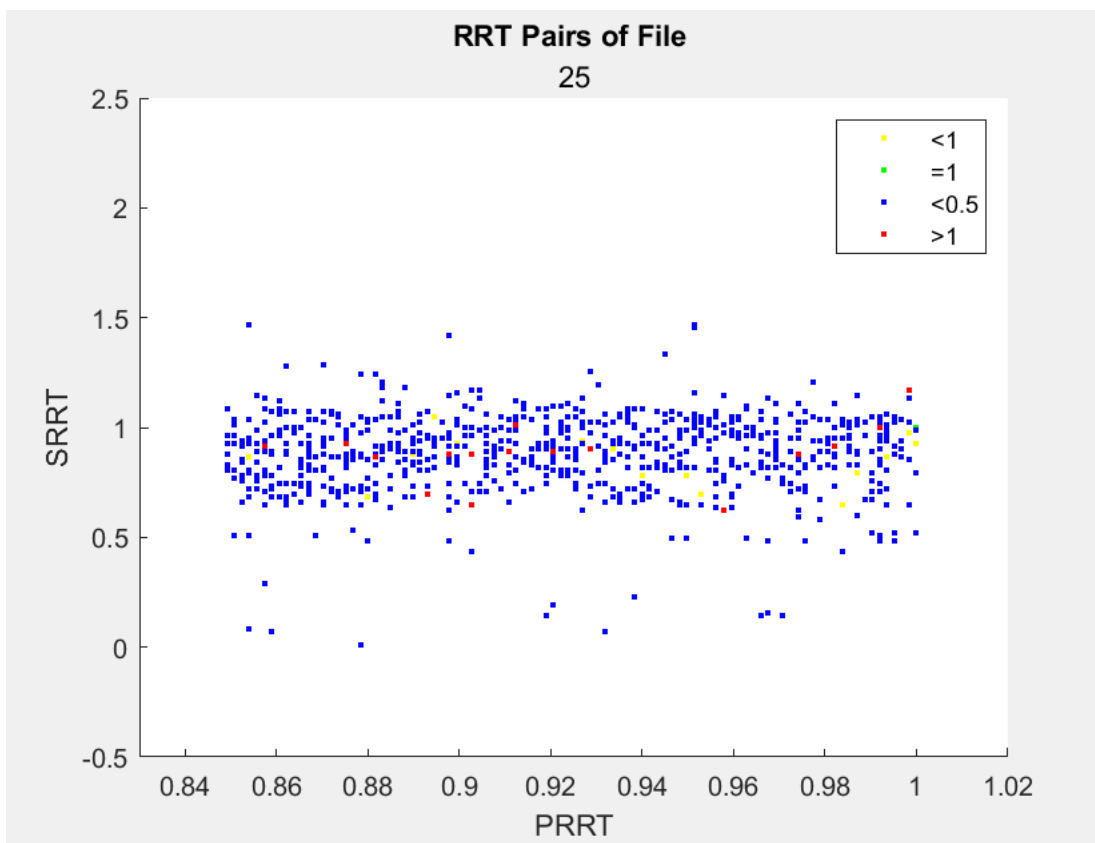


Figure B-129 The discrete GCxGC image of Sample 25 that belongs to cluster 7

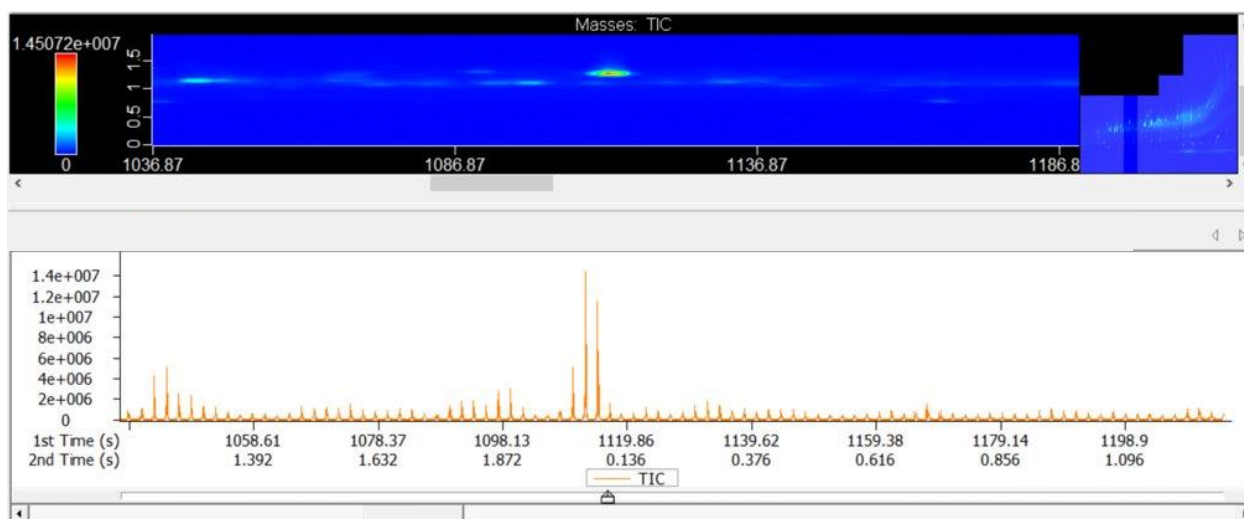


Figure B-130 The real GCxGC image of Sample 42 that belongs to cluster 7

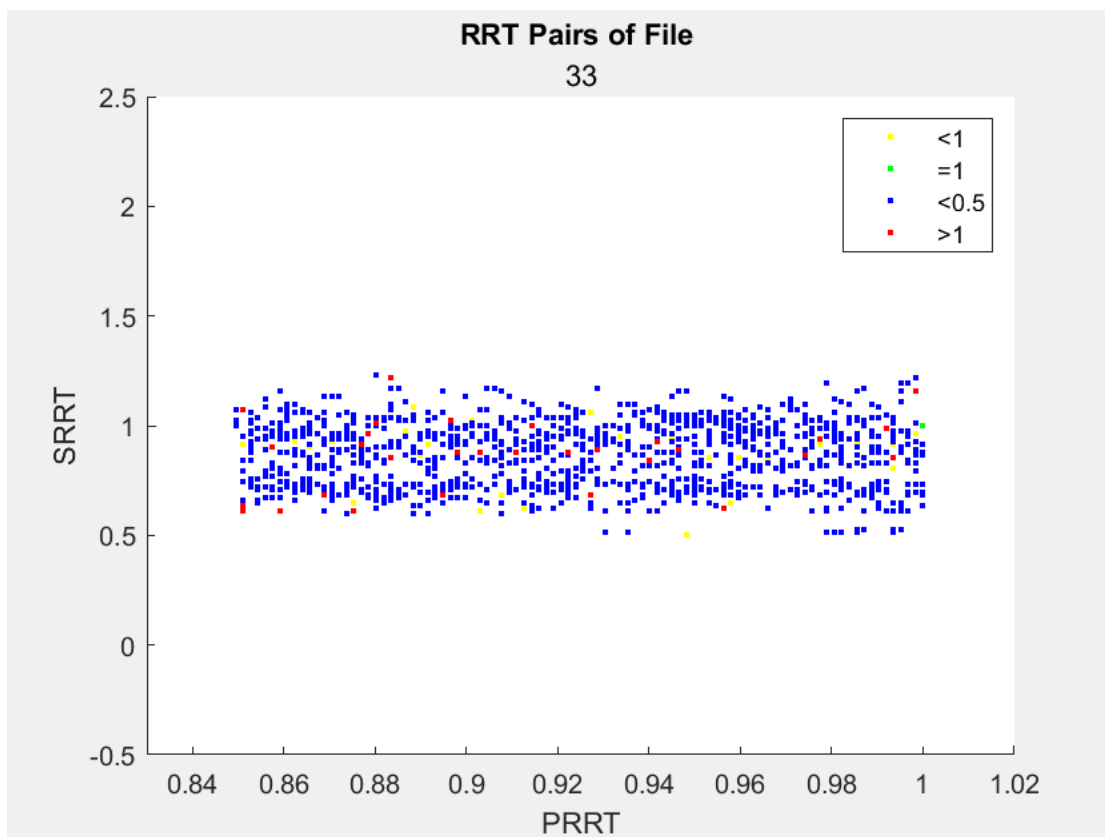


Figure B-131 The discrete GCxGC image of Sample 33 that belongs to cluster 7

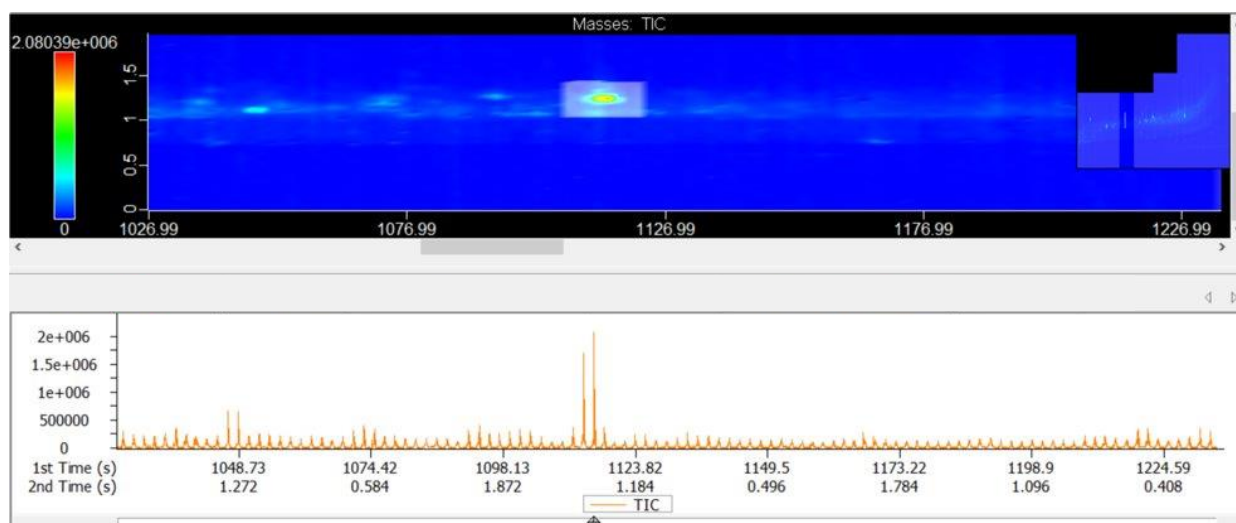


Figure B-132 The real GCxGC image of Sample 33 that belongs to cluster 7

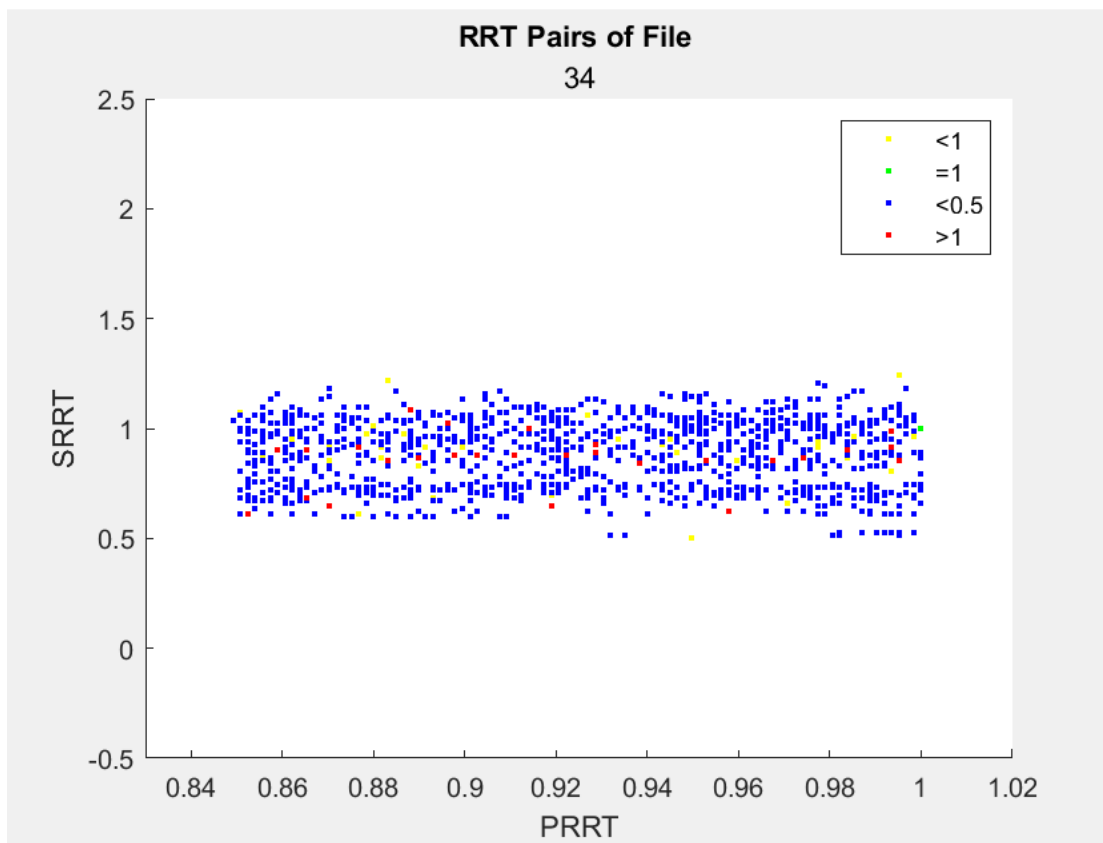


Figure B-133 The discrete GCxGC image of Sample 34 that belongs to cluster 7

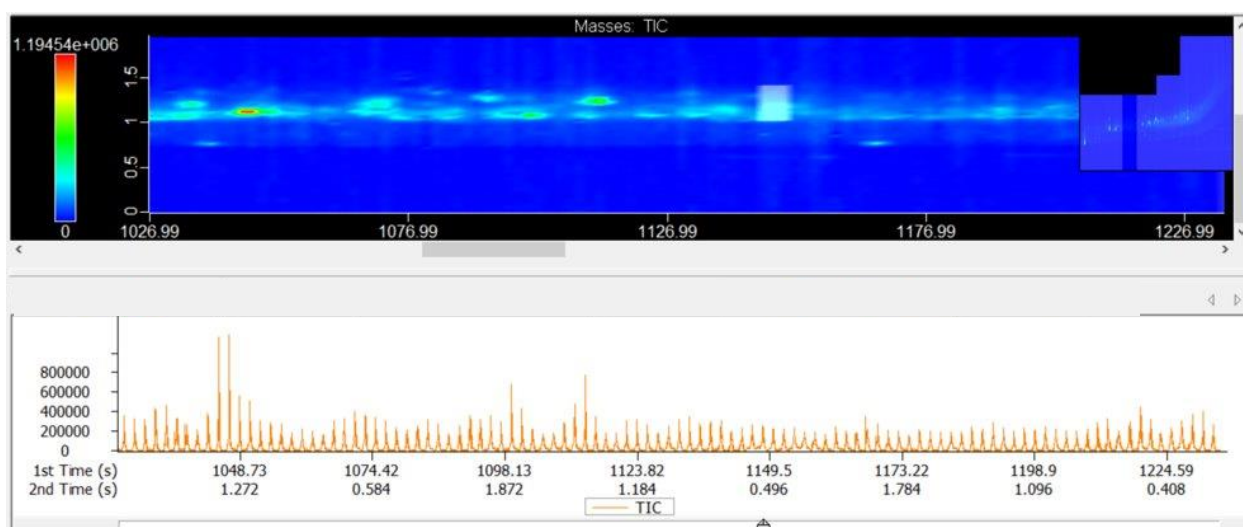


Figure B-134 The real GCxGC image of Sample 34 that belongs to cluster 7

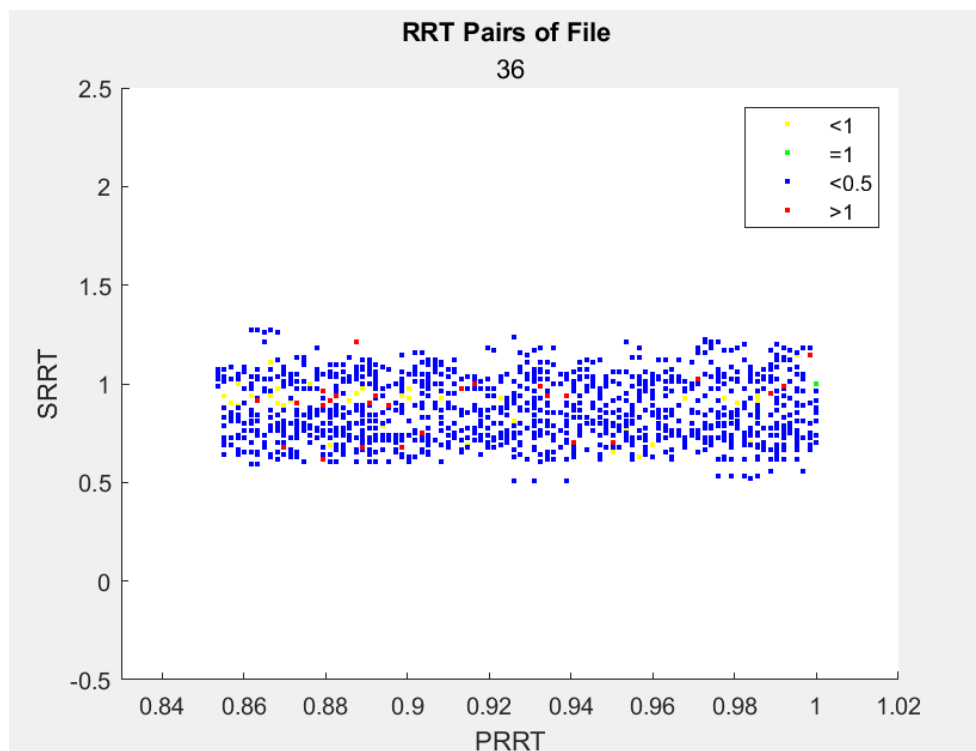


Figure B-135 The discrete GCxGC image of Sample 36 that belongs to cluster 7

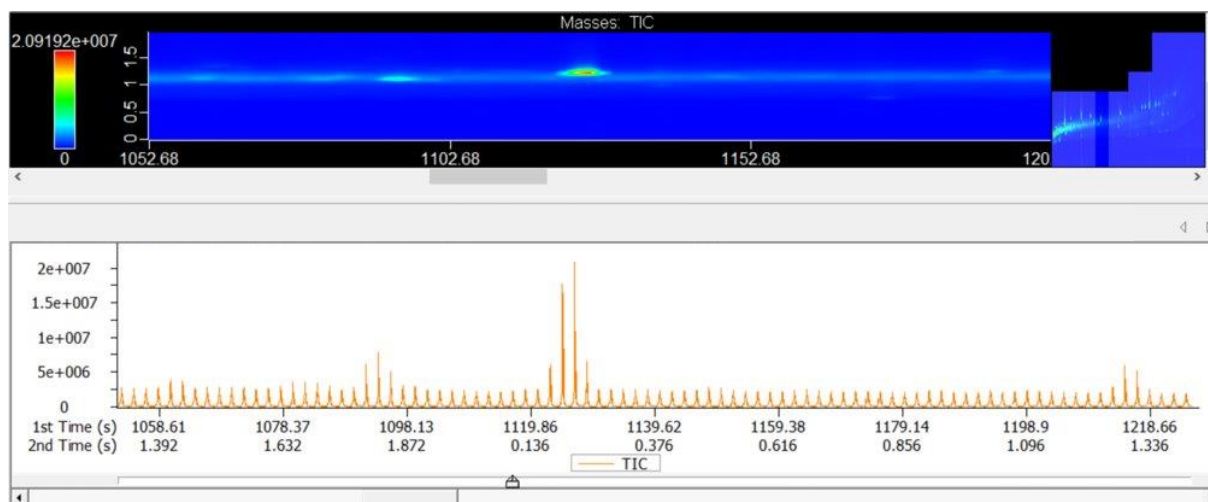


Figure B-136 The real GCxGC image of Sample 36 that belongs to cluster 7

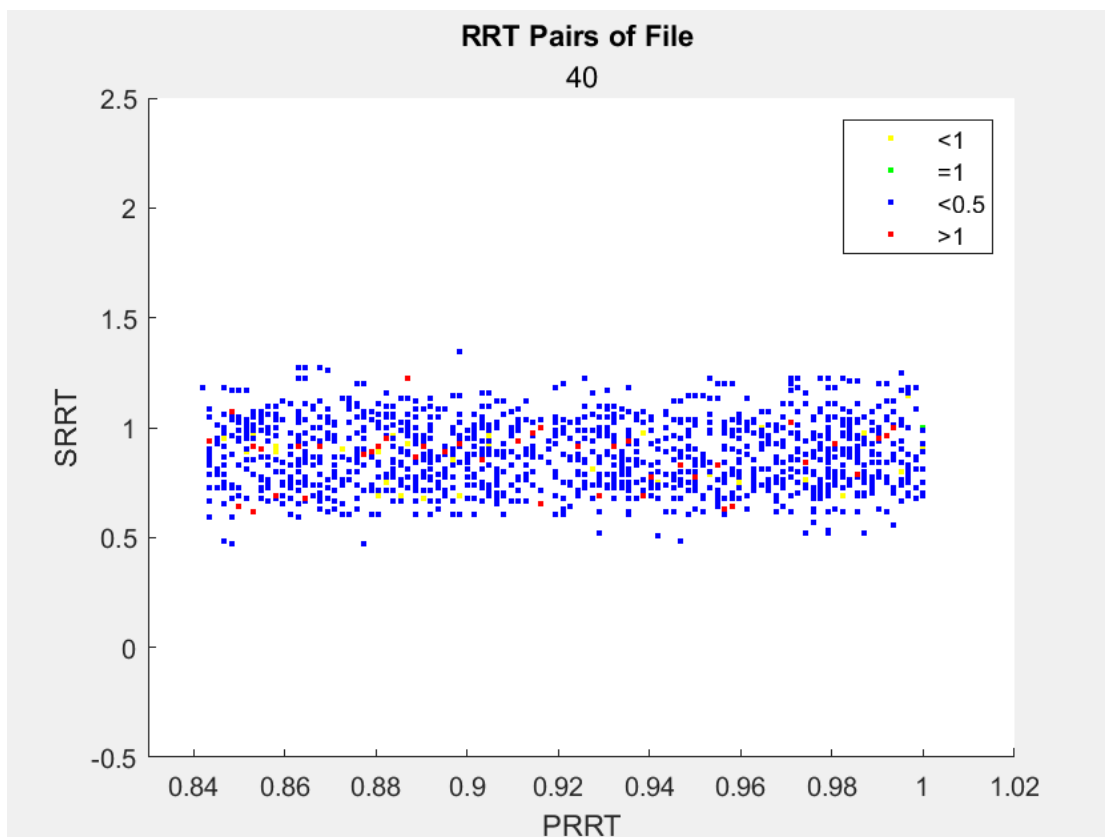


Figure B-137 The discrete GCxGC image of Sample 40 that belongs to cluster 7

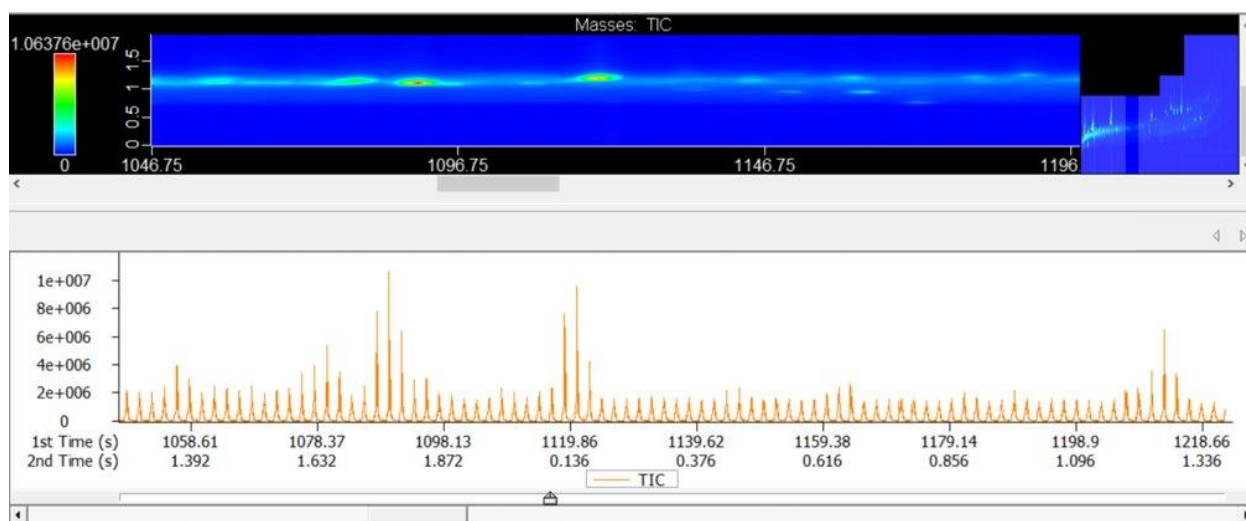


Figure B-138 The real GCxGC image of Sample 40 that belongs to cluster 7

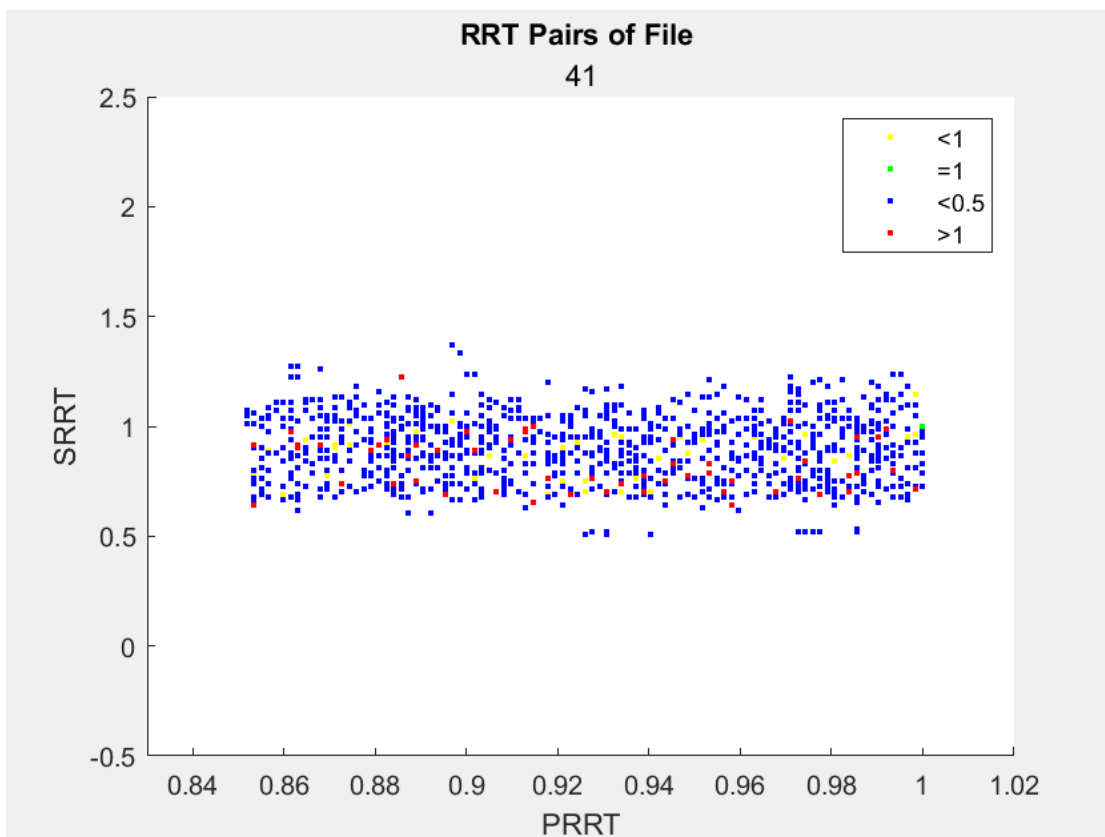


Figure B-139 The discrete GCxGC image of Sample 41 that belongs to cluster 7

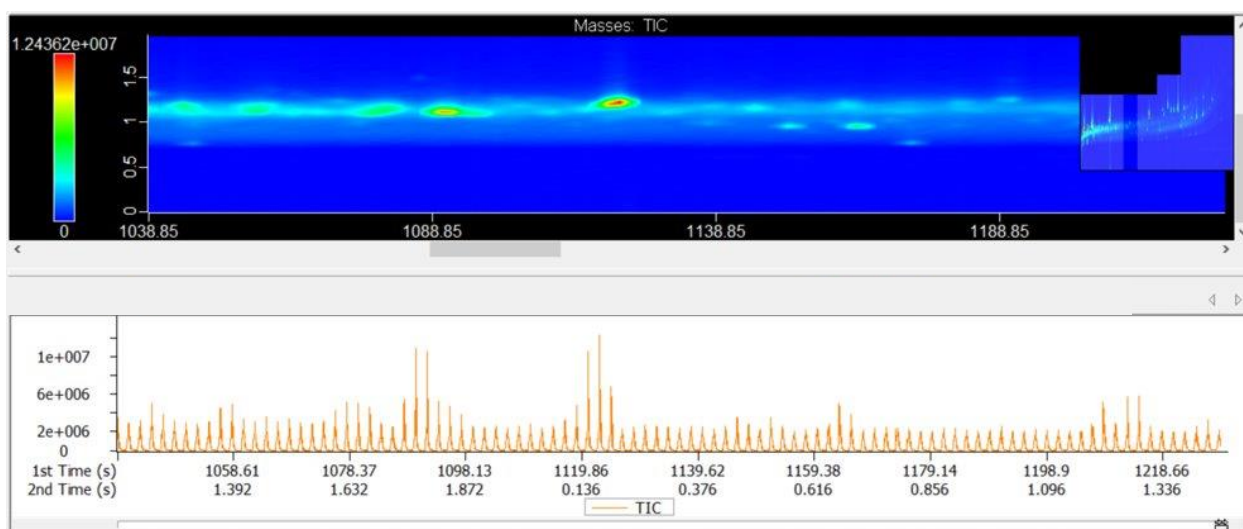


Figure B-140 The real GCxGC image of Sample 41 that belongs to cluster 7

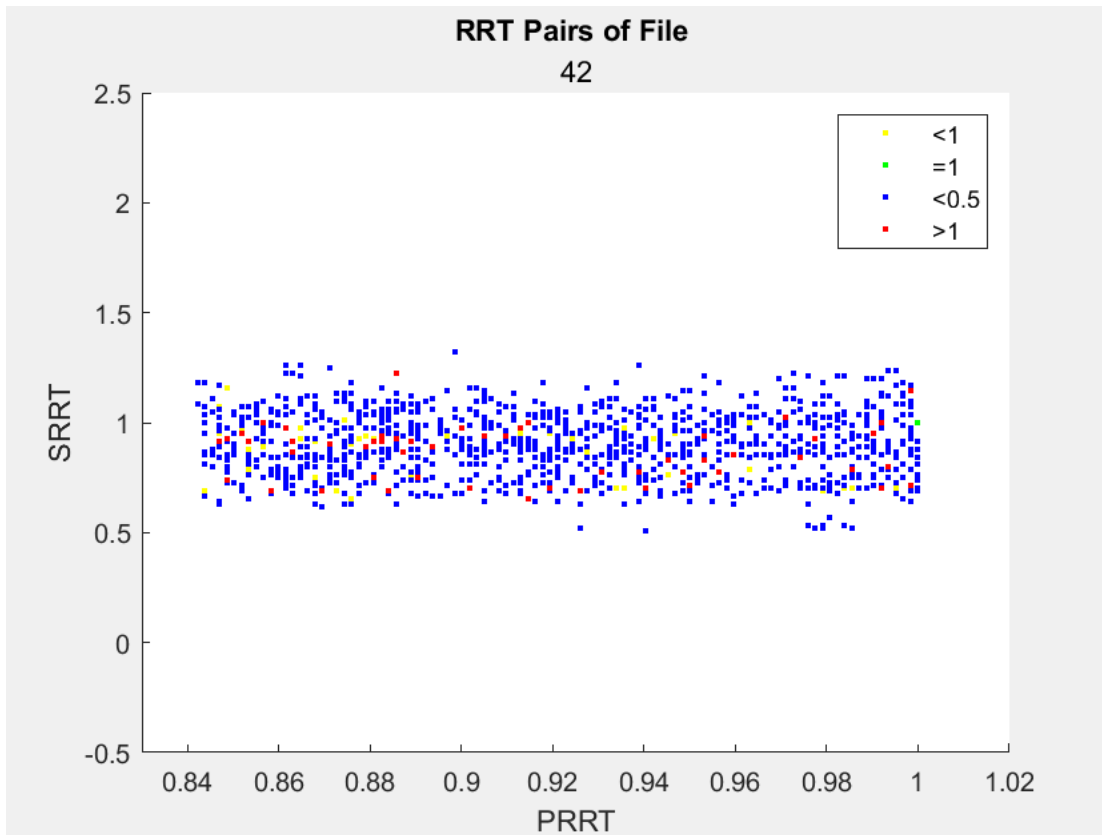


Figure B-141 The discrete GCxGC image of Sample 42 that belongs to cluster 7

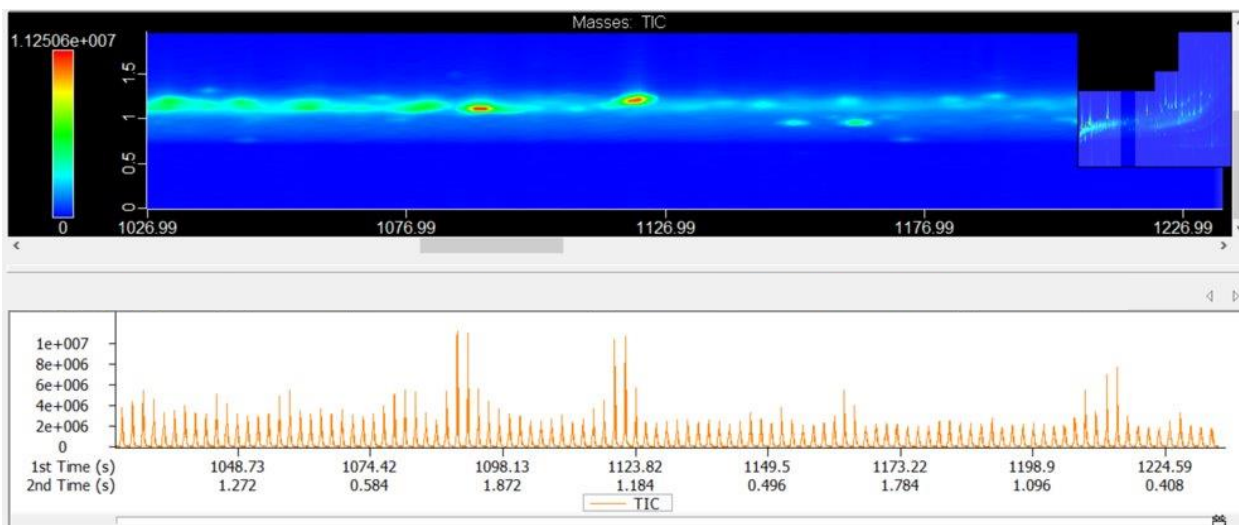


Figure B-142 The real GCxGC image of Sample 42 that belongs to cluster 7

## APPENDIX C: SCORES OF PRINCIPAL COMPONENTS

Figure C-1 Principal component 2 scores .....	158
Figure C-2 Principal component 3 scores .....	158
Figure C-3 Principal component 4 scores .....	158
Figure C-4 Principal component 5 scores .....	158
Figure C-5 Principal component 6 scores .....	159
Figure C-6 Principal component 7 scores .....	159
Figure C-7 Principal component 8 scores .....	159
Figure C-8 Principal component 9 scores .....	159
Figure C-9 Principal component 10 scores .....	159
Figure C-10 Principal component 11 scores .....	159
Figure C-11 Principal component 12 scores .....	160
Figure C-12 Principal component 13 scores .....	160
Figure C-13 Principal component 14 scores .....	160
Figure C-14 Principal component 15 scores .....	160
Figure C-15 Principal component 16 scores .....	160
Figure C-16 Principal component 17 scores .....	160
Figure C-17 Principal component 18 scores .....	161
Figure C-18 Principal component 19 scores .....	161
Figure C-19 Principal component 20 scores .....	161
Figure C-20 Principal component 21 scores .....	161
Figure C-21 Principal component 22 scores .....	161
