Summer 2016

Data analytics, interpretation and machine learning for environmental forensics using peak mapping methods

Hamidreza Ghasemi Damavandi
University of Iowa

DATA ANALYTICS, INTERPRETATION AND MACHINE LEARNING FOR ENVIRONMENTAL FORENSICS USING PEAK MAPPING METHODS

by

Hamidreza Ghasemi Damavandi

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

August 2016

Thesis Supervisor: Professor Ananya Sen Gupta

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Hamidreza Ghasemi Damavandi

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Electrical and Computer Engineering at the August 2016 graduation.

Thesis Committee: _____
Ananya Sen Gupta, Thesis Supervisor

_____
Erwei Bai

_____
Guadalupe Canahuate

_____
Keri C. Hornbuckle

_____
Anton Kruger

To my sister, Maedeh.

Everything you can imagine is real.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Ananya Sen Gupta for her immense help to bring this work till this phase and would extend my profound gratitude to her for guiding me passing it. I would also like to thank Dr. Reddy and Robert Nelson at Woods Hole Oceanographic Institution for providing the precious dataset of the petroleum samples. I would also like to extend a special thanks to Dr. Lehmler and Dr. Korwel for providing us with the breastmilk dataset and my PhD committee members, Dr. Bai, Dr. Hornbuckle, Dr. Canahuate and Dr. Kruger for their precious help and guide to bring this thesis to the end.

ABSTRACT

In this work our driving motivation is to develop mathematically robust and computationally efficient algorithms that will help chemists towards their goal of pattern matching. Environmental chemistry today broadly faces difficult computational and interpretational challenges for vast and ever-increasing data repositories. A driving factor behind these challenges are little known intricate relationships between constituent analytes that constitute complex mixtures spanning a range of target and non-target compounds. While the end of goal of different environment applications are diverse, computationally speaking, many data interpretation bottlenecks arise from lack of efficient algorithms and robust mathematical frameworks to identify, cluster and interpret compound peaks. There is a compelling need for compound-cognizant quantitative interpretation that accounts for the full informational range of gas chromatographic (and mass spectrometric) datasets. Traditional target-oriented analysis focus only on the dominant compounds of the chemical mixture, and thus are agnostic of the contribution of unknown non-target analytes. On the other extreme, statistical methods prevalent in chemometric interpretation ignore compound identity altogether and consider only the multivariate data statistics, and thus are agnostic of intrinsic relationships between the well-known target and unknown target analytes. Thus, both schools of thought (target-based or statistical) in current-day chemical data analysis and interpretation fall short of quantifying the complex interaction between major and minor compound peaks in molecular mixtures commonly encountered in

environmental toxin studies. Such interesting insights would not be revealed via these standard techniques unless a deeper analysis of these patterns be taken into account in a quantitative mathematical framework that is at once compound-cognizant and comprehensive in its coverage of all peaks, major and minor.

This thesis aims to meet this grand challenge using a combination of signal processing, pattern recognition and data engineering techniques. We focus on petroleum biomarker analysis and polychlorinated biphenyl (PCB) congener studies in human breastmilk as our target applications.

We propose a novel approach to chemical data analytics and interpretation that bridges the gap between target-cognizant traditional analysis from environmental chemistry with compound-agnostic computational methods in chemometric data engineering. Specifically, we propose computational methods for target-cognizant data analytics that also account for local unknown analytes allied to the established target peaks. The key intuition behind our methods are based on the underlying topography of the gas chromatigraphic landscape, and we extend recent peak mapping methods as well as propose novel peak clustering and peak neighborhood allocation methods to achieve our data analytic aims. Data-driven results based on a multitude of environmental applications are presented.

PUBLIC ABSTRACT

Most laboratory petroleum data contain significant overlap with regional fingerprints that mislead forensic apportioning of the environmental impact of major oil spills (e.g. the British Petroleum spill, Gulf of Mexico, April 2010). We develop methods to better distinguish between highly correlated petroleum sources that share unknown regional fingerprints along with their source-specific unique signatures. We also devise compression techniques that harness source-sensitive pattern recognition to drastically reduce the volume of gas chromatographic datasets, leading to efficient storage, indexing and querying of vast data repositories across environmental and petroleum laboratories.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

LIST OF ALGORITHMS

Algorithm

CHAPTER 1
INTRODUCTION

1.1 Problem Statement

Forensic distinction among the oil sources can be of paramount importance in apportioning the environmental impact of major oil spills such as the *Deepwater Horizon* disaster in the Gulf of Mexico (April 20, 2010) and the Refugio oil spill, California (May 19, 2015). The problem of oil source fingerprinting and identification has been well-studied and many experimental methods have been used to determine the petroleum fingerprint of different areas([1–7]). The current state-of-the-art in petroleum forensics typically employs one-dimensional gas chromatography combined with mass spectrometry (GC-MS) as well as two-dimensional gas chromatography ($GC \times GC$) as the separation technology, followed by peak-ratio analysis on target biomarkers ([8–11]). However, despite the success of target-driven analysis in existing art [12] key challenges remain when comparing crude oil from closely correlated sources, e.g. neighboring oil reservoirs in a petroleum-rich locale.

These challenges are discussed in depth in Section 1.2.4 and as such, are not limited to petroleum forensics alone. A broad range of applications in environmental chemistry, e.g. air quality monitoring and toxins ingested by newborns through breastmilk, face similar interpretational issues. The end goal in these environmental applications is generating a robust quantitative framework for comparing field samples that connect the role of unknown non-target analytes measured in the raw

instrument signal with labeled target analytes that are routinely identified in complex mixtures. Beyond statistical robustness of field samples against instrumental variability, the interpretation framework also needs to exhibit relative immunity to training biases when supervised methods [13] are used. Robust prior knowledge of the source (e.g. an oil reservoir) extracted from the experimental analyses in the chemical lab can be expensive to derive and reproduce in terms of instrument and personnel time. Target-based source differentiation also fails popular correlation tests (e.g. [14]) when a reference sample from a known source is compared against a closely correlated but unknown sample that may be from a neighboring source without an established fingerprint. Therefore, there is a compelling need to analyze environmental datasets, e.g. $GC \times GC$ images of pre and post-spill crude oil samples, through algorithmical methods that are not only target-cognizant but also driven by the nuances of unknown non-target analytes that span the complete dataset.

Beyond the compelling motivation from environmental chemistry, there is currently considerable enthusiasm around the study and analysis of "large data-sets". Large data-sets are popularly referred to as "Big Data" where the dimensionality of the data-set is intractably large. In the context of environmental chemistry, this can mean hundreds of thousands, if not millions of compound peaks that need to be accessed, compared, and indexed across the whole data repository. A vast and growing literature in data processing, knowledge extraction and data mining have been proposed to manage, interpret and classify the high volume of data being produced at a rapid pace daily. High-volume data analytic techniques include but are

not limited to numerous (i) data indexing and clustering methods [15], (ii) data compression schemes ([16], [17], [18]) and (iii) pattern recognition [19] and related signal processing techniques.

This thesis seeks to combine techniques from these related data engineering fields and propose novel solutions to chemometric data interpretation. The research objectives combine empirical data analytics as well as mathematical constructs which are summarized as three related thrusts:

(i) Discover underlying patterns within gas chromatographic datasets, with particular emphasis on petroleum biomarkers (hopanes and steranes) and polychlorobiphenyl (PCB) compounds found in air quality monitoring and human breastmilk;

(ii) Enable compact data representation schemes that enable reconstruction of relevant information;

(iii) Develop analyte-cognizant data indexing and querying techniques for high-volume chromatographic datasets.

1.1.1 Quantitative Source Differentiation in Chemical Forensics

Suppose we have a library of $GC \times GC$ images saving the $GC \times GC$ information of different geographical regions of the world. Suppose, the size of our library is large enough to have at least one sample from any geographical region. Our problem is to estimate the geographical region of a newly-extracted unknown test image based upon the information in the library. We perform the task, by comparing the test

image with each of the elements of the library and declare the one with the highest percentage of match to the test image as the potential sample with the same region. Therefore, we should try to come up with a valid similarity criterion to compare the images.

1.2 Background Motivation

1.2.1 Review of Analytical Hardware

Gas Chromatography(GC) is the process in which a chemical mixture is separated into its underlying chemical elements or compounds. Figure 1.1 shows the gas chromatography system components. Gas chromatography consists of two phases, the mobile phase and the stationary phase. The mobile phase is a gas carrier(usually helium or nitrogen) that carries the injection(entered at Injector) through the system. The stationary phase is a layer of liquid inside a glass called column. The compounds interact with the walls of the column differently according to their chemical properties, the boiling points, the polarity and the temperature of the oven. Then they are eluted and sensed by the detector at different times called their retention time. Different compounds are sensed at different retention times and the detector records these compounds as peaks. The thermostatic oven is set so that the temperature of the gas will be controlled for a precise work of separation.

Figure 1.2 shows the signal recorded by the detector in the system. Two-dimensional gas chromatography($GC \times GC$) is this process while being done in two steps, in order to well-separate the elements from each other. In the first stage, the

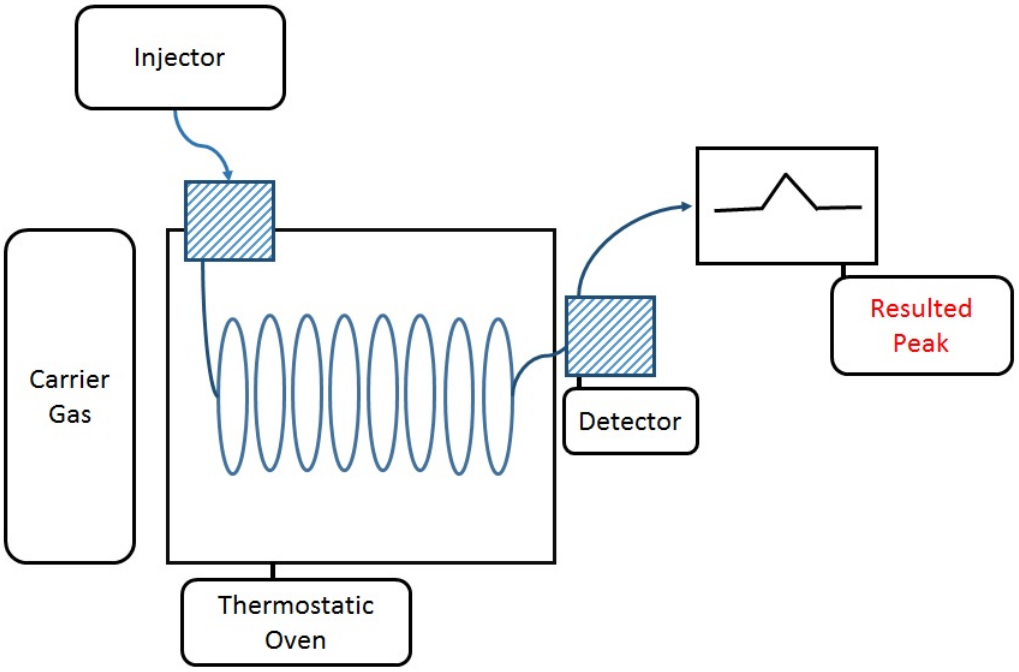

Figure 1.1: Gas Chromatography system.

chemical elements are separated and then sensed at the output of the first step, and they produce the corresponding peaks at different retention times (first dimension). Then these elements go through the second stage and are separated in another stage which leads to the second dimension of the retention time. So the final signal is a 3-dimensional image, where the first and the second stages correspond to the retention times and the third dimension is the amplitude of the corresponding peak for each element at the corresponding retention times(Figure 1.3).

Figure 1.3 shows the two-dimensional chromatography instrument and the final resulting signal.

Note that the time for the first stage is far more than the time for the second

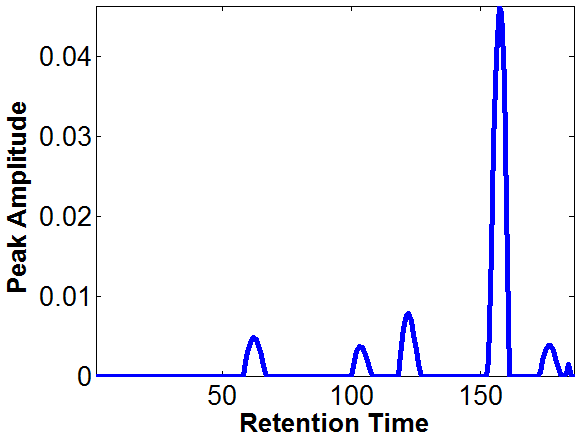

Figure 1.2: one-dimensional GC image.

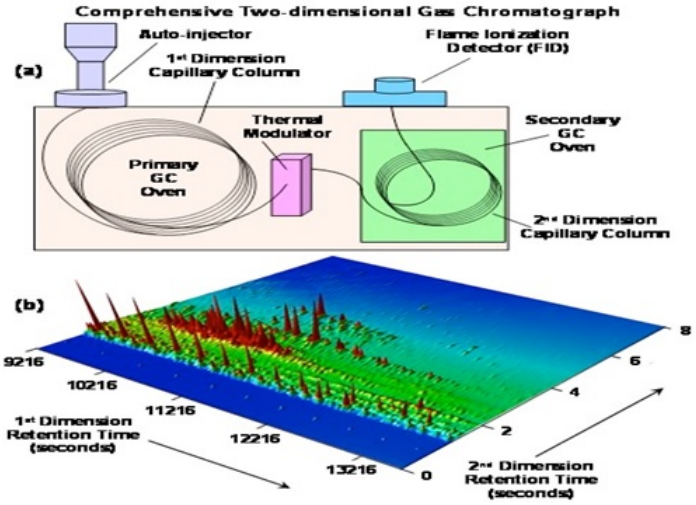

Figure 1.3: 2-dimensional $GC \times GC$ image. This image has been provided by Robert Nelson and Woods Hole Oceanography Institution.

stage, because the main separation is done in the first stage, while the second stage has the role of separation for those not separated well in the first stage.

1.2.2 Interdisciplinary Challenges Addressed in Proposed Research

In this work a compound analysis of the $GC \times GC$ images have been taken into consideration while any minor variations and compounds carry important information about the forensic property of the petroleum source. The current state of the art methods is divided into two general areas:

- Separating Technology for analyzing complex molecular mixtures(Compound Analysis): Analysis of the image based on the variations and target and non-target peak of the image. This method is effective once the little differences among different forensic sources should be captured. Section 1.2.1 reviews the analytical hardware of gas chromatographic setups.

- Infometric methods for interpreting GC image(Chemometric method): This method is based on the statistical property of the image. In this method an overall analysis is done through-out the image while many minor variations of the image is either ignored or averaged. Section 1.3.2 discusses chemometric state-of-the-art in detail.

Table 1.1 shows the comparison between the target-compound analysis of the $GC \times GC$ images and the commonly used chemometric methods. The complete description on difference between the two broad methods is comprehensively explained in chapter 5.

Table 1.1: Comparison between Target-Compound Analysis and Chemometric methods

Compound Analysis	Chemometric methods
Compounds with chemical meaning	Points have no underlying meaning
Main targets play the important role	Analysis is based on both main and minor analytes
The effect of non-targets are ignored	Comprehensive large-scale analysis
Highly sensitive to drift in retention time	Robust to retention time
Reliable source diagnosis, based on peaks	Requires many training data-sets

1.2.3 Peaks in the $GC \times GC$ image

Any $GC \times GC$ image is composed of couple of classes of biomarkers,say hopane and sterane biomarkers. Any biomarker of the image itself is composed of couple of target analytes which are analyzed by the chemists in order to identify the corresponding petroleum source. They are some peaks within the image which are not targets referred to as non-target analytes. These non-target analytes could carry significant information about the forensic source. Figure 1.4 shows one biomarker and its corresponding targets and non-target analytes. The main contribution of this work is to avail both the target and non-target analytes in order to have a better scheme to separate and identify the forensic source.

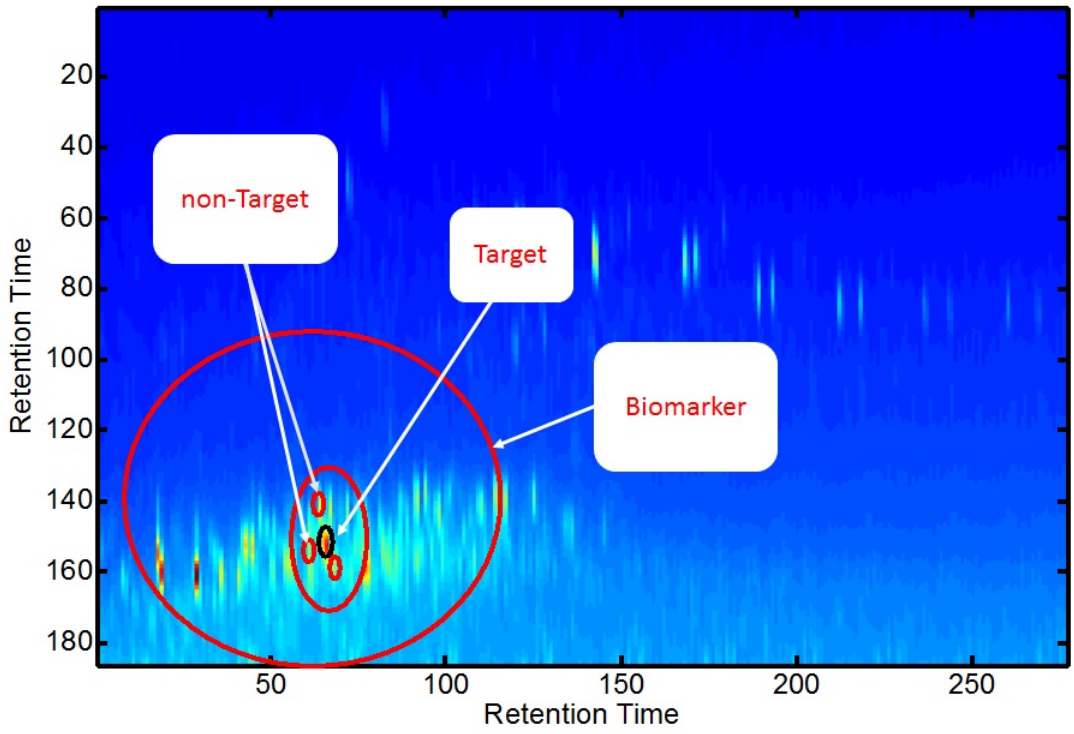

Figure 1.4: Visual illustration, target and non-target analytes within hopanes and strains $GC \times GC$ image. The target analytes dominate the biomarker topography and some of the non-target analytes are gathered around the target analytes.

1.2.4 Challenges in chemical data interpretation with focus on the $GC \times GC$ image

There are several challenges that underlie the analysis of the $GC \times GC$ image:

- Retention Time variability

  The $GC \times GC$ system is a non-ideal system, it means that a peak experiences a drift from its actual point. If a peak should arrive at the output of the second stage at (r_1, r_2), it will arrive at $(r_1 + \delta_1, r_2 + \delta_2)$, where it is assumed $|\delta| \leq \Delta$,

where Δ is the maximum variations of the peaks. Furthermore, the system is non-uniformly non-ideal, it means it may drift one peak with δ and drift the other by $\delta^{'}(\delta^{'} \neq \delta)$.

- Peak Baseline

  Aside from the drift, the peaks of the signal, which corresponds to chemical constituents in the sample, rise above a background level in the output. In other words, the peaks ride on a baseline bias, that changes the actual amplitudes of the peaks. This bias signal in GC technology is often called column bleed which is affected by some factors such as the temperature of the oven of the column. Several methods have been proposed as baseline correction schemes in the literature ([20–22]), which are based on either correction near any single peak or along the whole image. Of particular interest is the tophat filtering [23] used in chapter 2 as the baseline correction scheme. Tophat filtering is widely used in the image processing as a way to extract small elements and details from any given image. In $GC \times GC$ images, the small details are the small variations or the non-target analytes near the target analytes.

1.3 Cluster behaviour of $GC \times GC$ image

In any $GC \times GC$ image, there are some main target analytes and significant numbers of non-target analytes. The targets are indeed the representative of the corresponding oil sample, where these targets gather many non-target peaks around them. Figure 1.5 shows the clustering behaviour of the $GC \times GC$ image where any

target peaks have been surrounded by non-target points. Remember that in the $GC \times GC$ image, the retention time in which a peak is located is determined by its molecular mass. So, if two peaks are near to each other they probably share similar properties, e.g. they may be isomers. With this in mind, considering the points near to each other as one cluster may be a good idea to extract the forensic information locally in each cluster. We have availed this property of the $GC \times GC$ image in chapter 3 and then performed the pattern matching analysis on the clustered data-set.

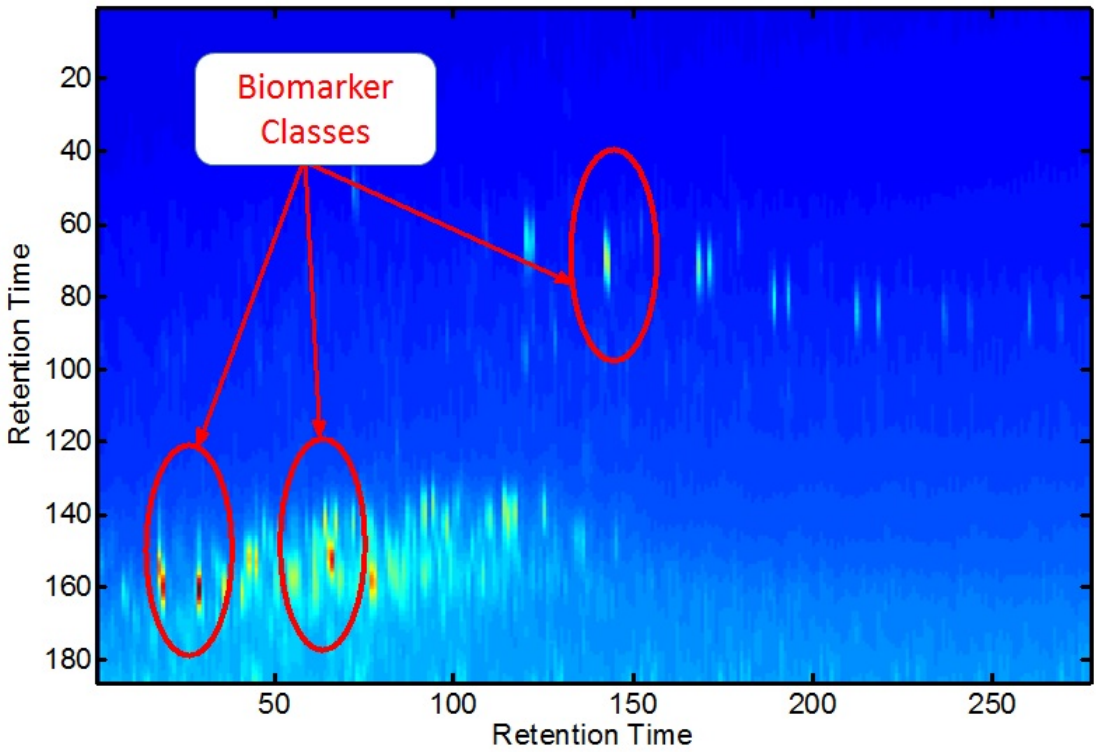

Figure 1.5: Clusters in $GC \times GC$ image.

1.3.1 Computational methods related to chemical fingerprinting

The end goal of chemical fingerprinting is to compare the chromatographic image of an unknown sample against a reference sample, e.g. a standard NIST sample from the region or a reference sample kept in laboratory repositories (library). To enable fair comparison between the samples, several computation steps need to be observed. The general approach introduced in the literature is:

- Correct the baseline in the images.
- Detect the targets and/or non-targets in each image.
- Map the targets and/or the non-targets from the reference image to those of the test image. These target or non-target points are the comparison units.
- Apply comparison metric between the corresponding comparison units in the reference and test image.

We perform the steps mentioned above for any of the to-be-tested samples against one reference sample in the library, then map those with the highest percentage of match to the reference sample in the library. In the next step we change the reference sample and extract the samples with the highest percentage of match to the new reference sample and then map these samples to the new reference sample.

For each of the steps several methods have been proposed in the literature. Reichenbach et al. [20] described a method for extracting the $GC \times GC$ baseline comprehensively. In this method a statistical method is built by tracking neighborhoods

around the smallest values as a function of time and the noise and then subtracts the background model from the data.

Detecting the peaks and mapping the corresponding peaks can be addressed as the problem of curve matching for each column of the $GC \times GC$ image. Because of the drift (variability) and baseline, the curves of the image are deviated, we may consider that the deviated curve is the transformed version of the actual curve. In this case by an inverse transform we may recover the original curve. In [24] affine transform has been proposed as a way for matching the curves. But because the $GC \times GC$ system is non-uniformly non-ideal, the transformation should be done for every peak of the image, which requires a lot of computational complexity, then by extracting the statistics of the parameters of affine transform and assigning a density function, the parameters are chosen from the constructed density function. But it is highly computationally complex, and the assumption of a linear transform may not be necessary true.

Generally, in order to have a comparison between two $GC \times GC$ images, we may first try to remove the background baseline of the image, detect the peaks or the targets and non-target analytes carrying the information about the image and then have a point to point comparison between the two images. In this comparison, we may define a similarity metric between the two images to test the match. Also the prior knowledge of the clusters within the $GC \times GC$ images could help us to introduce the target-cognizant clustering concept and look at the points occurring at each cluster.

1.3.2 Chemometric methods

Chemometric methods refer to sophisticated statistical techniques ([25] , [26]) applied to chemical data analytics and interpretation. The chemometric methods offered in the literature, cross-correlation, Principle Component Analysis (PCA) [14], Robust PCA [27], PARAFAC [28] , PARAFAC2 [29] , Independent Component Analysis (ICA) [30], WFA [31] and ITTFA [32] have been proposed in order to analyze the forensic properties, while these methods ignore the importance of the minor variations and little peaks of the $GC \times GC$ image. The minor variations is a key to identify the petroleum sources from the areas located closely to each other but from distinct sources. Chromatographic and spectroscopic signal processing techniques also fall under the realm of chemometric methodologies and are discussed in [33] and references within. In synopsis these methods include signal smoothing, signal de-noising, curve-fitting schemes, peak finding algorithms, resolution enhancement , convolution and de-convolution and other schemes commonly used in signal processing have been indicated as tools for signal analysis. Applying these methods on the chemical signal renders it ready for final analysis, where analysis usually is done for the purpose of recognizing the underlying constructive elements and thus expose hidden *patterns* within the chemical signal.

1.3.3 Principle Component Analysis

We provide a detailed synopsis of Principle Component Analysis (PCA) due to its prominence in the current state-of-the-art in chemometric analysis. PCA trans-

form uses orthogonal transformation to separate potentially correlated data points into linearly independent variables called principle components. This transformation performs such that the first principle component has the largest possible variance, and then the following principle component is constructed to be orthogonal to the previous principal component and along the direction of highest variance of the remaining data set. Mathematically speaking, suppose we have a m-dimensional feature space data-set X and we want to project it to a n-dimensional feature space data-set ($n \leq m$). We want to project the data-set into line $\vec{W}$ going through the origin. The goal is to project into the vector $\vec{W}$ in order to have the maximum variance along it, suppose $W^{(1)}$ refers to the first principle component of the dataset X, then by the definition of variance we will have:

$$var(W^{(1)T}X) = E(W^{(1)T}XX^TW^{(1)}) - E(W^{(1)T}X)E(X^TW^{(1)}) \tag{1.1}$$

$$= W^{(1)T}E(XX^T)W^{(1)} - E(W^{(1)T}X)E(X^TW^{(1)}) \tag{1.2}$$

$$= W^{(1)T}(E(XX^T) - E(X)E(X^T))W^{(1)} \tag{1.3}$$

$$= W^{(1)T}C_XW^{(1)} \tag{1.4}$$

Where C_X is the covariance matrix of the dataset X. We want to maximize the $var(W^{(1)T}X)$ under the constraints that $W^{(1)T}W^{(1)} = 1$. We construct the Lagrangian function and then calculate the optimal value for $W^{(1)}$. Taking $v(w) = var(W^{(1)T}X)$ under the constraint that $w(w) = W^{(1)T}W^{(1)} - 1 = 0$, we will have:

$$L(\lambda, W^{(1)}) = v(w) - \lambda(W^{(1)T}W^{(1)} - 1) \tag{1.5}$$

Where the λ is the Lagrangian multiplier.

Then by taking derivative with respect to w we will have:

$$\frac{\partial L}{\partial W^{(1)}} = \frac{\partial v}{\partial W^{(1)}} - \lambda \frac{\partial w}{\partial W^{(1)}} \tag{1.6}$$

From 1.5 we have: $L(\lambda, W^{(1)}) = v(w) - \lambda(W^{(1)T}W^{(1)} - 1) = W^{(1)T}W^{(1)} - \lambda(W^{(1)T}C_X W^{(1)} - 1)$, then we will have:

$$\frac{\partial u}{\partial W^{(1)}} = 2C_X W^{(1)} - 2\lambda W^{(1)}\lambda \tag{1.7}$$

which leads to

$$C_X W^{(1)} = \lambda W^{(1)} \tag{1.8}$$

or in another word the desired $W^{(1)}$ is the eigen-vector of C_x. The second principle components is calculated in the same way in which $W^{(2)}$ refers to the second largest eigenvalue of C_X and the same holds for the remaining principle components.

1.4 Existing Data Engineering Methods

In previous sections, we have provided a synopsis of existing signal processing and statistical methods related to chemical fingerprinting. However, a literature review of chemical forensics and associated data analytics is incomplete without a general review of data engineering methods. Given the data deluge in environmental

data libraries, we review the data engineering literature in three related categories: (i) Data compression, (ii) Pattern recognition and (iii) High-volume Data Querying and Indexing methods.

1.4.1 Data Compression

A vast literature exists in data compression techniques that exploit underlying patterns of the dataset or transform data to achieve compact (feature) representations and hence dimensionality reduction across high-volume datasets. Realizing the underlying patterns within chromatographic data helps to extract the most information-bearing part of the data, e.g. petroleum sample repositories may be compactly organized by extracting hydrocarbon biomarkers clusters relevant to the fingerprint of a petroleum source fingerprint. Broadly speaking, large dimensional data-sets are analyzed in two complementary ways: (i) through the analysis of the underlying distribution of the elements of the data-set, or (ii) estimating its distribution in case the distribution is not known([34, 35]). The emphasis of this thesis is more on the latter class of data analytical and compression techniques due to general dearth of knowledge of the underlying distribution of target and non-target analytes that make up the chromatographic signals.

Compression techniques such as principal component analysis (PCA) exploit principle components of data-sets to compress the data along its salient features ([36–38]) and as such, define the state-of-the-art in chemometric methods. Such compression along the principal feature spaces also serve as a mathematical framework

for fast data indexing and querying [Section 1.4.3]. A detailed discussion of PCA is given in Section 1.3.3.

Dimensionality reduction can also be posed as a machine learning problem where established supervised learning methods are employed to extract the most important data features (refer e.g. [39,40]) and threshold the others. This feature reduction ([41–43]) can be done through diverse methods ranging from simple thresholding , different quantization techniques ([44, 45]) to more sophisticated methods that exploit the internal correlation between data elements and map the whole data-set to its main components. Compact representation of any data-set and image has been offered through some popular techniques such as Karhunen-Loeve Transform([46]), Discrete Cosine Transform [17], Discrete Wavelet Transform [16] .etc.

1.4.2 Pattern Recognition

Pattern recognition schemes help the analytical chemists to find the underlying features and elements of the chemical observation which may not be easily seen just by measurement. As long as there exists a library of data-set of different chemical samples, pattern recognition helps to map a new sample to one of the entries of the library. Of main importance of the pattern recognition scheme is its capability of feature extraction where a subset of the data needed to have an accurate mapping. Hence, pattern recognition serves the chemist with the reduction of the data they need for the chemical analysis. Mapping a sample to one of the members of the library requires some sort of similarity between the tested sample and the entries

of the library. In [47] and [48] the similarity has been defined as the parameter of closeness, where a simple Euclidean distance serves as a similarity parameter. In case the points of the chemical sample is d-dimensional, then any dimension of the point can be considered as the similarity or dissimilarity factor individually [47].

1.4.3 Data Indexing and High-volume Data Query

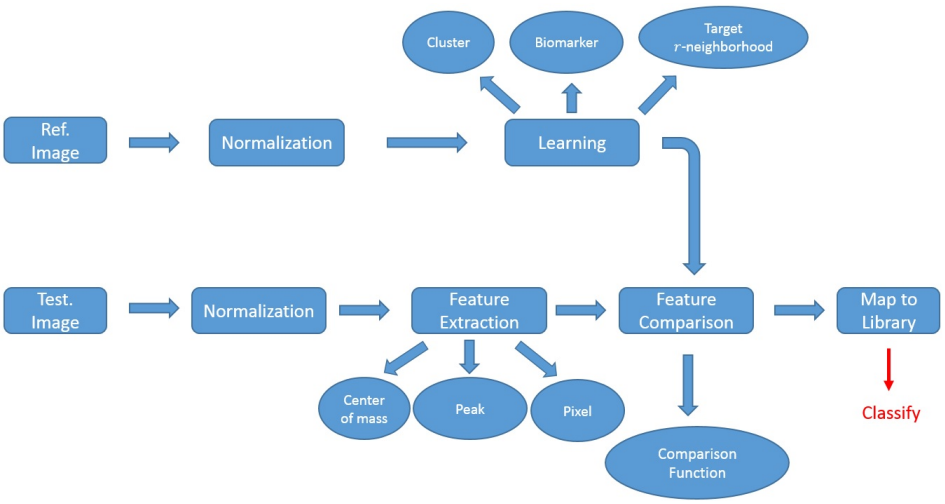

Figure 1.6: Block Diagram to classify the test image and map it to one of the entries in the library. Chemists use target biomarkers to interpret each $GC \times GC$ image, we attempt to classify the petroleum sources based upon their target and non-target analytes or clusters of targets.

Data Mining [15] can be interpreted as the process of turning the raw data into information. This information is then used for the purpose of further predic-

tion. The major task of data-mining has been addressed as six broad sub-areas [49] as class description, association, classification, clustering, time-series analysis and prediction. The class description governs the characteristics of the data. Any underlying correlation and inter-element relationship is analyzed through association method. Clustering helps to group the similar data-points of the given data-set for the purpose of classification. The prediction is done in order to find the class of a future under-test data-point and the time-series analysis is an effort to find successive measurements the sequential patterns. In this work we focus on the clustering and classification properties of data mining.

In Figure 1.6 the general approach for mining the forensic sources and then classifying them based upon their similarities to one member of the library is shown. First, from the reference image we learn biomarkers, cluster of biomarkers or the pixels within a neighborhood around the target of the image. Once a new test image is received, we extract the maximum possible features out of it and then compare its features against the features of the reference image. The features can be the center of masses of the clusters (will be covered in detail in chapter 2), or the peaks of the image or the pixels. We will further propose a comparing function in chapter 3 in order to compare the features of the test and reference image.

1.5 New Approach in Environmental Data Query and Indexing

Query in the environmental forensic sources have been done through the analysis of the biomarkers within their image. The target analytes are the analytes that

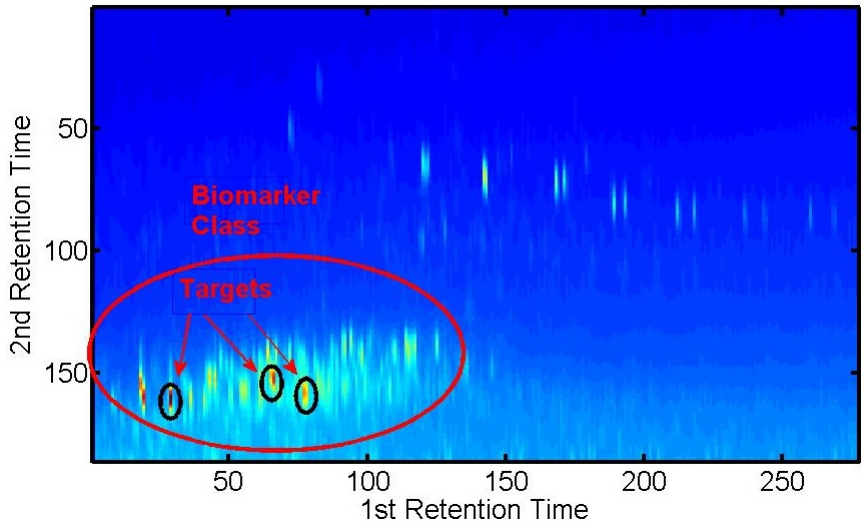

Figure 1.7: The figure is a two-dimensional $GC \times GC$ image regarding to one petroleum source from the Macondo area. The targets are a sub-set of the peaks in biomarker classes in the image. The characteristics of the petroleum source is carried by the targets.

the chemists look for them in order to distinguish the forensic sources. In a higher resolution scale we look at the both the target and the non-target analytes to well separate the forensic source. So query and indexing the forensic sources can be done in biomarker analysis level, target level and in a very high resolution level which is the target and non-target analysis and then index the forensic source based on these analytes. Indexing the forensic source based on targets and non-targets can be considered as one of the main contributions of this work.

1.6 Synopsis of Proposed Research

The organization of this thesis is as follows.

Chapter 2 discusses the similarity criterion for the recognition and comparison of chemical images and introduces a method to remove the baseline regarding the images. Chapter 3 discusses in detail the main algorithmic innovations in this thesis, which combines target cognizant clustering (TCC) methods [50], with local interpretation over the raw signal using Target Neighborhood Analyzer (TNA). PTM provides target cognizant interpretation as it retains peak information, TCC looks at the center of masses of the clusters of the targets, and the TNA method a target-centric method where neighborhoods within a definite radius of the main targets are seen. A combination of TCC and TNA method which has been addressed in the proposed work as another way of looking at the chemical images in which avail the benefits of both methods, TCC and TNA. These methods represent a target cognizant clustering, compression, indexing and querying of high volume data-set such as $GC \times GC$, and $GC - MS$ data-sets. Of particular interest, we have adopted linkage clustering scheme [51].

The proposed methods in Chapter 3 are focused on data compression where clustering may reduce the dimensionality of the data-set considerably. However, as detailed in Chapter 4, these methods may also be employed to achieve efficient target-cognizant data mining over high-volume environmental datasets. In chapter 4, we increase the analysis resolution and try to find a way to analyze the chemical images based upon individual elements or peaks. In chapter 5, we try to combine the already-

existent data compression methods like SAX and design a systematic method to go towards the pattern recognition goal. Finally, in chapter 6, we try to go deeply through the comprehensive study of the peaks topography map, a detailed method of chemical pattern recognition. We also introduce the notion of robust peaks in which the analysis of the chemical images can be done via these peaks and not the others, mainly because these peaks exist in all of the images extracted from one specific geographical region.

In summary, we propose a novel approach to chemical data analytics and interpretation that bridges the gap between target-cognizant traditional analysis from environmental chemistry with compound-agnostic computational methods in chemometric data engineering. Specifically, we propose computational methods for target-cognizant data analytics that also account for local unknown analytes allied to the established target peaks. The key intuition behind our methods are based on the underlying topography of the gas chromatographic landscape, and we extend recent peak mapping methods as well as propose novel peak clustering and peak neighborhood allocation methods to achieve our data analytic aims.

CHAPTER 2
TARGET COGNIZANT CLUSTERING AND INTERPRETING LOCAL NEIGHBORHOODS IN GAS CHROMATOGRAPHIC IMAGES

2.1 Introduction

In this chapter we study clustering behaviour among target and non-target analytes within gas chromatographic datasets, with special emphasis on $GC \times GC$ topography of petroleum biomarkers. As discussed in Section 1.3, the well-known target analytes of the $GC \times GC$ image dominate the topography but also cohabitate the chromatographic landscape with a larger group of unknown non-target analytes. In this chapter, we wish to study in particular, among target and non-target analytes as they may be indicators of a common factor, e.g. a region-specific fingerprint, co-indicative PCB congeners, ec. As also discussed in Chapter 1, the time in which a peak arrives at the end of the $GC \times GC$ system relates directly to the chemical properties of the corresponding chemical compound. With this in mind, peaks occurring near to each other may represent similar classes of compounds and therefore, may exhibit group behavior regarding their source of origin, impact on public health, environmental dependencies and other important factors.

2.2 Data Compression Challenges Unique to Environmental Forensics

We explore data compression as a grand challenge at the intersection of high-volume data analytics and human-environment interactions, with petroleum forensics as a relevant and exemplary field application. Our goal is to meet the growing practi-

cal necessity for reliable high-rate compression of environmental data-sets in a manner that yields robust yet compact forensic interpretation. To achieve this end, we derive compression techniques that are:

- compound-cognizant, thus preserving well-known forensic markers
- statistically robust, i.e., employs the full informative power of rich intricate datasets.

A practical example where such compression products are of critical need are coastal surveillance vessels that sample industrial spills and leaks. These surveillance missions are challenged with legal mandates to test for well-known labeled analytes and against closely related sources, and yet have no real-time access to large data libraries to ensure robust forensic match (or lack thereof). Furthermore, most data compression and chemometric techniques that provide fast indexing and querying of large environmental databases are, by design, agnostic of target analytes, which are labeled compounds (forensic markers) that dominate the data topography across diverse separation technologies, e.g. gas and liquid chromatography (GC, GC - GC, LC, LC-LC), mass spectrometry, and combinations thereof (GC-MS, GC-LC/MS, etc.). Our goal in this proposed work is two-fold: (i) Bridge the duality in environmental forensics literature between target-cognizant forensic chemistry and compound agnostic chemometrics; and (ii) Address the competing challenges of high-volume data interpretation and target-cognizant forensics from a compression perspective. Specifically, our intellectual contributions are summarized as follows: (i) We develop target-driven

local features that compresses high-volume environmental datasets along compound-cognizant clusters. An important benefit of this target-driven compression allows fast indexing and querying of high-volume data for a single biomarker class (e.g. hopanes, steranes, diasteranes) without needing to interpret the full data archive in a field test. (ii) We enable fast and potentially real-time forensic interpretation that can match an unknown sample against arbitrarily high-dimensional datasets, compressed along target-cognizant clusters. We achieve this by local interpretation of biomarker topography, across target and non-target peaks within the target's neighborhood.

2.3 Technical Approach

Our technical approach in this chapter may be summarized as two complementary algorithms: (i) Target-Cognizant Clustering algorithm, detailed in Section 2.3.1, and (ii) Target Neighborhood Analysis, detailed in Section 2.3.2. Our proposed approach has been detailed with two-dimensional gas chromatography ($GC \times GC$), but is easily extensible to other forms of environmental data that exhibit high-dimensional peak profiles (e.g. LC-LC, GC-MS, and others).

2.3.1 Target-cognizant Clustering

We introduce a target-cognizant compression technique that drastically reduces data dimensionality by clustering the data into primary target neighborhoods. The key idea is to cluster the information inherent in a $GC \times GC$ dataset along forensic markers, hopanes and steranes, which manifest as the dominant peaks in the biomarker topography. We algorithmically detect the primary targets within a given

$GC \times GC$ image I by employing peak-thresholding with a threshold τ. Mathematically, we construct the sub-image $I^{'}$ such that:

$$I^{'} = I(I \geq \tau)$$

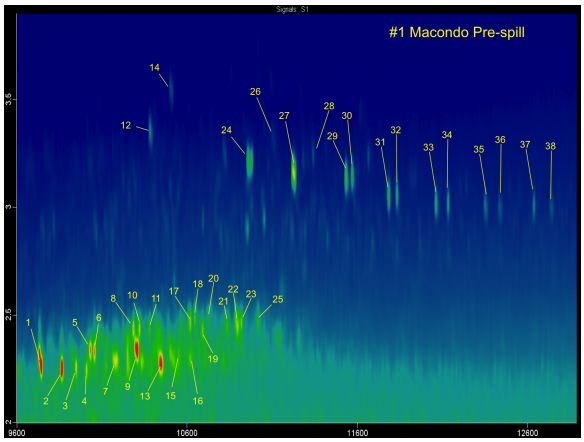

Figure 2.1: $GC \times GC$ separation of petroleum biomarkers (hopanes and steranes) for a pre-spill crude oil sample taken from the Macondo well, site of the *Deepwater Horizon* disaster in the Gulf of Mexico, April 2010. Target biomarkers are labeled numerically.

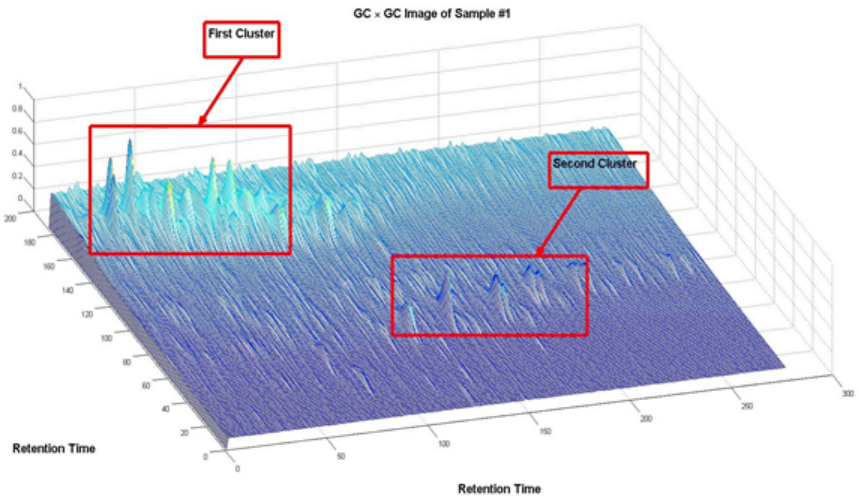

Figure 2.2: $GC \times GC$ image in Figure 1.5 split into two main clusters using single-linkage clustering. The chosen clusters roughly correspond to two categories of biomarkers, hopanes and steranes.

We note that this step may be replaced by manual labeling in the event that some targets may fall below the assigned peak threshold τ. However, for practical datasets, simple thresholding isolates a majority of well-known targets though it leaves out most non-targets crucial to robust forensic interpretation. We then employ single-linkage clustering [13] over the selected targets to discover target classes that group together in potentially overlapping neighborhoods. Our intuition is that target biomakers that share Euclidean proximity within the $GC \times GC$ topography likely share chemical inter-relationships (e.g. isomer groups) and thus designing feature compression based on target clusters will derive precise forensic information from a drastically reduced dataset. Figure 2.2 shows the two main clusters for Figure 2.1,

which as expected, fall along two primary classes of compounds, hopanes and steranes. Generally speaking,N clusters separated using single-linkage clustering based on the target biomakers folds the $GC \times GC$ topography into N sub-classes of target and non-target analytes, each of which are anchored to a specific target neighborhood. Mathematically, we may break up the $GC \times GC$ image I as a disjoint union of N target clusters.

$$I = \cup_{n=1}^{N} I_n \tag{2.1}$$

where I_n denotes the n^{th} cluster of target analytes, representing a local neighborhood for unique forensic interpretation. We note that Equation 2.1 also provides the basis for compression of a large $GC \times GC$ image into neighborhoods anchored to $\{I_n\}_{n=1}^{N}$ and local interpretation and querying is pivoted to individual neighborhoods (Section 2.3.2 for details). Specifically, we achieve a compression rate of C_n given by the ratio of pixels in I to the ratio of pixels in each target cluster I_n, i.e.

$$C_n = \frac{|I|}{|I_n|} \tag{2.2}$$

where $|\cdot|$ denotes number of pixels within the given image.

A related question that naturally arises is: *How to robustly compare between target clusters?* Accordingly, we define the $\Delta(I_j, I_k)$, the distance between the center of mass of the two clusters, as the geometric metric to facilitate inter-cluster comparisons. Mathematically, $\Delta(I_j, I_k)$ is given as:

$$\Delta(I_j, I_k) = ||c_j - c_k||_2^2, \tag{2.3}$$

where c_j and c_k are the centers of mass of the clusters[1] of I_j and I_k respectively. For forensic analysis, we may use $\Delta(\cdot)$ as the distance metric to match two samples over the same target cluster, and thus provide robust local interpretation. Our goals for local forensic interpretation are to meet the complimentary needs of target cognizance and high precision over highly compressed datasets. Specifically we pursue target-anchored forensic matching for an unknown field sample against large data archives that are compressed along the $\{I_n\}_{n=1}^{N}$ clusters. Figure 2.4 in Section 2.4 provides field data-driven results for the effectiveness of this method, both in terms of compression ratio and forensic precision. Pseudo-code for the target-cognizant clustering is given below.

---

[1]The center of mass of a cluster is given by:

$$gcm_j = \frac{\sum_{i=1}^{m} h_{i,j}.l_{i,j}}{\sum_{i=1}^{m} h_{i,j}}, j \in 1, 2, 3, ..., N \tag{2.4}$$

where N is the number of clusters, m is the number of points within a cluster, and h_i and l_i are the amplitude and the location of the i-th point in that cluster, respectively.

Algorithm 2.1 Target-Cognizant Clustering (TCC)

☐ **Input:**

∗ Reference Image I_{ref}.

∗ Test Image I_{test}.

∗ N (number of clusters).

∗ τ (threshold).

☐ **Output:**

∗ Difference between the two input images

☐ **Step0:**

∗ Apply a peak threshold on the reference image: $I^{'}_{ref} = I_{ref}(I_{ref} \geq \tau)$.

☐ **Step1:**

∗ Map the thresholded reference image into couple of clusters using single linkage clustering: $I^{'}_{ref} = \bigcup_{k=1}^{N} I^{'k}_{ref}$.

☐ **Step2:**

∗ Construct the clusters in the test image I_{test} at locations given by the clusters of I_{ref} (Look at Figure 2.3).

☐ **Step3:**

∗ Compute the center of mass of each cluster using Equation 2.4 in both I_{ref} and I_{test}.

☐ **Step4:**

∗ Calculate the Euclidean distance of center of masses of clusters as dissimilarity criterion (Implement equation 2.3):

Dissimilarity Score = $\sum_{i=1}^{N} \Delta(I^{i}_{ref}, I^{i}_{test})$,

∗ **return** Dissimilarity Score.

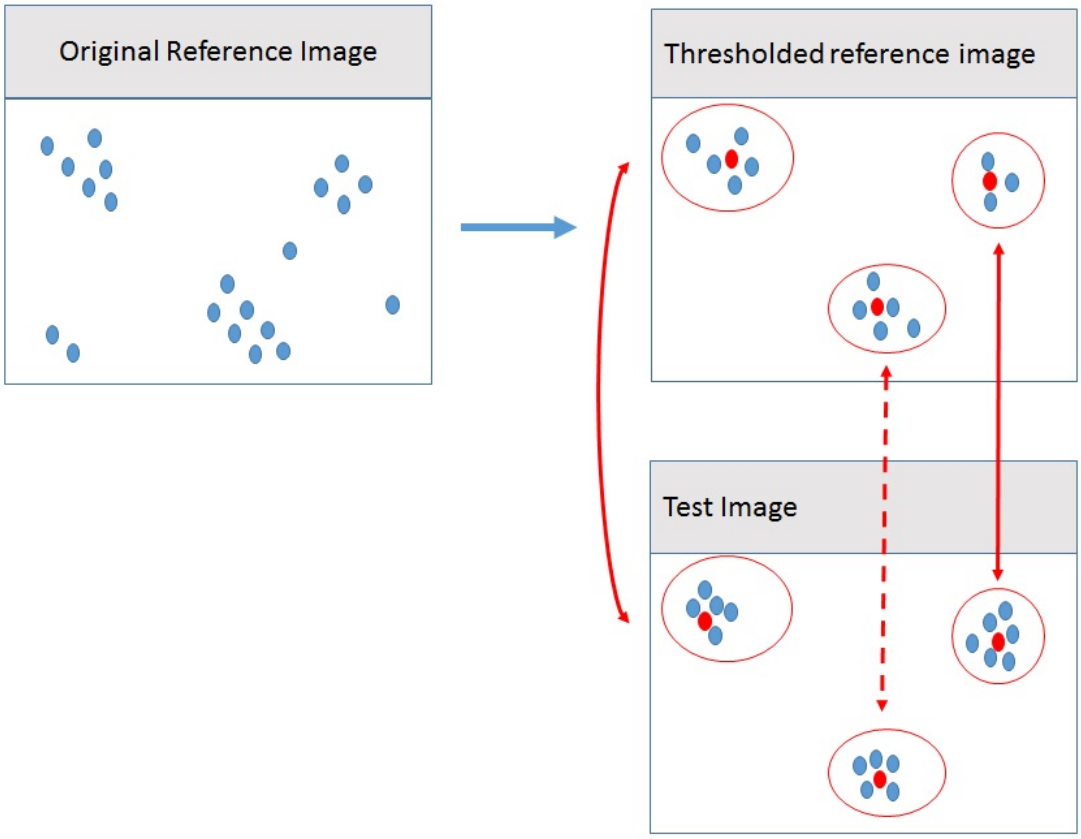

Figure 2.3: The top left figure shows the original reference image. First, the image is thresholded so that the clusters of the image are well-separated. After applying single-linkage clustering, the boundaries of the clusters are determined. In order to compare a new test image with reference image, TCC algorithm just looks at the location of clusters computed in the thresholded reference image and then maps the clusters to the center of mass. The red boxes show the center of masses of clusters.

2.3.2 Target Neighborhood Analyzer (TNA)

Despite representational simplicity of target-cognizant clustering, the high compression rate is achieved at a cost of information loss. This information loss is fundamentally due to completely ignoring the statistical significance of non-target analytes that constitute the majority of GC×GC topography. Therefore, forensic interpretation derived purely based on target-driven clustering inevitably suffers from high false alarm rates when comparing between closely correlated contaminant sources that share similar proportions of target biomarkers. To remedy the loss of non-target information, we augment target-based clustering with feature encoding along the local neighborhood of each target peak. The goal in this augmented clustering technique is to aid statistically robust indexing, querying and cross-sample comparisons albeit at a slightly higher compression rate. We also include traditional peak-ratio comparisons within each feature neighborhood to facilitate comparisons against the current state-of-the-art in environmental forensics.

Mathematically speaking, we construct the r-neighborhood around each chosen target T_i such that $T_i \in I_n$, where I_n is one of the target clusters chosen using target-cognizant clustering. Thus, the sub-image $\tilde{I}(r, n)$ considered for forensic interpretation will be given by the sub-set of the $GC \times GC$ image I that contains the n^{th} target cluster *as well* as the local topography around it within a neighborhood of radius r centered on the target peak. The compression ratio C_n for the n^{th} cluster is now given by:

$$C_n = \frac{|I|}{|\tilde{I}(r,n)|}, \tag{2.5}$$

where $|\cdot|$ denotes the number of pixels within the given image. The overall compression ratio is given by:

$$C_n = \frac{|I|}{|\cup_{n=1}^{N} \tilde{I}(r,n)|}, \tag{2.6}$$

where N is the total number of clusters obtained from target-cognizant clustering. To compare between two samples with GC×GC images $I^{(1)}$ and $I^{(2)}$, we proceed as follows:

1. **Step 1:** For each cluster n, consider the sub-images $\{I_n^{(m)}\}_{m=1}^2$. For each pixel location $(r_1, , r_2) \in \cap_{m=1}^2 I^{(m)}$, determine the peak-ratio

$$\rho(\{I_n^{(m)}\}_{m=1}^2, r_1, r_2) = \max\left(\frac{I_n^{(1)}(r_1,r_2)}{I_n^{(2)}(r_1,r_2)}, \frac{I_n^{(2)}(r_1,r_2)}{I_n^{(1)}(r_1,r_2)}\right), \tag{2.7}$$

where $\{I_n^{(m)}(r_1,r_2)\}_{m=1}^2$ are the pixel intensities at the location (r_1,r_2) for both sub-images $\{I_n^{(m)}\}_{m=1}^2$.

2. **Step 2:** Calculate the similarity factor $S(I_n^{(1)}, I_n^{(2)})$ as the forensic match metric of $I_n^{(1)}$ with respect to $I_n^{(2)}$ given as:

$$S(I_n^{(1)}, I_n^{(2)}) = \frac{\sum_{(r_1,r_2):\rho(\{I_n^{(m)}\}_{m=1}^2, r_1, r_2) \geq \rho_\tau} I_n^{(1)}(r_1,r_2)}{\sum_{(r_1,r_2)} I_n^{(1)}(r_1,r_2)}, \text{ where } (r_1,r_2) \in \cap_{m=1}^2 I_n^{(m)}, \tag{2.8}$$

where ρ_τ is the tolerance threshold.

$S(I_n^{(1)}, I_n^{(2)}) = 1$ denotes perfect match between ($I_n^{(1)}$ and $I_n^{(2)}$. It is easy to verify that $S(\cdot)$ is commutative, i.e., $S(I_n^{(1)}, I_n^{(2)}) = S(I_n^{(2)}, I_n^{(1)})$ since we only consider $(r_1, r_2) \in \cap_{m=1}^{2} I_n^{(m)}$ and $\rho(\{I_n^{(m)}\}_{m=1}^{2}, r_1, r_2)$ is commutative by design. It is also easy to verify that the number of pixels $\in \cap_{m=1}^{2} I_n^{(m)}$ considered to calculate $S(\cdot)$, denoted as $\kappa(I_n^{(1)}, I_n^{(2)})$, is upper-bounded by the total number of pixels for the r-neighborhood of each target peak, given by the $(2r+1)^2$-square which has an r-radius in-circle. Mathematically, this may be expressed as:

$$\kappa(I_n^{(1)}, I_n^{(2)}) \leq \Gamma(I_n^{(1)}, I_n^{(2)}) \times (2r+1)^2, \tag{2.9}$$

where $\Gamma(I_n^{(1)}, I_n^{(2)})$ is the number of target peaks common to $I_n^{(1)}$ and $I_n^{(2)}$.

The above algorithm is setup for target-cognizant matching between $I^{(1)}$ and $I^{(2)}$ across the n^{th} cluster. We can setup the target-cognizant clustering approach to potentially enable hierarchical clustering across classes and sub-classes of target analytes and perform forensic matching based on the similarity factor $S(\cdot)$ between any two clusters at a given level between two GC×GC images. However, the similarity factor $S(\cdot)$ is not limited by the cluster size and is scalable across the cluster hierarchy. To combine the forensic match score across N clusters at the same level of hierarchy, we simply change the scope of $(r_1, r_2) \in \cap_{m=1}^{2} I_n^{(m)}$ to $(r_1, r_2) \in \cup_{n=1}^{N} \left(\cap_{m=1}^{2} I_n^{(m)}\right)$.

The key benefit of this hierarchical target allocation is that it enables compression, indexing, querying and subsequently forensic interpretation at multiple scales within high-dimensional data. For this reason, the TNA method was originally presented as Hierarchical Target Allocation (HTA) in [52], the first publication related to this work. However, as we have been discovering over extensive data analysis, the

potential for hierarchical clustering, though computationally elegant, leaves out the important scenarios where target clusters representing different elements of the fingerprint may indeed overlap without constituting a hierarchy. Thus imposing artificial hierarchy across the target clusters not only limits the robustness of interpretation but may mislead the scientific investigation in key scenarios such as discovering the locally overlapping fingerprint of neighboring oil reservoirs, or say, the linking maternal age and environmental factors to PCBs in breastmilk. On the other hand, not imposing hierarchy between target clusters still leads to efficient not sub-optimal data compression, indexing and querying possibilities. Pseudo-code implementation of the TNA algorithm is given in the next page.

2.4 Results

We perform the comparison between different injections via the scores given by TCC and TNA algorithms. Figures 2.4 and 2.6 show the percentage of match, Cross-TCC score and Cross-TNA scores. As can be seen from Figure 2.6 the TNA algorithm has done a very good job in identifying the samples from Macondo area.

Algorithm 2.2 Target Neighborhood Analyzer (TNA)

☐ **Input:**

∗ Reference Image I_{ref}.

∗ Test Image I_{test}.

∗ r (neighborhood radius).

∗ ρ_τ (threshold).

☐ **Output:**

∗ Similarity between the input graphs

☐ **Step0:**

∗ Allocate the main targets of the reference image as the highest peaks (or manually label the N main targets):

$I^{'} = \cup_i T_i = I_{ref}(\text{top } N \text{ peaks}) \ , i = 1, 2, 3, ..., N.$

☐ **Step1:**

∗ Construct the r-neighbourhoods by a determined neighborhood within the main targets in the reference image:

$I^{''k}_{ref} = \bigcup_{i=r_1-r}^{r_1+r} \bigcup_{j=r_2-r}^{r_2+r} I^{'}(r_1+i, r_2+j)$ ($k^{th}(1 \leq k \leq N)$ r-neighborhood in I_{ref} using the given ρ_τ and r around the main target location with the retention time of (r_1, r_2)).

☐ **Step2:**

∗ Construct the r-neighborhoods in the test image I_{test} at locations given by the neighborhoods of I_{ref}.

☐ **Step3:**

∗ Compute the Similarity Score for each of the corresponding r-neighborhoods. (Implementing Equation 2.8), call it $SimScore$.

∗ **return** $SimScore$.