Figure C-22 Principal component 23 scores .....	161
Figure C-23 Principal component 24 scores .....	162
Figure C-24 Principal component 25 scores .....	162



Figure C-25 Principal component 26 scores .....	162
Figure C-26 Principal component 27 scores .....	162
Figure C-27 Principal component 28 scores .....	162
Figure C-28 Principal component 29 scores .....	162
Figure C-29 Principal component 30 scores .....	163
Figure C-30 Principal component 31 scores .....	163
Figure C-31 Principal component 32 scores .....	163
Figure C-32 Principal component 33 scores .....	163
Figure C-33 Principal component 34 scores .....	163
Figure C-34 Principal component 35 scores .....	163
Figure C-35 Principal component 36 scores .....	164
Figure C-36 Principal component 37 scores .....	164
Figure C-37 Principal component 38 scores .....	164
Figure C-38 Principal component 39 scores .....	164
Figure C-39 Principal component 40 scores .....	164
Figure C-40 Principal component 41 scores .....	164
Figure C-41 Principal component 42 scores .....	165
Figure C-42 Principal component 43 scores .....	165
Figure C-43 Principal component 44 scores .....	165
Figure C-44 Principal component 45 scores .....	165
Figure C-45 Principal component 46 scores .....	165
Figure C-46 Principal component 47 scores .....	165
Figure C-47 Principal component 48 scores .....	166
Figure C-48 Principal component 49 scores .....	166
Figure C-49 Principal component 50 scores .....	166

Figure C-50 Principal component 51 scores .....	166
Figure C-51 Principal component 52 scores .....	166
Figure C-52 Principal component 53 scores .....	166
Figure C-53 Principal component 54 scores .....	167
Figure C-54 Principal component 55 scores .....	167
Figure C-55 Principal component 56 scores .....	167
Figure C-56 Principal component 57 scores .....	167
Figure C-57 Principal component 58 scores .....	167
Figure C-58 Principal component 59 scores .....	167
Figure C-59 Principal component 60 scores .....	168
Figure C-60 Principal component 61 scores .....	168
Figure C-61 Principal component 62 scores .....	168
Figure C-62 Principal component 63 scores .....	168
Figure C-63 Principal component 64 scores .....	168
Figure C-64 Principal component 65 scores .....	168
Figure C-65 Principal component 66 scores .....	169
Figure C-66 Principal component 67 scores .....	169
Figure C-67 Principal component 68 scores .....	169
Figure C-68 Principal component 69 scores .....	169
Figure C-69 Principal component 70 scores .....	169
Figure C-70 Principal component 71 scores .....	169

Each of the principal component scores are graphed below. Each sample number has a score. The x-axis corresponds to the sample number and the y-axis is the  $L_2$  normalization score.

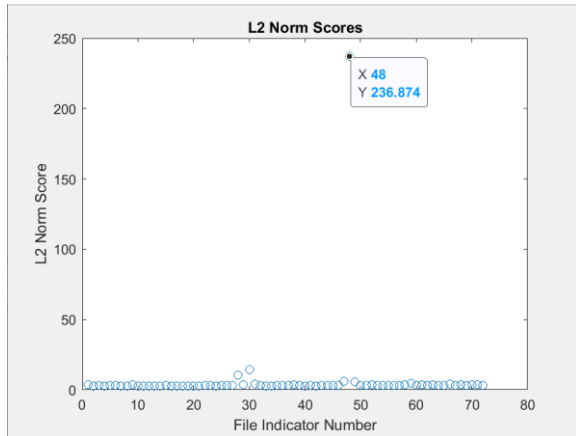


Figure C-1 Principal component 2 scores

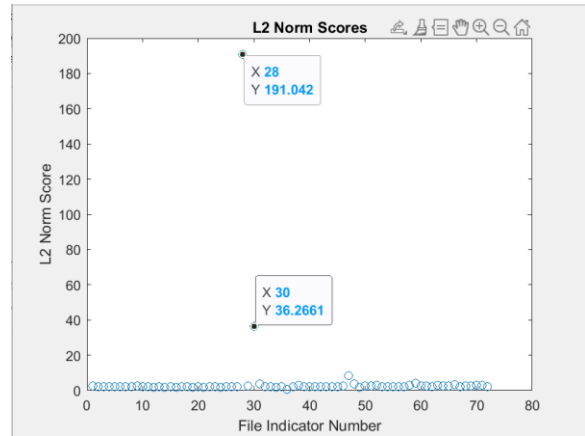


Figure C-3 Principal component 4 scores

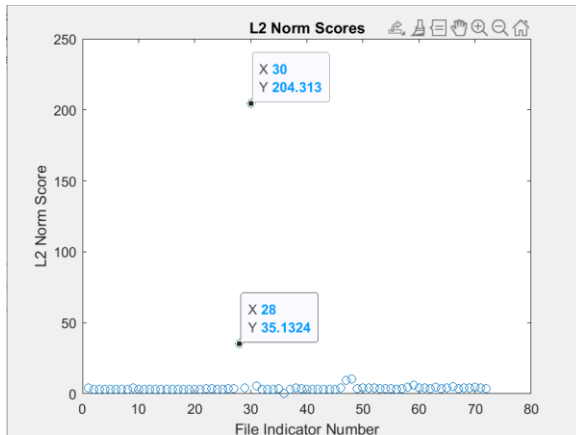


Figure C-2 Principal component 3 scores

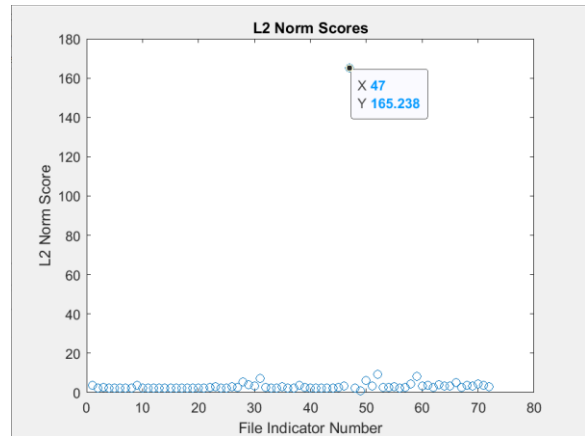


Figure C-4 Principal component 5 scores

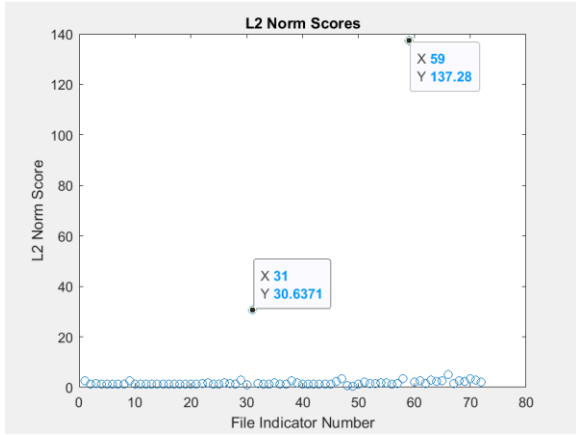


Figure C-5 Principal component 6 scores

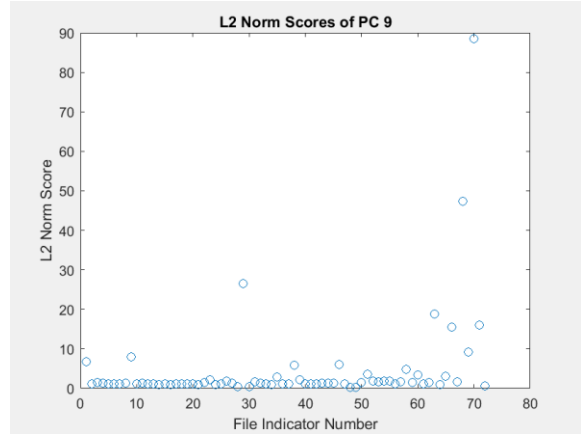


Figure C-8 Principal component 9 scores

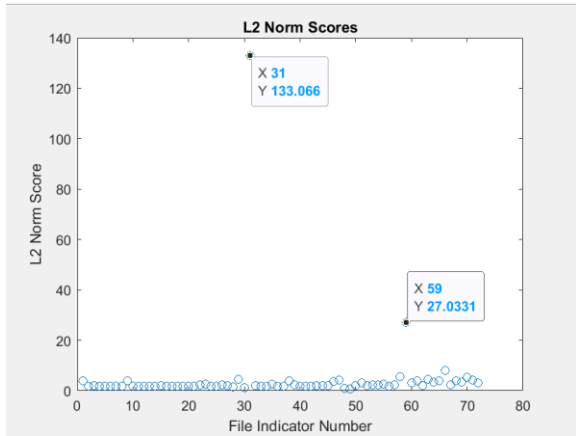


Figure C-6 Principal component 7 scores

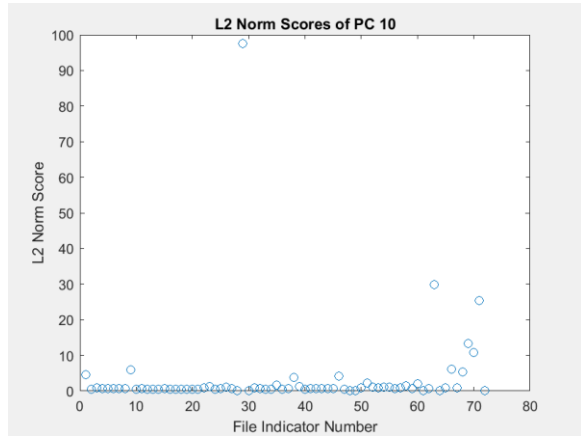


Figure C-9 Principal component 10 scores

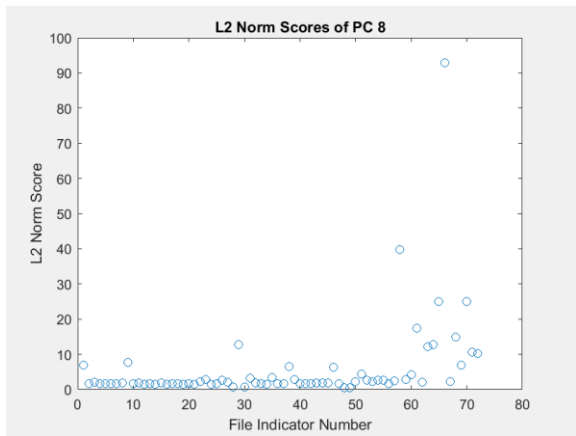


Figure C-7 Principal component 8 scores

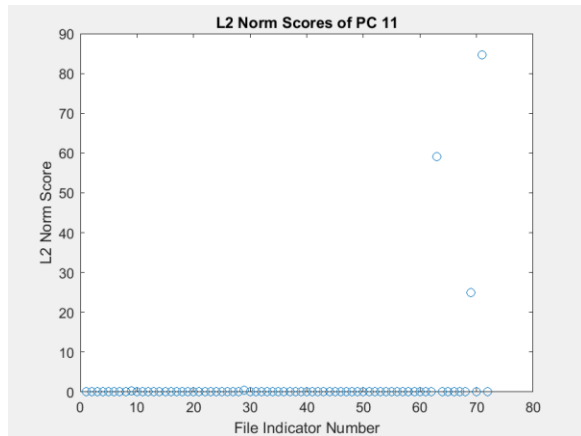


Figure C-10 Principal component 11 scores

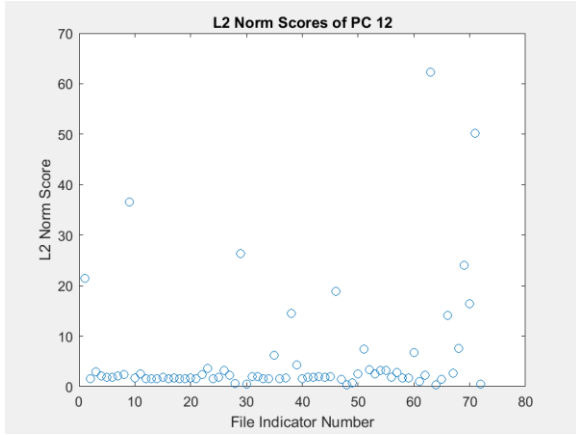


Figure C-11 Principal component 12 scores

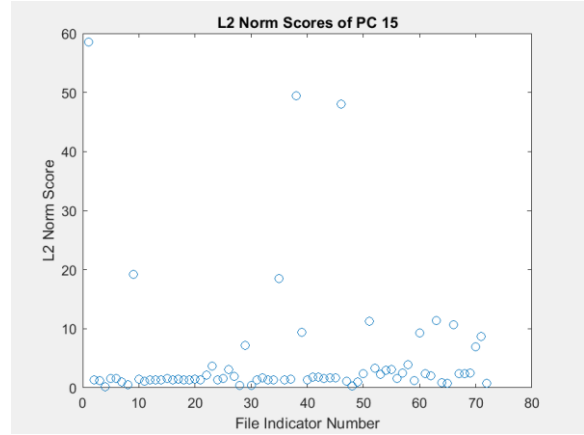


Figure C-14 Principal component 15 scores

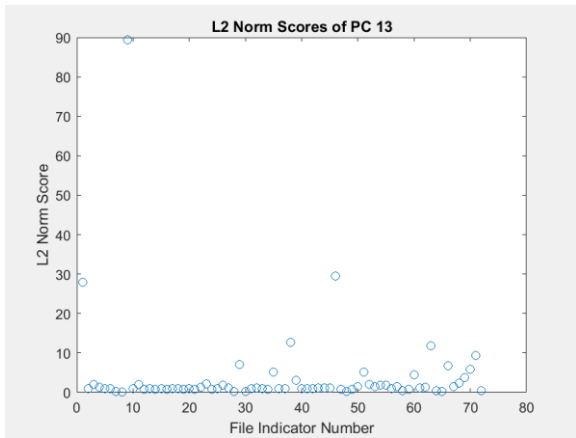


Figure C-12 Principal component 13 scores

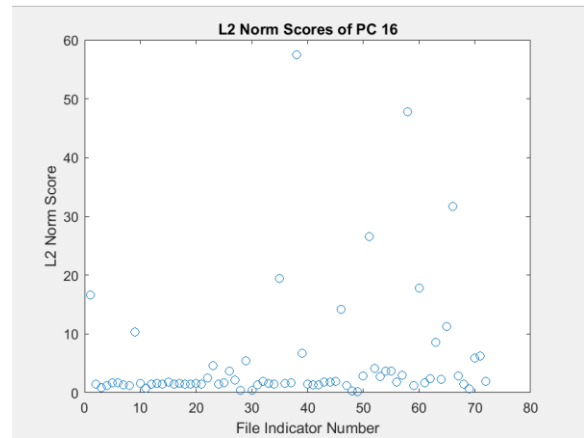


Figure C-15 Principal component 16 scores

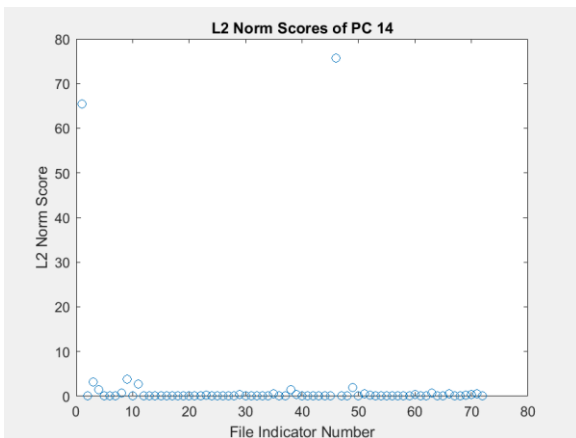


Figure C-13 Principal component 14 scores

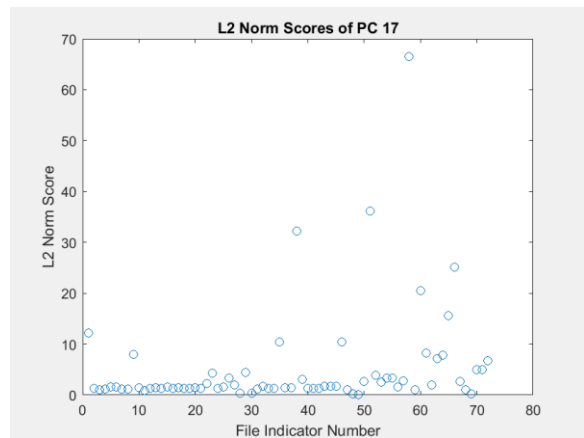


Figure C-16 Principal component 17 scores

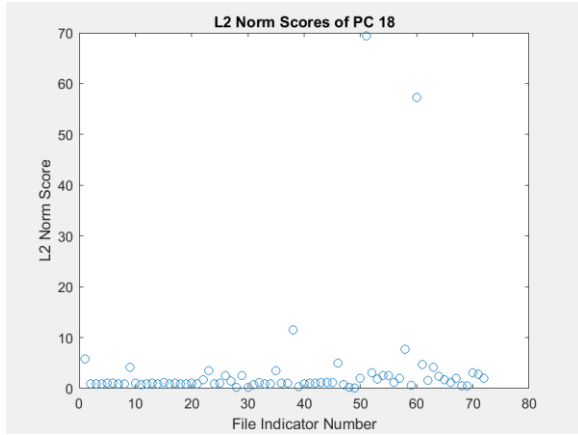


Figure C-17 Principal component 18 scores

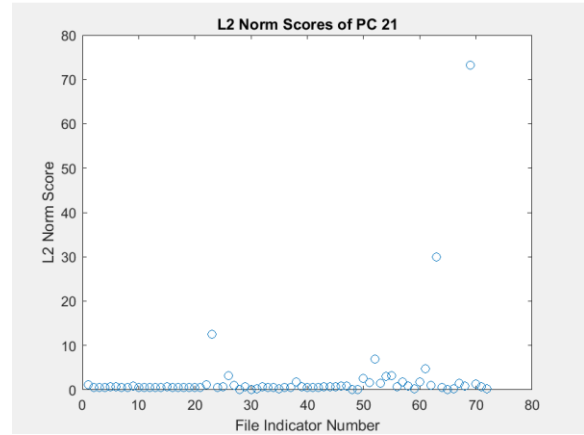


Figure C-20 Principal component 21 scores

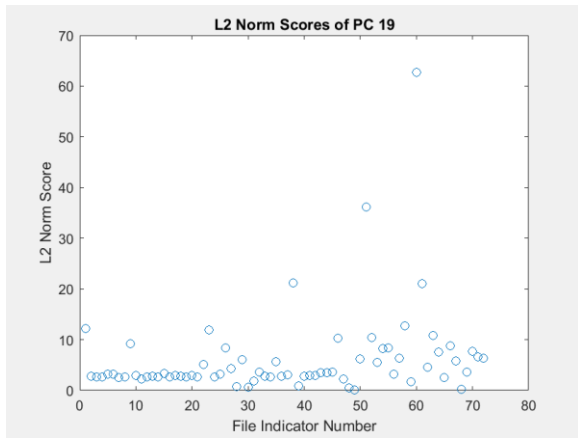


Figure C-18 Principal component 19 scores

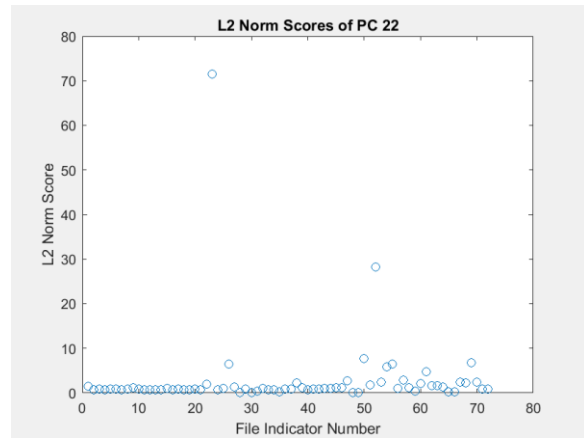


Figure C-21 Principal component 22 scores

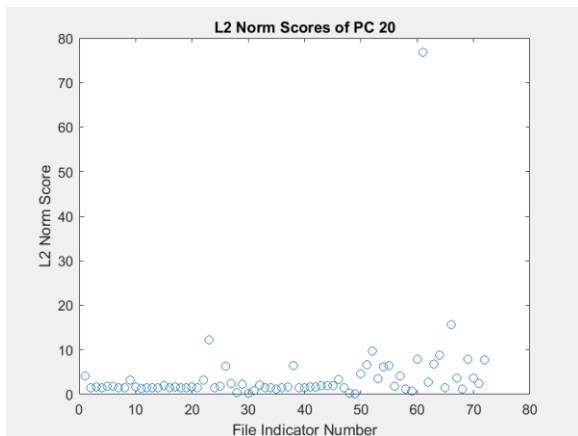


Figure C-19 Principal component 20 scores

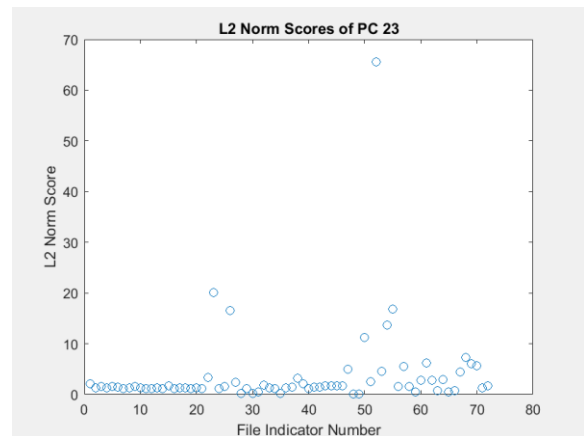


Figure C-22 Principal component 23 scores

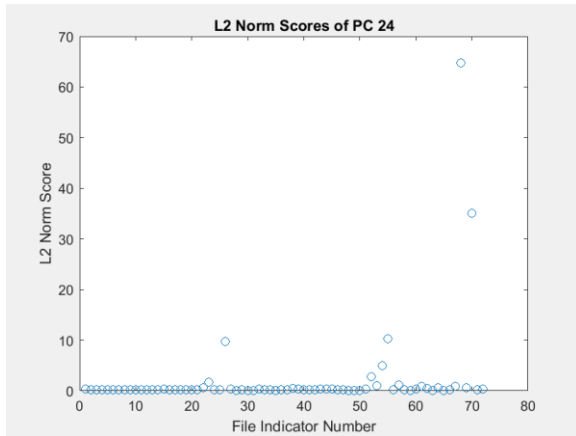


Figure C-23 Principal component 24 scores

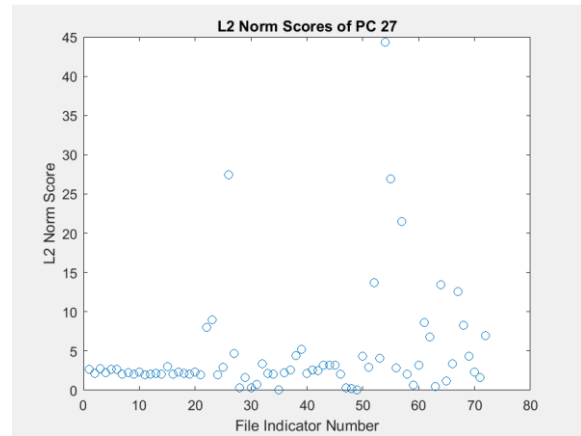


Figure C-26 Principal component 27 scores

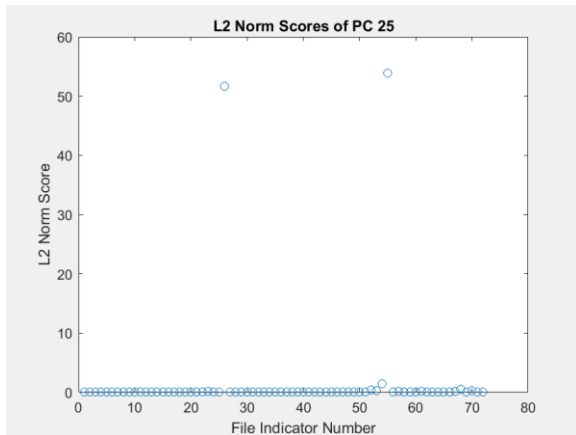


Figure C-24 Principal component 25 scores

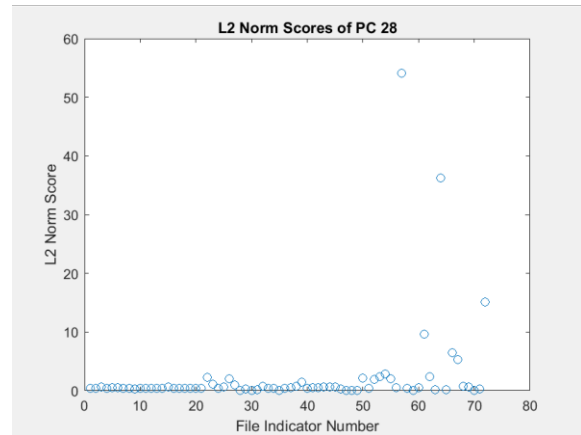


Figure C-27 Principal component 28 scores

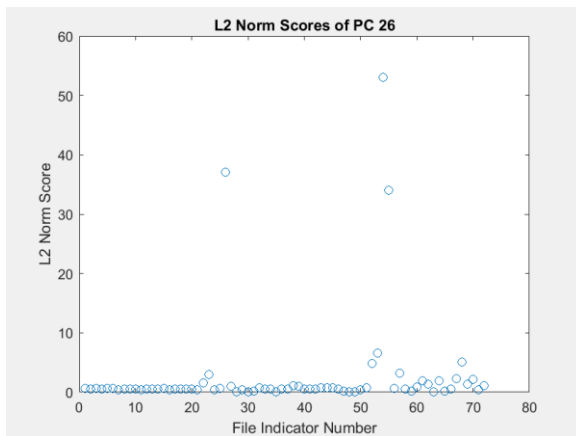


Figure C-25 Principal component 26 scores

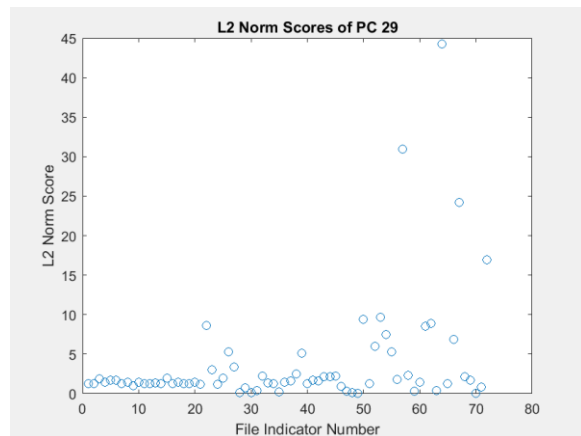


Figure C-28 Principal component 29 scores

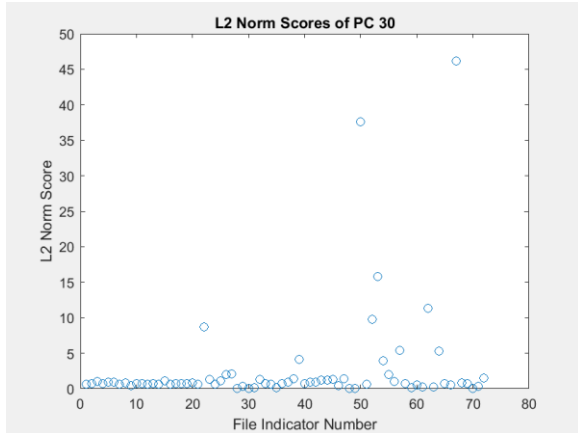


Figure C-29 Principal component 30 scores

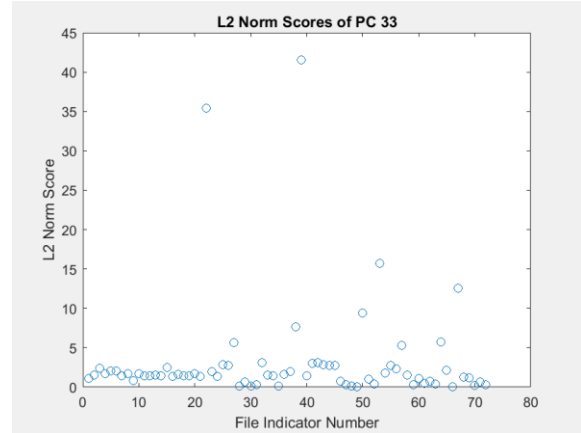


Figure C-32 Principal component 33 scores

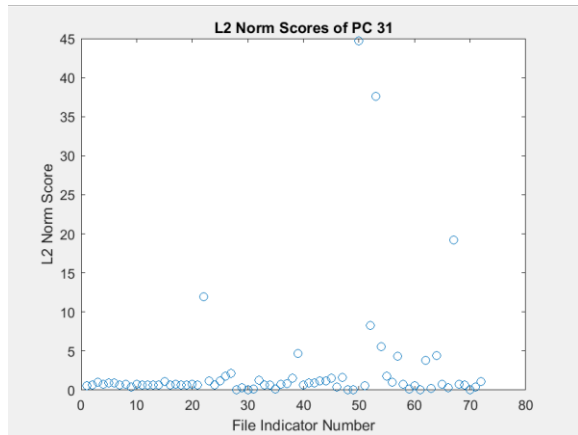


Figure C-30 Principal component 31 scores

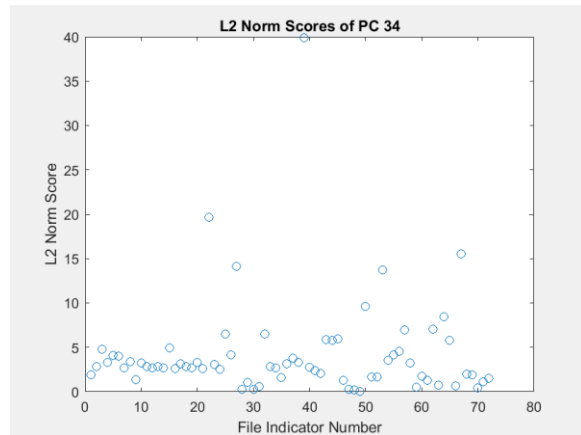


Figure C-33 Principal component 34 scores

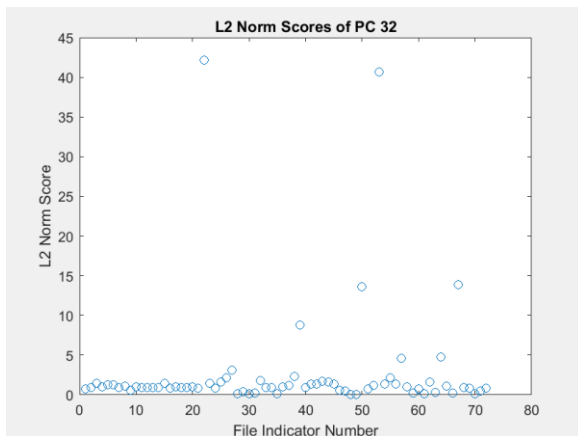


Figure C-31 Principal component 32 scores

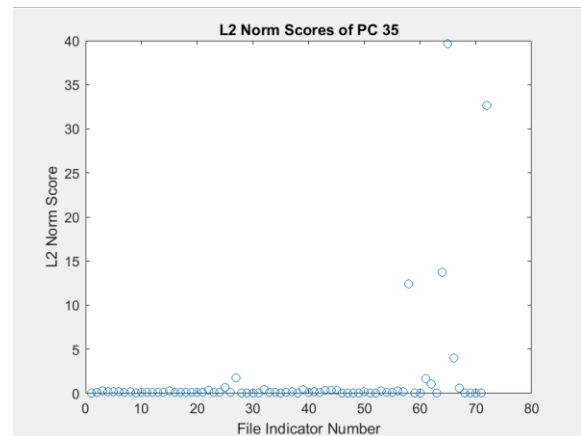


Figure C-34 Principal component 35 scores



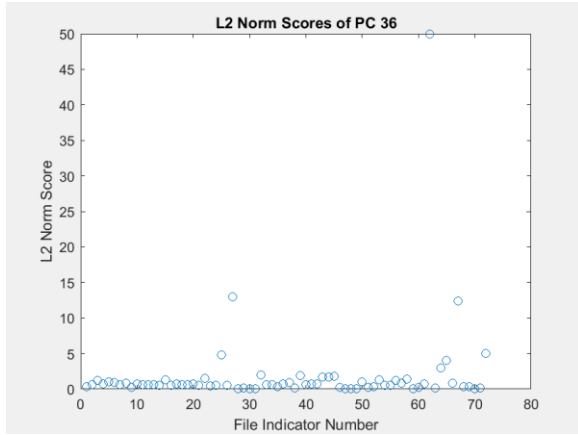


Figure C-35 Principal component 36 scores

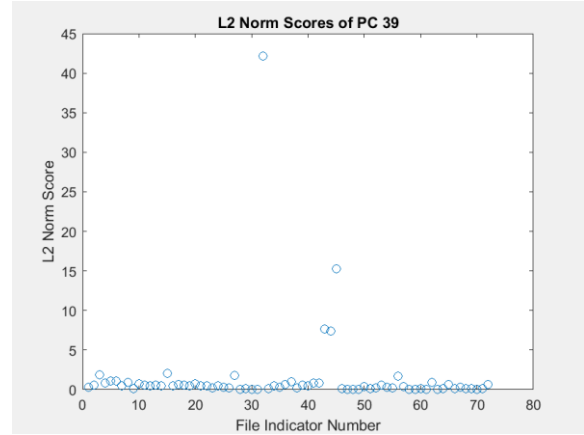


Figure C-38 Principal component 39 scores

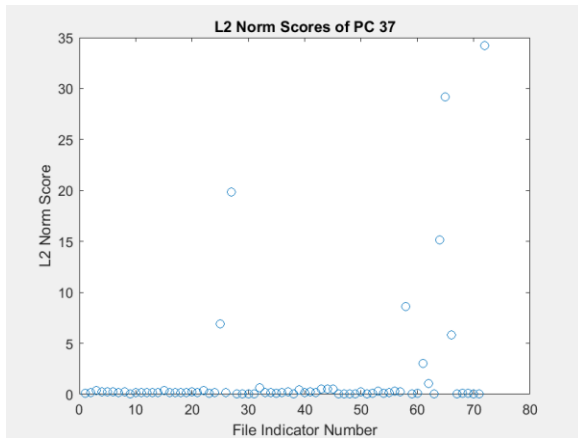


Figure C-36 Principal component 37 scores

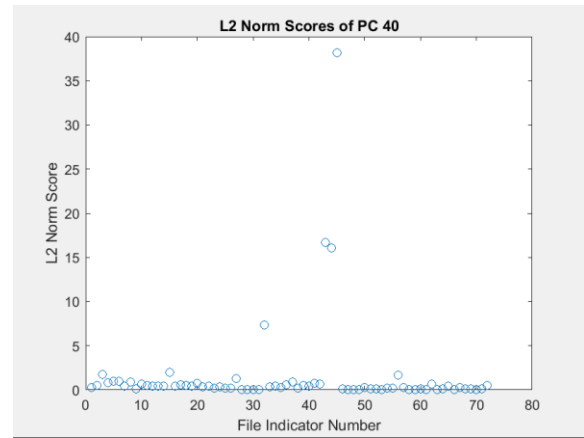


Figure C-39 Principal component 40 scores

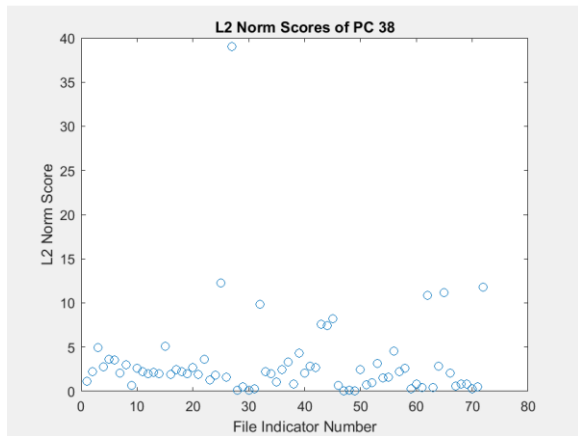


Figure C-37 Principal component 38 scores

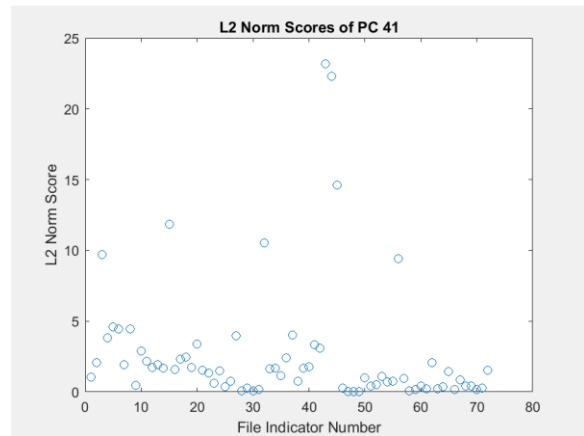


Figure C-40 Principal component 41 scores

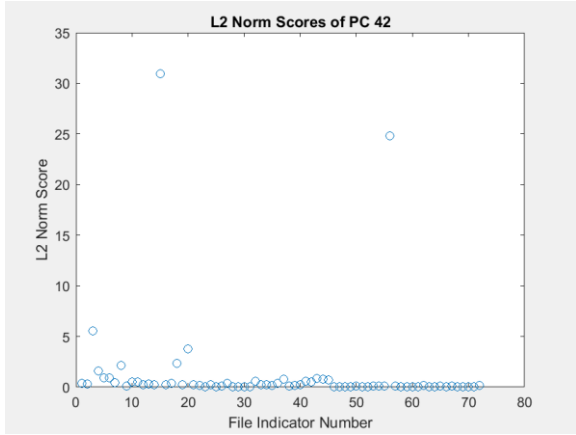


Figure C-41 Principal component 42 scores

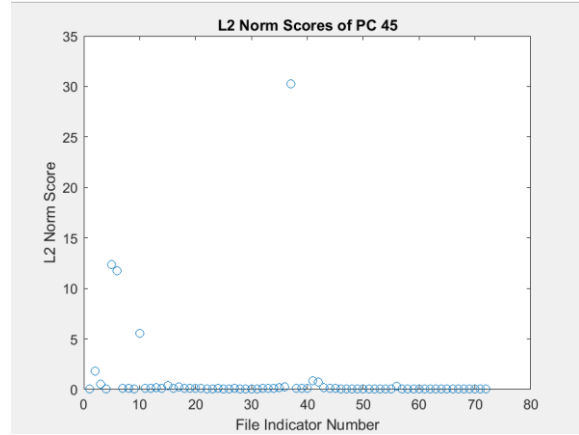


Figure C-44 Principal component 45 scores

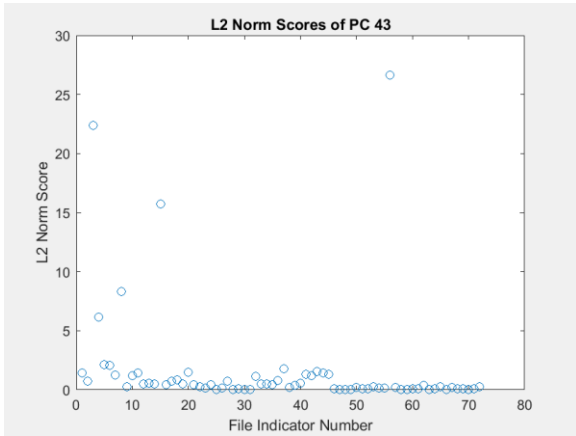


Figure C-42 Principal component 43 scores

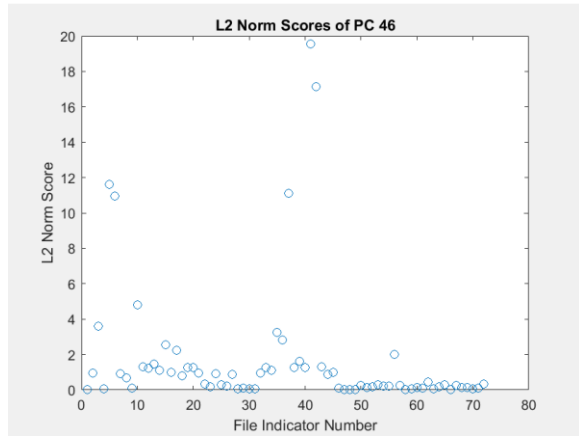


Figure C-45 Principal component 46 scores

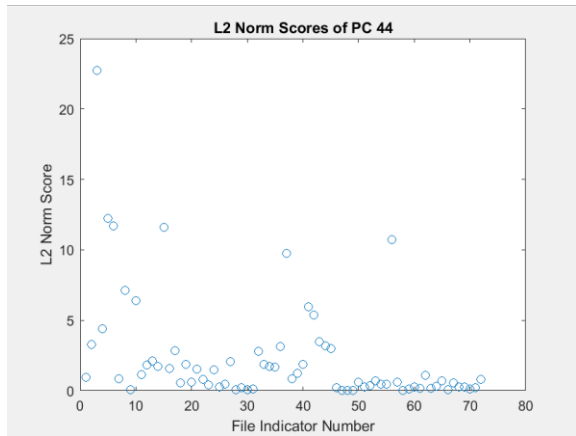


Figure C-43 Principal component 44 scores

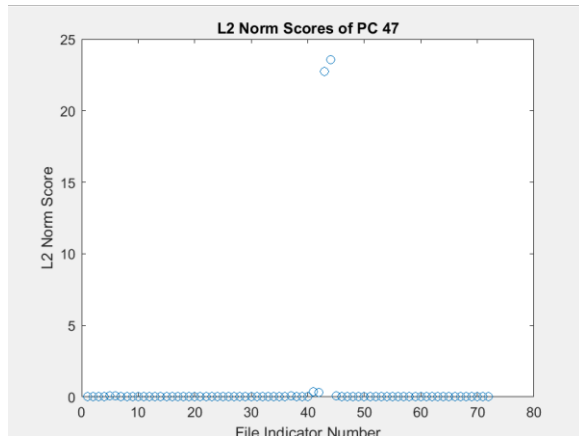


Figure C-46 Principal component 47 scores

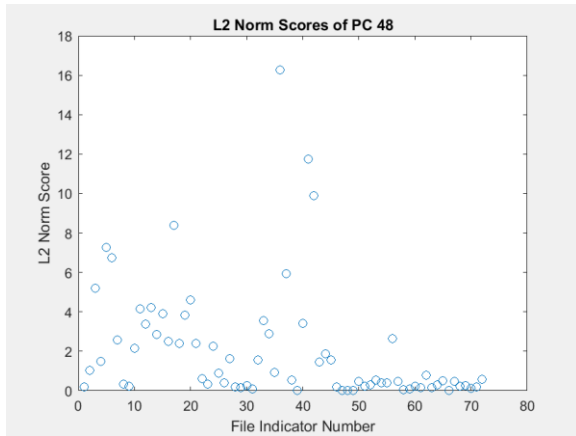


Figure C-47 Principal component 48 scores

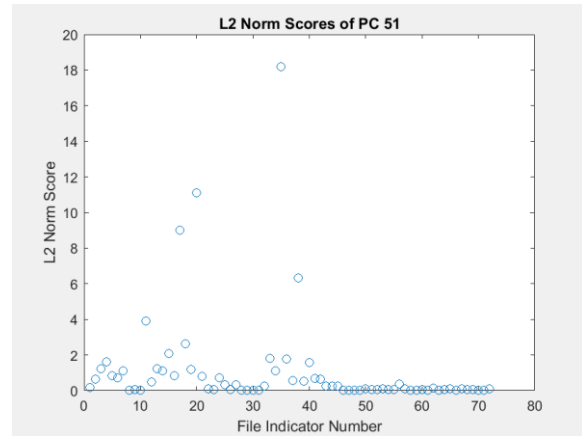


Figure C-50 Principal component 51 scores

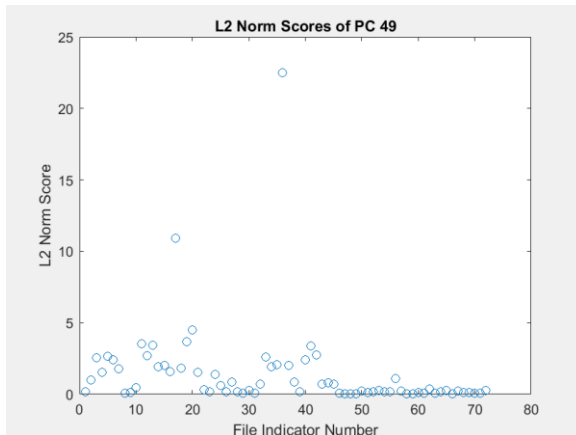


Figure C-48 Principal component 49 scores

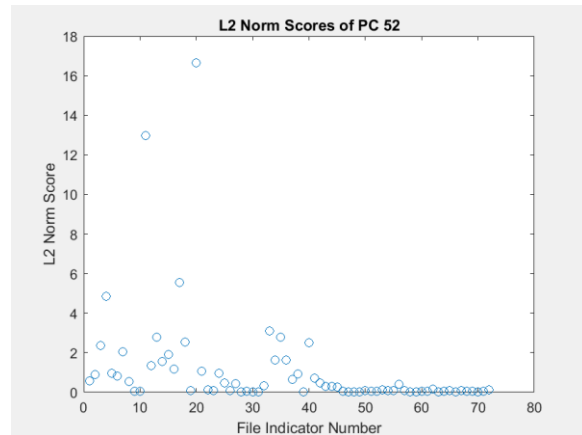


Figure C-51 Principal component 52 scores

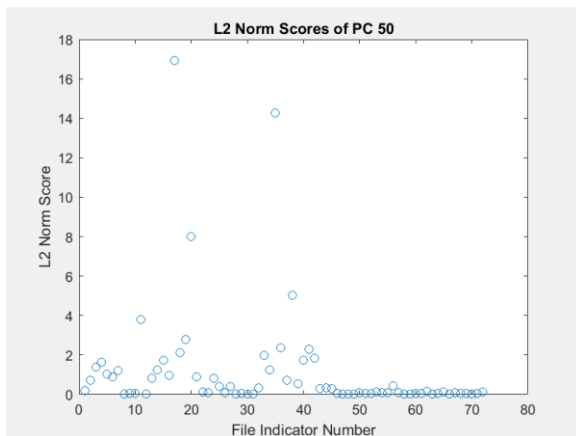


Figure C-49 Principal component 50 scores

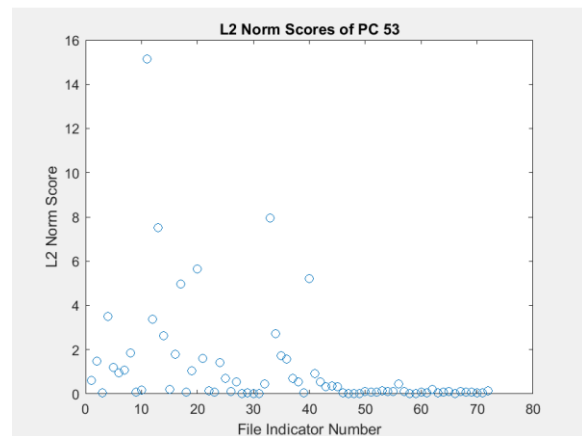