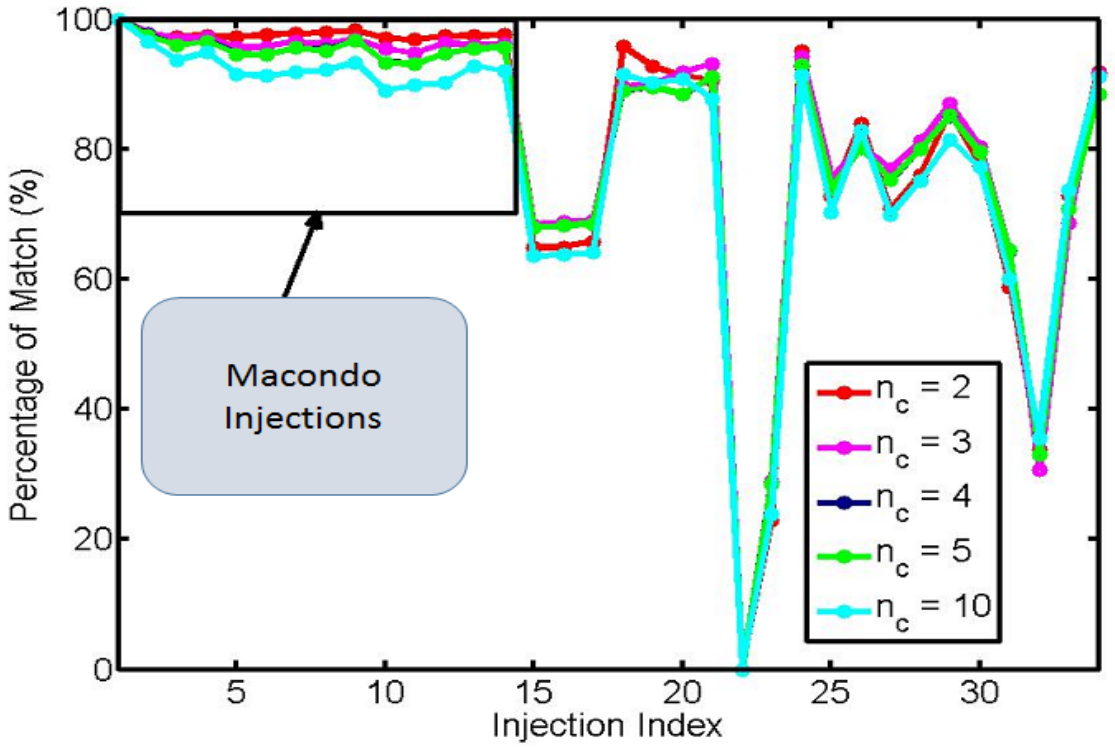

Figure 2.4: Cross-TCC score. In this figure Sample 1 from Macondo area has been set as the reference sample and the other samples have been compared against it. The plot is shown for different choices of number of clusters where clusters have been constructed using the single linkage clustering. The peak threshold is set to 0.2.

As Figure 2.5 suggests, the choice of two for the number of clusters is the best for the given dataset. The minimum percentage of match from one Macondo sample to the reference sample is 97.52% and the maximum percentage of match of one non-Macondo sample to the reference sample is 95.71%, so with choice of 2 for the number of clusters the Macondo samples can be detected from non-Macondo injections.

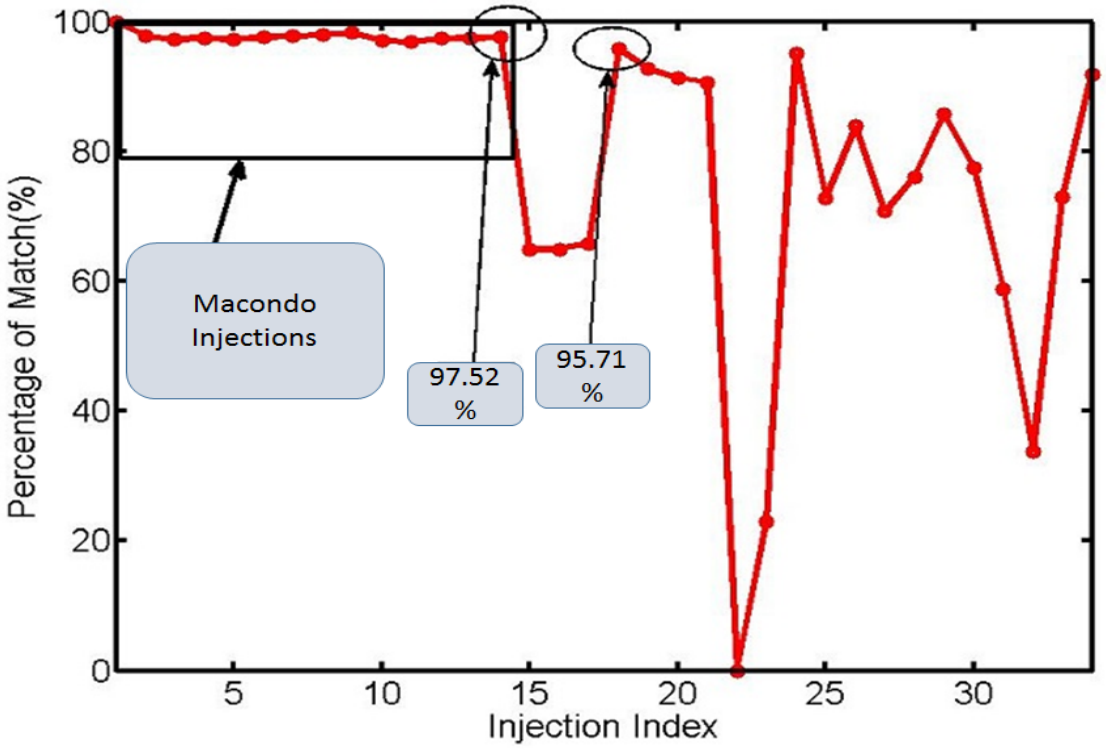

Figure 2.5: Cross-TCC score. In this figure Sample 1 from Macondo area has been set as the reference sample and the other samples have been compared against it. The plot is shown for different choices of number of clusters where clusters have been constructed using the single linkage clustering. The peak threshold is set to 0.2.

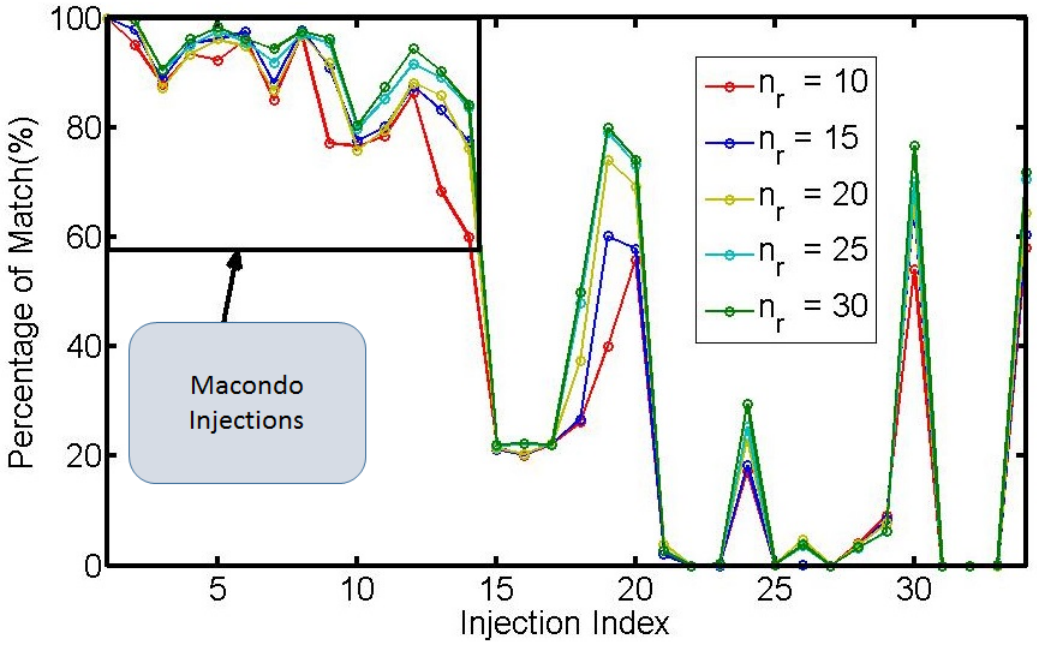

Figure 2.6: Cross-TNA score. In this figure Sample 1 from Macondo area has been set as the reference sample and the other samples have been compared against it. The plot is shown for different choices of number of $r - neighborhoods$ (n_r) where it has been evaluated at $r = 5$.

To have a better illustration of the TNA method, we can look (Figure 2.7) at the choice of thirty for the number of neighborhoods already shown with other choices of number clusters in Figure 2.6.

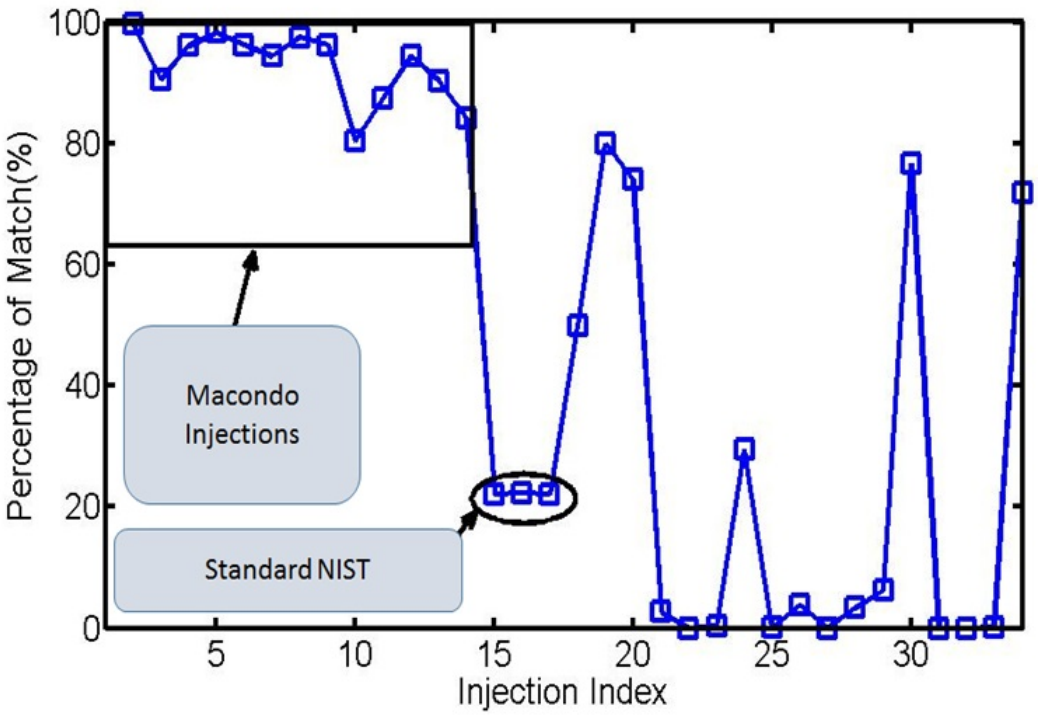

Figure 2.7: Cross-TNA score for the choice of thirty for the number of neighborhoods shown in Figure 2.6.

2.4.1 Space-saving achieved by TCC and TNA

We test the performance of our algorithm in terms of compression, i.e. the number of the points needed from the reference image to test against the test samples to detect those similar to it or from the same area. The number of the points used in the TCC algorithm is a function of the peak threshold imposed in the reference image at the start of the algorithm. As we apply higher thresholds we will use less number of the target and non-target analytes from the reference image in TCC algorithm, but the risk of miss-detection of the samples similar to the reference image increases.

In the TNA algorithm also the number of the points used from the reference image is proportional to the number of the neighborhoods and the radius of the neighborhoods. As we saw earlier, as the neighborhood radius increases in the TNA algorithm, the samples similar to the reference image increases and the distinction between samples are better seen. It is natural because as we use more number of target or non-target analytes we will take more pixels into consideration and this should lead to a better distinction among the forensic injections. If we could achieve our goal, which is the robust separation with less number of points for comparison, we have saved computational complexity where computation here means point to point comparison.

We define the amount of Space-saving as:

$$Space - saving = 1 - \frac{n}{N} \tag{2.10}$$

where n is the number of points needed for comparison from the reference image for the accurate separation and distinction of forensic sources with respect to it, and N is the total number of the points in the reference image. In Figure 2.8 we have plotted the amount of the space-saving achieved by TNA (on the left) and TCC (on the right). The number of the points needed for comparison in the TNA method is a function of the (number of neighborhoods)n_r and their radius (r). Hence, space-saving is a decreasing function with respect to these terms. In Figure 2.8 for TNA we have set the radius of neighborhood to five and plotted the amount of the space-saving with respect to the number of neighborhoods. The figure labeled as Upper-bound refers to the maximum number of points needed from Equation 2.9 and

Experiment which refers to the amount of unique points needed. The difference in the number of the points needed in the experiment case and the case given by the upper-bound is that many points are common between the neighborhoods that should be considered once. Additionally, some of the points have zero amplitude at both of the reference and test images that should not be considered as computation. In TCC the space-saving is a function of the peak threshold and in Figure 2.8 on the right we have plotted the space-savings for the values between 0.2 to 0.52. The value of 0.52 for the peak threshold is the maximum amount of the threshold that we can apply to the reference image and still have an accurate distinction between forensic injections by comparing the remaining points against the test image such that we record those samples similar to the reference accurately. Once we apply a higher peak threshold the separation and identification is not done accurately anymore. In the plot for the space-savings of TCC we have set the number of clusters equal to two.

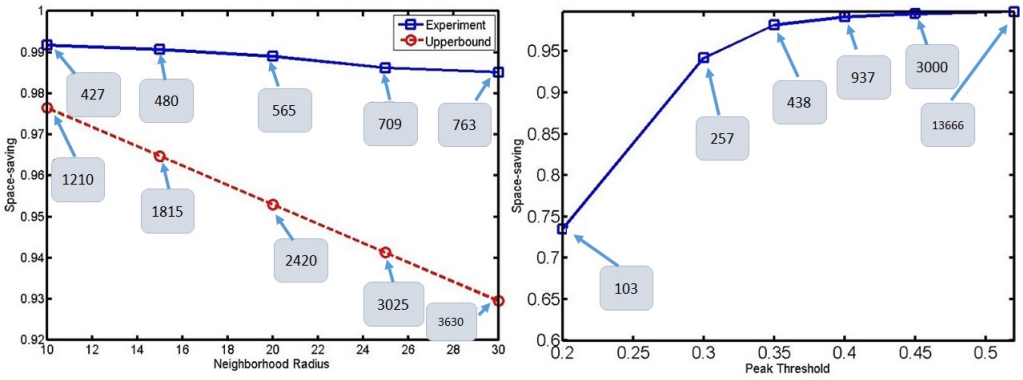

Figure 2.8: Amount of space-saving achieved by TCC and TNA. The amount of space-saving in TCC is a function of the peak threshold applied at the first stage of the algorithm to distinguish the clusters. The space-saving in TNA is a function of the r-neighborhood. The numbers recorded in the images refer to the number of points needed from the reference image to compare against test samples in order to accurately distinguish those that are the same as the reference image.

2.5 Creating simulated images in order to increase the dimensionality of the dataset

In order to test the robustness of the proposed TCC and TNA methods, we may need to have a dataset with a higher number of images. In fact as the proposed work in chapter 4, we will test the methods with different datasets, but we could also create simulated datasets just by adding random noise to one of the images in the dataset and create as many simulated images as we wish. We can add noise, both to the location and the amplitude of the peaks in the $GC \times GC$ image. The way we

tested the performance of our methods is that we first set one sample as the reference sample, say one sample from the Macondo injections. Then we take the average of the remaining thirteen samples from the Macondo area and then add noise to it in order to create as many simulated images as we wish, say two hundred samples with some noise standard deviation. Then we treat these simulated samples as simulated Macondo samples. Hence, we will have two hundred and thirteen Macondo samples to compute their TCC or TNA scores, and then record the average of the percentage of match of Macondo injections. In the next step, we set another sample from the Macondo area as the reference, say second sample, and go through the same steps while creating simulated images from the averaged images of samples one and samples three to fourteen. Again record the average of the percentage of match of Macondo injections. We will do the same steps for all of the fourteen injections, so we will have a vector with fourteen entries where the k^{th} entry of this vector is the average of the percentage of the Macondo injections to the k^{th} Macondo sample. Now, we change the value set for the noise standard deviation and compute the same vector for that standard deviation. Figure 2.9 and Figure 2.10 shows the average TCC, TNA and PCA scores, where in Figure 2.9 all of the three scores decrease as the standard deviation of the noise increases but in Figure 2.10 the average TCC score increases as the standard deviation of noise increases. In the proposed work we will also analyze this situation where TCC increases in the percentage of match even if the standard deviation of noise increases.

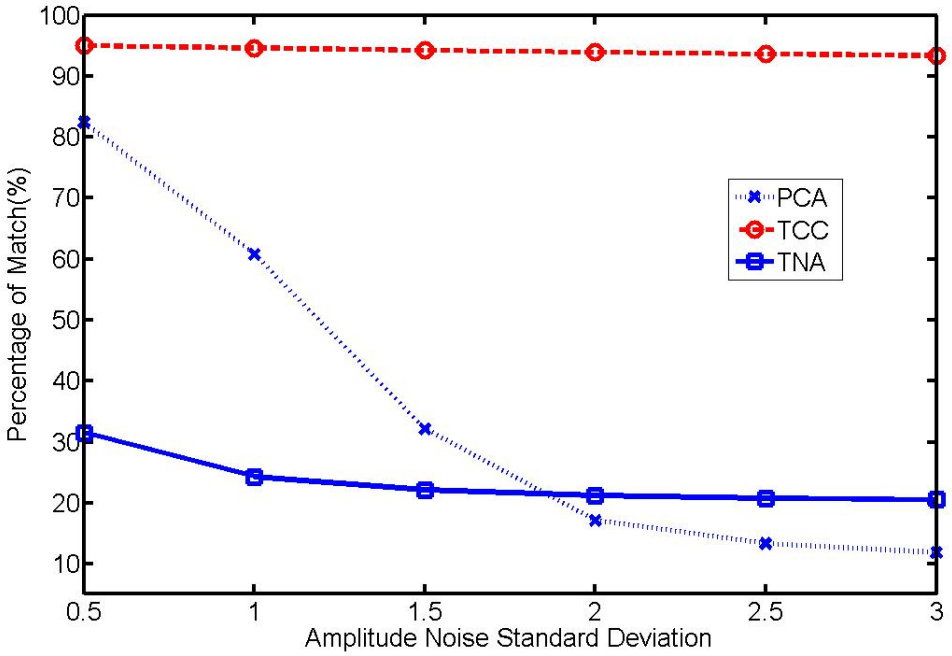

Figure 2.9: Average percentage of match of the Macondo samples using simulated samples by adding noise to the amplitude of the test images and creating a larger data-set. The number of injections created by the simulation is two hundred images.

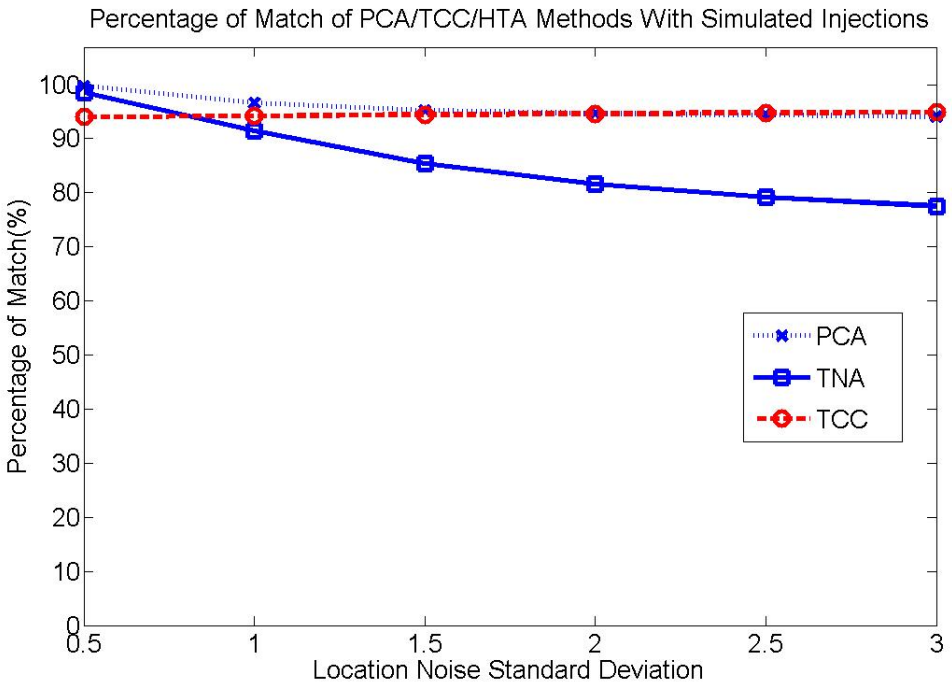

Figure 2.10: Average percentage of match of the Macondo samples using simulated samples by adding noise to the location of the test images and creating a larger data-set. The number of injections created by the simulation is two hundred images.

CHAPTER 3
COMPRESSED FORENSIC SOURCE IMAGE USING SOURCE PATTERN MAP

3.1 Background Motivation

Forensic source differentiation between marine oil samples after major oil spills (e.g. Deepwater Horizon spill, Gulf of Mexico, April 2010) is not only fundamental to environmental monitoring, but also a daunting data compression and signal processing challenge. This is primarily due to three related factors: (i) High-volume data generated from analyzing petroleum samples from different industrial oil reservoirs, (ii) Petroleum fingerprinting relying heavily on interpreting joint biomarker distributions that carry overlapping fingerprints specific to the locale and the individual reservoir, (ii) Lack of robust disambiguation techniques between the highly correlated fingerprints of neighboring reservoirs. As such, no known method exists to robustly disambiguate source-specific information against correlated interference from regional characteristics. Therefore, a vast amount of data repository space is wasted across numerous commercial and national petroleum laboratories storing redundant region-specific biomarker information that is shared across thousands of oil reservoirs common to a locale. Beyond highly inefficient data storage, source-agnostic data analytics (i) mislead forensic interpretation in the aftermath of major oil spills, (ii) render offline in-situ comparison of field samples against known sources impossible due to the highly redundant data volume, and (iii) renders high-speed indexing and querying across large petroleum databases impractical. Harnessing pattern discov-

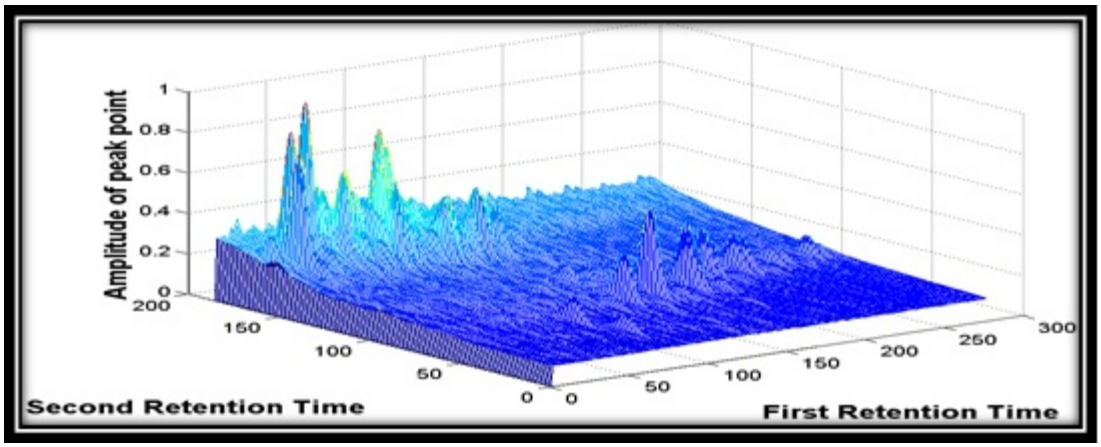

Figure 3.1: Two-dimensional Gas Chromatography related to one oil sample.

ery, data compression and associated learning techniques to drastically compact these highly redundant datasets, while preserving key chemical signatures of the petroleum biomarkers in the motivation and scope of this work [53].

Two-dimensional Gas Chromatography ($GC \times GC$) provides the current state-of-the-art in analytical resolution of a complex molecular mixture such as petroleum into constituent components, of which the biomarkers (hopanes and steranes) are special compounds that host the regional and source-specific fingerprint. This technology generates a two-dimensional time-series of peaks, each time-series representing compound resolution along independent instrument columns, which are jointly rendered as a high-resolution peak distribution image or a three-dimensional surface with the two time-series as the two dimensions, and peak magnitude constituting the third dimension (refer Figure 3.1).

3.1.1 Data Compression Challenges

The $GC \times GC$ signal for each oil sample consists of an overlapping spread of hundreds, if not thousands, of biomarker peaks, which roughly follow a locally Gaussian peak shape. When we consider the reality that one oil reservoir (source) can have hundreds of thousands of samples in one data repository, and that hundreds of data repositories that share data across hundreds of highly correlated neighboring sources, the magnitude of data deluge and the compelling need for compact data representation, signal separation, and efficient storage, analysis, indexing and querying is apparent. However, blind application of existing data compression techniques (e.g. PCA, ICA) will not achieve the data engineering goals as it is critical to preserve the individual identity of the biomarker compounds in the compressed and classified end-product to allow a human expert (e.g. an EPA agent) to physically interpret and validate the source-specific data fingerprint.

3.1.2 Key contributions

One of the most interesting problems related to these $GC \times GC$ images are learning the common pattern between those images corresponding to the samples extracted from the same area, and then use the common pattern to compare it against the patterns of the other areas. A $GC \times GC$ image can have a pretty large size, if we can learn the common pattern specific to one source, then we can just save the common pattern, as the features of the area and then throw away the remaining peaks. So this process can lead to a compressed version of the image, which is the

basis for this work. Our objective is to compact high-resolution intricate $GC \times GC$ images along domain-specific (source or locale) class boundaries. Accordingly, we propose two key innovations:

- Derive and localize biomarker-cognizant source-specific features: We achieve this by localizing biomarker peaks in peak topography maps using peakratio thresholding.

- Compact source information along compound-cognizant peak dictionaries: We achieve this by constructing and classifying peakratio threshold maps that compact the $GC \times GC$ image along source-specific patterns, represented as highly compact peak dictionaries uniquely specifying a source. We distinguish between biomarker and compound cognizance as not all classified compounds fall into well-established biomarker dictionaries.

3.2 Compression as a pattern recognition problem

For a library $D = \{I_1^{k_1}, I_2^{k_2}, I_3^{k_3}, \ldots, I_K^{k_K}\}$, where $I_i^k (1 \leq i \leq K)$ means there are k number of images for the $GC \times GC$ image of the region indexed by i, available in the library. Note that these k images are not exactly the same. This can have a couple of reasons, first the $GC \times GC$ images are achieved after injecting oil sample through $GC \times GC$ system. This $GC \times GC$ system carries some noise within its internal physical elements. This noise can potentially have a random behavior, so if we inject the same samples at two different times, we may get two images at the output which may not be exactly the same. Secondly, oil samples can be extracted from the same

area in different time intervals, for example in different years, and the area could have potentially been affected by some phenomenon like oil spill and this may affect the sample and consequently the $GC \times GC$ image at the output of the $GC \times GC$ system. Two-dimensional $GC \times GC$ image related to a petroleum source from a petroleum-rich region consist of source-specific and sub-regional peaks. Sub-regional peaks constitute the common $GC \times GC$ topography of all of the sources from the same region, and source specific peaks exhibit topography specific to the particular source. One of the main tasks of the pattern recognition is to disentangle the common regional fingerprint from the topographic fingerprint specific to the source. Mathematically, an oil sample generated by the $GC \times GC$,$GC - MS$ etc, denoted as I_c, can be represented as a union of overlapping fingerprints images, i.e.

$$I_c = \left\{ \bigcup_{m=1}^{M} I_{r_m} \right\} \cup I_s \tag{3.1}$$

Where I_s is the source specific fingerprint and $I_(r_m)$ represents the fingerprint of the r^{th} sub-region, among M potentially over-lapping regions. If we could successfully disentangle $\left\{ \bigcup_{m=1}^{M} I_{r_m} \right\}$ and I_s, then we have learnt the source pattern of the oil sample and can use I_s instead of I_c as the $GC \times GC$ representative of the region when the problem is to compare between two $GC \times GC$ images related to two regions not closely located to each other. In other words, we can assume the pixels in I_s are the features we have extracted from I_c, and then would treat I_s as the compressed version of I_c. Needless to say that I_s is the result of the union of I_s and the interference terms, so the intensity of the pixels of I_s at any location is less than or equal to

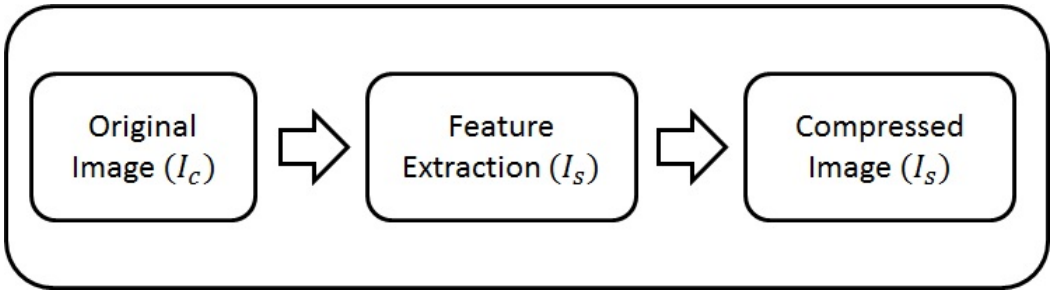

Figure 3.2: A compressed image is achieved by constructing a new image based upon the extracted features as opposed to using the whole original image.

that of I_c. The term $\left\{\bigcup_{m=1}^{M} I_{r_m}\right\}$ is used as a comparison between the samples with the same source fingerprint (are from the same region) but with different sub-regional effects. *Once we have learnt the pattern, the coding complexity of the image reduces considerably, while coding the image I_c will reduce to the coding of I_s.*

3.3 Method

As discussed in Section 3.2, the problem is to learn the common pattern, or equivalently the feature(s) of the images related to one source (one element in library D) and then save the pattern as the compressed version of the source, which requires an algorithm to find the pattern. We adopt the local PTM algorithm for this purpose, where the PTM algorithm has successfully worked and comprehensively been explained in [54].

Now, the problem reduces to the following: Given, for example, the first element of the library $I_1^{k_1}$, which is a family with k_1 members (images) for the first source, find the pattern common between these images using an algorithm, with the

knowledge that each of these images is a $GC \times GC$ image.