Figure C-52 Principal component 53 scores

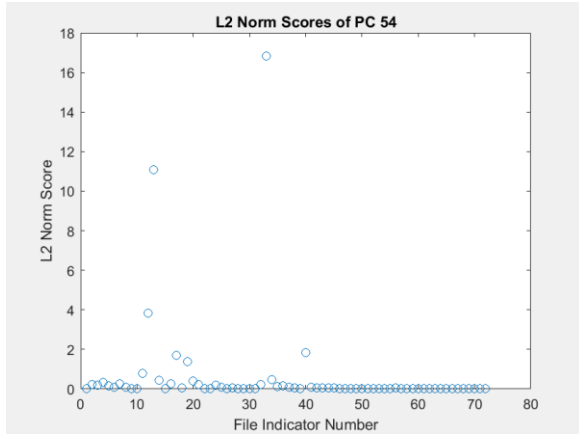


Figure C-53 Principal component 54 scores

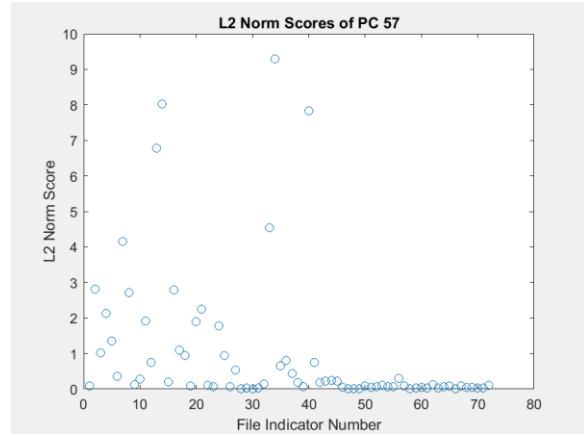


Figure C-56 Principal component 57 scores

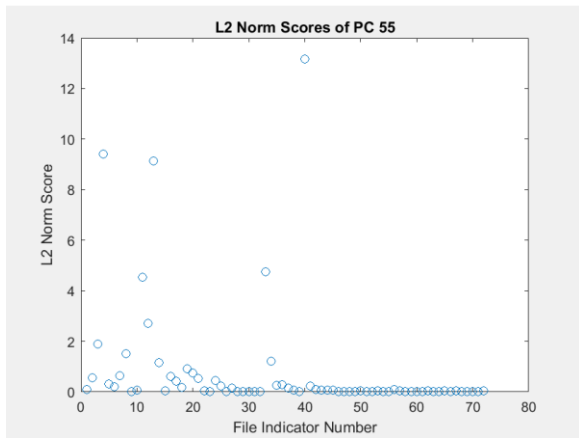


Figure C-54 Principal component 55 scores

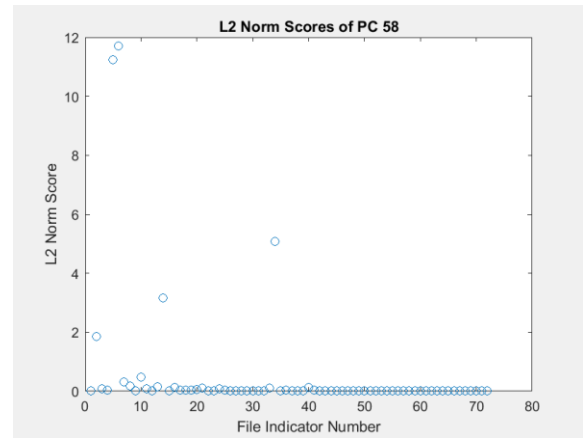


Figure C-57 Principal component 58 scores

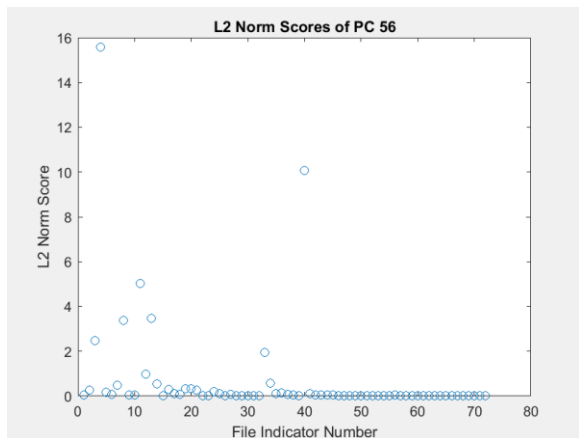


Figure C-55 Principal component 56 scores

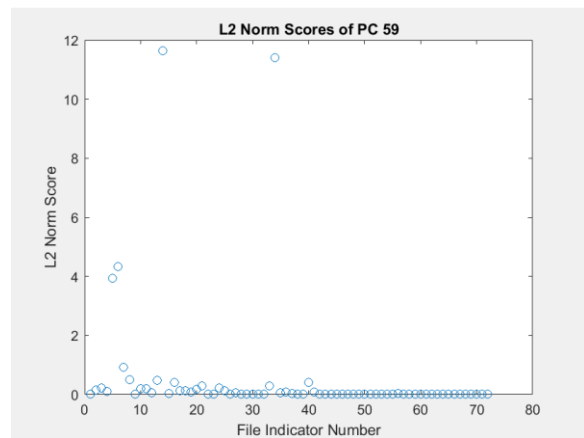


Figure C-58 Principal component 59 scores

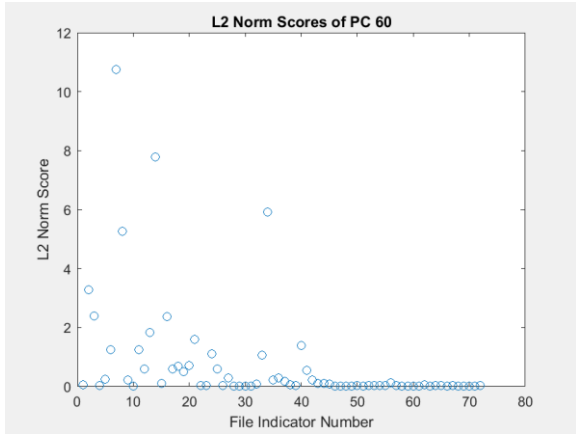


Figure C-59 Principal component 60 scores

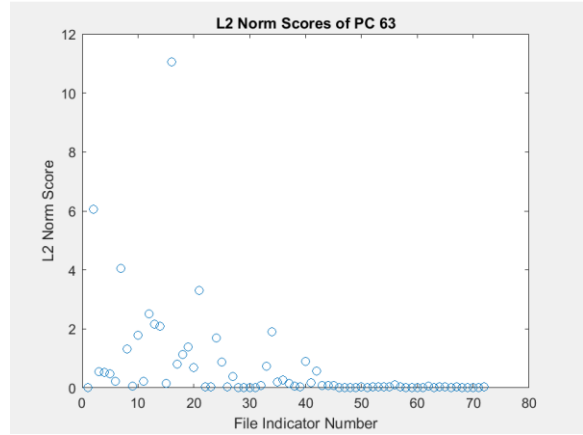


Figure C-62 Principal component 63 scores

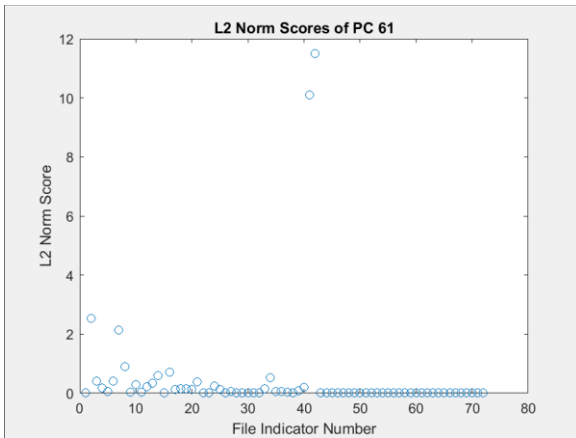


Figure C-60 Principal component 61 scores

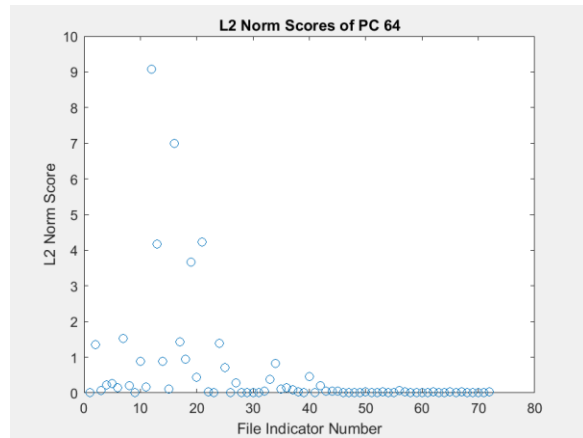


Figure C-63 Principal component 64 scores

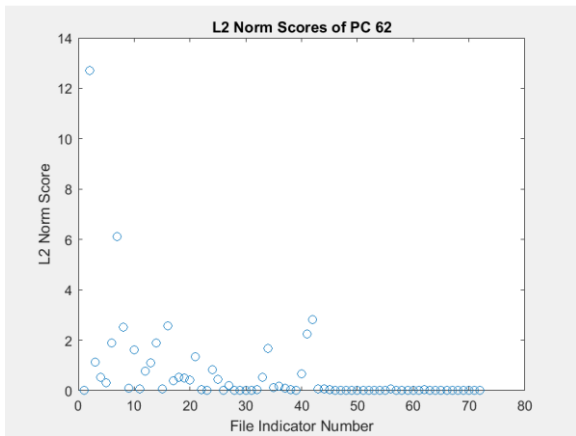


Figure C-61 Principal component 62 scores

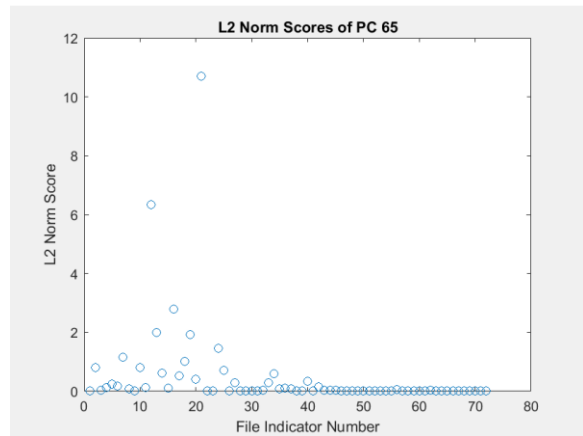


Figure C-64 Principal component 65 scores

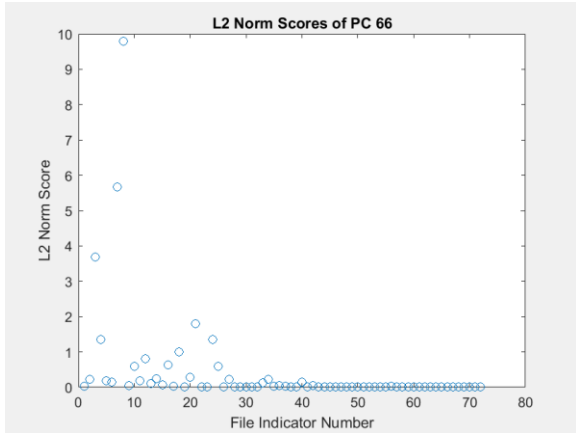


Figure C-65 Principal component 66 scores

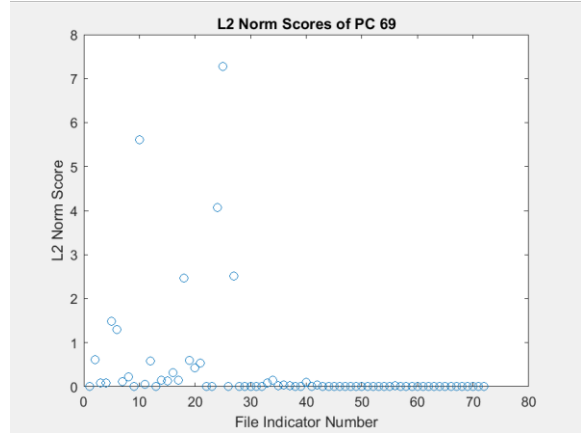


Figure C-68 Principal component 69 scores

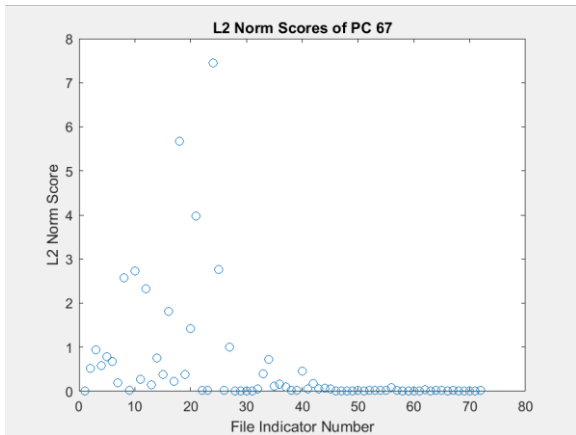


Figure C-66 Principal component 67 scores

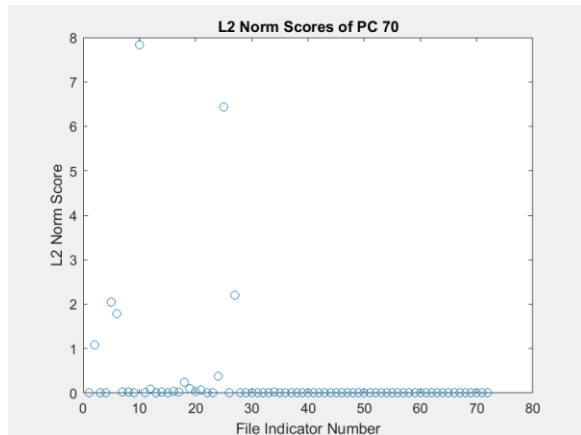


Figure C-69 Principal component 70 scores

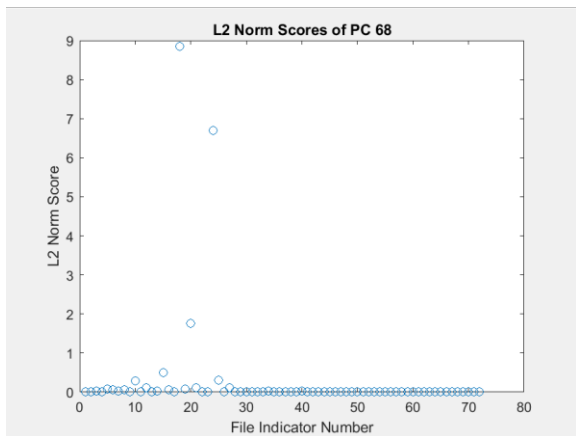


Figure C-67 Principal component 68 scores

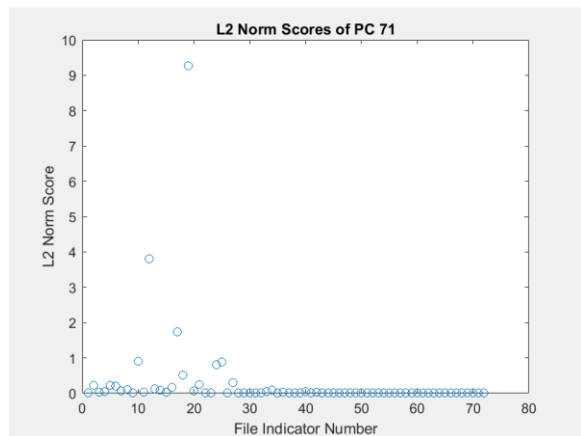


Figure C-70 Principal component 71 scores

APPENDIX D: COMPONENT WEIGHTS

Figure D-1 Weights of each relative retention time location of principal component 1 ..... 173

Figure D-2 Weights of each relative retention time location of principal component 2 ..... 173

Figure D-3 Weights of each relative retention time location of principal component 3 ..... 173

Figure D-4 Weights of each relative retention time location of principal component 4 ..... 173

Figure D-5 Weights of each relative retention time location of principal component 5 ..... 174

Figure D-6 Weights of each relative retention time location of principal component 6 ..... 174

Figure D-7 Weights of each relative retention time location of principal component 7 ..... 174

Figure D-8 Weights of each relative retention time location of principal component 8 ..... 174

Figure D-9 Weights of each relative retention time location of principal component 9 ..... 175

Figure D-10 Weights of each relative retention time location of principal component 10 ..... 175

Figure D-11 Weights of each relative retention time location of principal component 11 ..... 175

Figure D-12 Weights of each relative retention time location of principal component 12 ..... 175

Figure D-13 Weights of each relative retention time location of principal component 13 ..... 176

Figure D-14 Weights of each relative retention time location of principal component 14 ..... 176

Figure D-15 Weights of each relative retention time location of principal component 15 ..... 176

Figure D-16 Weights of each relative retention time location of principal component 16 ..... 176

Figure D-17 Weights of each relative retention time location of principal component 17 ..... 177

Figure D-18 Weights of each relative retention time location of principal component 18 ..... 177

Figure D-19 Weights of each relative retention time location of principal component 19 ..... 177

Figure D-20 Weights of each relative retention time location of principal component 20 ..... 177

Figure D-21 Weights of each relative retention time location of principal component 21 ..... 178

Figure D-22 Weights of each relative retention time location of principal component 22 ..... 178

Figure D-23 Weights of each relative retention time location of principal component 23 ..... 178

Figure D-24 Weights of each relative retention time location of principal component 24 ..... 178

Figure D-25 Weights of each relative retention time location of principal component 25 .....	179
Figure D-26 Weights of each relative retention time location of principal component 26 .....	179
Figure D-27 Weights of each relative retention time location of principal component 27 .....	179
Figure D-28 Weights of each relative retention time location of principal component 28 .....	179
Figure D-29 Weights of each relative retention time location of principal component 29 .....	180