Let's assume one of the members of this family, lets say the first element,I_1^1 (reference sample), is the image extracted in a pretty much ideal case ($I_c \cong I_s$); without noise, without the interference of other closely located sample ($\left\{\bigcup_{m=1}^{M} I_{r_m}\right\} = \emptyset$). Then finding the common pattern among all of the family members, is equivalent to finding the common pattern of them with respect to the first member (reference sample (I_{ref})). Without losing generality, from now onwards, we develop our method on this family, called $I_{(1,gcxgc)}$. Before going through the algorithm lets have a look at the $GC \times GC$ image:

A $GC \times GC$ image is an $M \times N$ image consists of M rows and N columns: $I_{1,gcxgc} = I^{M \times N}$, where each column of the image is composed of a couple of intertwined Gaussian functions:

$$I_{1,gcxgc(i)} = \delta(x - i) \times \sum_{j=1}^{\kappa} G(w_j, \mu_j, \sigma_j) \tag{3.2}$$

$$G(w, \mu, \sigma) = w.e^{-(y-\mu)^2/2\sigma^2}$$

Where x and y denote the location for the row and column respectively and κ is the number of Gaussian functions in the corresponding column. Assuming the is pretty close to zero, then $G(w, \mu, \sigma) \cong w.(y - \mu)$. Hence:

$$I_{1,gcxgc(i)} \cong \delta(x - i) \times \sum_{j=1}^{\kappa} w_j.\delta(y - \mu) \tag{3.3}$$

So along one column of the image, *the peaks of the signal*, represent the information bearing part of the image.

3.3.1 A Brief Overview on Peak Topography Map (PTM)

PTM [54] explains the procedure pretty comprehensively, here we just touch upon it briefly as we will use the local-PTM score later. Btw, the comprehensive detalis about the Peak Topography Map has been studied in chapter 5. As discussed earlier, the $GC \times GC$ image consists of couple of peaks, $w'_j s$ located at $\mu'_j s$, so the problem of comparing between two $GC \times GC$ images will turn into the problem of comparing their amplitudes at the same location. Suppose image I_1 has a peak amplitude of w_1 and image I_2 has a peak amplitude of w_2 at the same location. PTM [54] introduces a metric to compare between these peaks as following:

$$sim(w_1, w_2, loc(x, y)) = max\left\{\frac{w_1(x, y)}{w_2(x, y)}, \frac{w_2(x, y)}{w_1(x, y)}\right\} \tag{3.4}$$

In case $sim(\cdot)$ for this location has a value of 1, then $w_1 = w_2$. But as discussed in Section 3.2, because of some issues like the baseline noise , two peaks can be considered as equal or matched even if the function $max(\cdot, \cdot)$ has a value rather than 1 and as we know this function has a value greater or equal to 1. Therefore, we accept two peaks as being matched if this function has a value of $\rho_\tau = 1 + \epsilon(\epsilon > 0)$. So we claim two peaks w_1 and w_2 at the location of (x, y) are matched if:

$$sim(w_1, w_2, loc(x, y)) = max\left\{\frac{w_1(x, y)}{w_2(x, y)}, \frac{w_2(x, y)}{w_1(x, y)}\right\} \leq \rho_\tau \tag{3.5}$$

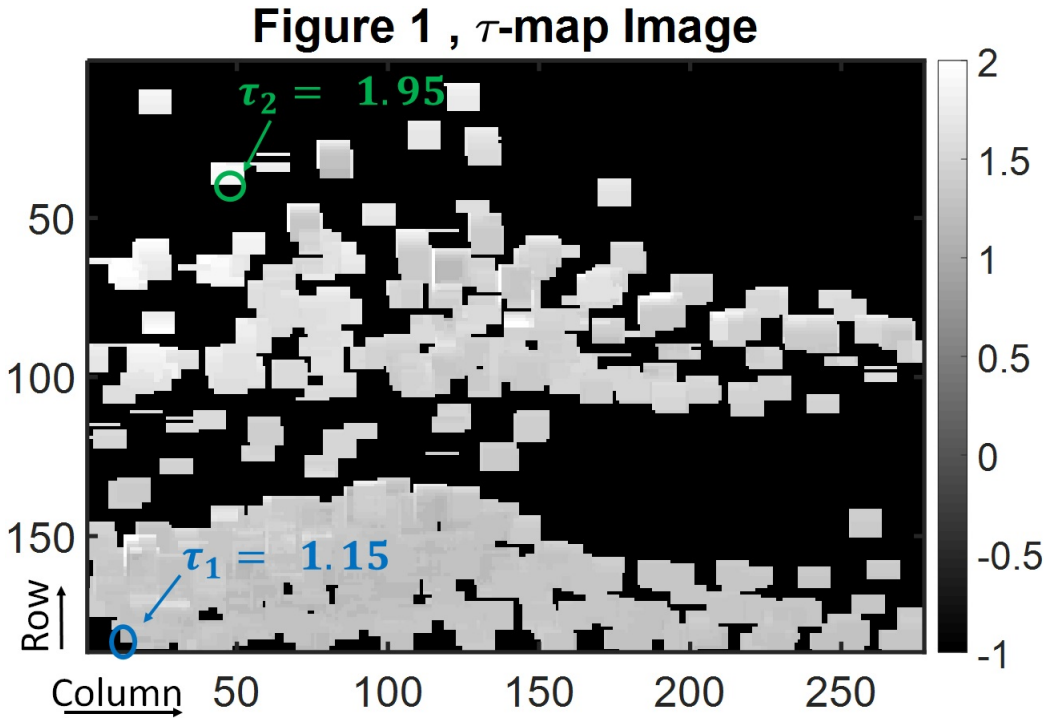

Figure 3.3: The $\tau - map$ image for the 14 injections of Macondo well for $r = 5$ and $M_p = 95\%$.

Clearly setting ρ_τ to ∞ will lead to a complete match between any two arbitrary peak amplitudes. Therefore, we say similarity between two $GC \times GC$ using PTM is a function of ρ_τ. The local-PTM score is the PTM score while comparing the two images locally around each of the pixels of the image.

3.3.2 What is a $\tau - map$?

$\tau - map$ image is a 3d image, $\tau(x, y, r)$, depending on the parameter r, where $\tau(x, y, r)$ is the minimum ρ_τ for which M_1, the mean of the percentage of match of $I_1^m (m \neq$ reference index) to the reference image(I_{ref}), is greater or equivalent to $M_p\%$

match for the local PTM centered at $I(x, y)$ for a given radius of r. The parameter r, represents the local PTM neighborhood, which means when we are to compare two image, like I_{ref} and I_{test} via PTM algorithm, we construct the r-neighborhood around the location denoted by (x, y) in both of the images and the compute the PTM score fir these neighborhoods. The r-neighborhood around a location (x, y) in an image I is defined as:

$$I_{rn}^{(2r+1)\times(2r+1)}(r) = \bigcup_{i=-r}^{r} \bigcup_{j=-r}^{r} I(x+i, y+j) \tag{3.6}$$

M_p is a parameter controlling the common pattern in a way that as M_p increases, the common part gets more more specific and small, because as we need a more percentage of match between the members in a family, the common part will decrease in size.

3.3.3 Model Petroleum Dataset

For demonstration purposes, we adopt a public-domain dataset from the Reddy laboratory, which has 34 petroleum $GC \times GC$ injections extracted from samples collected across different parts of the world, with a focus (14 samples) on the Macondo well, Gulf of Mexico, the source of the Deepwater horizon spill. The first family, $I_1^{k_1}(k_1 = 14)$ are these 14 samples and the reference sample, I_{ref} is the first sample of these 14 injections, and define the case study for a particular oil source. In Figure 3.3 the $\tau - map$ image has been plotted for this dataset and the corresponding family. Any image in this dataset is a 186×277 matrix (total of 51522 pixels).

3.4 Creating a Compressed Image Using $\tau - map$ Image

Figure 3.3 shows the -map image, where the color-bar on the right side shows the numbers between 0 to 2. Note that PTM algorithm depends upon ρ_τ and as we increase this parameter, more and more peaks can match to each other, so in an extreme case setting $\rho_\tau = \infty$ will match all of the peaks to each other, hence a large choice of ρ_τ is not good and we need to set an upper bound for ρ_τ. In Figure 3.3, we have set this upper bound $\rho_{\tau_{up}} = 2$. For two peaks which are matched with a choice of $\rho_\tau > 2$, we simply consider them as non-matched and set a value of 0 for the corresponding location in $\tau - map$ image.

Algorithm 3.1, shows the process to construct a $\tau - map$ image:

Algorithm 3.1 $\tau - map$ Image

□ **Input:**

* The family images $I_1^{k_1}$. Suppose the first image of the family is the reference image,I_{ref}. Let the rest of the images in the family be $I_{test}^m (2 \leq m \leq k_1)$.
* r, the local PTM-neighborhood radius.
* P_τ , the valid ρ_τ values vector, for example: $P_\tau = [1 : 0.05 : 2]$.

□ **Output:**

* $\tau - map$ image

□ **Initialization:**

* Start from the first element of image; ($x = 0$ and $y = 0$).

□ **Step0:**

* $\rho_{index} = 1$. (the index choosing one choice of ρ_τ form P_τ).
* $m_1 = 2$.(the index choosing one image from the test images of the family).
* $sum = 0$.(the parameter counting the sum of percentages of match of the test images against the reference image).

□ **Step1:**

* Let $I_T = \{I_{test}^m\} | m = m_1$
* Construct the r-neighborhood image around the location (x, y) for both I_{ref} and I_T.

$$I_{ref_{local}} = \bigcup_{i=-r}^{r} \bigcup_{j=-r}^{r} I_{ref}(x+i, y+j)$$

$$I_{T_{local}} = \bigcup_{i=-r}^{r} \bigcup_{j=-r}^{r} I_T(x+i, y+j)$$

□ **Step2:** $sum+ = PTM(I_{ref_{local}}, I_{T_{local}}, \rho_\tau)$

□ **Step3:**

* If $m < k_1$: Increment m_1 by 1 ,go to Step1.
* Else: $sum = \frac{sum}{k_1 - 1}$ (computing the average of percentages of match), go to Step4.

□ **Step4:**

* If $sum > M_p$: $\tau(x, y, r) = \rho_\tau$, proceed to the next location of the image ,go to Step0.
* Else: Increment ρ_{index} by 1,go to Step1.

In Figure 3.3, the values of τ for two points of the image have been shown in green and blue. The location related to the $\tau_2 = 1.95$ is $(x_g, y_g) = (42, 36)$ and that of

the blue is $(x_b, y_b) = (14, 181)$. Therefore, the region with the radius of $r_0 = 5$ around the blue point requires a very low $\rho_\tau = 1.15$, so this biomarker region shared the fingerprint pattern with all the members of the source family (Macondo well) of $I_1^{k_1}$, and defines the source-specific dictionary. On the other hand, the biomarker region around the green location requires a very high $\rho_\tau = 1.95$, i.e., the peakratio threshold constraint needs to be significantly laxed to match any patterns between the images. Therefore this part of the image is not representative of the images in the source (Macondo well) family. **The final compressed image of the Macondo family of images of $I_1^{k_1}$, consists only the pixels for which the corresponding $\tau - map$ image is less than a threshold** $\rho_{\tau_{up}}$. (The value for $\rho_{\tau_{up}}$ is a tunable parameter that could be chosen based upon the application of the chemical image, in Figure 3.3 and 3.5, we have set this value $\rho_{\tau_{up}} = 2$ hence in our figures we will not sweep over the values of $\rho_{\tau_{up}} > 2$). Note that a family has been learnt well and the values for the corresponding $\tau - map$ image has been selected optimally, once adding a new member to the family will not change their -map parameters. Suppose for a given choice of ρ_{τ_0}, the x-locations and y-locations of all the points having ρ_{τ_0} in $\tau - map$ are denoted by X_0 and Y_0, respectively. (x_0, y_0) is one location from these location vectors:

The local compressed image around the location (x_0, y_0) is:

$$I_{1,gcxgc(local-compressed)}(r,\rho_{\tau_0},\rho_{\tau_{up}}) = \begin{cases} \bigcup_{i=-r}^{r}\bigcup_{j=-r}^{r} I_{ref}(x_0+i,y_0+j) & \text{if } \rho_{\tau_0} \leq \rho_{\tau_{up}} \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

Therefore the final compressed image will be:

$$I_{1,gcxgc(compressed)}(r,\rho_{\tau_0},\rho_{\tau_{up}}) = \bigcup_{x_0\in X_0, y_0\in Y_0} I_{1,gcxgc(local-compressed)}(r,\rho_{\tau_0},\rho_{\tau_{up}}) \tag{3.8}$$

And the compression rate at the peak level will be as following where this metric is novel as it considers the compression along the peaks, as the information bearing part of $GC \times GC$ image:

$$cr(r,\rho_{\tau_0},\rho_{\tau_{up}}) = \frac{\sharp peaks\ of\ I_{ref}}{\sharp peaks\ of\ I_{1,gcxgc(compressed)}} \tag{3.9}$$

3.4.1 Good Choice of ρ_τ: A trade-off Between the Compression Ratio and Common Part Maximization

The compressed image depends highly on the value of ρ_τ, a very low choice of ρ_τ may just extract a handful of the peaks of the image as the representative of the family. But it may not be a very good choice because then any little jitter or noise injected to the sample during the process of generating the $GC \times GC$ image via $GC \times GC$ system can rule out a peak as the common peak among the family members. Therefore, we should also take the noise into consideration and have a reasonable choice for this parameter. One good choice is to construct the histogram

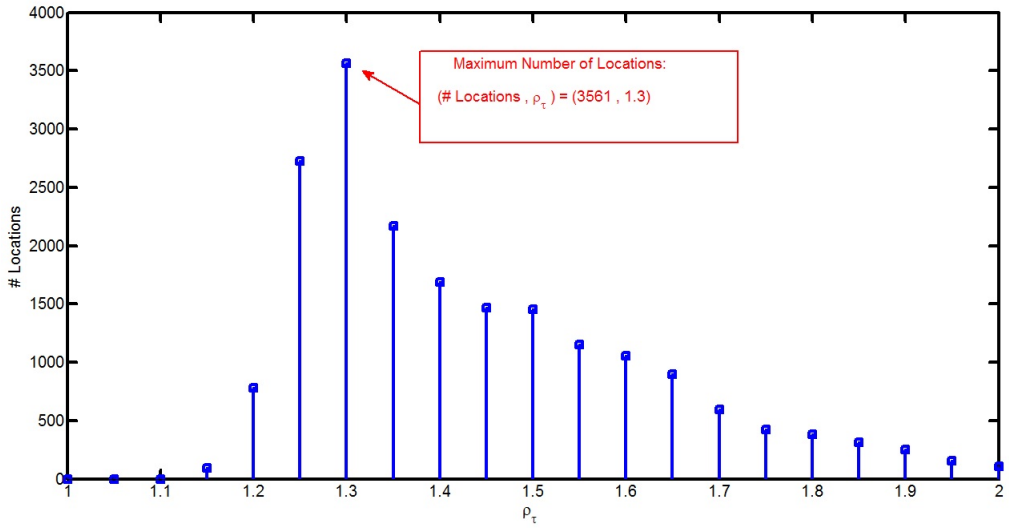

Figure 3.4: Histogram of $\tau - map$ image for $r = 5$ and $M_p = 95\%$.

of -map image and observe the different number of locations with the same choice of ρ_τ and set the optimal value of ρ_τ where this number is maximum, the reason is that the choice of ρ_τ^* which maximizes the histogram indicates that most of locations of the image agree upon the fact that this choice of ρ_τ will match the images within the family the most.

$$\rho_\tau^* = argmax(Hist(\tau - map, \rho_\tau)) \tag{3.10}$$

$$1 \leq \rho_\tau < \rho_{\tau_{up}}$$

Where $Hist(\alpha, \beta)$ denotes the histogram of image for the pixel value of . Figure 3.4 shows this histogram, it can be seen that the choice of $\rho_\tau = 1.3$ will lead to the maximum number of locations within the family, so we will set $\rho_\tau = \rho_\tau^* = 1.3$. On the other hand the choice of ρ_τ affects the achievable compression ratio, a choice

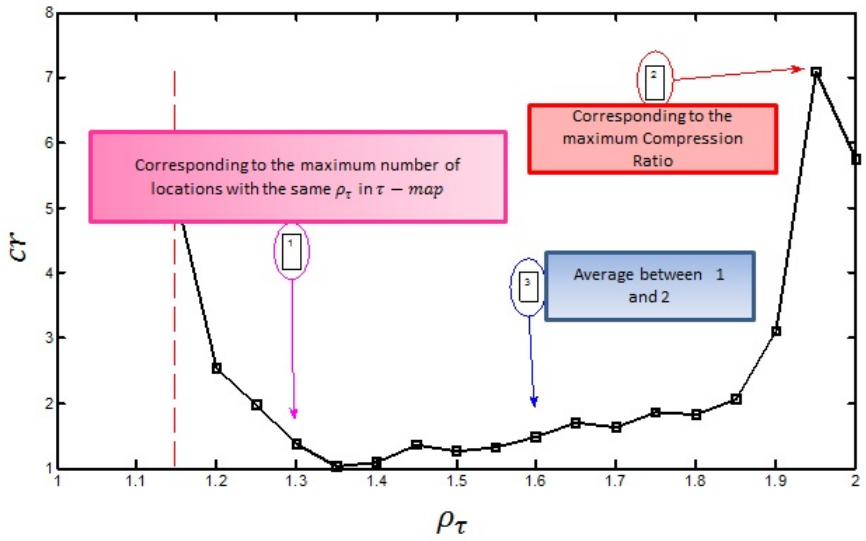

Figure 3.5: Compression ratio achieved by different choices of ρ_τ for the model petroleum dataset.

of ρ_τ that maximizes the above-mentioned histogram will lead to a compressed image which has a high non-zero elements, and this will lead to a low compression ratio. Therefore, the choice of ρ_τ^* should also take the compression ratio into account.

3.4.2 Compression Ratio

As discussed in Section 3.4, the compression ratio is given by:

$$cr(r, \rho_{\tau_0}, \rho_{\tau_u p}) = \frac{\sharp peaks\ of\ I_{ref}}{\sharp peaks\ of\ I_{1,gcxgc(compressed)}} \tag{3.11}$$

As can be seen from Figure 3.5, the compression ratio values exist with the choice of $\rho_\tau \geq 1.15$, this is because in a value lower than 1.15 the number of peaks in the compressed image is zero, so we are just interested in $\rho_\tau \geq 1.15$. Remember

the choice of $\rho_\tau = 1.3$ led to the maximum of locations with the same ρ_τ where the achieved compression ratio for this choice is about 1.37. The maximum achieved compression ratio has been achieved at $\rho_\tau = 1.95$ but this choice of ρ_τ may not be a very good idea because it will introduce some peaks as the common peaks within the family which may not be the case; so a number between these two extremes can work fine ($1.3 < \rho_\tau^* < 1.95$).

3.5 Conclusion

In this chapter we introduced a novel representational map, called $\tau - map$, for one family of the $GC \times GC$ images. We discussed how we could use this map in order to achieve a compact representative of the family and come to a compressed image for the family. This map is actually a function of the peak-ratio threshold ρ_τ which can be tuned in order to achieve the desired compressed image.

CHAPTER 4
LEARNING FORENSIC PATTERNS WITHIN A NEURAL NETWORK FRAMEWORK

4.1 Introduction

Separation and classification of $GC \times GC$ images can be done by analyzing the whole chromatogram's pixels and peaks. As the number of the peaks in a $GC \times GC$ image is large, the analysis and interpretation requires high computational complexity and the analysis time may be considerable. Therefore, a need for the analysis using a subset of the peaks or a reduced version of the image is needed. In this chapter, we apply the SAX algorithm as a method to achieve a compressed version of the image in a neural network framework. We then compare this method against the other traditionally used methods to gauge its performance.

4.2 Problem Statement

As discussed earlier, suppose we have a dictionary D, with K $GC \times GC$ samples, $D = \{I_1, I_2, , I_K\}$ where each of the members of the library illustrates the $GC \times GC$ pattern of one unique geographical region. The geographical regions of these $GC \times GC$ images are saved in a set $R = \{r_1, r_2, .., r_K\}$ where the i^{th} element in R indicated the geographical region of i^{th} image of D. Lets assume K is large enough, so that we have the $GC \times GC$ patterns of all possible geographical regions. Now, for a newly-extracted unknown $GC \times GC$ test image, I_{test}, we are to determine the geographic region of I_{test} by the information in D. If K is large enough, the

geographical region of I_{test} will be the same as or pretty close to the region of the member of the library in which I_{test} has the most similarity with, by choosing a suitable choice of similarity criterion, which will be discussed later. Therefore, we should compare I_{test} with all of the members of D one by one. Mathematically, if we define a similarity criterion like, $\mathcal{S}$, then:

$$i = \underset{1 \leq ref \leq K}{\text{argmax}}\ \mathcal{S}(\mathrm{I_{ref}}, \mathrm{I_{test}}). \tag{4.1}$$

$$\textit{Geographical Region of } I_{test} = R(i).$$

Where in equation 4.1 ref indicates the index of the image in D. Needless to say, once we have a dataset of test images with more than one image under test, we can use the network for each of the samples from the dataset and realize its geographical region.

4.3 Technical Approach

We present the solution to the problem stated in Section 4.2 with the network shown in Figure 4.1. As can be seen the network is similar to a neural network with three layers of input, hidden and output layer. As opposed to the normal neural network, we don't learn the weights of the network through an iterative learning process, but we set these weights according to our proposed method.

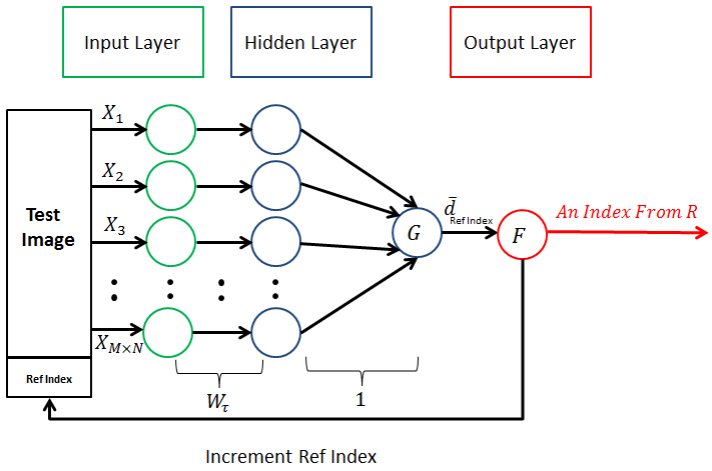

Figure 4.1: Illustrative model of the proposed network.

4.3.1 Some Notes on the proposed network

Let's quickly take a note on different parts of the network in Figure 4.1. The box on the left, has two parts, first the $GC \times GC$ image of the image under the test (I_{test}) and a block for the index of the reference samples from the library D which this index is initialized as one (Ref Index=1). The test image is an $M \times N$ matrix, so in the input layer we have $M \times N$ inputs, shown as X_1 to $X_{M\times N}$ where each of these inputs indicates one pixel of the test image. These inputs are transmitted by the weights W_τ to the hidden layer. In Section 4.3.3 we will discuss on the algorithm to construct W_τ. The hidden layer nodes are also carried by all one vector, $1_{(M\times N)\times 1}$, to the next hidden layer. In this layer the G function is implemented between I_{test} and the reference image. Finally, the function F has the role to determine the geographical region of I_{test}. F has also K memories. First, the Ref Index is one, so I_{test} is compared

against the first element of D and the corresponding d_{min} is saved in one memory of F. Once this comparison is done completely and d_{min} is saved, F sends an Increment Ref Index command and in the next clock the test should be compared against the second element of D and so on. Finally once Ref Index exceeded the size of the dictionary, K, F outputs the index having the lowest d_{min}, which is the index of the geographical region of I_{test} from the set R.

4.3.2 Similarity Criterion

As discussed comprehensively in [55] and chapter 5, comparing two $GC \times GC$ images is the comparison between their corresponding peaks at the same location of their image. A $GC \times GC$ image is an $M \times N$ image having $M \times N$ pixels. We sweep the image by going along each of its N columns and save the local maxima. We call these local maxima along each of the columns as peaks. Suppose we are to compare two images, I_{ref} and I_{test}. At one location, like (x_0, y_0) the amplitude of the peaks in I_{ref} and I_{test} are $p_{ref}(x_0, y_0)$ and $p_{test}(x_0, y_0)$. In [55] the similarity between these two peaks are defined as:

$$Sim(p_{ref}, p_{test}, x_0, y_0) = max(\frac{p_{ref}}{p_{test}}, \frac{p_{test}}{p_{ref}}) \tag{4.2}$$

As known, the function $max(\alpha, \alpha^{-1})$ has a value greater or equal to one for all $\alpha \in \mathcal{R}$. Therefore, if the function $Sim(\cdot)$ has a value of one, then $\alpha = \alpha^{-1}$ or in our case $p_{ref} = p_{test}$. Any difference between the values of p_{ref} and p_{test} causes in a value greater than one for $Sim(\cdot)$. We can call two peaks as similar once their the $Sim(\cdot)$

for their corresponding location is one, but there exists potential experimental noise during the process of producing the $GC \times GC$ image in the lab. Hence, we define a peak-ratio parameter τ as:

$$\tau = 1 + \epsilon \ (\epsilon > 0) \tag{4.3}$$

And claim two peaks as similar, once the $Sim(\cdot)$ function for their corresponding location is less than or equal to $(Sim(\cdot) \leq \tau)$. The parameter ϵ indicates the amount of deviation that is acceptable for us to consider two peaks as similar, the more the value of ϵ is, the less strict we are in the definition of similarity between two peaks. In an extreme case, if we set $\epsilon = \infty$, all of the peaks will be considered as similar because $Sim(\cdot)$ is always less than or equal to ∞.

4.3.3 How to set W_τ

There are $M \times N$ inputs, X_1 to $X_{M \times N}$, but as discussed in the previous section, just the local maxima of the image are saved and used as the comparison. Therefor, the weights for the non-peaks are set to zero. For calculating the weight of one peak like p_{test} at the location (x', y'), once the peak in the same location in the reference image is p_{ref}, we set an upper bound for the acceptable deviation, say $_{up}$ and the corresponding peak-ratio, $\tau_{up} = 1 + \epsilon_{up}$. Finally, we perform the following two-stage process:

Stage1: Calculate the $Sim(\cdot)$ for the location of the peaks as in Equation 4.2.

Stage2:

* If $Sim(\cdot)\rho_{up}$ then replace both p_{ref} and p_{test} by a common peak, p_{common}:

$$p_{common} = min(p_{ref}, p_{test}).$$

so the weight of p_{test} will be:

$$w_{test,x',y'} = \frac{p_{common}}{p_{test}}.$$

* Otherwise, keep the exact values of p_{ref} and p_{test} which means $w_{test}(x', y') = 1$.

The reason which we replace p_{ref} and p_{test} by p_{common} is that, we have assumed once $Sim(\cdot)$ is less than equal to τ for that location, these two peaks are similar, and therefore we manually assign an equal value to them.

4.3.4 How to set ϵ

Suppose we have the library $R = \{I_1^{k_1}, I_2^{k_2}, \ldots, I_K^{k_N}\}$, where the i^{th} element of the library, $I_i^{k_i}$, or i^{th} family, means we have k_i number of $GC \times GC$ image for the region indexed by i. The $\epsilon(i)$, the ϵ for the i^{th} family is disproportional to the deviation, or d_{min} between the members of the family. In Figure 4.2 we have plotted the total MSE of the PAA's or, d_{min}, for the family of images from Macondo region with seven members. We have considered these seven samples as training samples. As can be seen, the minimum deviation between the PAA's of the family occures at $\epsilon_{up} = 0.4$, therefore we use this value for ϵ_{up}.

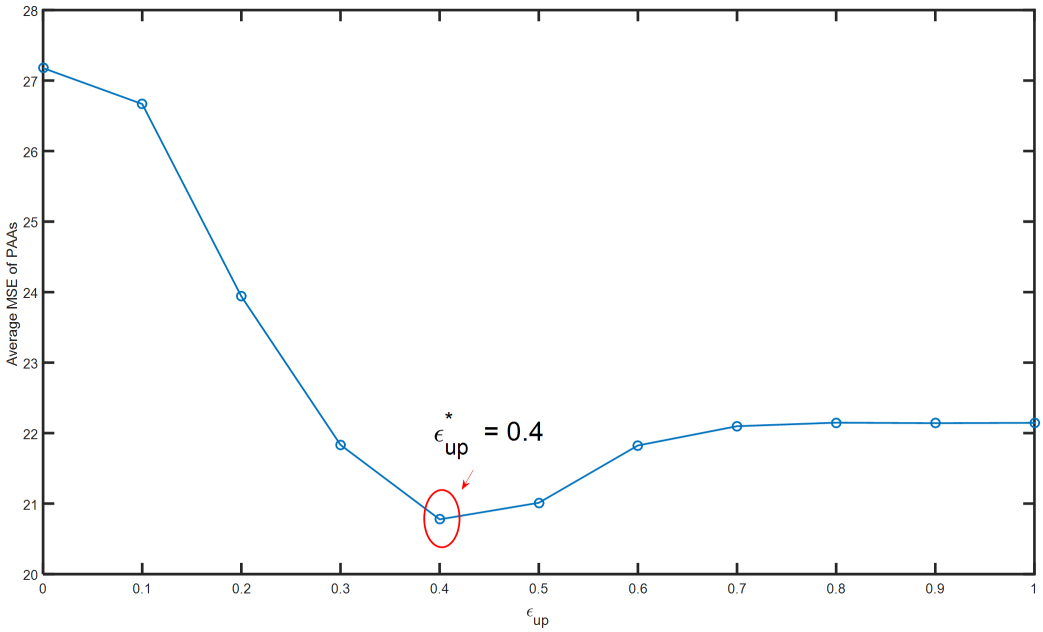

Figure 4.2: Optimal choice of ϵ

4.4 Function G

Function G can compare the reference (I_{ref}) and the test (I_{test}) images based upon one of the following six schemes:

- Direct Euclidean distance I_{ref} and I_{test}
- Correlation between I_{ref} and I_{test}
- Difference between the SAX representation of I_{ref} and I_{test}
- Difference between the PAA representation of I_{ref} and I_{test}
- Evaluating the maxima within each interval(w) of the time series and calculating the Euclidean ditance between the maxima

- Similarity between (I_{ref}) and (I_{test}) via computing PTM score

The comprehensive study on SAX algorithm can be found in [2]. Here, we touch upon it briefly, as we will use it in our proposed network in Figure 4.1. Symbolic representation of time series (SAX) proposes a method to represent time series of size m to a string of arbitrary size of $w(w < n)$. In this case, we will have dimensionality reduction which can be a big deal once the size of the time series are large. As formally defined in [2], a time series C of length n can be represented in a w-dimensional space by a vector $\bar{C} = \bar{c_1}\bar{c_2}\bar{c_3}\ldots\bar{c_w}$. The i^{th} element of C is calculated by the following equation:

$$\bar{c_i} = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{j=\frac{n}{w}i+1} c_j \tag{4.4}$$

In other words, the time series is divided into w intervals of the same size, then the data in each of these intervals is replaced by the mean of the data. We call this representation of the times series, the Piecewise Aggregate Approximation (PAA) representation of the time series. After applying the SAX algorithm the final symbol representation of the times series C will be:

$$\hat{C} = \hat{c_1}\hat{c_1}\ldots\hat{c_w} \tag{4.5}$$

Note that C is the original times series, $\bar{C}$ is its PAA representation and $\hat{C}$ is its SAX representation.

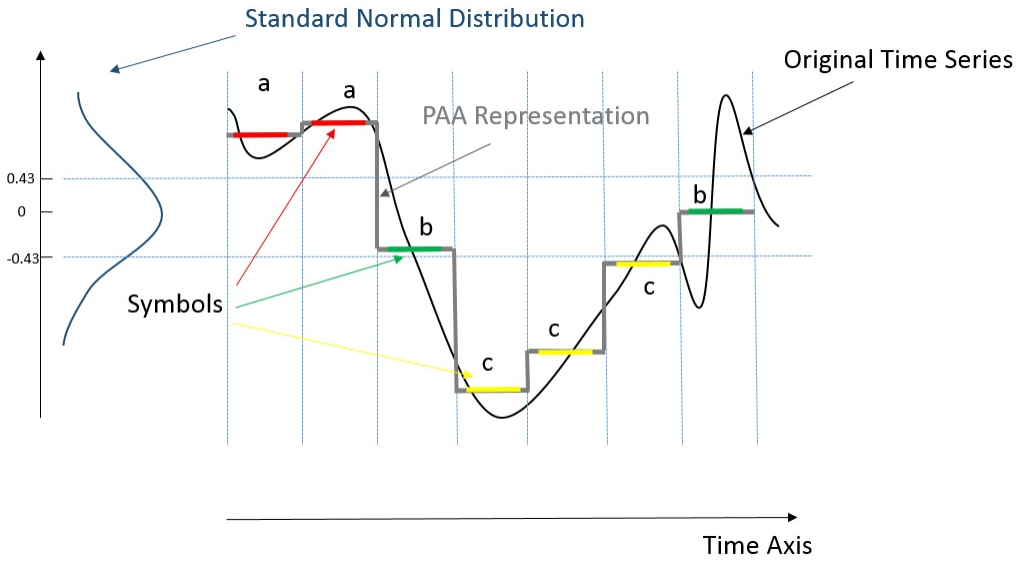

Figure 4.3: The PAA and SAX representation of a model time series. In this figure, the there are three symbols, a,b and c. The time axis has been sliced into seven intervals. The SAX representation of the time series in this case would be $\hat{C} = aabcccb$.

4.4.1 Calculating the distance between two time series using the six different schemes

Suppose we have two time series C_1 and C_2, the Euclidean distance between them is:

$$d = \sqrt{\sum_{i=1}^{n}(c_1^i - c_2^i)^2} \tag{4.6}$$

Where c_1^i and c_2^i mean the i^{th} element of C_1 and C_2. After transforming them into their PAA representation and constructing their symbol representation we should calculate the symbol distance of their representations. Therefore, we need a look-up table in order to have the distance between the symbols. Such a table is given in Table 4.2. Then, The Euclidean distance between the two SAX representation of C_1 and C_2 is given as [2]:

$$\hat{d} = \sqrt{\frac{n}{w}\sum_{i=1}^{w}[dist(\hat{c}_1^i - \hat{c}_2^i)]^2} \tag{4.7}$$

Where $\hat{c}_1^i$ and $\hat{c}_2^i$ mean the i^{th} element of the $\hat{C}_1$ and $\hat{C}_2$, respectively.

And the Euclidean distance between the PAA representation of the two time series is :

$$\bar{d} = \sqrt{\sum_{i=1}^{w}(\bar{c}_1^i - \bar{c}_2^i)^2} \tag{4.8}$$

We can also take the maximum along each of the intervals (w) as the representative of the interval and compute the Euclidean distance between the maxima as the comparison metric:

$$c_{max}(i) = \underset{i \in \{1,2,\ldots,\frac{n}{w}\}}{\text{argmax}}\ C(i{:}\, i+w). \tag{4.9}$$

And the Euclidean distance between the maximum-along-interval representation of the two time series is :

$$d_{max} = \sqrt{\sum_{i=1}^{w} (c^i_{max,1} - c^i_{max,2})^2} \tag{4.10}$$

The two-dimensional correlation between the two images I_{ref} and I_{test} images are computed as following:

$$Corr2(I_{ref}, I_{test}) \quad = \quad \frac{\sum_m \sum_n [(I^{m,n}_{ref} - \bar{I}_{ref}) \times (I^{m,n}_{test} - \bar{I}_{test})]}{\sqrt{\sum_m \sum_n (I^{m,n}_{ref} - \bar{I}_{ref})^2 \times \sum_m \sum_n (I^{m,n}_{test} - \bar{I}_{test})^2}} \tag{4.11}$$

where m and n represent the first and the second dimension of the image, respectively. $\bar{I}_{ref}$ and $\bar{I}_{test}$ also represent the mean value of the reference and test images, respectively.

Now, let's look at the initial network in Figure 4.1, the output of the G node is d which is calculated for the comparison of I_{test} and any of the reference images from D. Function F, then saves all of these values for $\hat{d}$ and outputs the index of these values of $\bar{d}$ to determine the geographical region of I_{test} from R.

$$i = \underset{1 \leq RefIndex \leq K}{\text{argmin}}\ d_{min}. \tag{4.12}$$

$$\textit{Geographical Region of } I_{test} = R(i).$$

4.4.2 Why SAX helps us in solving our problem

The main usage of SAX is that it reduces the dimensionality but the other reason we have used it in solving our problem is that it is pretty robust to the noise and peak shifts. As discussed, the $GC \times GC$ images are prone to noise and this causes a change to their amplitudes. On the other hand, due to the noise, peaks experience some shifts in their location of occurrence. A peak should occur at the retention time of (r_1, r_2) but it shows up at $(r_1 + \delta_1, r_2 + \delta_2)$. By slicing the time into couple of intervals and averaging the data, we have implicitly applied a moving average filter on the data which can work as a noise filter.

4.5 Result

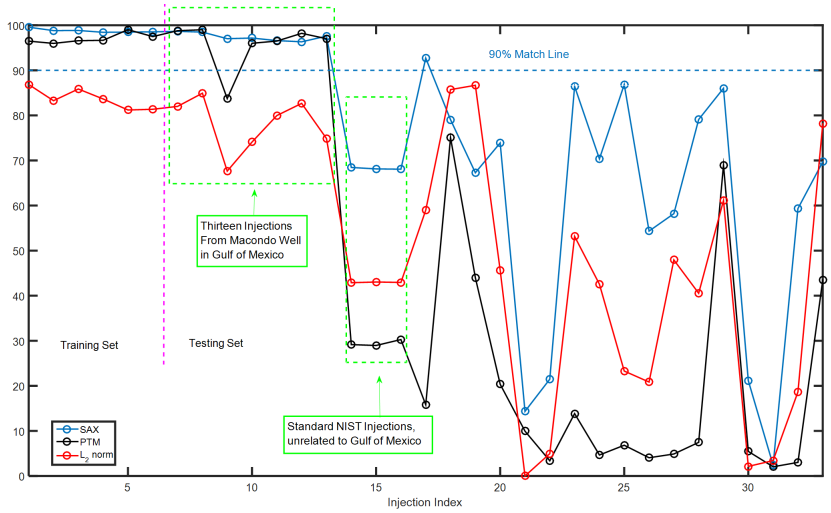

Figure 4.4: The percentage of similarity of thirty three samples from the model dataset to the first reference sample from Macondo well(1).

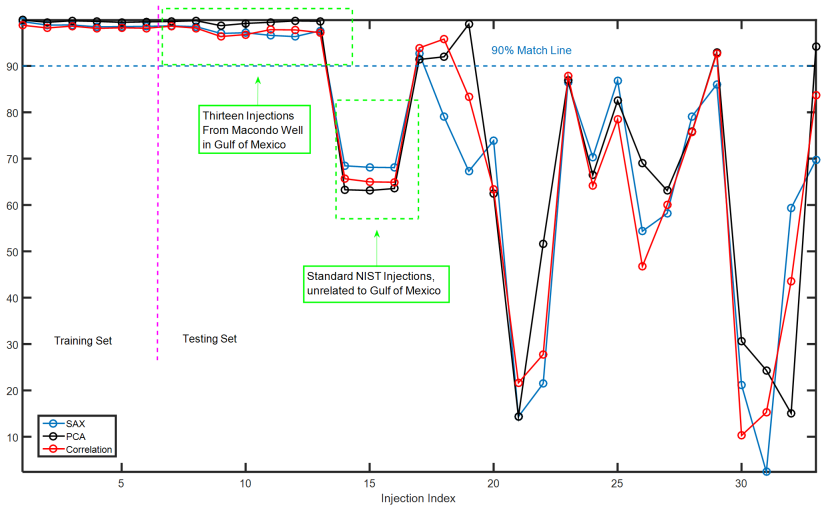

Figure 4.5: The percentage of similarity of thirty three samples from the model dataset to the first reference sample from Macondo well(2).

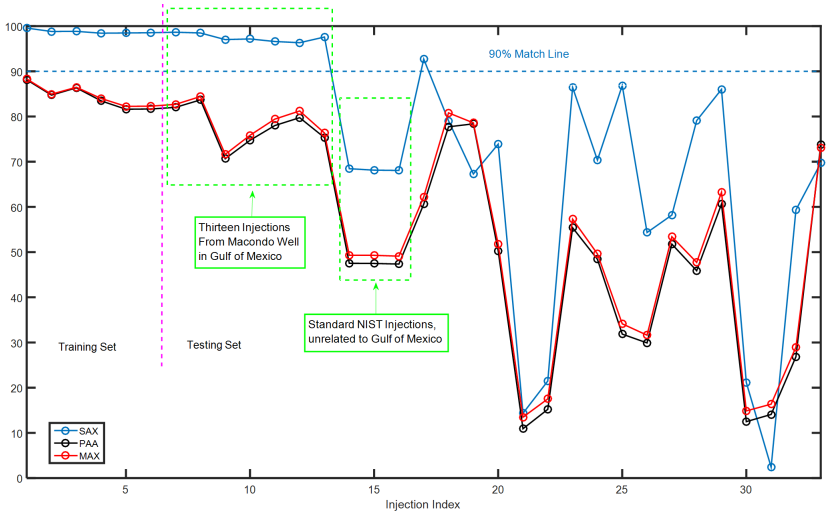

Figure 4.6: The percentage of similarity of thirty three samples from the model dataset to the first reference sample from Macondo well(3).

Table 4.1: Percentage match between different Gulf of Mexico sources against Macondo injections.

Method	Mac vs. Mac	EI vs. Mac	SLC vs. Mac	Nseep vs. Mac
SAX	$98.88 \pm 0.67\%$	$77.86 \pm 0.28\%$	$91.69 \pm 0.49\%$	$67.70 \pm 0.23\%$
PTM	$94.52 \pm 6.16\%$	$72.82 \pm 2.75\%$	$24.68 \pm 8.71\%$	$38.64 \pm 2.99\%$
PCA	$99.83 \pm .16\%$	$91.97 \pm .1\%$	$91.8 \pm .09\%$	$97.84 \pm .51\%$
Correlation	$98.7 \pm .82\%$	$94.08 \pm .82\%$	$92.37 \pm 1.18\%$	$83.39 \pm .24\%$
L_2 norm	$92.53 \pm 3.95\%$	$87.71 \pm 2.53\%$	$81.38 \pm 5.11\%$	$75.06 \pm 4.53\%$
MAX	$92.37 \pm 3.03\%$	$84.78 \pm 1.45\%$	$80.08 \pm 4.06\%$	$70.75 \pm 2.94\%$
PAA	$92.62 \pm 3.046\%$	$84.52 \pm 1.06\%$	$79.74 \pm 4\%$	$71.74 \pm 2.81\%$

4.5.1 Model Dataset

We verify our method with a dataset of thirty four injections with $GC \times GC$ pattern from different parts of the world. Of particular interest, there are fourteen samples from the Macondo well in Gulf of Mexico, and three standard NISt samples unrelated to Gulf of Mexico.

4.5.2 Discussion on the result

We have tested our proposed method using $\alpha = 10$, $w = 31$ and $\epsilon_{up} = 0.4$. We set aside the first sample from the Macondo well as one the elements of D and test the remaining dataset against it. As can be seen in Figure 4.6, the first thirteen samples have a pretty high percentage of matches to the reference sample. These samples

Table 4.2: The distance table between the SAX symbols.

Alphabet	a	b	c	d	e	f	g	h	i	j
a	0	0	0.19	0.57	1.06	1.63	2.34	3.24	4.49	6.55
b	0	0	0	0.1	0.34	0.7	1.18	1.84	2.82	4.49
c	0.19	0	0	0	0.07	0.27	0.59	1.08	1.84	3.24
d	0.57	0.1	0	0	0	0.06	0.25	0.59	1.18	2.34
e	1.06	0.34	0.07	0	0	0	0.06	0.27	0.7	1.63
f	1.63	0.7	0.27	0.06	0	0	0	0.07	0.34	1.06
g	2.34	1.18	0.59	0.25	0.06	0	0	0	0.1	0.57
h	3.24	1.84	1.08	0.59	0.27	0.07	0	0	0	0.19
i	4.49	2.82	1.84	1.18	0.7	0.34	0.10	0	0	0
j	6.55	4.49	3.24	2.34	1.63	1.06	0.57	0.19	0	0

are actually from the Macondo well which shows that our method has successfully worked. The three NIST samples are also shown in figure. As can be seen, they have the same percentage of similarity to the reference sample. The 90% match line also separates the Macondo samples from the others.

CHAPTER 5
DETAILED ANALYSIS OF PEAK TOPOGRAPHY MAPS FOR FORENSIC INTERPRETATION

5.1 Introduction

Comprehensive two-dimensional gas chromatography ($GC \times GC$) provides high-resolution separations across hundreds of compounds in a complex mixture, thus unlocking unprecedented information for intricate quantitative interpretation. We exploit this compound diversity across the ($GC \times GC$) topography to provide quantitative compound-cognizant interpretation beyond target compound analysis with petroleum forensics as a practical application. We focus on the ($GC \times GC$) topography of biomarker hydrocarbons, hopanes and steranes, as they are generally recalcitrant to weathering. We introduce peak topography maps (PTM) and topography partitioning techniques that consider a notably broader and more diverse range of target and non-target biomarker compounds compared to traditional approaches that consider approximately twenty biomarker ratios. Specifically, we consider a range of 33-154 target and non-target biomarkers with highest-to-lowest peakratio within an injection ranging from 4.86-19.6 (precise numbers depend on biomarker diversity of individual injections). We also provide a robust quantitative measure for directly determining match between samples, without necessitating training datasets. We validate our methods across thirty-four ($GC \times GC$) injections from a diverse portfolio of petroleum sources, and provide quantitative comparison of performance against established statistical methods such as principal components analysis (PCA).

Our dataset includes a wide range of samples collected following the 2010 Deepwater Horizon disaster that released approximately 160 million gallons of crude oil from the Macondo well. Samples that were clearly collected following this disaster exhibit statistically significant match $(99.55 \pm 0.96)\%$ using PTM-based interpretation against other closely related sources. PTM-based interpretation also provides higher differentiation between closely correlated but distinct sources than obtained using PCA-based statistical comparisons. We provide a peak-cognizant informational framework for quantitative interpretation of $GC \times GC$ topography. Proposed topographic analysis enables $GC \times GC$ forensic interpretation across target petroleum biomarkers, while including the nuances of lesser-known non-target biomarkers clustered around the target peaks. This allows potential discovery of hitherto unknown connections between target and non-target biomarkers [56].

5.2 Background

Comprehensive two-dimensional gas chromatography ($GC \times GC$) provides high-resolution separation across hundreds, sometimes thousands, of crude oil hydrocarbons, thus unlocking unprecedented information for intricate quantitative interpretation. The broad objective of this work is to exploit this rich compound diversity and provide compound-cognizant quantitative interpretation of $(GC \times GC)$ peak topography that bridges the gap between target-driven analysis and statistical methods. We propose peak topography maps that extend individual $(GC \times GC)$ peak analysis beyond the well-known target peaks that dominate the $(GC \times GC)$ image,

and present techniques for interpreting ($GC \times GC$) topography that provide nuanced quantitative peak-based comparisons between ($GC \times GC$) images. While we present our results in the context of petroleum forensics as a practical application of interest, the scope of our work applies generally to quantitative ($GC \times GC$) interpretation and as such, goes beyond the stated application.

A key distinction of our technique against multi-variate statistical methods [57] is compound-cognizant interpretation that preserves the identity of individual target peaks while extending the scale of peak-level interpretation to all peaks, target and non-target, within the ($GC \times GC$) topography. This allows nuanced ($GC \times GC$) distinction between closely related yet different complex mixtures, e.g. crude oil from neighboring oil sources, which share the regional fingerprint, and therefore, difficult to differentiate robustly using purely statistical methods.

5.2.1 Current state-of-the art in chromatographic interpretation: challenges and opportunities

Many separation technologies routinely filter out non-target analytes, thus eliminating possibility of understanding their connection to dominant target analytes in an environmental sample. More comprehensive datasets recording the joint contributions of target and non-target analytes may be enabled through comprehensive two-dimensional gas chromatography ($GC \times GC$), liquid chromatography ($LC \times LC$), mass spectrometry (MS) and combinations thereof. However, despite the informational richness of these comprehensive datasets, non-target analytes are traditionally

ignored in sample analysis in preference to peakratio comparisons between the target chemicals. Although non-target chemicals are empirically considered in the chemometric literature, their role is typically limited to the major statistical loadings in multi-variate distributions [58–60]. Thus, current state-of-the-art in environmental forensics and analytical chemistry are broadly divided into two complementary approaches:

- Target-based analysis [59–70]: Focuses on the target chemicals (well-known hopanes, steranes, diasteranes in petrochemicals) that dominate the analytical landscape as the major peaks in a chromatogram or a GC-MS image. This includes statistical methods employed towards target-based analysis [68, 71].

- Target-agnostic analysis [72–78]: Statistical pattern-recognition techniques that analyze comprehensive separation datasets using different forms of multi-variate analysis.

Table 5.7 (in Section 5.15) provides a point-by-point comparison between the two approaches in the context of environmental forensics.

5.2.2 Petroleum forensics using $GC \times GC$ separation of crude oil samples

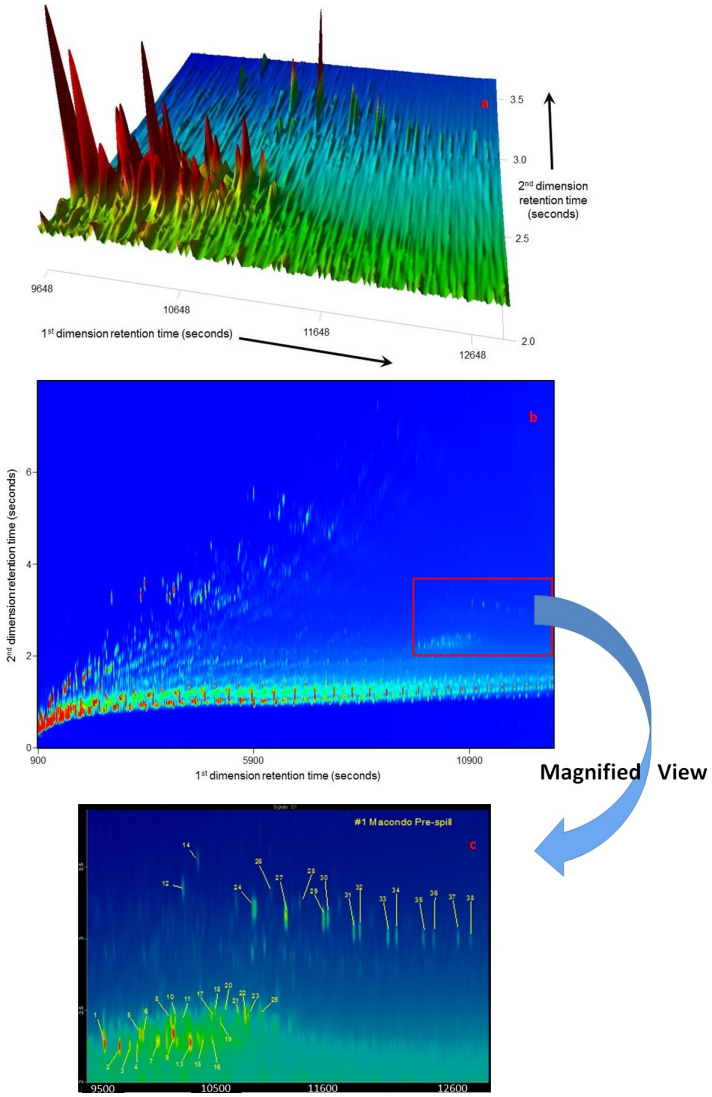

Figure 5.1: **a.** The three-dimensional view of $GC \times GC$ image of crude oil pre-spill sample from Macondo well, site of Deepwater Horizon spill disaster, Gulf of Mexico, 2010. **b.** The two-dimensional view of a.**c.** Detailed topography of biomarker region (hopanes and steranes) marked as red box in b.Target biomarkers are labeled and itemized in Table 5.6.

Reliable fingerprinting of petroleum and its weathered products has been an important field of study in the last four decades [58–66, 79–87]. Forensic analysis techniques fingerprinting crude oil samples in the ocean typically interpret the $GC \times GC$ peak profiles of biomarker hydrocarbons (hopanes and steranes), as they are generally recalcitrant against environmental weathering [60, 63, 67, 81–87]. Figure 5.2.2 shows the $GC \times GC$ biomarker topography of a pre-spill crude oil sample taken from the Macondo well, source of the Deepwater Horizon disaster, spanning over a hundred compounds across a relative scale of 1 to 14.53 between the lowest and highest summits (peaks occupying lowest 5% of the $GC \times GC$ peak magnitude profile were rejected as baseline noise). Traditional analysis employs approximately forty target biomarker compounds (refer to labeled compounds in Figure 5.2.2 and Table 5.9 (in Section 5.9)), which occur as major peaks dominating the $GC \times GC$ topography, and about twenty well-known peak ratios [81] based on these target compounds.

5.2.3 Background motivation: Peak-cognizant interpretation beyond target biomarkers

Target biomarkers are generally abundant within a sample, robust to chromatographic variability, and therefore, provide a well-established basis to compare two oil samples [59,60,62,63,81]. However, the interpretation power of target analysis can be magnified significantly if we harness the full informational potential $GC \times GC$: combining the well-known characteristics of target biomarkers (major peaks) with the lesser known nuances of non-target biomarkers (minor peaks), which occupy the

breadth of the intricate $GC \times GC$ topography. More recently, chemometric interpretations of $GC \times GC$ datasets have been proposed that adopt a multi-variate statistical approach to forensic interpretation [71–74, 88–90]. While these statistical approaches exploit the data variance of the $GC \times GC$ topography beyond the target peaks, they are typically agnostic of the target biomarkers and the dominant role they play in forensic interpretation [59, 60, 62–66, 81]. We harness the rich compound diversity across the $GC \times GC$ biomarker (hopanes and steranes) topography to provide potentially transformative compound-cognizant interpretation beyond target compound analysis. Our objective is to extend the scope of target-centric standards [64–66] to include non-target biomarkers within a compound-cognizant framework, and thus bridge the gap between target-based forensics (e.g. [59, 60, 62, 63, 81] and references therein) and existing target-agnostic statistical approaches [71–73, 88–90]. We achieve source-specific and regional fingerprints by mapping connections between target and non-target biomarkers within the $GC \times GC$ topography. While the established target peaks dominate forensic interpretation, and can be individually identified in the topography map proposed, the unutilized contribution of the minor (non-target) peaks (e.g. the 73 unlabeled non-target peaks in Figure 5.2.2) are also employed to distinguish closely related samples. Furthermore, we propose partitioning techniques that enable discovery of peak clusters connecting known targets to unknown non-target biomarkers, and thus derive common regional characteristics of petroleum-rich areas.

5.2.4 Key innovation and contributions

Our motivation in this work is to achieve robust forensic distinction between closely related oil sources by utilizing rich peak information diversity in two-dimensional gas chromatography. In this thesis, we significantly enhance seminal ideas introduced in [54] through extensive data validation. Specifically, we validate our peak topographic methods across a set of thirty-four $GC \times GC$ injections from a diverse portfolio of petroleum sources, including a wide range of samples collected from the Macondo well, the source of the Deepwater Horizon disaster in the Gulf of Mexico, April 2010. The Macondo samples exhibit statistically significant match ($99.55 \pm 0.96\%$) against other closely related sources (refer Table 5.1). We build upon peak mapping and partitioning techniques introduced in [54] that combine source-specific and regional characteristics manifested through the $GC \times GC$ topography of neighboring oil sources. We also provide a robust quantitative measure for directly determining match between samples, without necessitating training datasets. This is a key distinction against supervised learning techniques [75–78] that necessitate strong ground truths derived from large training databases that may be difficult to avail in the event of localizing a natural seep or surveying connectivity between newly discovered oil prospects. Our contribution is summarized in three novel concepts introduced in [54] and expanded through extensive data validation in this thesis:

- Peak topography map (PTM), a feature representation that collectively captures $GC \times GC$ topography derived from the $GC \times GC$ chromatogram,

- Topography partitions, a threshold-based partitioning technique for discovering

source-specific and regional characteristics, and

- cross-PTM analysis, mathematical technique for directly determining match between two $GC \times GC$ separations without needing training datasets.

A natural outcome of PTM-based analysis is the discovery of topographic clusters (closely eluting groups of target and non-target biomarkers), which are key to understanding the regional and source-specific fingerprint.

5.3 Experimental Data Description

Table 5.5 (in Section 5.8) lists the thirty-four injections along with the corresponding details on sample identity and geographic origin. The injections may be classified into three groups:

- Fourteen injections clearly originating from the Macondo well, source of the Deepwater Horizon disaster;
- Three injections from non-Macondo well oil originating from three different sources in the Gulf of Mexico; and,
- Seventeen injections from diverse oil sources outside the Gulf of Mexico region.

In particular, injections 1 and 2 correspond to independent injections of a pre-spill sample taken directly from the Macondo well during normal operations before the disaster; injection 3 corresponds to a surface slick sample from the Macondo well collected after the spill; injection 4 is a post-spill sample collected directly from the broken riser pipe on June 21, 2010 [84, 91]; injections 5 through 14 correspond to

ten separate oil samples that were obviously from the Macondo well spill collected from grass blades along the Louisiana Gulf of Mexico coast; injections 15 and 16 are from two other crude oil sources from northern Gulf of Mexico and were collected before the Deepwater Horizon disaster, and injection 17 is collected from a natural oil seep in the Gulf of Mexico in 2006. The remaining injections correspond to distant sources unrelated to the Gulf of Mexico. For example, injections 18, 19 and 20 are independent consecutive injections of the National Institute of Standards and Technology (NIST) Standard Reference Material 1582 (its characteristics suggest it is derived from Monterey Shale and likely a California crude similar to injection 21).

5.3.1 $GC \times GC$-Flame ionization detector (FID) analysis

The samples were analyzed on a $GC \times GC$-FID system equipped with a Leco dual stage cryogenic modulator installed in an Agilent 7890A gas chromatograph configured with a 7683 series split/splitless auto-injector, two capillary columns, and a flame ionization detector. Samples were injected in splitless mode, and the split vent was opened at 1.0 minutes. The inlet temperature was 300 $^{\circ}C$. The first-dimension column and the dual stage cryogenic modulator reside in the main oven of the Agilent 7890A gas chromatograph. The second-dimension column is housed in a separate oven installed within the main GC oven. With this configuration, the temperature profiles of the first-dimension column, dual stage thermal modulator, and the second-dimension column can be independently programmed. The first-dimension column was a Restek Rtx^{-1}, (30 m, 0.25 mm I.D., 0.25 μm film thickness)

that was programmed to remain isothermal at 45 $^{\circ}C$ for 10 minutes and then ramped from 45 to 315 $^{\circ}C$ at 1.2 $^{\circ}C$ min^{-1}. Compounds eluting from the first dimension column were cryogenically trapped, concentrated, and re-injected (modulated) onto the second dimension column. The modulator cold jet gas was dry nitrogen, chilled with liquid nitrogen. The thermal modulator hot jet air was heated to 45 $^{\circ}C$ above the temperature of the main GC oven (thermal modulator temperature offset = 45 $^{\circ}C$). The hot jet was pulsed for 1.0 second every 12 seconds with a 5.0 second cooling period between stages. Second-dimension separations were performed on a SGE BPX50 (1 m, 0.10 mm I.D., 0.1 μm film thickness) that remained at 75 $^{\circ}C$ for 10 minutes and then ramped from 75 to 345 $^{\circ}C$ at 1.2 $^{\circ}C$ min^{-1}. The carrier gas was hydrogen at a constant flow rate of 1.1 mL min^{-1}. The FID signal was sampled at 100 data points sec^{-1}.

5.3.2 Methods

We introduce the Peak Topography Map (PTM) representation of $GC \times GC$ data as an informational method that characterizes the peak information across the $GC \times GC$ biomarker topography as a connected graph. Wherever applicable in this work, peak refers to a single second-dimension peak, and $GC \times GC$ region of interest (ROI) refers to the biomarker sub-region (hopanes and steranes) of a two-dimensional gas chromatogram.

5.3.2.1 Peak Topography Map (PTM) Representation

PTM is a scalable node-based representation computed over a pre-selected $GC \times GC$ ROI representing the biomarker compounds. The PTM representation is scalable because: (i) PTM computation can be scoped to a smaller sub-region within the chosen $GC \times GC$ ROI, and (ii) PTMs computed across disjoint $GC \times GC$ ROIs can be combined to construct the PTM across the union of these regions, e.g. PTMs for the hopanes and steranes can be computed separated and then combined to give the PTM over both hopanes and steranes. Each PTM consists of a two-dimensional node structure that preserved peak characteristics, e.g. peak height, peak location and order of elution.

Mathematically, each peak collapses into a single PTM node that stores two attributes: (i) the magnitude at the peak summit, and (ii) peak location. We represent information at a PTM node (denoted as η) with the value assignment $\eta = \{p, m, n\}$, where p denotes the peak summit value, and m and n respectively denote the first and second dimension retention time indices for the particular peak in the $GC \times GC$ image.