Figure D-30 Weights of each relative retention time location of principal component 30 .....	180
Figure D-31 Weights of each relative retention time location of principal component 31 .....	180
Figure D-32 Weights of each relative retention time location of principal component 32 .....	180
Figure D-33 Weights of each relative retention time location of principal component 33 .....	181
Figure D-34 Weights of each relative retention time location of principal component 34 .....	181
Figure D-35 Weights of each relative retention time location of principal component 35 .....	181
Figure D-36 Weights of each relative retention time location of principal component 36 .....	181
Figure D-37 Weights of each relative retention time location of principal component 37 .....	182
Figure D-38 Weights of each relative retention time location of principal component 38 .....	182
Figure D-39 Weights of each relative retention time location of principal component 39 .....	182
Figure D-40 Weights of each relative retention time location of principal component 40 .....	182
Figure D-41 Weights of each relative retention time location of principal component 41 .....	183
Figure D-42 Weights of each relative retention time location of principal component 42 .....	183
Figure D-43 Weights of each relative retention time location of principal component 43 .....	183
Figure D-44 Weights of each relative retention time location of principal component 44 .....	183
Figure D-45 Weights of each relative retention time location of principal component 45 .....	184
Figure D-46 Weights of each relative retention time location of principal component 46 .....	184
Figure D-47 Weights of each relative retention time location of principal component 47 .....	184
Figure D-48 Weights of each relative retention time location of principal component 48 .....	184
Figure D-49 Weights of each relative retention time location of principal component 49 .....	185



Figure D-50 Weights of each relative retention time location of principal component 50 .....	185
Figure D-51 Weights of each relative retention time location of principal component 51 .....	185
Figure D-52 Weights of each relative retention time location of principal component 52 .....	185
Figure D-53 Weights of each relative retention time location of principal component 53 .....	186
Figure D-54 Weights of each relative retention time location of principal component 54 .....	186
Figure D-55 Weights of each relative retention time location of principal component 55 .....	186
Figure D-56 Weights of each relative retention time location of principal component 56 .....	186
Figure D-57 Weights of each relative retention time location of principal component 57 .....	187
Figure D-58 Weights of each relative retention time location of principal component 58 .....	187
Figure D-59 Weights of each relative retention time location of principal component 59 .....	187
Figure D-60 Weights of each relative retention time location of principal component 60 .....	187
Figure D-61 Weights of each relative retention time location of principal component 61 .....	188
Figure D-62 Weights of each relative retention time location of principal component 62 .....	188
Figure D-63 Weights of each relative retention time location of principal component 63 .....	188
Figure D-64 Weights of each relative retention time location of principal component 64 .....	188
Figure D-65 Weights of each relative retention time location of principal component 65 .....	189
Figure D-66 Weights of each relative retention time location of principal component 66 .....	189
Figure D-67 Weights of each relative retention time location of principal component 67 .....	189
Figure D-68 Weights of each relative retention time location of principal component 68 .....	189
Figure D-69 Weights of each relative retention time location of principal component 69 .....	190
Figure D-70 Weights of each relative retention time location of principal component 70 .....	190
Figure D-71 Weights of each relative retention time location of principal component 71 .....	190

Each graph below represents which of the 51,143 relative retention time locations are most important when describing the principal components. The weight of each location is normalized by the maximum value as described in Equation 4-14. The y-axis represents the significance of each location.

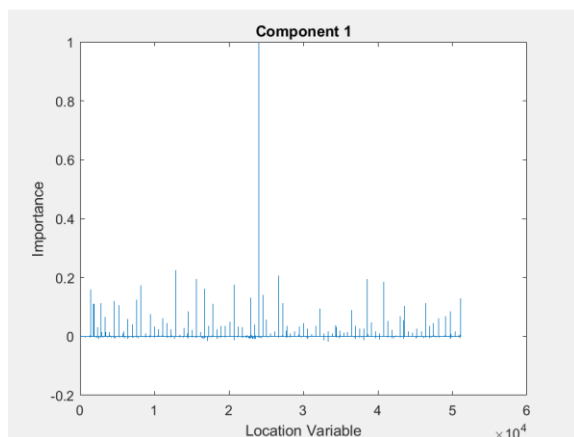


Figure D-1 Weights of each relative retention time location of principal component 1

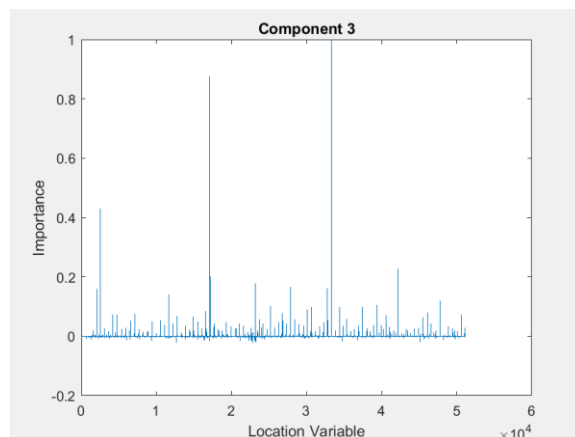


Figure D-3 Weights of each relative retention time location of principal component 3

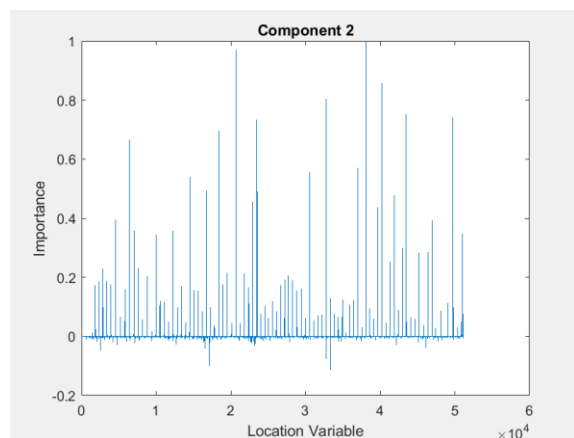


Figure D-2 Weights of each relative retention time location of principal component 2

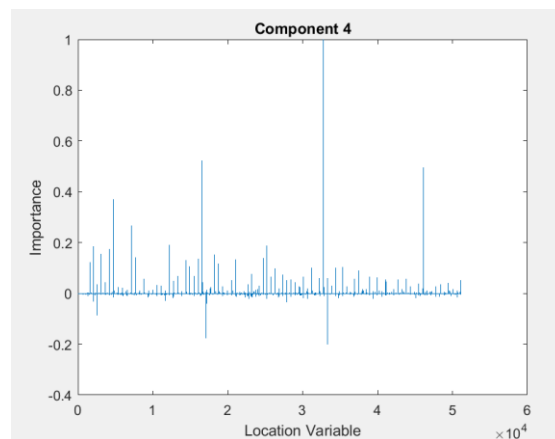


Figure D-4 Weights of each relative retention time location of principal component 4

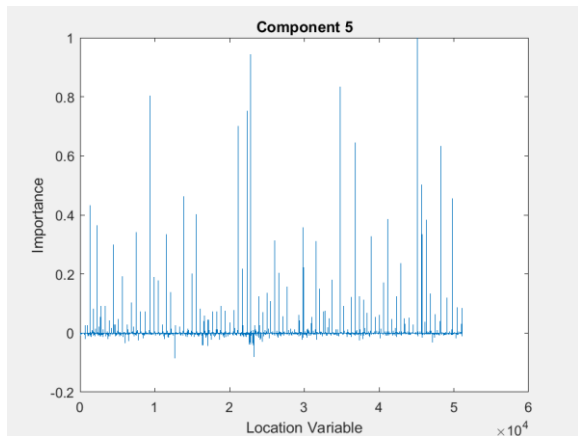


Figure D-5 Weights of each relative retention time location of principal component 5

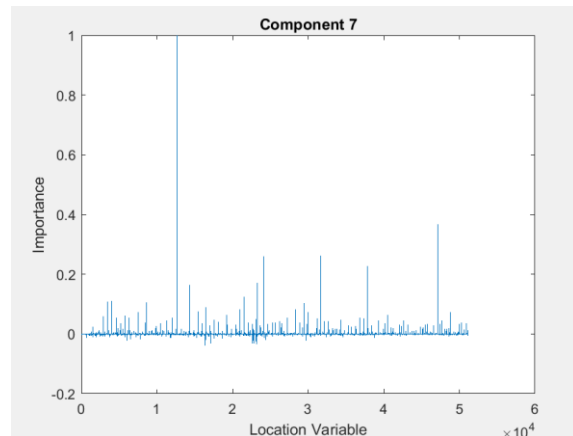


Figure D-7 Weights of each relative retention time location of principal component 7

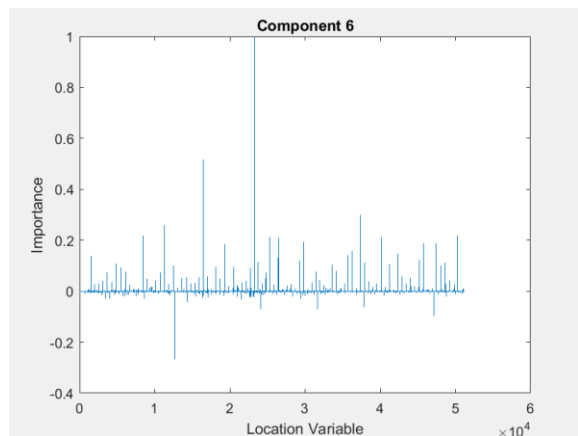


Figure D-6 Weights of each relative retention time location of principal component 6

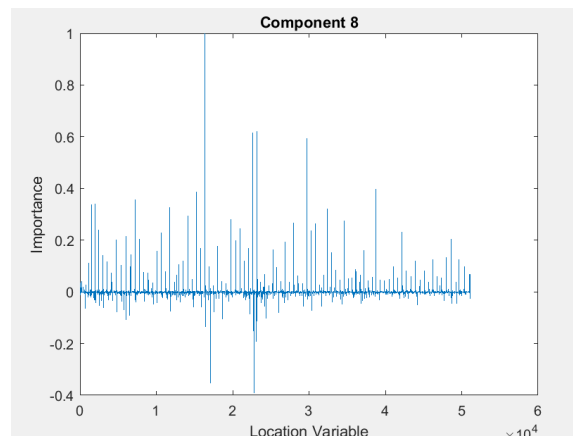


Figure D-8 Weights of each relative retention time location of principal component 8

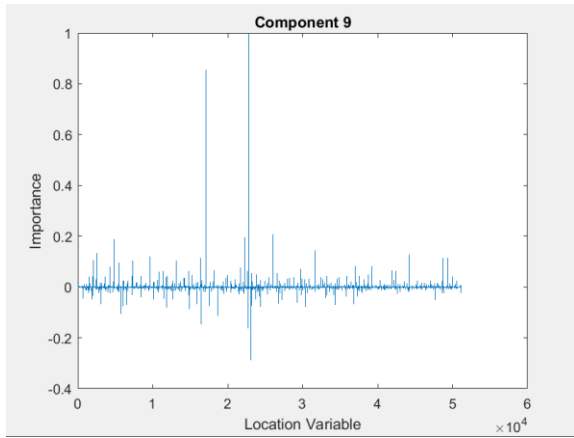


Figure D-9 Weights of each relative retention time location of principal component 9

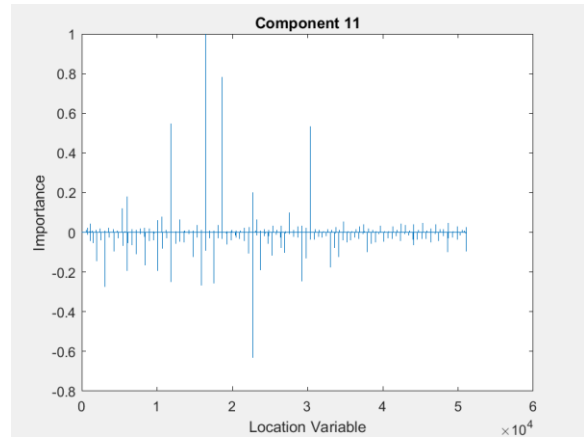


Figure D-11 Weights of each relative retention time location of principal component 11

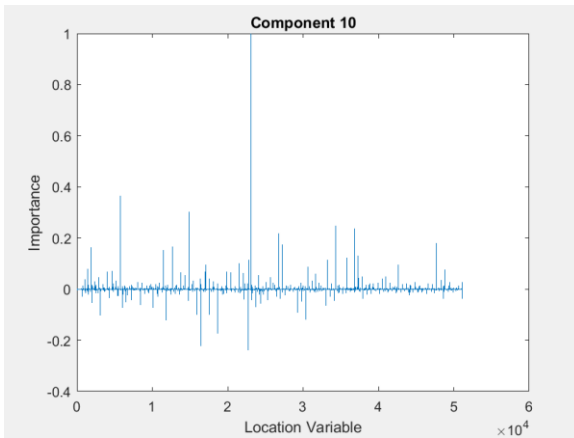


Figure D-10 Weights of each relative retention time location of principal component 10

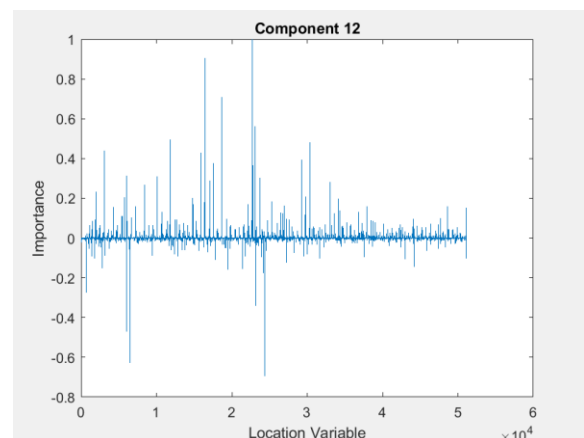


Figure D-12 Weights of each relative retention time location of principal component 12

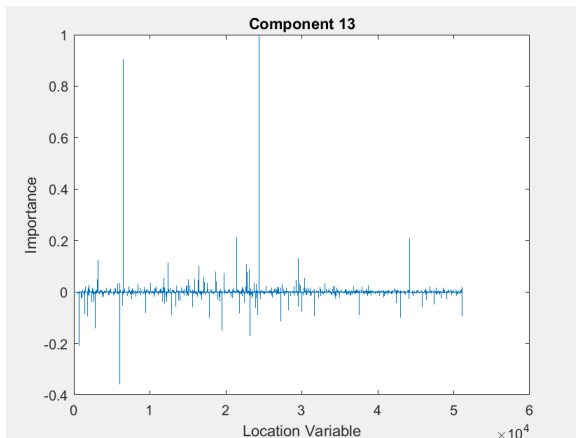


Figure D-13 Weights of each relative retention time location of principal component 13

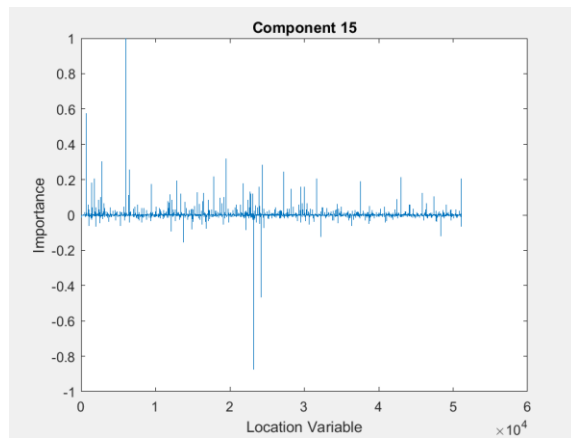


Figure D-15 Weights of each relative retention time location of principal component 15

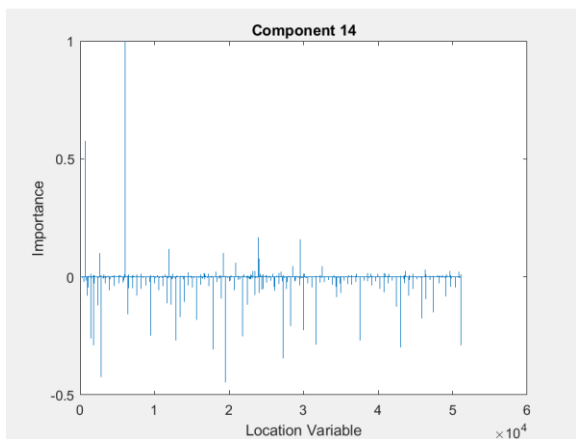


Figure D-14 Weights of each relative retention time location of principal component 14

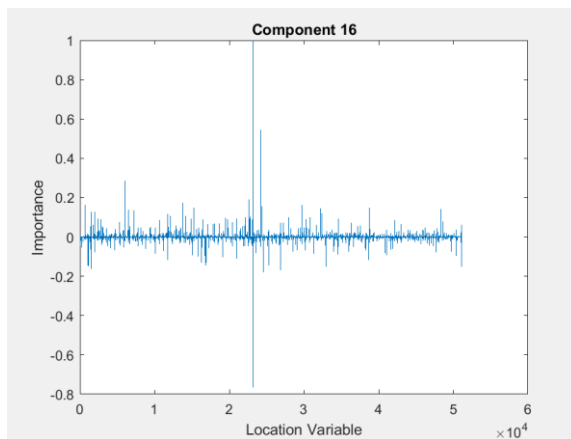


Figure D-16 Weights of each relative retention time location of principal component 16

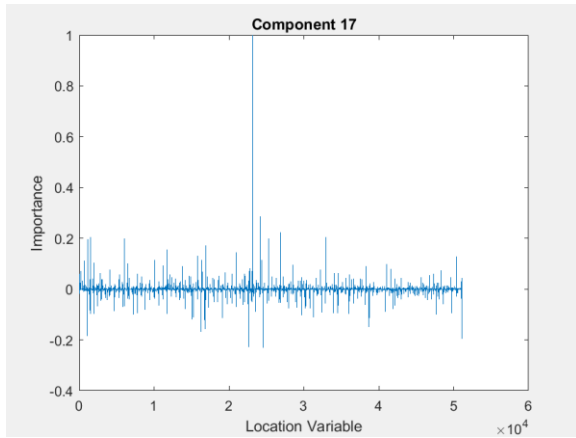


Figure D-17 Weights of each relative retention time location of principal component 17

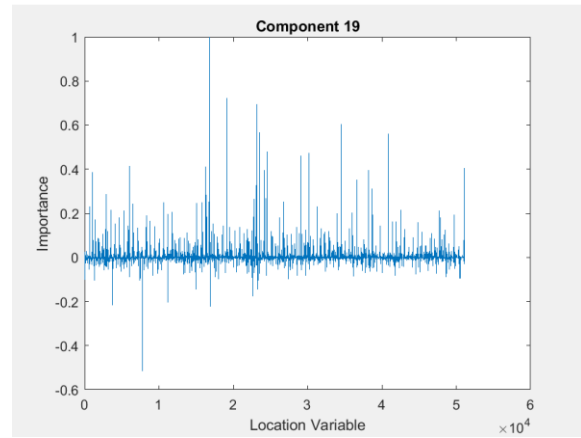


Figure D-19 Weights of each relative retention time location of principal component 19

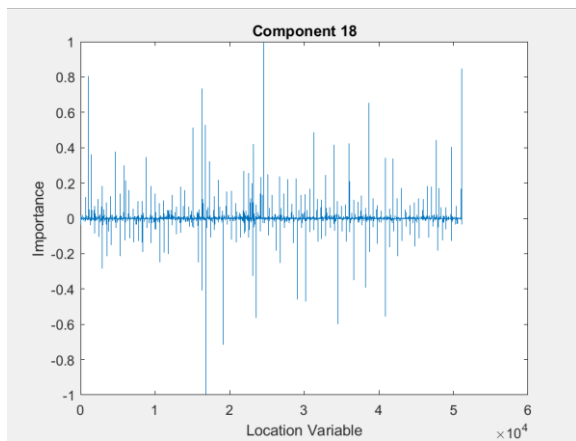


Figure D-18 Weights of each relative retention time location of principal component 18

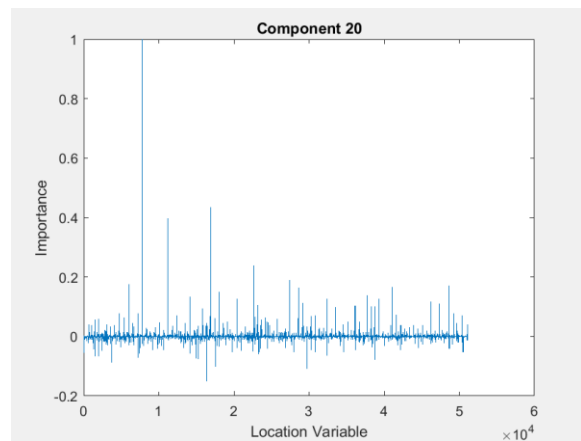


Figure D-20 Weights of each relative retention time location of principal component 20

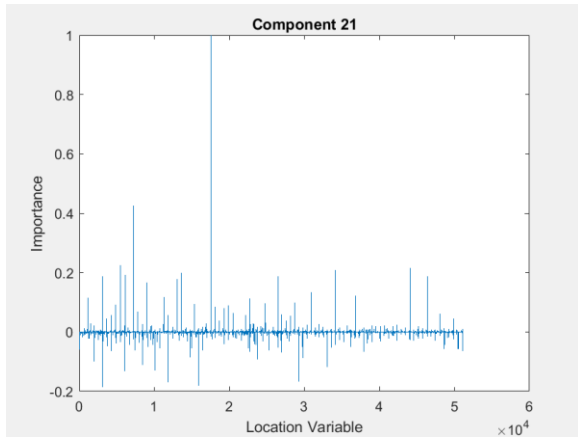


Figure D-21 Weights of each relative retention time location of principal component 21

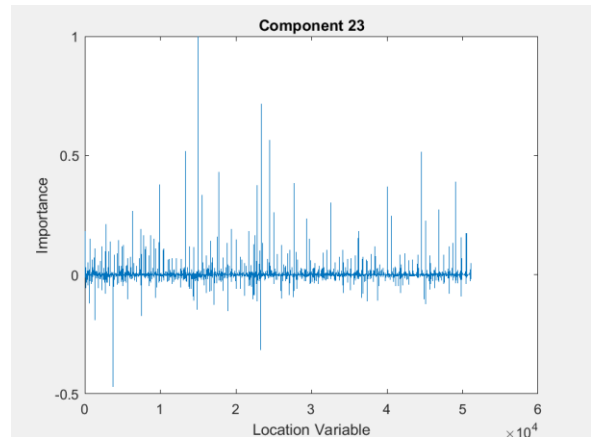


Figure D-23 Weights of each relative retention time location of principal component 23

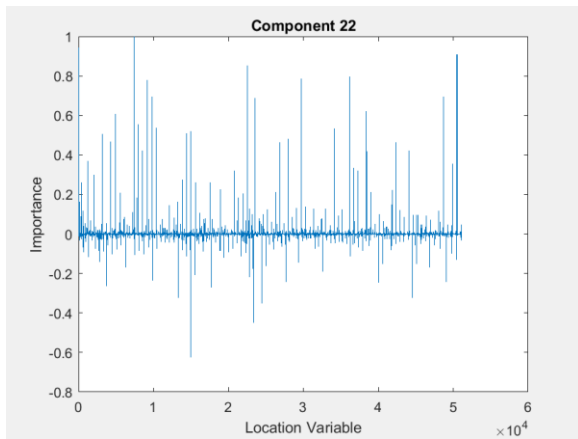


Figure D-22 Weights of each relative retention time location of principal component 22

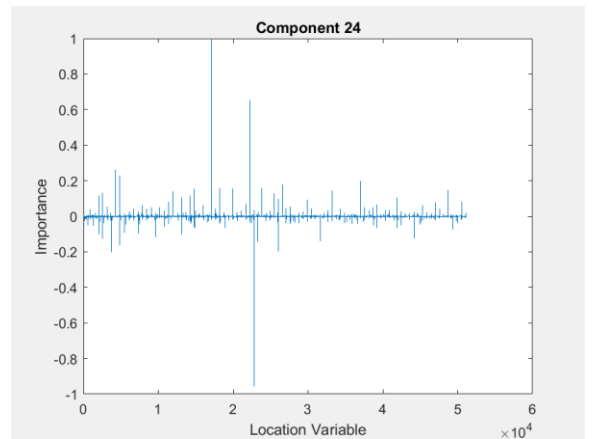


Figure D-24 Weights of each relative retention time location of principal component 24

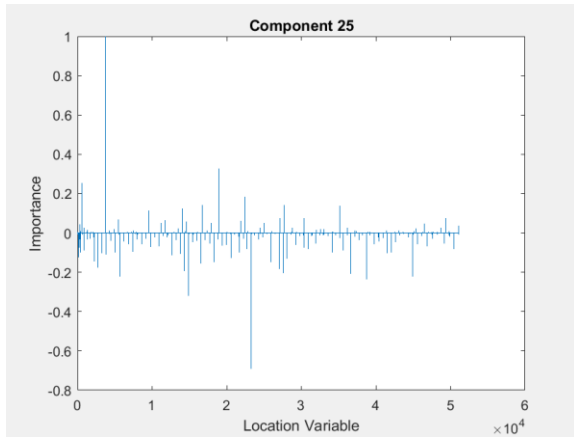


Figure D-25 Weights of each relative retention time location of principal component 25

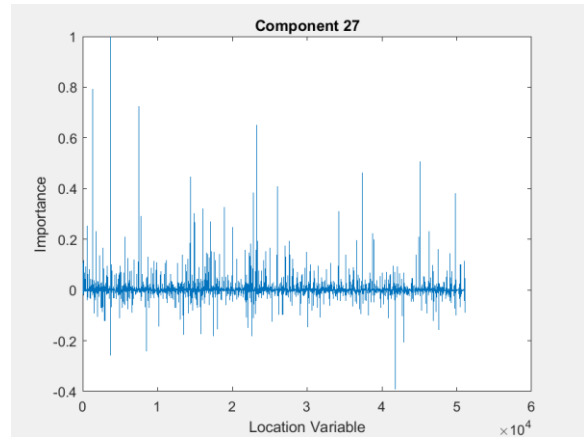


Figure D-27 Weights of each relative retention time location of principal component 27