The nodes are stored as an ordered two-dimensional matrix, with the first dimension coinciding with the first dimension retention time indices and the second dimension storing the PTM nodes in the consecutive order of elution of peaks along the second dimension. Thus the $[q, m]$-th element of the PTM matrix with node value $\eta = \{p, m, n\}$ stores the q^{th} compound with peak height p, eluting along the second dimension with peak location $[n, m]$ in the $GC \times GC$ image. The number

of columns N of the PTM matrix represents the total number of first dimension modulations for the $GC \times GC$ ROI. The number of rows Q represents the maximum number of peaks eluting along the second dimension within the $GC \times GC$ ROI. The maximum number of peaks is computed across all second dimension indices within the $GC \times GC$ ROI. A PTM matrix column with fewer peaks than Q stores the PTM nodes in ascending order of peak locations, and populates the remaining entries with zeros to denote absence of a peak in those PTM nodes. We will henceforth refer to these entries in the PTM matrix that do not have a peak as blank nodes. To compute the PTM of a $GC \times GC$ ROI we normalize the PTM against the maximum value of the peaks. This normalization nullifies the effect of variable signal strengths between different injections by measuring all peak heights relative to the maximum signal strength within each $GC \times GC$ ROI. We locate all peaks within this ROI by employing a gradient-based maxima search (ref. Section 5.12). Peaks that fall below 5% of the maximum peak height within the $GC \times GC$ ROI are rejected as baseline noise. Mathematically, suppose the n^{th} column of a $GC \times GC$ image has κ_n number of peaks. The amplitudes and the locations of the peaks in this column can be stored in $Peak_n = \{p_{1,n}, p_{2,n}, \ldots, p_{\kappa_n,n}\}$ and $Loc_n = \{m_{1,n}, m_{2,n}, \ldots, m_{\kappa_n,n}\}$. We construct the $(l,n)^{th}$ element of its PTM representation matrix as:

$$PTM[l,n] = \begin{cases} p_{l,n} + j \times m_{l,n} & \text{if } 1 \leq l \leq \kappa_n \\ 0 & \text{if } l > \kappa_n \end{cases} \tag{5.1}$$

In other words, if l corresponds to a peak location along the n^{th} column of

the $GC \times GC$ image, then the $(l, n)^{th}$ node of the PTM is a complex number with its real part as the amplitude of the peak and the imaginary part as its location. In case l does not correspond to a peak, $(l, n)^{th}$ node will be zero. Therefore, the problem of comparing two $GC \times GC$ image, like I_{test} and I_{ref} will turn into the problem of comparing the nodes at the same location in their PTM representation matrices. Figure 5.2 provides a visual representation of PTM computation for two crude oil samples from the Gulf of Mexico (injections 1 and 16 in Table 5.5). Figure 5.3 shows the PTM corresponding to Figure 5.2.2, with the thirty-eight target PTM nodes labeled for identification with the target compounds in Figure 5.2.2. We note that the target compounds align according to their order of elution along the second dimension rather than absolute coordinates by design, thus rendering them robust against variability. The Procedure (in Section 5.10) details computational methods for ensuring PTM nodes compared across injections store the same compound within a pre-selected variability threshold.

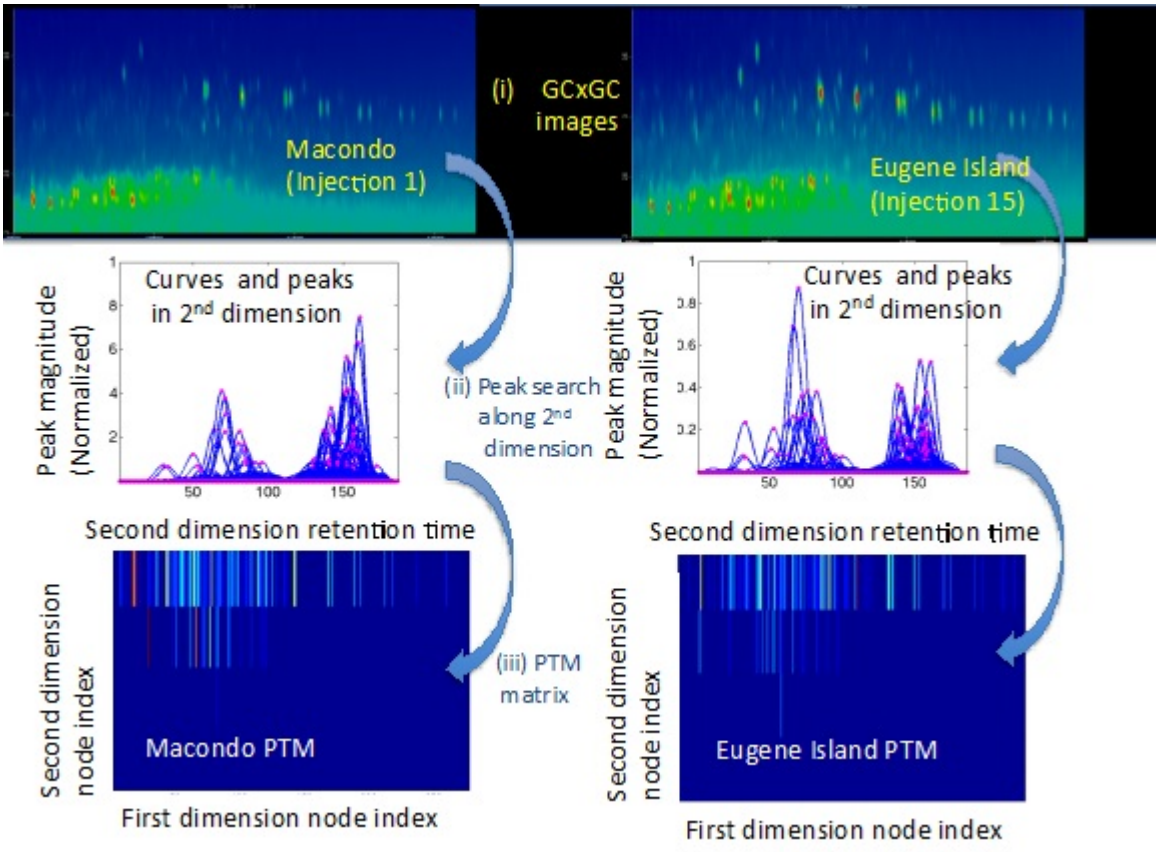

Figure 5.2: Sep-by-step PTM construction. Target biomarkers are labeled and itemized in Table 5.6. Total number of detected biomarker peaks (target and non-target) = 111, after removing peaks occupying lowest 5% of the $GC \times GC$ peak magnitude profile as baseline noise. Range of considered peak summits (highest:lowest) = 14.53:1.

5.3.2.2 Topography Partitioning: Direct $GC \times GC$ comparisons based on aligned PTMs

We introduce topography partitioning as a visual quantitative informational method to facilitate direct comparison between two $GC \times GC$ ROIs. Topography partitions provide intricate cross-comparison between oil samples highlighting nuances of their biomarker topographies. Topography partitions also form the basis for the cross-PTM score: a novel threshold-driven quantitative metric that provides a single numerical score for determining whether the two samples are a match. The key idea is to partition the $GC \times GC$ biomarker topography of a test sample based on which peaks, target and non-target, match against that of a reference sample using their respective PTM representations.

5.3.2.2.1 Mathematical computation of topography partitions

The peak-level match is determined using a peak ratio metric (ref. Equation in the procedure in 5.10). This peak ratio metric is calculated at the granularity of individual PTM nodes and assessed against a pre-selected threshold to decide a match between the test and reference samples for a given compound. These individual match assessments are then conducted across peak profiles spanning the $GC \times GC$ ROI.

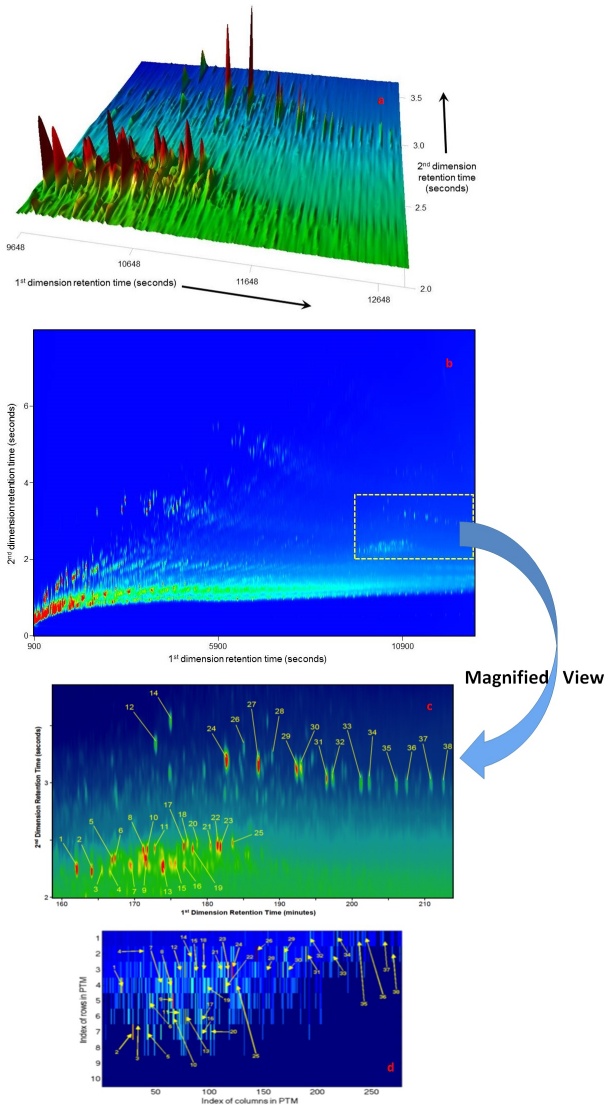

Figure 5.3: **a.** The three-dimensional view of $GC \times GC$ image of crude oil sample from Eugene Island, Gulf of Mexico, about 50 miles southwest of Macondo well, the oil source of the Deepwater Horizon disaster. **b.** The two-dimensional view of a. **c.** Detailed topography of biomarker region (hopanes and steranes) marked as yellow box in b.Target biomarkers are labeled and itemized in Table 5.6. **d.** PTM representation of Figure a and b. Thirty-eight target biomarkers are allocated to the numerically labeled PTM nodes.

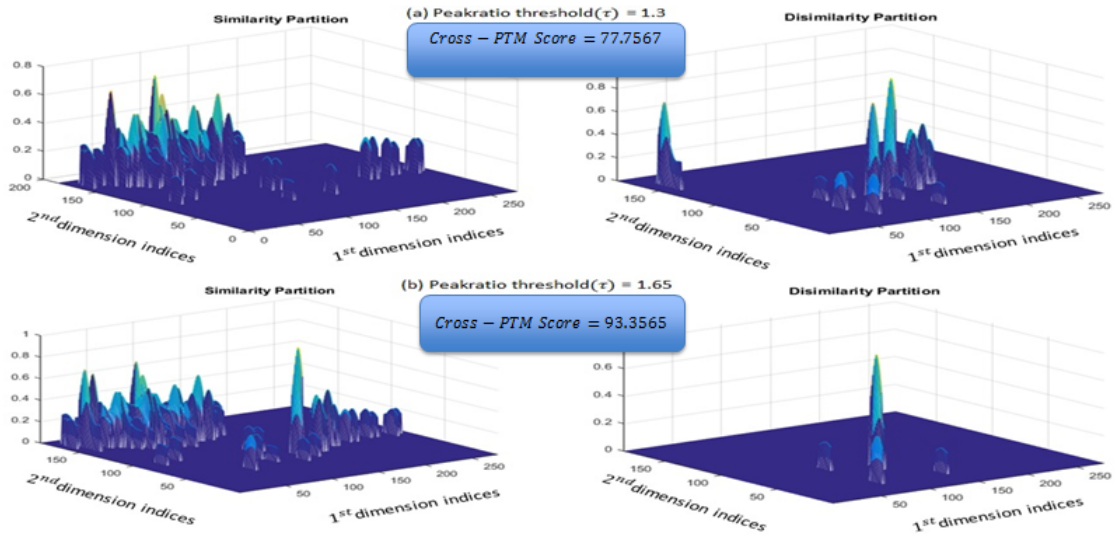

Figure 5.4: Topography partitioning of injection 15 (Eugene Island, Gulf of Mexico) with reference injection 4 (post-spill sample taken from the broken riser pipe of Macondo well) for peakratio threshold **a.** $\tau = 1.3$ and **b.** $\tau = 1.65$.

The topography is partitioned into similar and dissimilar peaks that meet or fall below the match threshold. The percentage of peaks in the similar topography generates the cross-PTM score. The two partitions are called similarity and dissimilarity partitions, where similarity indicates the partition of the test $GC \times GC$ ROI that matches that of the reference sample, and vice versa. Procedure of cross-PTM score provides a flowchart for determining the topography partitions of a test $GC \times GC$ ROI against a reference using PTM nodes. In the procedure of cross-PTM score, we have used a similarity criterion ρ between two nodes at the same location. As known, the function $max(a, a^{-1})$ has a value greater than or equal to one, with a value of one

when $a = a^{-1}$. In our case, a will be the ratio of the peaks at one location($a = \frac{p_{ref}}{p_{test}}$). In case $a = a^{-1} = 1$ (the peakratio is unity), the peaks at that location have exactly the same amplitudes($p_{ref} = p_{test}$) which we call them as equivalent nodes. But because of the baseline noise, column bleed and other chromatographic variability, the amplitudes of peaks may not be identical during the process of constructing the $GC \times GC$ image, and this requires a need to consider two peaks as being equivalent even if the value of the function $max(\cdot)$ deviates a little bit from unity. We then define a peak-threshold metric with a relaxing parameter ϵ which takes care of the noise deviation as following:

$$\tau = 1 + \epsilon, \qquad \epsilon > 0 \tag{5.2}$$

And claim two peaks as equivalent if the function for those peaks is less than or equal to τ(e.g. in Table 5.1 the results are shown for $\tau = 1.65$). Figure 5.4 illustrates the topography partitions of two Gulf of Mexico injections, which originate in distinct sources, but share regional characteristics that are captured in the similarity partitions. Similarity partition represents the common characteristics between the $GC \times GC$ topography between the two injections. Dissimilarity partition iterates the differences between the two $GC \times GC$ topographies. Therefore, topography partitions provide a threshold-dependent separation between the regional characteristics and source-specific features of a crude oil fingerprint.

When the peakratio τ is increased, less peaks between the injections are classified as dissimilar, as evidenced in Figures 5.4(a) and 5.4(b). We also note from

Figure 5.4(a) and 5.4(b) that both similarity and dissimilarity partitions consist of clusters of target (major peaks) and non-target (minor peaks) biomarkers. Thus the topography partitioning method allows discovery of important clusters of unknown non-target biomarkers around known target biomarkers that dominate the source-specific and regional fingerprints. We denote the $GC \times GC$ ROI of the test and reference samples as I_{ref} and I_{test}, the corresponding PTM matrices as PTM_{test} and PTM_{ref} , and the PTM nodes as η_{test} and η_{ref} respectively. To compare the PTMs, we follow the algorithm procedure of cross-PTM score. We denote the modified PTM_{test} procedure of cross-PTM scoreafter node insertions for alignment with PTM_{ref} as $PTM_{test,aligned}(PTM_{ref})$. The topography partitions are set up as a threshold classification of the test $GC \times GC$ ROI into two disjoint classes:

- Similarity partition: Portions of I_{test} corresponding to test PTM nodes (originally present or inserted) that meet the peakratio threshold τ (refer Step 3, procedure of cross-PTM score). We denote the similarity partition as $I_{test,similar}$.

- Dissimilarity partition: Portions of I_{test} corresponding to test PTM nodes (originally present or inserted) that does not meet the peakratio threshold τ (refer Step 3, procedure of cross-PTM score). We denote the dissimilarity partition as $I_{test,dissimilar}$.

We note that either partition not only includes the peak summits, but also the region under a peak. In the scenario where a node was inserted in the test PTM (refer Step 2b: Case 2, procedure of cross-PTM score) the I_{test} partition will include

the same peak sub-region corresponding to the equivalent peak region of η_{ref}, the ref PTM node.

5.3.2.2.2 Cross-PTM score calculation

The cross-PTM score, denoted as $S_\tau(I_{test}, I_{ref})$, is a PTM-based threshold-driven numerical comparison between the test and reference $GC \times GC$ ROIs. Mathematically, it is derived as the percentage of nodes in $PTM_{test,aligned}(PTM_{ref})$ that meet the threshold τ and therefore, belong in $I_{test,similar}$, i.e.,

$$S_\tau(I_{test}, I_{ref}) = \frac{|\eta_{test} \in PTM_{test,aligned}(PTM_{ref}) : \rho(m,n) \geq \tau|}{|\eta_{test} \in PTM_{test,aligned}(PTM_{ref})|} \tag{5.3}$$

Figure 5.4 illustrates topography partitioning for injection 4 (post-spill sample from Macondo well) in Table 5.5 using injection 15 (from Eugene Island, Gulf of Mexico) as the reference for direct cross-PTM comparison for different thresholds. We note that the higher value of τ selects more of the topography into the similar partition, as is to be expected.

5.4 Results and discussion

PTMs derived from $GC \times GC$ biomarker ROIs corresponding to thirty-four injections (refer Table 5.5 for details on origin) were compared pairwise against each other based on the threshold-based cross-PTM score. The thirty-four injections compared span across thirty-one distinct oil samples that originate from nineteen distinct sources. Fourteen samples originate from the Macondo well, source of the Deepwater

Horizon disaster, including two pre-spill samples, and twelve post-spill samples collected at diverse locations after the Deepwater Horizon disaster, e.g. the plume at the base of the Macondo well, grass blades on the Louisiana coastline, and oil slicks collected kilometers away from the disaster site (details provided in Table 5.5). These samples were collected in areas well documented [67, 81] to be heavily contaminated by the Deepwater Horizon disaster compared to the background.

We evaluate the cross-PTM score as a function of the peakratio threshold across a diverse selection of injection pairs. We examine the robustness of intra-class match between injections of same origin against inter-class distinction between injection pairings from different origins. Specifically, we compare the fourteen Macondo injections (injections 1-14 in Table 5.5) against each other and against other sources within and outside the Gulf of Mexico region. We also compare the strength of Macondo vs. Macondo match against three other Gulf of Mexico injections (injections 15-17 in Table 5.5): (i) Eugene Island, (ii) Southern Louisiana Crude (SLC) and (iii) a Gulf of Mexico natural seep. Three consecutive injections from a non-Gulf of Mexico NIST sample originating in the Monterey area are also analyzed as an ideal intra-class case study, independent of any co-provenance bias with the Gulf of Mexico samples.

Figure 5.4 plots the average cross-PTM score as a function of peakratio threshold across important comparison classes. Figure 5.17 in Section 5.13 provides the statistical performance of the cross-PTM score for matching Gulf of Mexico injection pairs, with emphasis on distinguishing the fourteen Macondo injections against non-

Macondo Gulf of Mexico injections. We note that consistently the intra-class match between Macondo injections is statistically higher than the inter-class score between Macondo and other Gulf of Mexico injections. In Figure 5.6, the cross-PCA score as a function of the number of principle components have been plotted. The statistical performance of the cross-PCA score for matching Gulf of Mexico injection pairs has been shown in Figure 5.18 in Section 5.13.

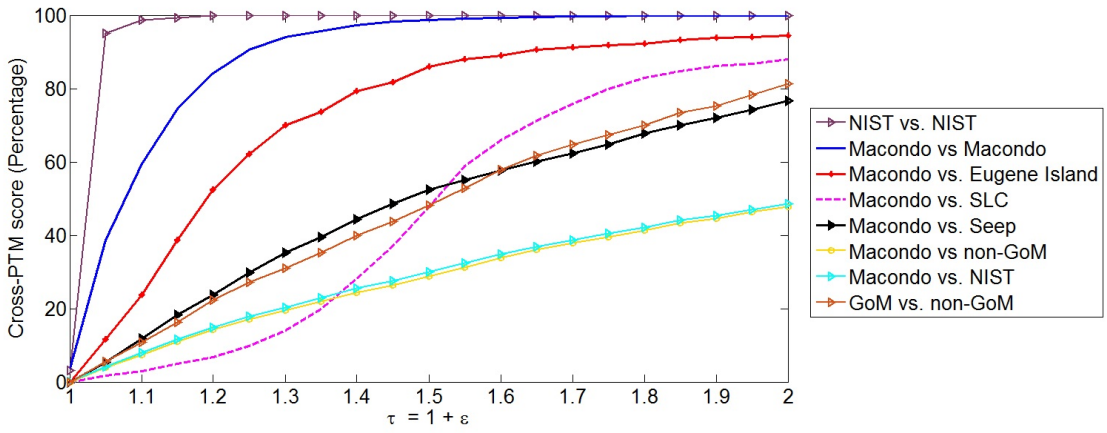

Figure 5.5: Mean cross-PTM scores plotted as a function of the peakratio threshold τ for important intra-class (same source) and inter-class (distinct sources) comparisons. Each plot shows the average cross-PTM score taken over all possible pairings of injections for the corresponding comparison class (e.g. NIST vs. NIST plot shows the average cross-PTM score for three possible parings between the three NIST injections).

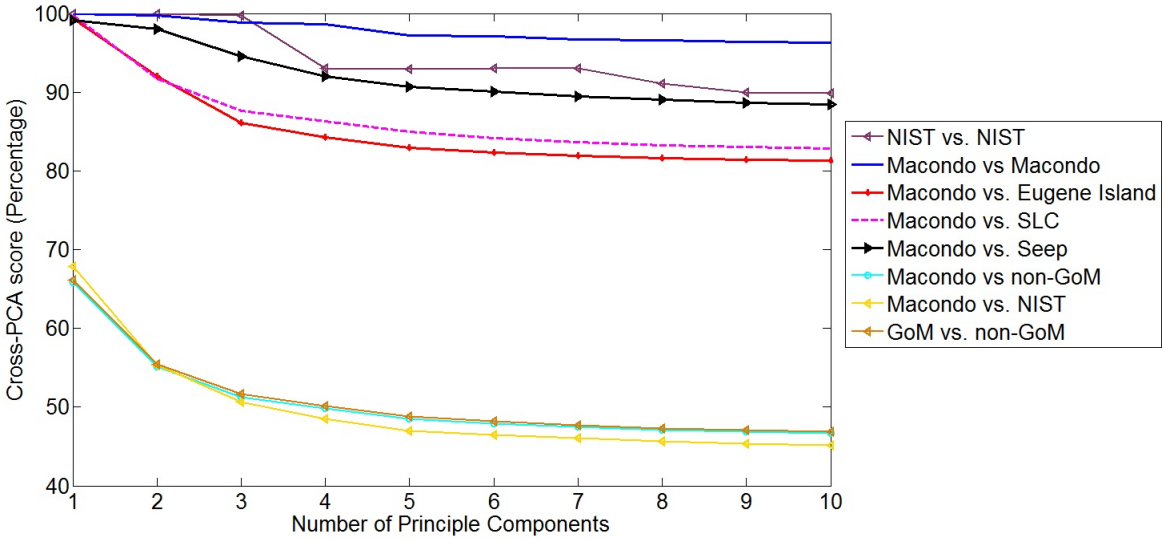

Figure 5.6: Mean cross-PCA scores plotted as a function of the peakratio threshold τ for important intra-class (same source) and inter-class (distinct sources) comparisons. Each plot shows the average cross-PCA score taken over all possible pairings of injections for the corresponding comparison class (e.g. NIST vs. NIST plot shows the average cross-PCA score for three possible parings between the three NIST injections).

5.4.1 Best-case scenario for same-source match: NIST vs. NIST

To provide a neutral baseline for best-case performance, we compare three NIST injections (injections 19-21 in Table 5.5), all of which were taken from the same sample of non-Gulf of Mexico origin. The NIST injections were run consecutively under practically identical experimental conditions. We observe in Figure 5.4 that the NIST vs. NIST cross-PTM score rapidly reaches 100% match with increasing

peakratio threshold. This is to be expected as the $GC \times GC$ biomarker topographies of injections run consecutively from the same sample are expected to be very similar, if not identical. In reality, cross-comparisons for source determination are made between injections from different samples that may have same origin but are not consecutive runs from the same physical sample. $GC \times GC$ topographies for same-source injections from different samples are therefore, bound to exhibit more variation due to shifting of minor peaks, co-elution of different biomarkers, as well as baseline variability. Thus we expect the NIST vs. NIST cross-PTM performance to provide an idealized upper bound to measure cross-PTM score performance.

5.4.2 Comparison between Macondo injections from fourteen distinct samples

The fourteen Macondo injections exhibit a range of 105-131 detected peaks spanning target and non-target $GC \times GC$ biomarkers with highest-to-lowest peakratio within an injection ranging from 14.27-16.22. Majority of the peaks considered are non-target biomarkers (only 38 target biomarkers present among over 100 biomarkers considered) and thus offer a nuanced cross-PTM interpretation that accounts for both target and non-target contributions to an oil fingerprint. From Table 5.1 we observe that the inter-class match between Macondo injection pairings is well within statistical range, i.e., within one standard deviation (σ) of the statistical mean (μ), for robust ($\mu \pm \sigma$) differentiation against other Gulf of Mexico injections.

Specifically, at the choice of $\tau = 1.65$ the Macondo injections exhibit ($99.55 \pm 0.96\%$,$Median$: 100%) intra-class match, which is sufficient to distinguish against

inter-class cross-PTM score with other Gulf of Mexico injections.

This choice of peakratio, τ, was empirically selected at $\tau = 1.65$ which was observed to give the best distinguishment between the Macondo and other Gulf of Mexico sources.

Table 5.1: Percentage match between different Gulf of Mexico sources against Macondo injections for PTM with the optimal choice of $\tau = 1.65$ and for PCA with two principle components.

Method	Mac vs. Mac	EI vs. Mac	SLC vs. Mac	Ns vs. Mac
PTM	$99.55\% \pm 0.96\%$	$90.66\% \pm 2.096\%$	$71.28\% \pm 11.03\%$	$60.12\% \pm 3.064\%$
PCA	$99.76\% \pm 0.26\%$	$91.98\% \pm 0.14\%$	$91.71\% \pm 0.24\%$	$98.01\% \pm 0.51\%$

5.4.3 Comparison between Gulf of Mexico injections and injections outside the region

We observe from Table 5.1 and Figure 5.4 that using $(\mu \pm \sigma)$ differentiation the Gulf of Mexico injections are robustly differentiated against each other and also exhibit considerable distinction against sources outside the Gulf of Mexico region. In conclusion, we observe that the mean and median performance of the cross-PTM score is highly robust in source distinction and worst-case performance is sensitive to choice of peakratio τ and number of detected peaks. Thus, the PTM approach combines target and non-target analysis to address multi-layered forensic questions regarding

whether the injections are from the same sample, from different samples of same origin, from samples of different origin but similar locale, and so on as demonstrated above in our analysis based on a unique and diverse set of oil samples.

5.4.4 Differentiation between PTM and PCA in scope and performance

As indicated earlier the proposed methods in chemometrics such as PCA can be applied towards quantitative $GC \times GC$ interpretation. However, purely statistical methods limit interpretation to peak aggregates, and as such, cannot provide peak-level interpretation. Therefore, by design PCA and similar multivariate statistical methods are compound-agnostic and cannot provide quantitative comparison based on relative compound concentrations in two complex mixtures. In particular, PCA analysis projects the $GC \times GC$ image along the main directions of data variance and therefore, is well-suited to application scenarios where the incentive is dimensionality reduction and compound-agnostic comparison between weakly correlated sources.

The primary aim of this work is to provide quantitative peak-level interpretation beyond target biomarkers, with the end goal of robust differentiation between petroleum sources that share regional commonalities, and therefore, have highly correlated $GC \times GC$ fingerprints. So, even minor nuances between two sources can carry important information to help us separate them once they are extracted from two closely located regions.

This differentiation between the two interpretation methods can be easily seen in Table 5.1, where We compare the best performance for differentiating between

GoM oil sources using PTM ad PCA cross-comparison scores. The optimal parameter choice for each method is provided (number of components for PCA and peakratio threshold for PTM).

The intra-class match (Macondo vs. Macondo) is slightly higher using PCA than PTM but the inter-class differentiation (Macondo vs. other local sources) is significantly more robust using PTM over PCA. This is to be expected as PCA is biased towards the common regional fingerprint of the Gulf of Mexico locale, which constitutes the dominant component of data variance of $GC \times GC$ separations of crude oil collected in this region.

Mathematically, we can perform PCA cross-comparison between these correlated courses based on the non-dominant components, but these are typically vulnerable to baseline noise and other uncertainties, and as such, not reliable for robust source differentiation. This is evident in Figure 5.6, where increasing the number of components increases gap between inter-class scores but also reduces the intra-class (Macondo vs. Macondo) match. On the other hand, cross-PTM match scores (Figure 5.4) consistently provide high intra-class and considerably lower inter-class match scores over a wide range of the peakratio threshold.

In summary, PCA enables statistical distinction between two $GC \times GC$ separations which have been extracted from geologically unrelated sources far apart from each other, but falls short of robust differentiation between strongly correlated sources located within the same region. PTM analysis provides peak-cognizant quantitative interpretation that can robustly differentiate between $GC \times GC$ separations between

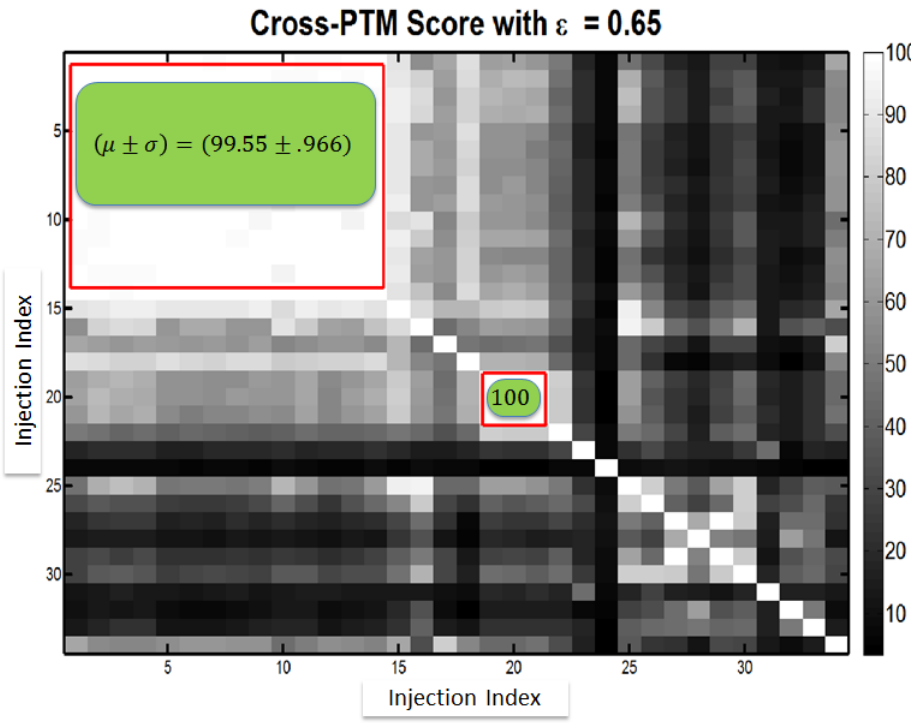

Figure 5.7: Cross-PTM score for the petroleum dataset.

strongly correlated but distinct sources that share the regional fingerprint.

We have also plotted the cross-PTM score for the thirty-four injections against each other in Figure 5.7. As can be seen the diagonal elements are 100% match, because samples are being tested against each other. The first fourteen samples have Macondo samples which their mean and the standard deviation is $99.55 \pm .96$ which is a pretty high match. The NIST samples have the perfect match of 100% which has been shown in the figure.