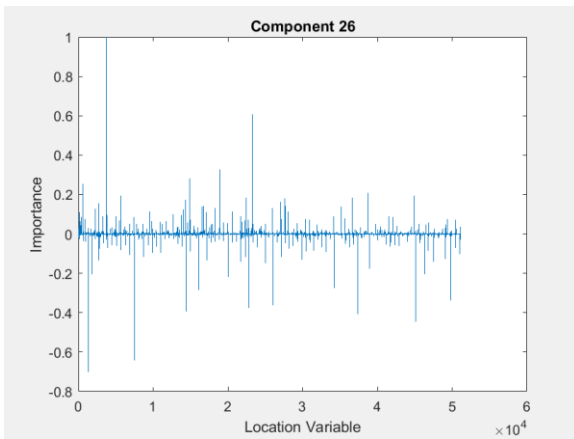


Figure D-26 Weights of each relative retention time location of principal component 26

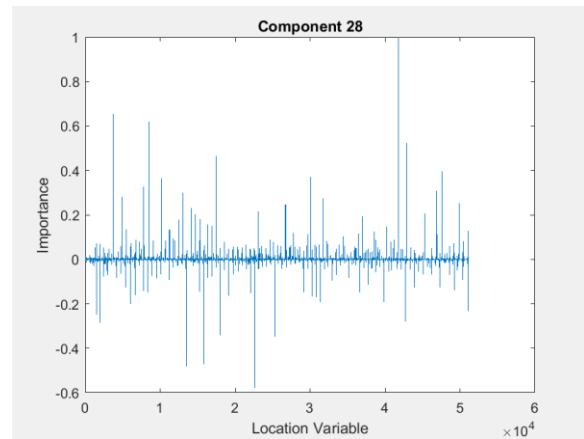


Figure D-28 Weights of each relative retention time location of principal component 28



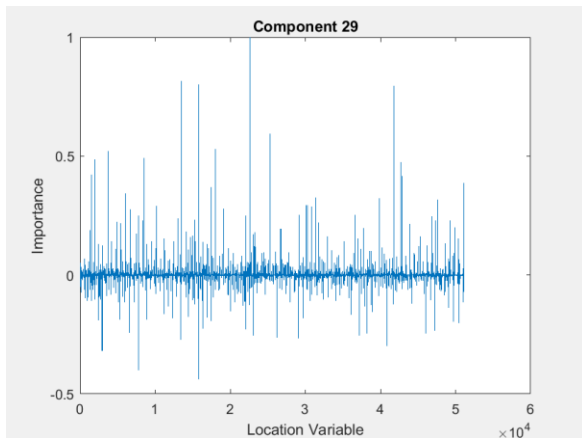


Figure D-29 Weights of each relative retention time location of principal component 29

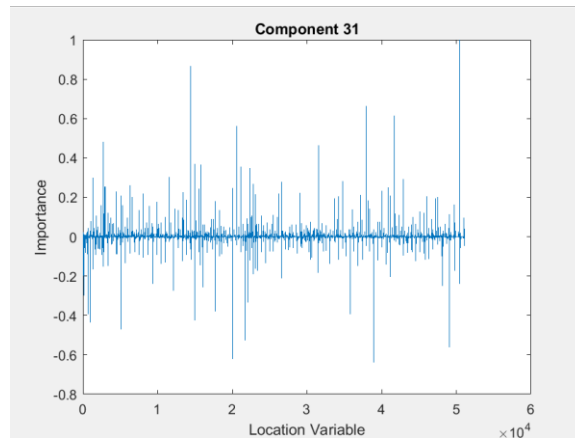


Figure D-31 Weights of each relative retention time location of principal component 31

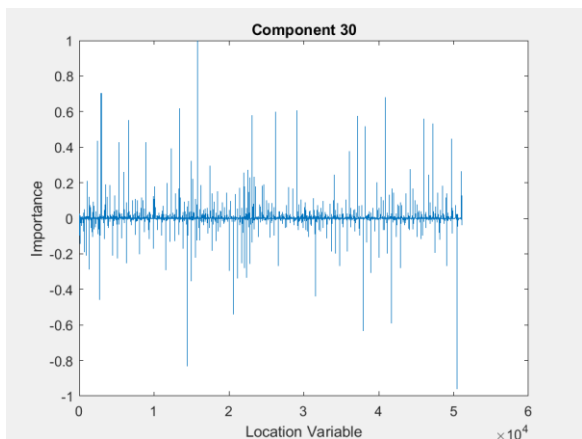


Figure D-30 Weights of each relative retention time location of principal component 30

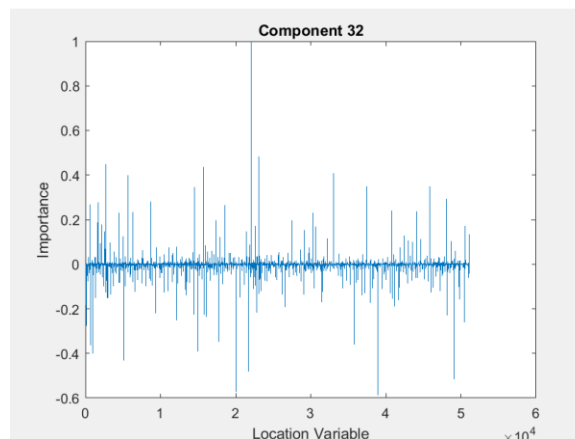


Figure D-32 Weights of each relative retention time location of principal component 32

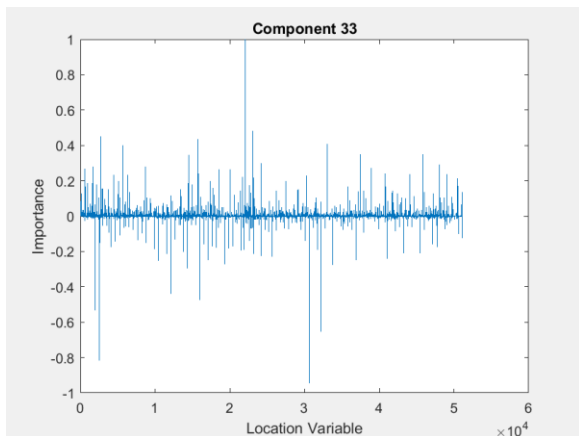


Figure D-33 Weights of each relative retention time location of principal component 33

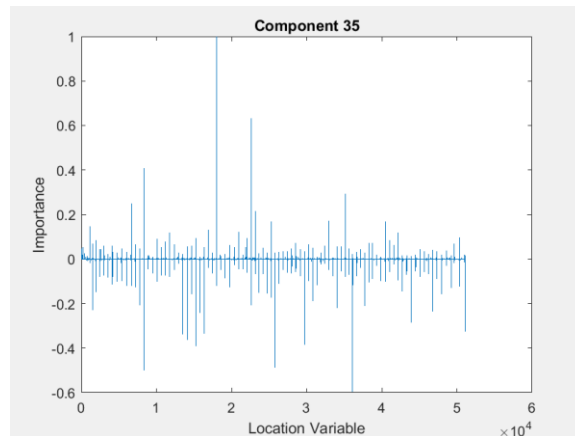


Figure D-35 Weights of each relative retention time location of principal component 35

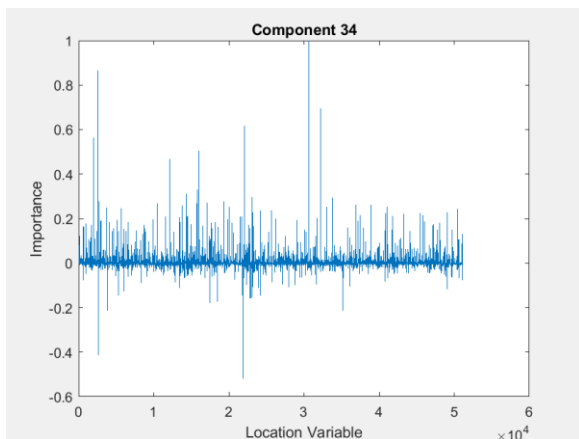


Figure D-34 Weights of each relative retention time location of principal component 34

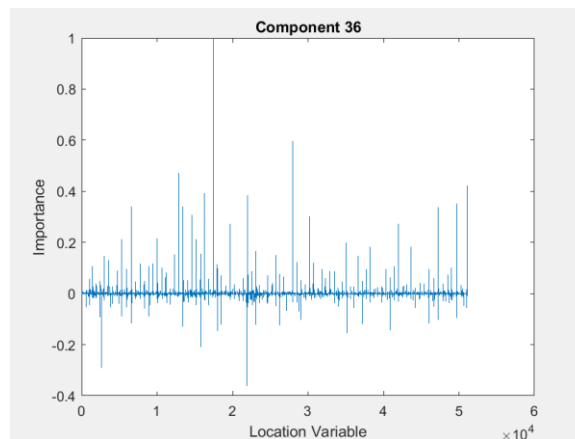


Figure D-36 Weights of each relative retention time location of principal component 36

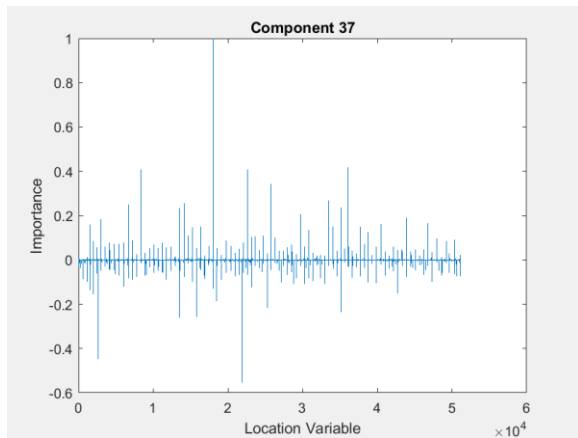


Figure D-37 Weights of each relative retention time location of principal component 37

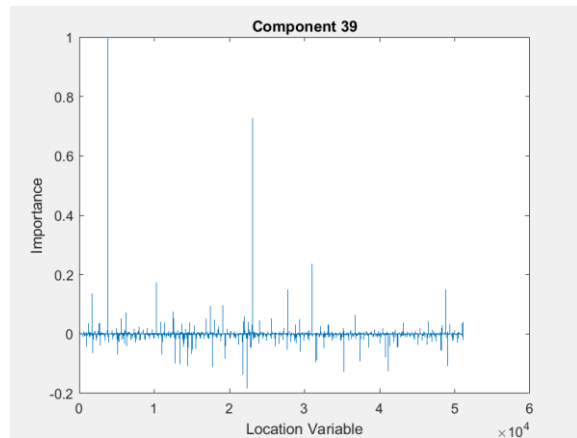


Figure D-39 Weights of each relative retention time location of principal component 39

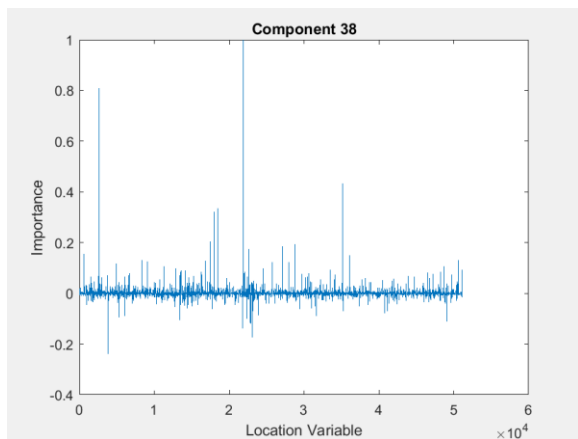


Figure D-38 Weights of each relative retention time location of principal component 38

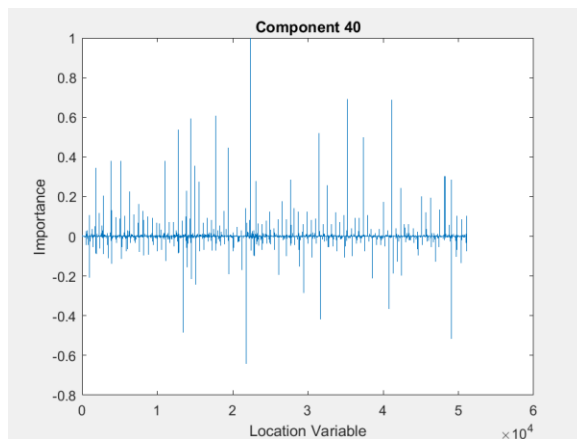


Figure D-40 Weights of each relative retention time location of principal component 40

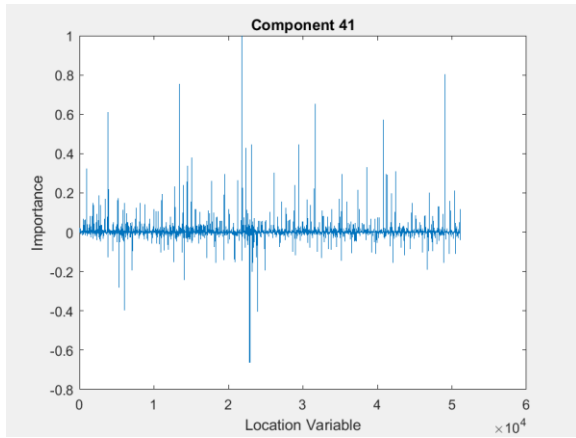


Figure D-41 Weights of each relative retention time location of principal component 41

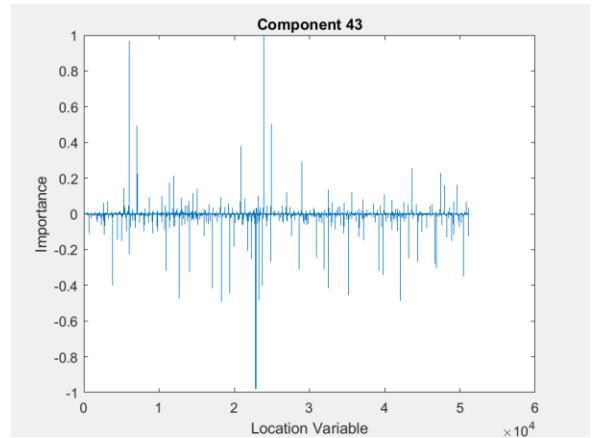


Figure D-43 Weights of each relative retention time location of principal component 43

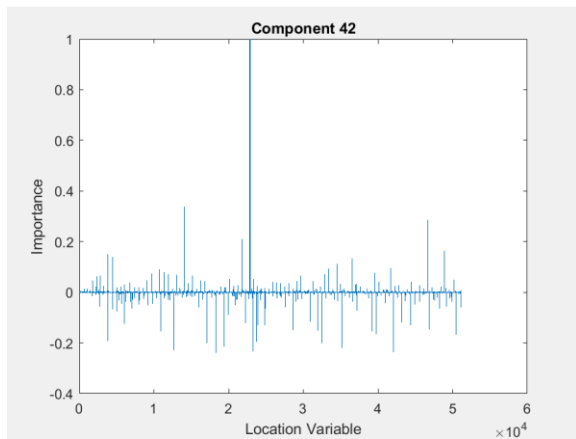


Figure D-42 Weights of each relative retention time location of principal component 42

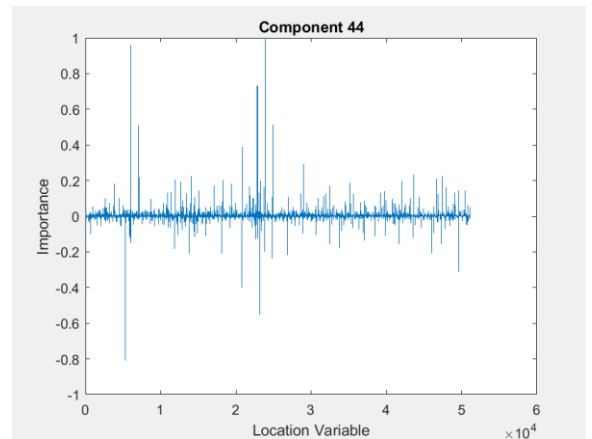


Figure D-44 Weights of each relative retention time location of principal component 44

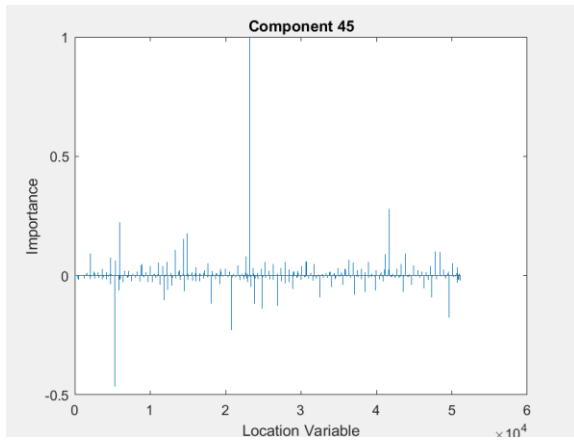


Figure D-45 Weights of each relative retention time location of principal component 45

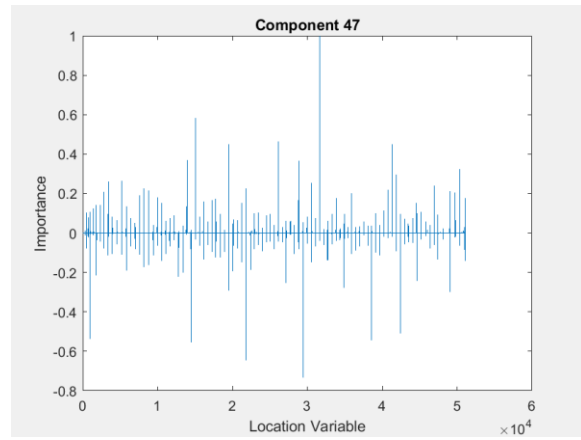


Figure D-47 Weights of each relative retention time location of principal component 47

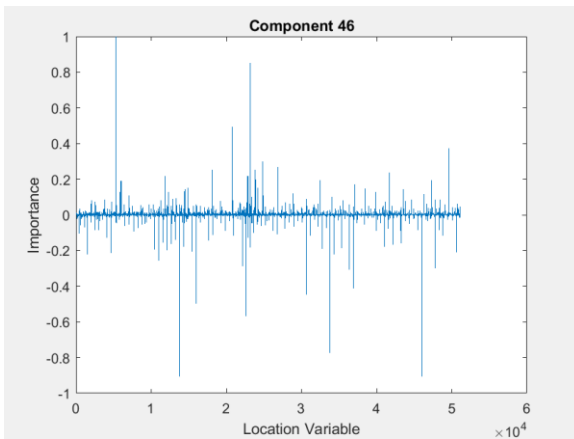


Figure D-46 Weights of each relative retention time location of principal component 46

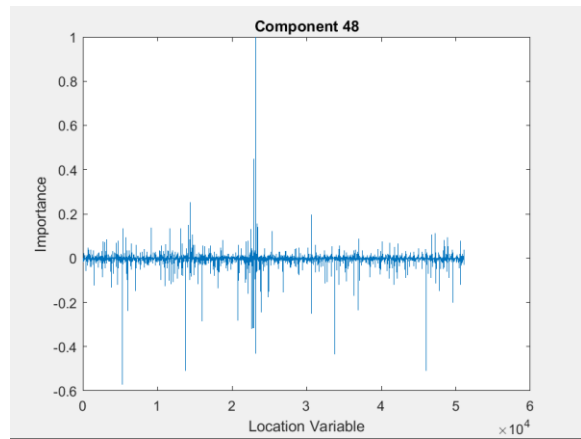


Figure D-48 Weights of each relative retention time location of principal component 48

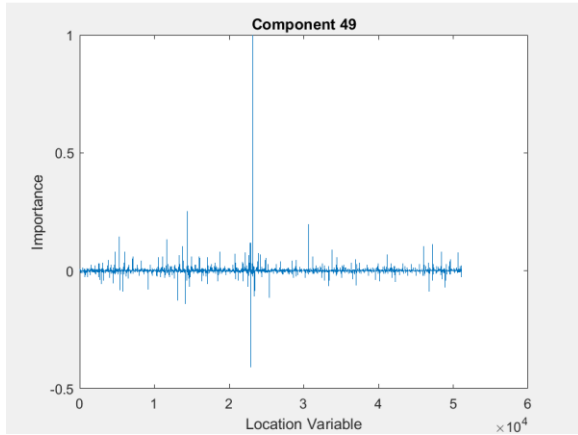


Figure D-49 Weights of each relative retention time location of principal component 49

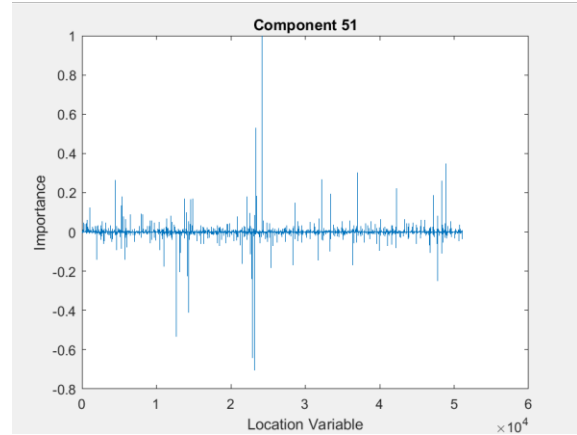


Figure D-51 Weights of each relative retention time location of principal component 51

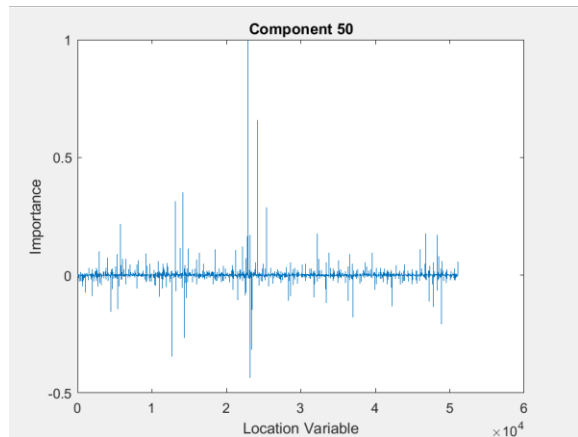


Figure D-50 Weights of each relative retention time location of principal component 50

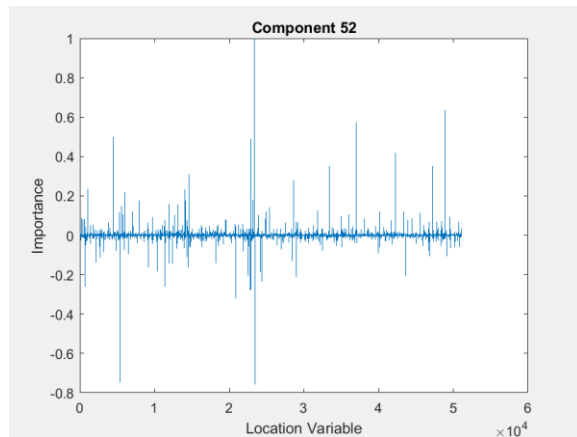


Figure D-52 Weights of each relative retention time location of principal component 52

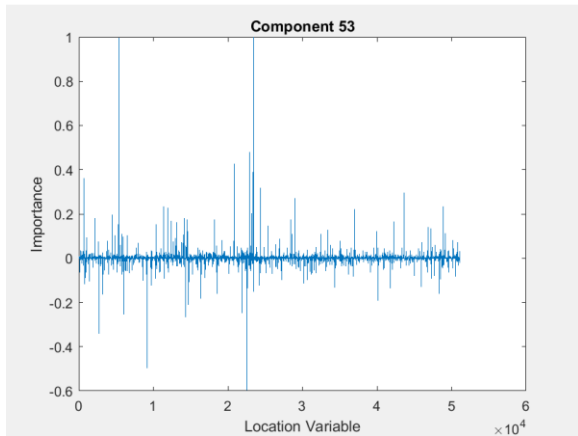


Figure D-53 Weights of each relative retention time location of principal component 53

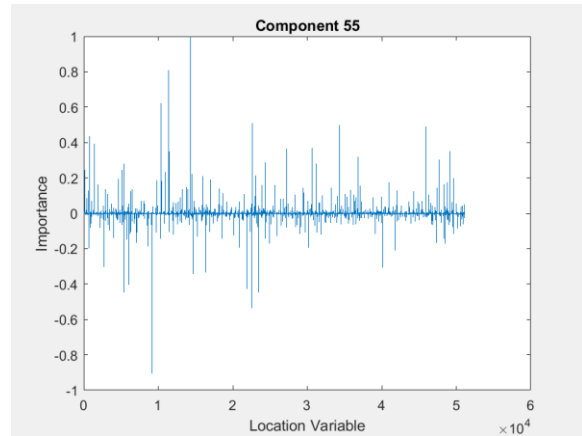


Figure D-55 Weights of each relative retention time location of principal component 55

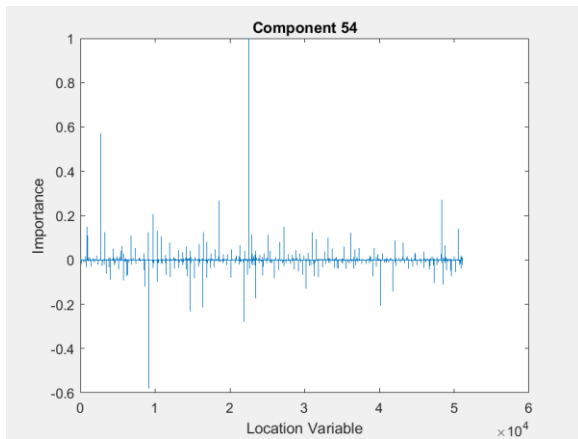


Figure D-54 Weights of each relative retention time location of principal component 54

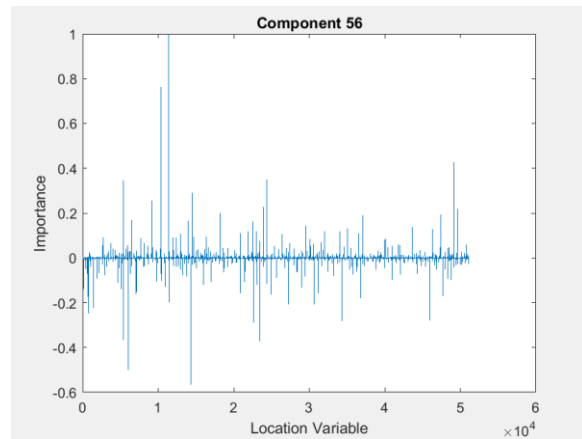


Figure D-56 Weights of each relative retention time location of principal component 56

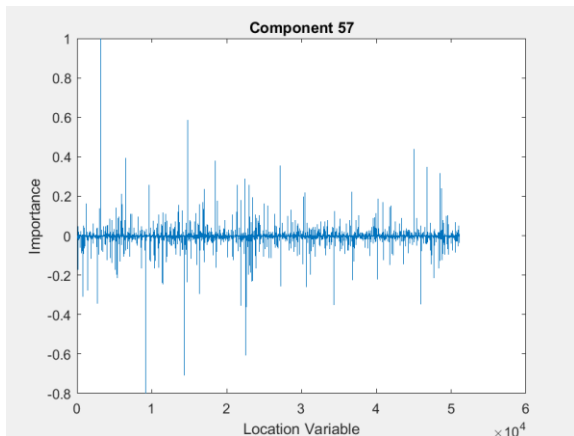


Figure D-57 Weights of each relative retention time location of principal component 57

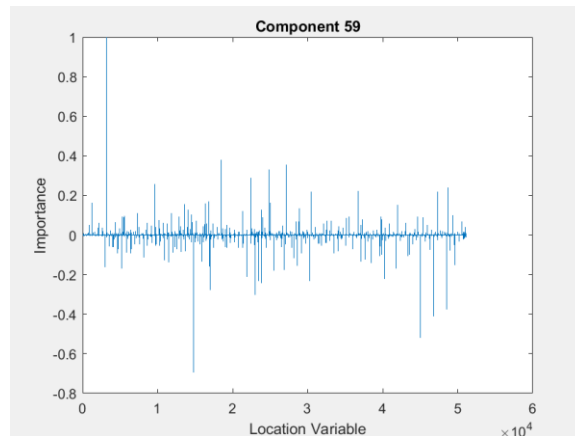


Figure D-59 Weights of each relative retention time location of principal component 59

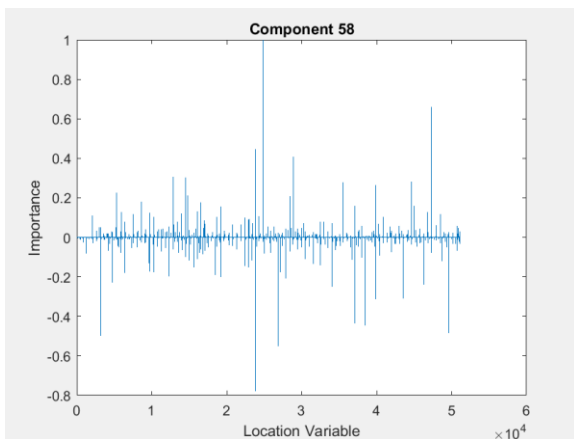


Figure D-58 Weights of each relative retention time location of principal component 58

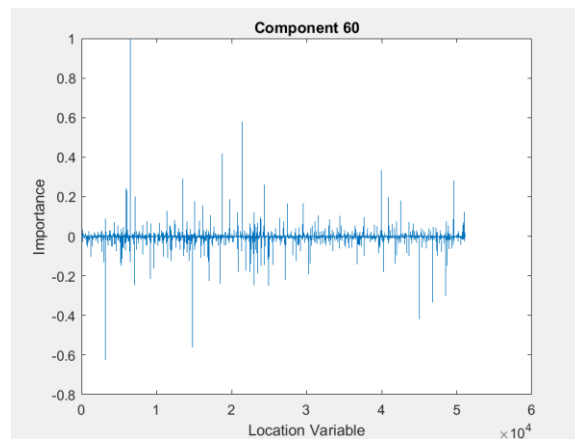


Figure D-60 Weights of each relative retention time location of principal component 60



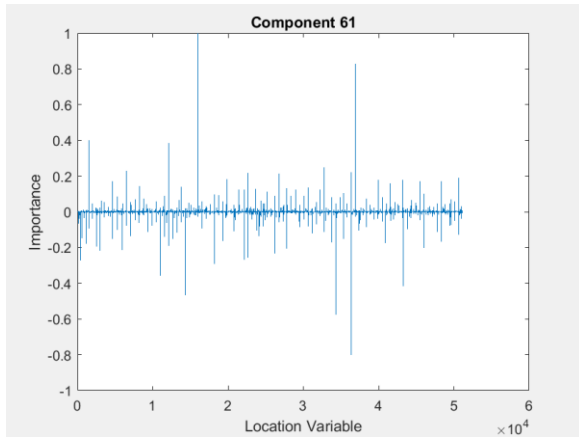


Figure D-61 Weights of each relative retention time location of principal component 61

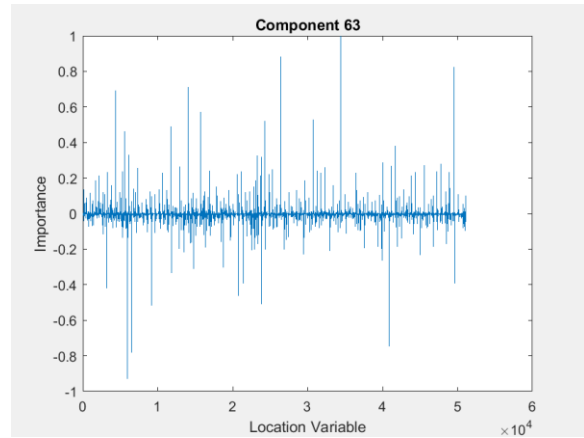


Figure D-63 Weights of each relative retention time location of principal component 63

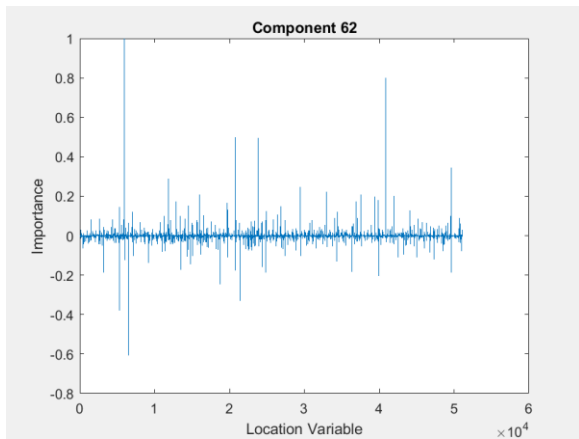


Figure D-62 Weights of each relative retention time location of principal component 62

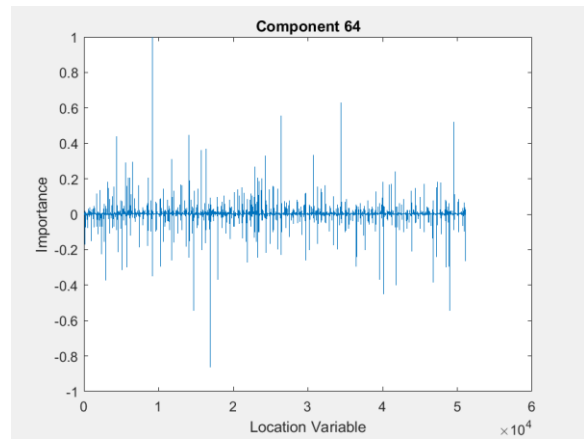


Figure D-64 Weights of each relative retention time location of principal component 64

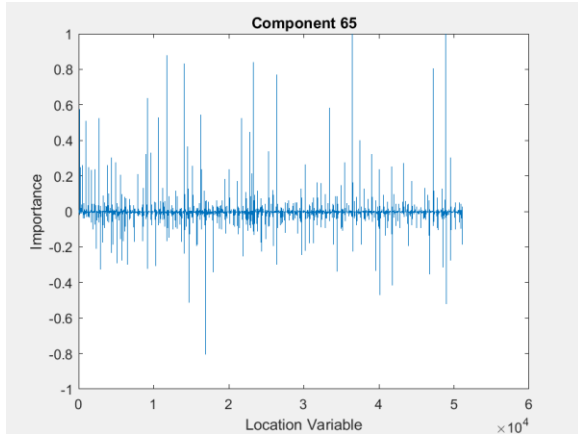


Figure D-65 Weights of each relative retention time location of principal component 65

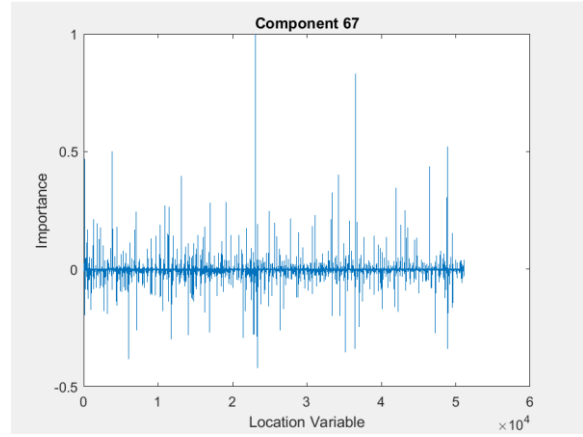


Figure D-67 Weights of each relative retention time location of principal component 67

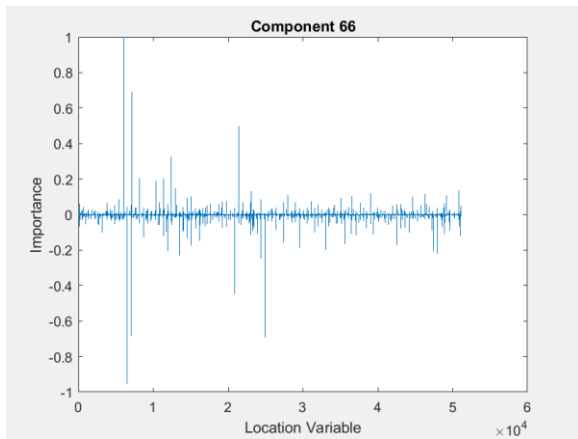


Figure D-66 Weights of each relative retention time location of principal component 66

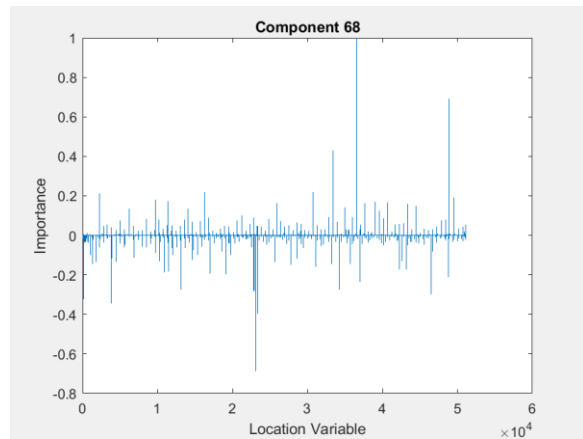


Figure D-68 Weights of each relative retention time location of principal component 68

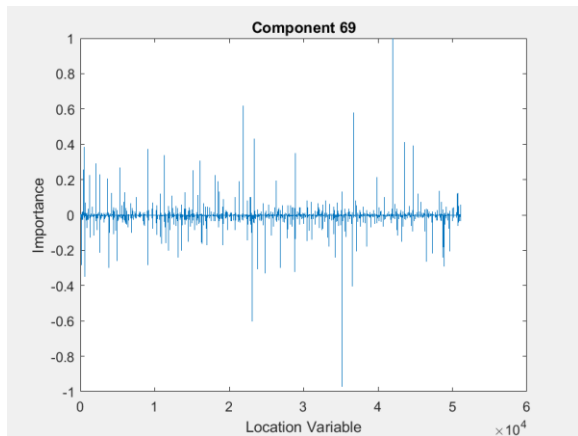


Figure D-69 Weights of each relative retention time location of principal component 69

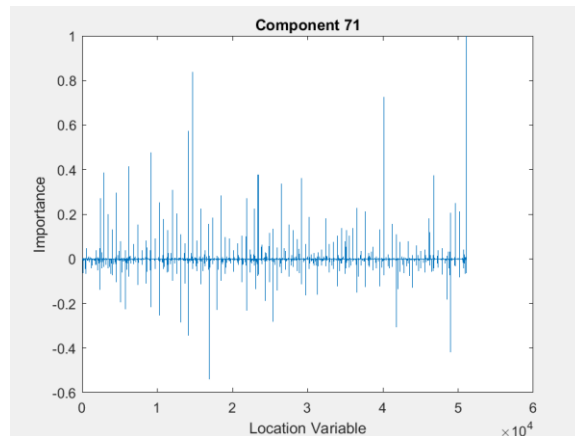


Figure D-71 Weights of each relative retention time location of principal component 71

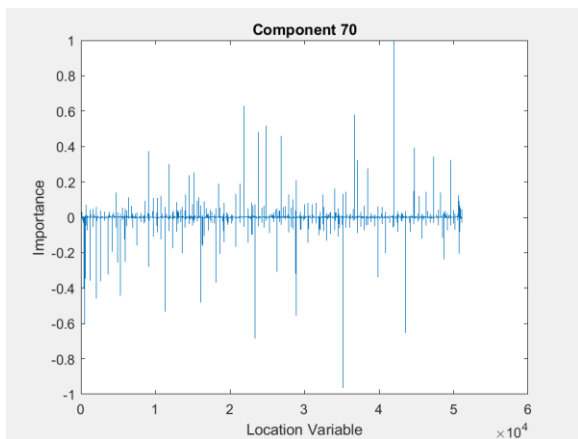


Figure D-70 Weights of each relative retention time location of principal component 70

## REFERENCES

1. Jolliffe, I.T., *Principal Component Analysis, Second Edition*. Springer Series in Statistics. Springer-Verlag New York, Inc.
2. *kmeans*. 2022 04/11/22]; Available from: <https://www.mathworks.com/help/stats/kmeans.html#buefs04-Distance>.
3. *silhouette plot*. 2022 04/11/22]; Available from: <https://www.mathworks.com/help/stats/silhouette.html>.
4. Kubicek, B., et al. *Peak-cognizant Signal Processing of Raw Instrument Signals to Quantify Environmental Weathering of Contaminants from the Deepwater Horizon Spill*. in *Global Oceans 2020: Singapore – U.S. Gulf Coast*. 2020.
5. McCarthy, R.A., et al., *Signal Processing Methods to Interpret Polychlorinated Biphenyls in Airborne Samples*. IEEE Access, 2020. **8**: p. 147738-147755.
6. Sengupta, A., et al. *Raw signal processing and graph-based visualization to autonomously interpret large repositories of GC-MS data: applications to oil spill weathering studies*. 2020.
7. Mccarthy, R.A. and A.S. Gupta, *Employing and Interpreting a Machine Learning Target-Cognizant Technique for Analysis of Unknown Signals in Multiple Reaction Monitoring*. IEEE Access, 2021. **9**: p. 24727-24737.
8. Ghasemi Damavandi, H., et al., *Interpreting comprehensive two-dimensional gas chromatography using peak topography maps with application to petroleum forensics*. Chemistry Central Journal, 2016. **10**(1): p. 75.
9. EPA. *Persistent Organic Pollutants: A Global Issue, A Global Response*. 2002 December 2009 05/16/22]; Available from: <https://www.epa.gov/international-cooperation/persistent-organic-pollutants-global-issue-global-response>.
10. Convention, S.o.t.S. *The POPs*. 2019 2019 05/16/22]; Available from: <http://chm.pops.int/TheConvention/ThePOPs/tabid/673/Default.aspx>.
11. EPA. *Learn about Polychlorinated Biphenyls (PCBs)*. 2022 05/16/22]; Available from: <https://www.epa.gov/pcbs/learn-about-polychlorinated-biphenyls-pcbs>.
12. EPA. *Polybrominated Diphenyl Ethers (PBDEs)*. 2022 05/16/22]; Available from: <https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/polybrominated-diphenyl-ethers-pbdes>.
13. Gupta, D.A.S. and D.K. Hornbuckle, *CDSE Grant Proposal*. 2017, National Science Foundation.
14. Archer, J., J. Bruce, and A. Adeuya, *The design and development of novel mathematical algorithms to discover nontargeted contaminants that co-occur with targeted persistent organic pollutants determined from food samples analyzed by GCxGC-ToFMS instrumentation*. 2021.
15. Center, N.C.f.B. 05/23/22]; Available from: <https://pubchem.ncbi.nlm.nih.gov/>.