5.5 Robust Peaks

Due to the noise, experimental errors and the environmental events the forensics undergo some deviation. So, if we have extracted two petroleum sources from the same geographical region, there might be some difference in their patterns. That would be nice to develop a method to see if there are some peaks that have remained the same in the sample extracted from the same region. These peaks will not change, therefore we call them robust peaks. Hence, we say the main role of a sample to be called to be from a specific region is on the shoulders of some of the peaks, and not all.

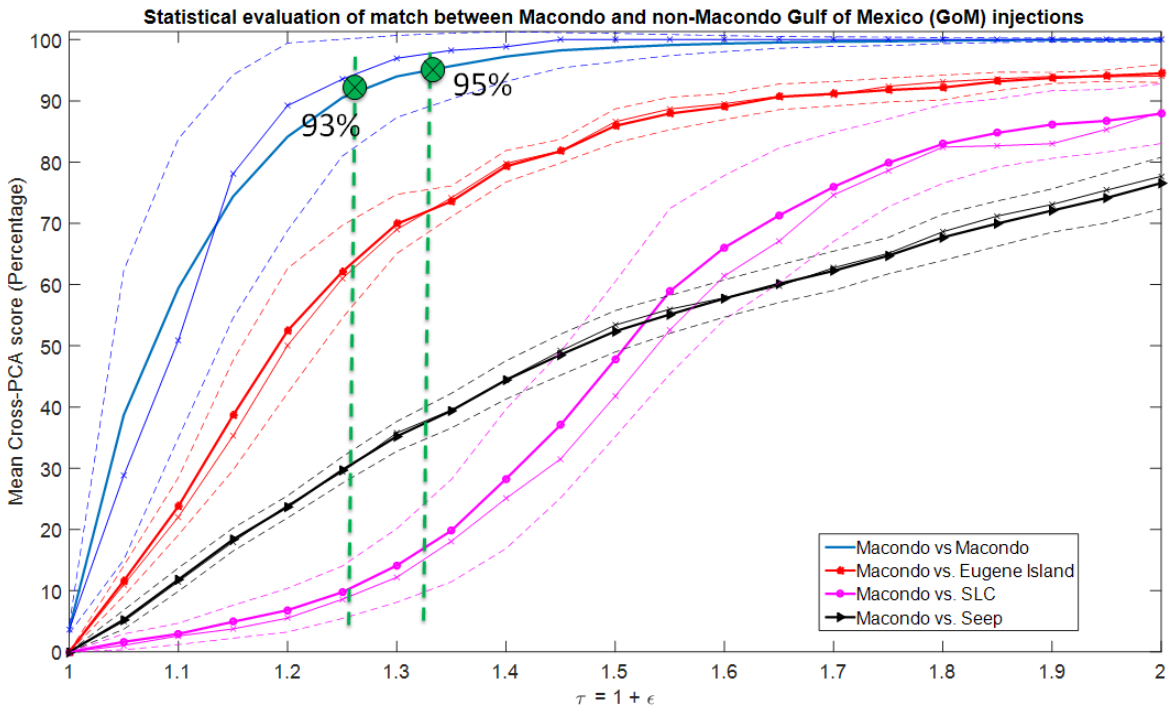

Figure 5.8: Statistical evaluation of match between the Macondo and non-Macondo.

The statistical evaluation of match between the Macondo and non-Macondo have been shown in Figure 5.8. The maximum amount of differentiation between the Macondo versus Macondo and Macondo versus Eugene Island occurs between the two bars shown in the Figure. So we can set the parameter M_p introduced in Section 3.3.2 somewhere between the points shown in Figure 5.8. In order to recap, the block diagram shown in 5.9 depicts the procedure to construct the $\tau - map$ for a family of images from one specific geographical region. We first set a reference image as the representative sample of the family. Then, we go over each of the peaks of the reference image and compute the cross-PTM score for the local neighborhood around the peak and save the minimum τ in which the mean cross-PTM score for that class is at least M_p. In case, there is a missing peak in at least one of the images from the family, the corresponding peak should be claimed as *non-robust* peaks and its τ will be saved as $\tau = -1$. In case there is peak in all of the images but the mean cross-PTM score for that peak is less than M_p will be a τ out of the feasible set of τ's. We call all of the peaks in which their corresponding τ is any number in the valid range ,and of course not equal to -1, *robust* peaks. In this work,the valid range of τ's is $1 \leq \tau \leq 2$.

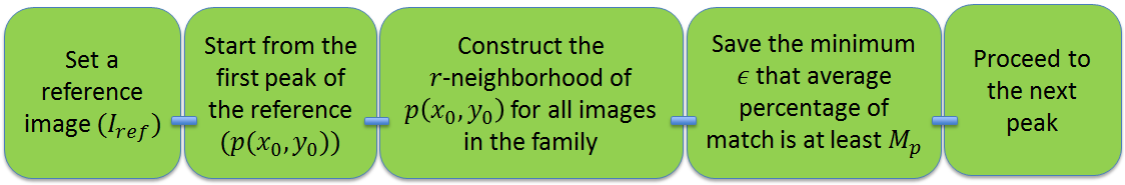

Figure 5.9: The block diagram of the $\tau - map$ image introduced in 3.3.2.

Figure 5.10 shows the $\tau - map$ image using the first pre-spill sample from the Macondo well with the other members of the family from Macondo.

As shown in Figure 5.10 most of the peaks in the reference image need $\tau = 1.3$ to be able to have the percentage of match of at least $M_p = 94\%$. There is just one peak at the location of $(x_0, y_0) = (72, 133)$ which does not exist in all of the samples, hence its corresponding τ will be -1.

In figure 5.11 the histogram of the peaks have been shown where the left image relates to the first pre-spill Macondo sample and the right one relates to one of the post-spill sample, grass blade. As can be seen the number of the peaks having $\tau = 1.3$ in the first sample is 49 and that of the grass blade is 41. The total number of peaks in the first and the grass blade samples are 111 and 106, respectively. Therefore, for the first pre-spill sample $49/111 = 44.14\%$ of the peaks agree that the appropriate τ is 1.3, where in the grass blade case this number is $41/106 = 38.68\%$. It shows that the agreement between the peaks of pre-spill is more than the post-spill sample, as expected. Although, having 5 more peaks has led the first pre-spill sample to have a non-robust peak.

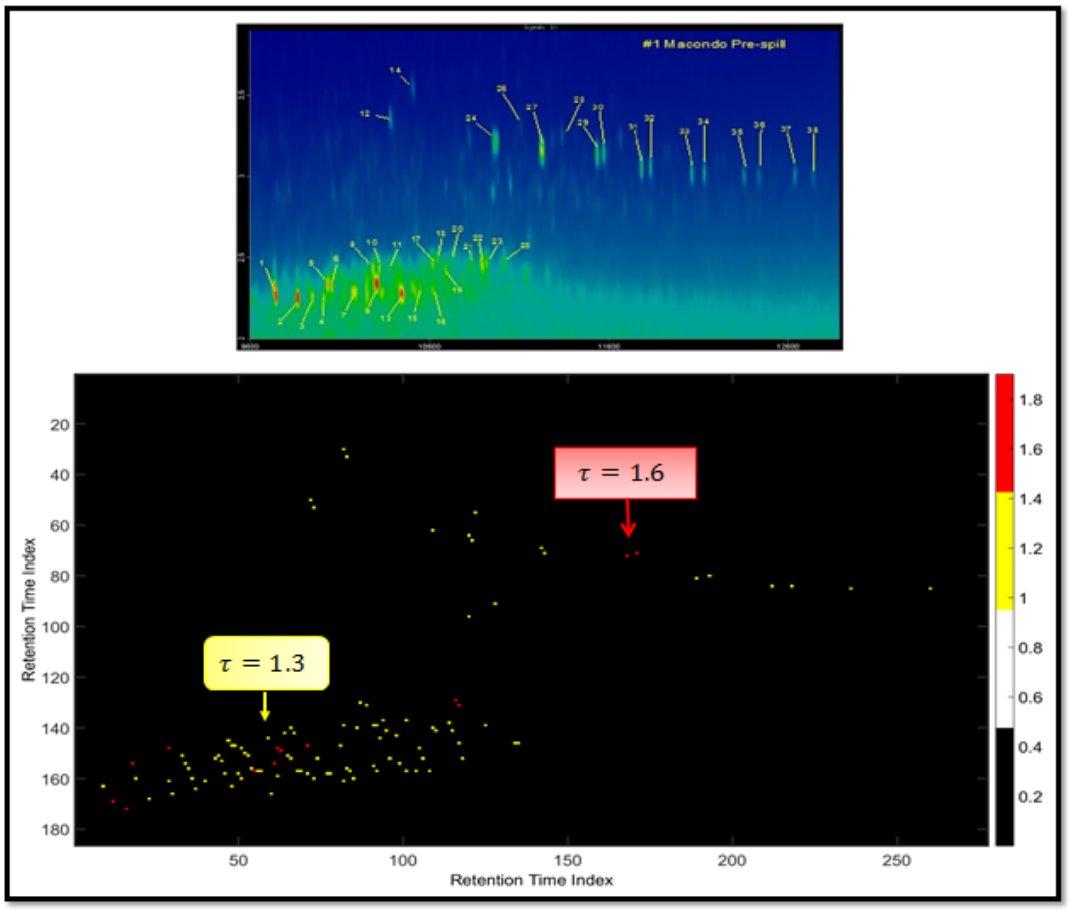

Figure 5.10: $\tau - map$ image for the Macondo well family of images. The figure on the top shows the target peaks selected by the chemists in the lab.

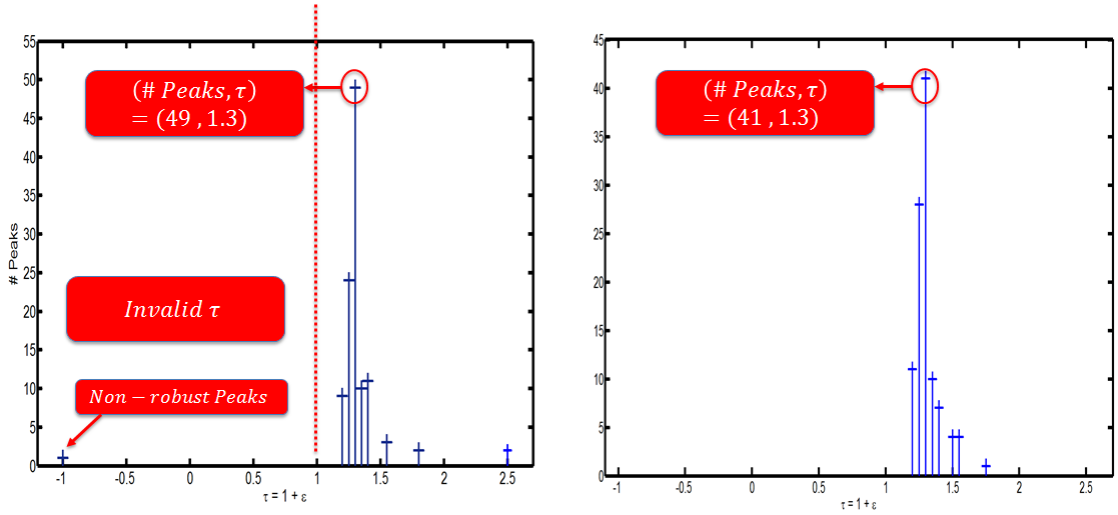

Figure 5.11: Histogram of the number of peaks occurring at different values of τ for a pre-spill (left) and post-spill (right) Macondo sample.

The values of the means are shown in Table 5.2. As can be seen the highest average of number of peaks occurs at $\tau = 1.3$. Remember, we chose $M_p = 94\%$ where the maximum differentiation occurred and the corresponding τ was 1.3. Here, we have come to a very important observation that most of the peaks agreed that in order to have a percentage of match of at least M_p, they need $\tau = 1.3$ and the whole image also needed this value for having the percentage of match of $M_p = 94\%$.

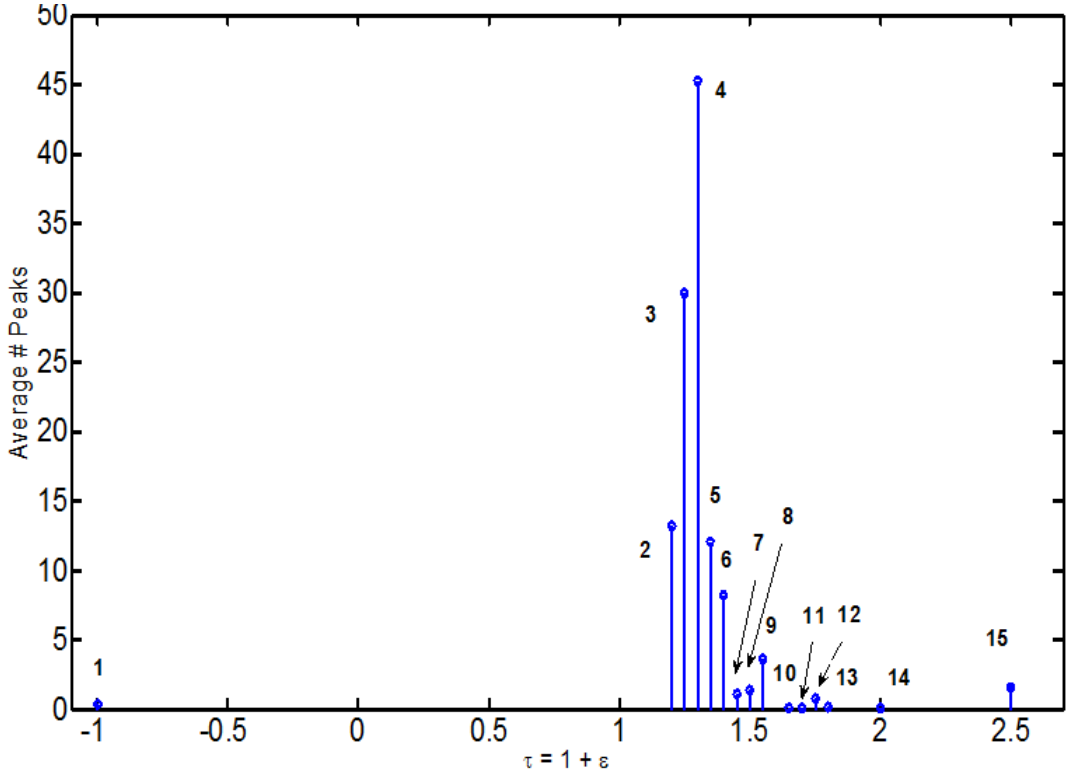

Figure 5.12: Histogram of the average number of peaks occurring at different values of τ.

Table 5.2: Average number of peaks as a function of τ.

Peak NO.	τ	**Average number of points**
1	-1	$0.3571 \pm 0.4972\%$
2	1.2	$13.2143 \pm 3.3092\%$
3	1.25	$30 \pm 3.8829\%$
4	1.3	$45.2143 \pm 3.5121\%$
5	1.35	$12.0714 \pm 2.4326\%$
6	1.4	$8.2143 \pm 1.9682\%$
7	1.45	$0.0714 \pm 1.0995\%$
8	1.5	$1.3571 \pm 1.2774\%$
9	1.55	$3.6429 \pm 0.9288\%$
10	1.65	$0.0714 \pm 0.2673\%$
11	1.7	$0.0714 \pm 0.2673\%$
12	1.75	$0.7857 \pm 0.4258\%$
13	1.8	$0.1429 \pm 0.5345\%$
14	2	$0.0714 \pm 0.2673\%$
15	2.5	$1.5714 \pm 1.4525\%$

Where $\tau = -1$ corresponds to the case where there is at least one figure in which there is a missing peak. We have plotted the out-of-range τ by $\tau = 2.5$.

5.6 Applying PTM on Breastmilk Dataset

The breastmilk dataset has been provided by Dr. Hans Lehmler and Iza Korwel where the information regarding the dataset is given in Table 5.3. The result of applying PTM on the dataset has been shown in Table 5.4. As can be seen the PTM method performs reasonably well in distinguishing the samples from the same family.

Table 5.3: Breastmilk Dataset Sheet

Injection Number	**Sample name**	**Sample Type**
1	0REF03	milk reextracted
2	A006	milk
3	B013	milk
4	B014	milk
5	C022	milk
6	C023	milk
7	D045	milk
8	D049	milk
9	D052	milk
10	D059	milk

Table 5.3 Continued.

Injection Number	**Sample name**	**Sample Type**
11	E086	milk
12	hexane1	hexane
13	hexane2	hexane
14	BLANK05	blank
15	BLANK06	blank
16	NIST01	NIST reference material
17	OPR03	OPR
18	REF03	authentic standard
19	SPREF03	spiked reextracted milk
20	0REF02	milk reextracted
21	A003	milk
22	B010	milk
23	D035	milk
24	D046	milk
25	D051	milk

Table 5.3 Continued.

Injection Number	**Sample name**	**Sample Type**
26	D060	milk
27	E072	milk
28	E074	milk
29	E076	milk
30	E083	milk
31	hexane3	hexane
32	hexane4	hexane
33	LRM02	reference material
34	BLANK03	blank
35	BLANK04	blank
36	OPR02	OPR
37	REFSTD0331	authentic standard
38	SPREF02	spiked reextracted milk
39	0REF01	milk reextracted
40	A005	milk
41	A008	milk

Table 5.3 Continued.

Injection Number	**Sample name**	**Sample Type**
42	B009	milk
43	D037	milk
44	D042	milk
45	D048	milk
46	E050	milk
47	E054	milk
48	E058	milk
49	hexane5	hexane
50	hexane6	hexane
51	LRM01	reference material
52	BLANK01	blank
53	BLANK02	blank
54	OPR01	OPR
55	REFSTD0323	authentic standard
56	SPREF01	spiked reextracted milk
57	0REF04	milk reextracted

Table 5.3 Continued.

Injection Number	**Sample name**	**Sample Type**
58	A002	milk
59	B011	milk
60	C024	milk
61	C027	milk
62	D047	milk
63	E062	milk
64	E063	milk
65	E065	milk
66	E066	milk
67	hexane7	hexane
68	LRM03	reference material
69	BLANK07	blank
70	BLANK08	blank
71	OPR04	OPR
72	REFSTD0507	authentic standard
73	SPREF04	spiked reextracted milk

Table 5.3 Continued.

Injection Number	**Sample name**	**Sample Type**
74	0REF05	milk reextracted
75	A001	milk
76	B016	milk
77	B017	milk
78	C020	milk
79	C025	milk
80	C028	milk
81	D054	milk
82	D033	milk
83	D035	milk
84	D039	milk
85	E061	milk
86	E069	milk
87	hexane9	hexane
88	LRM05	reference material
89	BLANK09	blank
90	BLANK10	blank

Table 5.3 Continued.

Injection Number	**Sample name**	**Sample Type**
91	REFSTD0513	authentic standard
92	SPREF05	spiked reextracted milk
93	A007	milk
94	B017	milk
95	C018	milk
96	C019	milk
97	D032	milk
98	D031	milk
99	D041	milk
100	D045	milk
101	D067	milk
102	D079	milk
103	D078	milk
104	hexane10	hexane
105	hexane11	hexane
106	LRM04	reference material
107	BLANK11	blank

Table 5.3 Continued.

Injection Number	**Sample name**	**Sample Type**
108	OPR5	OPR
109	0REF06	milk reextracted
110	REFSTD0515	authentic standard
111	REFSTD0519	authentic standard
112	SPREF06	spiked reextracted milk
113	hexane8	hexane
114	0REF08	milk reextracted
115	B010	milk
116	B012	milk
117	D032	milk
118	D044	milk
119	D050	milk
120	D052	milk
121	D053	milk
122	E068	milk
123	E077	milk

Table 5.3 Continued.

Injection Number	**Sample name**	**Sample Type**
124	E084	milk
125	MBLNK15	blank
126	NIST02	NIST reference material
127	OPR08	OPR
128	REFSTD0602	authentic standard
129	SPREF08	spiked reextracted milk
130	MBLNK17	blank
131	MBLNK18	blank
132	0REF09	milk reextracted, different milk
133	A004	milk
134	B015	milk
135	C026	milk
136	D036	milk
137	D043	milk
138	E071	milk

Table 5.3 Continued.

Injection Number	**Sample name**	**Sample Type**
139	E075	milk
140	E082	milk
141	hexane12	hexane
142	LRM08	reference material
143	OPR09	OPR
144	0REF07	milk reextracted, different milk
145	A003	milk
146	A005	milk
147	B009	milk
148	C027	milk
149	D031	milk
150	D037	milk
151	D048	milk
152	D055	milk
153	D057	milk
154	E064	milk

Table 5.3 Continued.

Injection Number	**Sample name**	**Sample Type**
155	E070	milk
156	E081	milk
157	hexane13	hexane
158	LRM07	reference material
159	MBLK13	blank
160	OPR07	OPR
161	REF0527	authentic standard
162	SPREF07	spiked reextracted milk
163	E079	milk
164	E073	milk
165	D040	milk
166	D034	milk
167	D029	milk
168	0REF10	milk reextracted, different milk
169	D057	milk

Table 5.3 Continued.

Injection Number	**Sample name**	**Sample Type**
170	LRM09	reference material
171	MBLK19	blank
172	OPR10	OPR
173	REFSTD0709	authentic standard
174	SPREF09	milk
175	SPREF10	milk
176	E080	milk
177	E085	milk
178	REFSTD0709	authentic standard

Table 5.4: Percentage match for breast milk dataset with $\tau = 1.7$.

Sample	Number of Samples	$\mu \pm \sigma$
Milk	99	$96.84\% \pm 4.97\%$
authentic standard	11	$97.5\% \pm 3.9\%$
Hexane	13	100%
OPR	9	$99.31 \pm .6\%$
Blank	16	$99.63 \pm .24\%$
NIST reference material	2	98.81%
reference material	8	$96.77 \pm 4.85\%$
spiked reextracted milk	10	$97.76 \pm 2.53\%$
milk reextracted	7	$99.26 \pm 0.49\%$
milk reextracted, different milk	3	$99.93 \pm 0.07\%$

5.7 Conclusions

We introduce three novel concepts in this work: (i) Peak topography map (PTM), a feature representation that collectively captures the $GC \times GC$ topography, (ii) PTM-based topography partitions, a threshold-based visualization technique for direct cross-sample comparisons, and (iii) cross-PTM analysis technique based on a quantitative score and topography partitions. Specifically, we address the broader question of what aspects of two oil samples are similar, and where do they differ, based on the molecular fossil (biomarker) topography of their $GC \times GC$ separations.

Our methodology provides a mathematical framework for quantitative visualization of $GC \times GC$ at the granularity of individual peaks across target and non-target compounds as well as groups of peaks connected by topographic proximity. Such multi-scale interpretation is enabled by the combination of individual peakratio evaluation between equivalent nodes, topography partitioning, and cross-PTM score spanning the collective topography of $GC \times GC$ ROI. Thus the PTM method enables $GC \times GC$ forensic interpretation across well-known target biomarkers, while including the nuances of lesser-known non-target compounds clustered around the target peaks. This allows potential discovery of hitherto unknown connections between biomarkers that are related through topographic similarity between samples.

5.8 Tables of injections and target biomarkers

All of the samples were collected without any necessary legal/operation permission or impacted endangered or protected species. However as part of the response to the Deepwater Horizon and a request from the official response, we collected the sample on June 21, 2010 at the Macondo well with assistance from the United States Coast Guard. This field sample is considered one of the most important from the Deepwater Horizon and eventually was involved in the Federal decision on the volume of oil released. Refer to [84, 91] for more information on its collection and usage in flow rate calculations. These samples were collected in areas well documented ([67, 81]) to be heavily contaminated by the Deepwater Horizon disaster compared to the background.

Table 5.5: List of Thirty-four injections across thirty-one samples from nineteen distinct sources

Injection Number	**Sample name**	**Sample description based on origin**
1,2	Macondo well oil	Sampled from the Macondo well before the Deepwater Horizon disaster, which occurred on April 20, 2010, as part of normal petroleum operations. It is often called the pre-spill.
3	Surface sample	Oil droplet collected near the Deepwater Horizon blowout during the spill (June 2010)
4	Macondo well oil	Collected directly from the broken riser pipe at the Macondo well 6/21/2010. (Referred as MW-2 in Reddy et al 2011)

Table 5.5 Continued.

Injection Number	**Sample name**	**Sample description based on origin**
5	Grass blade-1	First distinct and separate sample scraped from one blade of marsh grass on May 30, 2010 about 200 km from the Deepwater Horizon blowout This sample and the following were clearly from the disaster based on tracking of surface slicks to this location.
6	Grass blade-2	Second distinct and separate sample scraped from one blade of marsh grass on May 30, 2010 about 200 km from the Deepwater Horizon blowout This sample and the following were clearly from the disaster based on tracking of surface slicks to this location.

Table 5.5 Continued.

Injection Number	**Sample name**	**Sample description based on origin**
7	Grass blade-3	Third distinct and separate sample scraped from one blade of marsh grass on May 30, 2010 about 200 km from the Deepwater Horizon blowout This sample and the following were clearly from the disaster based on tracking of surface slicks to this location.
8	Grass blade-4	Fourth distinct and separate sample scraped from one blade of marsh grass on May 30, 2010 about 200 km from the Deepwater Horizon blowout This sample and the following were clearly from the disaster based on tracking of surface slicks to this location.

Table 5.5 Continued.

Injection Number	**Sample name**	**Sample description based on origin**
9	Grass blade-5	Fifth distinct and separate sample scraped from one blade of marsh grass on May 30, 2010 about 200 km from the Deepwater Horizon blowout This sample and the following were clearly from the disaster based on tracking of surface slicks to this location.
10	Grass blade-6	Sixth distinct and separate sample scraped from one blade of marsh grass on May 30, 2010 about 200 km from the Deepwater Horizon blowout This sample and the following were clearly from the disaster based on tracking of surface slicks to this location.

Table 5.5 Continued.

Injection Number	**Sample name**	**Sample description based on origin**
11	Grass blade-7	Seventh distinct and separate sample scraped from one blade of marsh grass on May 30, 2010 about 200 km from the Deepwater Horizon blowout This sample and the following were clearly from the disaster based on tracking of surface slicks to this location.
12	Grass blade-8	Eighth distinct and separate sample scraped from one blade of marsh grass on May 30, 2010 about 200 km from the Deepwater Horizon blowout This sample and the following were clearly from the disaster based on tracking of surface slicks to this location.

Table 5.5 Continued.

Injection Number	**Sample name**	**Sample description based on origin**
13	Grass blade-9	Nineth distinct and separate sample scraped from one blade of marsh grass on May 30, 2010 about 200 km from the Deepwater Horizon blowout This sample and the following were clearly from the disaster based on tracking of surface slicks to this location.
14	Grass blade-10	Tenth distinct and separate sample scraped from one blade of marsh grass on May 30, 2010 about 200 km from the Deepwater Horizon blowout This sample and the following were clearly from the disaster based on tracking of surface slicks to this location.
15	Eugene Island crude	Collected from a drilling rig in the Eugene Island block 330, Gulf of Mexico.

Table 5.5 Continued.

Injection Number	**Sample name**	**Sample description based on origin**
16	Southern Louisiana Crude	SRM prepared by the US Environmental Protection agency (WP681). Collected in the 1970s from the Gulf of Mexico.
17	Gulf of Mexico seep	Natural oil seep (Collected in 2006 - 560 miles SW of the Deepwater Horizon disaster in the Gulf of Mexico).
18, 19, 20 (Injections from same sample analyzed consecutively)	NIST SRM-1582	Standard reference material (SRM) National Institute of Standards and Technology (NIST), likely from Monterey Shale.
21	Monterey crude	Crude oil collected off the coast of Santa Barbara, CA.
22	Kamchatka crude	Crude oil collected from Russia.

Table 5.5 Continued.

Injection Number	**Sample name**	**Sample description based on origin**
23	Ardjuna basin crude	Crude oil collected off the coast of Indonesia.
24	Exxon Valdez	Collected from the Exxon Valdez cargo after the March 1989 grounding.
25	Permean Basin	Crude oil collected in West Texas.
26	Arabian light crude	SRM prepared by the US Environmental Protection agency.
27	Kuwait export crude	First of three cargoes spilled from the MT Hebei Spirit (occurred 12/2007).
28	UAE - Upper Zakum crude	Second of three cargoes from the MT Hebei Spirit (occurred 12/2007).
29	Iranian heavy crude	Third of three cargoes from the MT Hebei Spirit (occurred 12/2007).
30	PetroEcuador crude	Crude oil collected off the coast of Ecuador.
31	Green River shale	Crude oil produced from the Green River Shale.
32	Texas crude	Collected from south central Texas.

Table 5.5 Continued.

Injection Number	**Sample name**	**Sample description based on origin**
33	Nigerian crude	Sample collected from Nigeria.
34	Angola Crude	Sample collected from Angola.

5.9 List of hydrocarbon biomarkers labeled as targets in the manuscript

Table 5.6: List of compounds labeled in Figures 5.1(c), 5.3(c) and 5.3(d)

Injection Number	**Sample name**	**Sample description based on origin**
1	DiaC27Ba-20S	13β(H),17α(H)-20S-diacholestane
2	DiaC27Ba-20R	13β(H),17α(H)-20R-diacholestane
3	DiaC27aB-20S	13α(H),17β(H)-20S-diacholestane
4	DiaC27aB-20R	13α(H),17β(H)-20R-diacholestane
5	DiaC28Ba-20S(24X)	24X-methyl-13β(H),17α(H)-20S-diacholestane
6	DiaC28Ba-20S(24Y)	24Y-methyl-13β(H),17α(H)-20S-diacholestane
7	DiaC28Ba-20R(24S&R)	24S&R-methyl-13β(H),17α(H)-20R-diacholestane

Table 5.6 Continued.

Injection Number	**Sample name**	**Sample description based on origin**
8	C27aBB-20R	5α(H),14β(H),17β(H)-20R-cholestane
9	DiaC29Ba-20S(24S&R)	24S&R-ethyl-13β(H),17α(H)-20S-diacholestane
10	C27aBB-20S	5α(H),14β(H),17β(H)-20S-cholestane
11	C27aaa-20R	5α(H),14α(H),17α(H)-20R-cholestane
12	Ts	18α(H)-22,29,30-trinorneohopane
13	DiaC29Ba-20R(24S&R)	24S&R-methyl-13α(H),17β(H)-20R-diacholestane
14	Tm	17α(H)-22,29,30-trinorhopane
15	DiaC29aB-20S(24S&R)	24S&R-ethyl-13α(H),17β(H)-20S-diacholestane
16	DiaC29aB-20R(24S&R)	24S&R-ethyl-13α(H),17β(H)-20R-diacholestane
17	C28aBB-20R	24-methyl-5α(H),14β(H),17β(H)-20R-cholestane
18	C28aBB-20S Unknown sterane mass 400	24-methyl-5α(H),14β(H),17β(H)-20S-cholestane
19	(C29)	Unknown sterane mass 400 (C29)

Table 5.6 Continued.

Injection Number	**Sample name**	**Sample description based on origin**
20	C28aaa-20R	24-methyl-5α(H),14α(H),17α(H)-20R-cholestane
21	C29aaa-20S	24-ethyl-5α(H),14α(H),17α(H)-20S-cholestane
22	C29aBB-20R	24-ethyl-5α(H),14β(H),17β(H)-20R-cholestane
23	C29aBB-20S	24-ethyl-5α(H),14β(H),17β(H)-20S-cholestane
24	NH	17α(H),21β(H)-30-norhopane
25	C29aaa-20R	24-ethyl-5α(H),14α(H),17α(H)-20R-cholestane
26	NM	17β(H),21α(H)-30-norhopane
27	H	17α(H),21β(H)-hopane
28	M	17β(H),21α(H)-hopane
29	HH(S)	17α(H),21β(H)-22S-homohopane
30	HH(R)	17α(H),21β(H)-22R-homohopane
31	2HH(S)	17α(H),21β(H)-22S-bishomohopane
32	2HH(R)	17α(H),21β(H)-22R-bishomohopane

Table 5.6 Continued.

Injection Number	Sample name	Sample description based on origin
33	3HH(S)	17α(H),21β(H)-22S-trishomohopane
34	3HH(R)	17α(H),21β(H)-22R-trishomohopane
35	4HH(S)	17α(H),21β(H)-22S-tetrakishomohopane
36	4HH(R)	17α(H),21β(H)-22R-tetrakishomohopane
37	5HH(S)	17α(H),21β(H)-22S-pentakishomohopane
38	5HH(R)	17α(H),21β(H)-22R-pentakishomohopane

5.10 Procedure for cross-PTM comparison and related equations

Procedure for cross-PTM comparison

□ **Initialization:**

Start at the top left node of each PTM matrix, i.e., choose $\eta_{test} = PTM_{test}[1,1]$ and $\eta_{ref} = PTM_{ref}[1,1]$.

If the nodes are equivalent, then proceed to Step 2a. If they are not equivalent, then proceed to Step 2b.

☐ **Step 2a:**

Measure peakratio between equivalent nodes.

$$\rho(m_1, n_1) = max(\frac{p_{ref}}{p_{test}}, \frac{p_{test}}{p_{ref}}) \tag{5.4}$$

where the peakratio is indexed by the location of the peak at the test PTM node.

☐ **Step 2b:**

Determine the PTM node with the lower peak location in the second dimension, i.e., select η_{min}= $\arg_{\{n_1,n_2\}} min\{\eta_{test},\eta_{ref}\}$.

∗ ***Step 2b - Case 1:*** $(\eta_{min} = \eta_{test})$

In this scenario, test $GC \times GC$ ROI has a peak at $[m_1, n_1]$ while the reference $GC \times GC$ has none within the (Θ_1, Θ_2)-neighborhood of $[m_1, n_1]$.

To compensate for the missing peak in the reference sample, we insert a new reference PTM node $\tilde{\eta}_{ref} = \{m_1, n_1, \tilde{p}_{ref}\}$ *preceding* the current reference node at η_{ref}=$\{m_2, n_2, p_{ref}\}$.

We evaluate $\tilde{p}_{ref}$ as the maximum value within a (Θ_1, Θ_2)-vicinity of $[m_1, n_1]$ for the reference $GC \times GC$ ROI, i.e.,

$$\tilde{p}_{ref} = argmax\, I_{ref}(m_1 \pm \Theta_1, n_1 \pm \Theta_2) \tag{5.5}$$

The peakratio is evaluated as $\rho(m_1, n_1) = \max(\frac{\tilde{p}_{ref}}{p_{test}}, \frac{p_{test}}{\tilde{p}_{ref}})$ between equivalent nodes η_{test}=$\{m_1, n_1, p_{test}\}$ and the inserted reference PTM node $\tilde{\eta}_{ref} = m_1, n_1, \tilde{p}_{ref}$. The peakratio is indexed by the location of the existing peak at the test node.

∗ ***Step 2b - Case 2:*** $(\eta_{min} = \eta_{ref})$

In this other possible scenario, reference $GC \times GC$ ROI has a peak at $[m_2, n_2]$ while the test $GC \times GC$ ROI has none within the (Θ_1, Θ_2)-neighborhood of $[m_2, n_2]$.

We insert a new test PTM node $\tilde{\eta}_{test} = \{m_1, n_1, \tilde{p}_{test}\}$ where $\tilde{p}_{test}$ denotes the maximum value within the (Θ_1, Θ_2)-neighborhood of the test $GC \times GC$ ROI, i.e.,

$$\tilde{p}_{test} = argmax\, I_{test}(m_2 \pm \Theta_1, n_2 \pm \Theta_2) \tag{5.6}$$

The peakratio is evaluated as $\rho(m_2, n_2)$= $\max(\frac{p_{ref}}{\tilde{p}_{test}}, \frac{\tilde{p}_{test}}{p_{ref}})$ between the equivalent nodes η_{ref}=$\{m_2, n_2, p_{ref}\}$ and the inserted test PTM node $\tilde{\eta}_{test} = \{m_2, n_2, \tilde{p}_{test}\}$.

In this case, the peakratio is indexed by the location of the existing

peak at the reference PTM node.

☐ **Step 3:**

We threshold the peakratio $\rho(m, n)$ indexed by the peak location at either or both PTM nodes by a pre-selected threshold τ. Each peak (in test sample, reference sample, or both) is classified as:

1. "Similar" if $\rho(m, n) \leq \tau$, or
2. "Dissimilar" if $\rho(m, n) > \tau$.

☐ **Step 4:**

Increment the row index along the PTM matrix column (i.e., increment along the second $GC \times GC$ dimension) for the PTM node that did not have a node insertion. This reduces to three possibilities:

1. Increment both PTM nodes for Step 2a,
2. Increment test PTM node for Step 2b: Case 1, and
3. Increment ref PTM node for Step 2b: Case 2.

☐ **Terminate and move to next PTM matrix column:**

If both PTMs reach the last entry in the PTM matrix column, i.e, all remaining nodes in each PTM matrix column are blank nodes.

5.11 Cross-PTM Score, similarity as a percentage of match

After aligning the peaks of the $GC \times GC$ images we can use any desirable criterion to compare the two peak points at the same location. The criterion we have

used in this method to compare the two points p_{test} and p_{ref} is:

$$\rho(p_{test}, p_{ref}) = max(\frac{p_{test}}{p_{ref}}, \frac{p_{ref}}{p_{test}}) \tag{5.7}$$

5.12 Peak Detection using Maxima search

As stated in the manuscript, we employ a maxima finder using gradient computations to detect the peaks in a chromatogram. The key idea is to compute the gradient of the signal along the second dimension and locate the points where the gradient is zero with negative second derivative, indicating a maxima. Figure 5.14 demonstrates the performance of the maxima finder over different peaks, major and minor, for the chromatogram in Figure 5.13.

The peaks in Figure 5.14 have been plotted together as an overlapped collection. The overlapped visualization is intended to highlight the presence of hundreds of minor peaks besides the visible major peaks, and represents the same higher-dimensional information in the $GC \times GC$ image, and therefore, should not be treated as a collapsed GC plot. The first dimension retention times are for this box representing the biomarker region within a larger $GC \times GC$ chromatogram. The magenta dots at the bottom do not denote peaks but zero values. To provide resilience against noise, several measures may be taken to select a detected peak. We selected to keep the peaks detected by the maxima finder based on thresholding a ratio-driven measure. The metric chosen is sum of the absolute ratios of the slope to the peak width in the direction of rise and fall, as given in Equation 5.8. For the data analysis pre-

sented in Figure 5.3 in the manuscript, the peaks were thresholded to $\lambda \leq 0.01$. The variables used in Equation S4 below are defined in Figure 5.15.

$$\lambda = |\frac{h}{d_1}| + |\frac{h}{d_2}| \tag{5.8}$$

5.12.1 Selection of the values of d1, d2 and the threshold for λ

Choice of d_1 and d_2 dictate the ratio $\lambda = |\frac{s_1}{d_1}| + |\frac{s_2}{d_2}|$, which jointly considers the four design parameters s_1, s_2, d_1 and d_2, where s_1 and s_2 are dependent on the choice of d_1 and d_2 and the peak maxima. Choosing too high a value for d_1 and d_2 can lead to erroneously counting several minor peaks as one major peak, and too low a value can lead to regarding noise bumps as peaks. Therefore, high values for d_1 and d_2 will also lower lambda and lead to higher vulnerability to noise if the lambda threshold is small, and the possibility of lumping several small peaks into one. Choosing too high a threshold for lambda leaves out many of the smaller minor compounds. To avoid these scenarios, we tested a range of values of d_1 and d_2 across a random sample of well-detected peaks across the chromatogram for several samples in our dataset. For simplicity, we chose $d_1 = d_2$ and $s_1 = s_2$. Based on our empirical observations, we chose the value to be $d_1 = d_2 = 5$, and lambda threshold to 0.001 in the cross-PTM analysis to best capture most of the topography without vulnerability to noise, and choose lower values of the parameters to highlight more minor peaks in Figure 5.2.2(c). For confirmation that the choice of the parameters captured most of the peaks in the dataset, we applied the peak detection algorithm to

every chromatogram and generated Figure 5.14 for visual confirmation that all major peaks and considerable spread of minor peaks were detected.

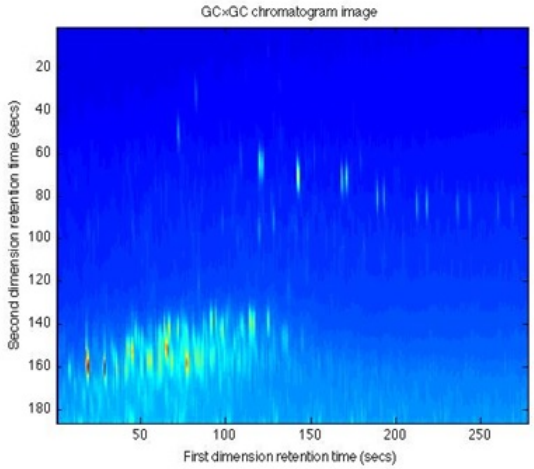

Figure 5.13: $GC \times GC$ chromatogram image of the sterane (lower left) and hopane (upper right) regions of an oil sample.The first dimension retention times are for this box within a larger $GC \times GC$ chromatogram.

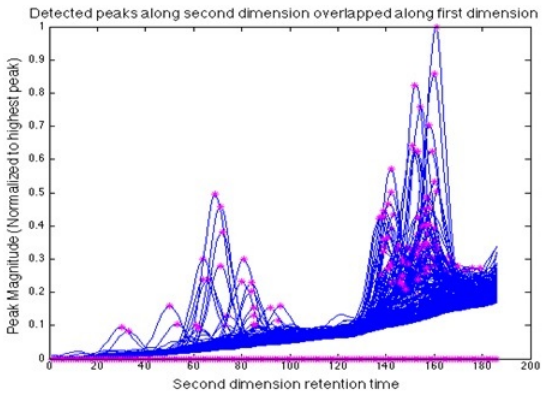

Figure 5.14: The peak shapes (in blue) and summit values (magenta stars on the peaks) detected along the 186 points along 2nd dimension) for each of the 277 points in the 1st dimension) of the $GC \times GC$ plot in 5.13.

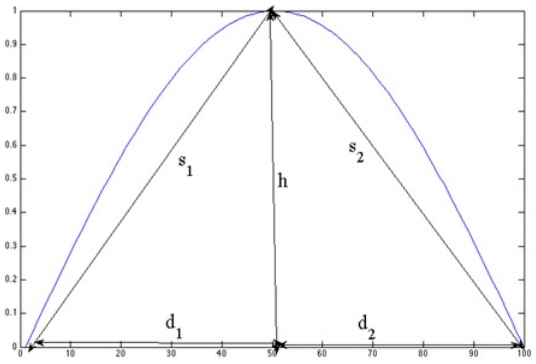

Figure 5.15: Peak parameters illustrated using a cosinusoidal peak.

5.12.2 Baseline correction

The effect of column bleed, is compensated for using first-order interpolation between the feet of each peak. Figure 5.16 below shows the original and corrected baseline for one column within the $GC \times GC$ image.

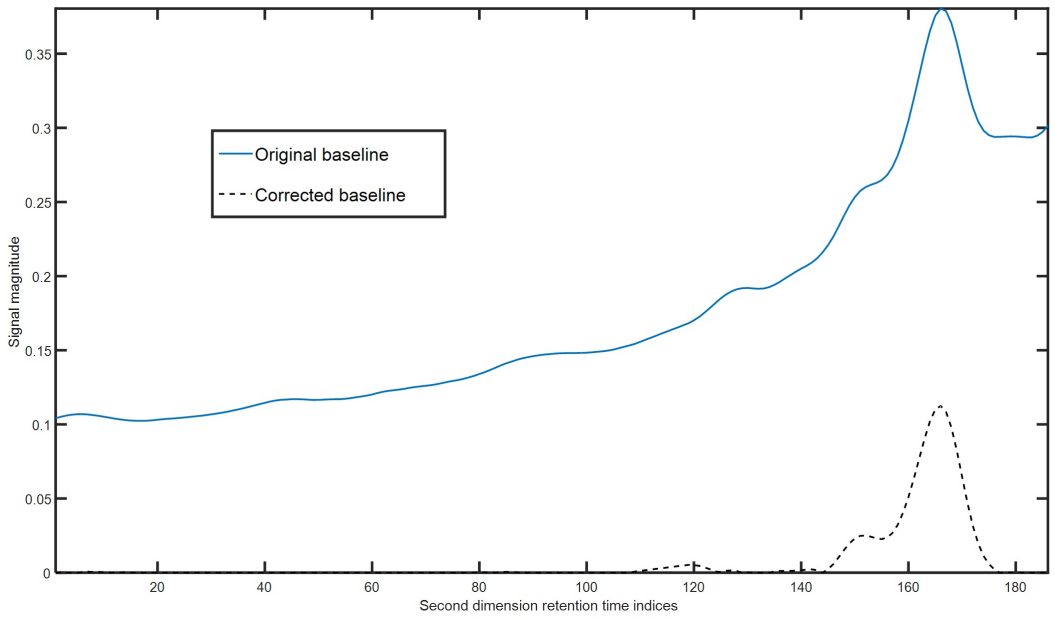

Figure 5.16: Original and corrected baseline for one column within the $GC \times GC$ image. The baseline is corrected for column bleed by estimating the local column bleed using a simple linear estimator and subtracting its effect from the original curve. Visually speaking, this has the effect of calculating the local gradient between the feet (estimated using close-to-zero gradient search before and after the peak maxima) of the peak and then subtracting its effect from the original curve.

5.13 Statistical boundaries for Cross-comparison scores for PTM and PCA

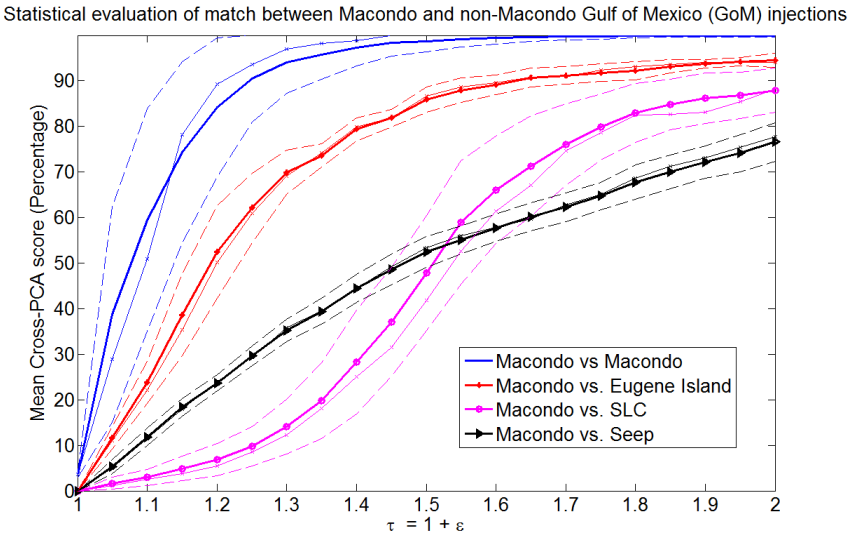

Figure 5.17: Statistical comparison; $(\mu \pm \sigma)$, when μ denotes the means and σ denotes the standard deviation of cross-PTM match between Macondo and other Gulf of Mexico injections: Eugene Island, Southern Louisiana Crude (SLC) and Gulf of Mexico natural seep.

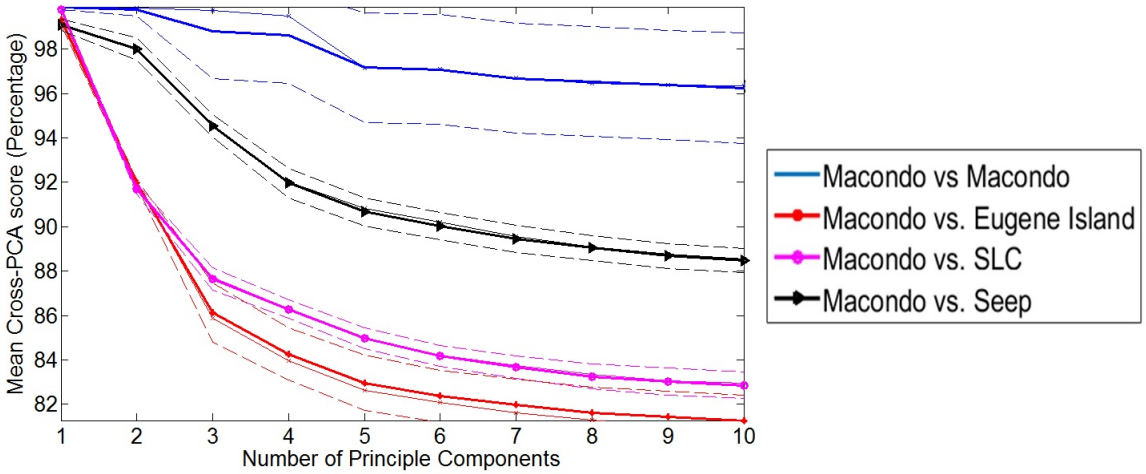

Figure 5.18: Statistical comparison; $(\mu \pm \sigma)$, when μ denotes the means and σ denotes the standard deviation of cross-PCA match between Macondo and other Gulf of Mexico injections: Eugene Island, Southern Louisiana Crude (SLC) and Gulf of Mexico natural seep.

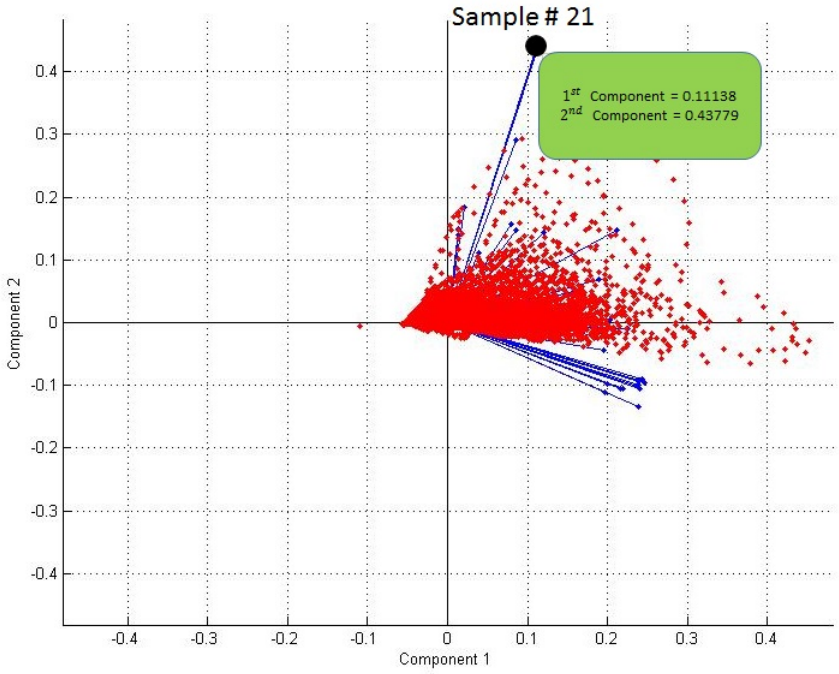

Figure 5.19: Projection score of sample 21 on the two different principle components.

5.14 Applying PCA on the model dataset

In figure 5.19 we have shown the biplot of the model petroleum dataset with two different principle components. The scores on the first and second dimensions are 0.11138 and 0.43779, respectively. Using these scores, we are unable to perform robust source classification using the PCA biplot itself, as most of the samples do not separate into source-specific clusters. We have nonetheless shown the scores of one of the outlier samples, sample #21, which belongs to the coast of Santa Barbara region. Therefore, have performed PCA directly on the $GC \times GC$ images for a fair comparison to the PTM method.

5.15 Comparison between two broad analytic approaches to environmental forensics

Table 5.7: Comparison between two broad analytic approaches to environmental forensics

Target-based analysis(Peak-ratio analysis between well-known analytes)	**Target-agnostic analysis (Statistical pattern recognition)**
Focuses on individual nuances of well-known target compounds, which manifest as major peaks in chromatograms.	Focuses primarily on the statistical properties of the multi-variate chromatographic data.
Assign forensic interpretation based on the relative proportions of target biomarkers, typically using peak ratio measurements.	Forensic diagnosis based on large-scale empirical differentiations between the data distribution of specimens sampled from known sources.
Ignores the effect of (potentially hundreds of) non-target compounds, which occur in relatively minor proportions in the complex mixture	Does not distinguish between target (big peaks) and non-target compounds (minor peaks).

Table 5.7 Continued.

Target-based analysis(Peak-ratio analysis between well-known analytes)	**Target-agnostic analysis (Statistical pattern recognition)**
Relatively immune to retention time variability, and robust across different specimens analyzed under diverse experimental conditions.	Vulnerable to retention time shifts which significantly shift the relative locations of minor peaks, and hence non-target compounds.
Robust identification of target compounds that belong to source fingerprint.	Agnostic of forensic signatures of individual target compounds.
Does not necessitate large-scale training data sets, few reliable source specimens may suffice.	Heavily dependent on training specimen libraries, reliably sampled from known source(s).
Provides reliable source diagnosis when the two sources exhibit distinct distributions across the major peaks.	Provides reliable source diagnosis when sufficient training samples are available to generate robust source ground truths.

Table 5.7 Continued.

Target-based analysis(Peak-ratio analysis between well-known analytes)	**Target-agnostic analysis (Statistical pattern recognition)**
Best-suited for direct comparison between two or more specimens based on their target compound distribution.	Best suited for comparing samples with reliable ground truths (e.g. oil samples from industrial oil reservoirs, transformer storage sites vs. pigment manufacturing).
Essential for scientific understanding of well-known compounds in environmental forensics.	Essential for broad statistical distinction between well-known sources with reliable ground truths.
Most chemists and EPA standards follow this approach due to higher scientific understanding and reliability of dominant compounds. Peak-ratio analysis of target chemcials dominate forensic analysis in environmental chemistry.	Most pattern-recognition techniques applied to environmental forensics fall in this domain, due to available pattern classification templates based on data statistics when reliable source ground truths are present.

CHAPTER 6
CONCLUDING REMARKS

6.1 Conclusion

In this research we worked on one-dimensional and two-dimensional petroleum forensic signals and images and the methods to study them better. These methods help us to learn the patterns of the petroleum forensics of miscellaneous geographical regions and use them for better distinction against that of the other region. The forensic pattern are extracted through the chromatography procedure. Chromatographic interpretation for petroleum forensics is challenging due to:

- Lack of robust ground truths
- Chromatographic variability
- Significant correlation between neighboring sources
- No quantitative method to reconcile target-based (peak-cognizant) interpretation with (peak-agnostic) statistical techniques

6.2 Contributions

- **Peak-based feature engineering:** Quantitative compound-cognizant interpretation for raw signal datasets.
  - Target Cognizant Clustering (TCC): Determine key clusters of target analytes that influence the fingerprint

  - Target Neighborhood Analysis (TNA): Local interpretation of raw signal around a target
  - Local PTM interpretation through τ-map

- **Connect two disparate branches of chromatographic interpretation:** Bridge peak-level interpretation (target analysis) with statistical (chemo-metric) interpretation through a combination of peak topography mapping, partitioning and clustering techniques
- **Quantify Sensitivity to variability:** Perturbation analysis of match sensitivity to peak height and location uncertainties
- **Identify robust peaks across a training set:** Empirical assessment of compounds that exhibit resilience to chromatographic variability
- **Localized Calibration:** Calibrate raw signal locally using τ-map framework

REFERENCES

[1] Z. Wang and S. Stout. Oil spill environmental forensics: Fingerprinting and source identification. *Academic Press*, 33(12):2106, 1999.

[2] R. Gaines, G. Frysinger, M. Hendrick-Smith, and J. Stuart. Oil spill source identification by comprehensive two-dimensional gas chromatography. *Environ. Sci. and Technology*, 2010.

[3] C. Venkatramani and J. Phillips. *Comprehensive two-dimensional gas chromatography applied to the analysis of complex mixtures.* J. Microcolumn Separations, 1993.

[4] J.Arey, R.Nelson, L.Xu, and C.M. Reddy. Using comprehensive two-dimensional gas chromatography retention indices to estimate environmental portioning properties for a complete set of diesel fuel hydrocarbons. *Anal. Chemistry*, 77(12):7172, 2005.

[5] J. Arey, R. Nelson, and C.M. Reddy. Disentangling oil weathering using gcxgc. 1. chromatogram analysis. *Environ. Sci. and Technology*, 41(12):5738, 2007.

[6] G. W. Johnson and R. Ehrlich. State of the art report on multivariate chemometric methods in environmental forensics. *Environ. Sci. and Technology*, 3(1):59, 2002.

[7] Grady Hanrahan. *Environmental Chemometrics, Principles and Modern Applications.* CRC Press, 2008.

[8] Wolfgang Bertsch. Two-dimensional gas chromatography. concepts, instrumentation, and applications–part 1: Fundamentals, conventional two-dimensional gas chromatography, selected applications. *Journal of High Resolution Chromatography*, 22(12):647–665, 1999.

[9] Wolfgang Bertsch. Two-dimensional gas chromatography. concepts, instrumentation, and applications–part 2: Comprehensive two-dimensional gas chromatography. *Journal of High Resolution Chromatography*, 23(3):167–181, 2000.

[10] Jens Dallüge, Jan Beens, and A Udo. Comprehensive two-dimensional gas chromatography: a powerful and versatile analytical tool. *Journal of Chromatography A*, 1000(1):69–108, 2003.

[11] Miroslav Kubat, Robert C Holte, and Stan Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine learning*, 30(2-3):195–215, 1998.

[12] Standard practice for oil spill source identification by gas chromatog- raphy and positive ion electron impact low resolution mass spectrometry, developed by astm international. *developed by ASTM International*.

[13] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, Thme Cog, Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, quipes-projets Willow, and Ecole Normale Suprieure. Supervised dictionary learning, 2008.

[14] I.T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

[15] Fayyad, Usama M., Piatetsky-Shapiro, Gregory, and Padhraic Smyth. Advances in knowledge discovery and data mining. chapter From Data Mining to Knowledge Discovery: An Overview, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.

[16] M. Shensa. The discrete wavelet transform: wedding the a trous and mallat algorithms. *Signal Processing, IEEE Transactions on*, 40(10):2464–2482, Oct 1992.

[17] N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transfom. *IEEE Trans. Comput.*, 23(1):90–93, January 1974.

[18] Syed Ali Khayam. The discrete cosine transform (dct): theory and application. *Michigan State University*, 2003.

[19] Richard Brereton. *Chemometrics for pattern recognition*. John Wiley & Sons, 2009.

[20] Stephen E Reichenbach, Mingtian Ni, Dongmin Zhang, and Edward B Ledford. Image background removal in comprehensive two-dimensional gas chromatography. *Journal of Chromatography A*, 985(1):47–56, 2003.

[21] Alexander Chernobelsky, Oleg Shubayev, Cindy R Comeau, and Steven D Wolff. Baseline correction of phase contrast images improves quantification of blood flow in the great vessels. *Journal of Cardiovascular Magnetic Resonance*, 9(4):681–685, 2007.

[22] Anne C Sauve and Terence P Speed. Normalization, baseline correction and

alignment of high-throughput mass spectrometry data. *Proceedings Gensips*, 2004.

[23] Edward R Dougherty, Roberto A Lotufo, and The International Society for Optical Engineering SPIE. *Hands-on morphological image processing*, volume 71. SPIE press Bellingham, 2003.

[24] Mingtian Ni, Stephen E Reichenbach, Arvind Visvanathan, Joel TerMaat, and Edward B Ledford. Peak pattern variations related to comprehensive two-dimensional gas chromatography acquisition. *Journal of Chromatography A*, 1086(1):165–170, 2005.

[25] Richard Kramer. *Chemometric techniques for quantitative analysis*. CRC Press, 1998.

[26] Jane Charlotte Miller and James N Miller. Statistics for analytical chemistry. 1988.

[27] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011.

[28] Rasmus Bro. Parafac. tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(2):149–171, 1997.

[29] Rasmus Bro, Claus A Andersson, and Henk AL Kiers. Parafac2-part ii. modeling chromatographic data with retention time shifts. *Journal of Chemometrics*, 13(3-4):295–309, 1999.

[30] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[31] Kathleen J Schostack and Edmund R Malinowski. Investigation of window factor analysis and matrix regression analysis in chromatography. *Chemometrics and intelligent laboratory systems*, 20(2):173–182, 1993.

[32] Yongnian Ni, Jieling Bai, and Ling Jin. Multicomponent chemometric determination of colorant mixtures by voltammetry. *Analytical letters*, 30(9):1761–1777, 1997.

[33] Tom O'Haver. A pragmatic introduction to signal processing with applications in scientific measurement. *A tutorial in Chemical Signal Processing*, 2015.

[34] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.

[35] Bernard W. Silverman. *Density estimation for statistics and data analysis*. Monographs on statistics and applied probability. Chapman and Hall Boca Raton, London, Glasgow, Weinheim, 1986. Titre sur le dos du livre : Density estimation.

[36] Allen Y Yang. Unsupervised segmentation of natural images via lossy data compression. *Computer Vision and Image Understanding*, 110(2):212, 2008.

[37] Luczak, Tomasz, and Wojciech Szpankowski. A suboptimal lossy data compression based on approximate pattern matching. *Information Theory*, 43(5):1439, 1997.

[38] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schtze. *Introduction to information retrieval*.

[39] Mark Nixon. *Feature extraction & image processing*. Academic Press, 2008.

[40] Isabelle Guyon. *Feature extraction: foundations and applications*, volume 207. Springer Science & Business Media, 2006.

[41] Luis O Jimenez and David A Landgrebe. Hyperspectral data analysis and supervised feature reduction via projection pursuit. *Geoscience and Remote Sensing, IEEE Transactions on*, 37(6):2653–2667, 1999.

[42] Soumya Raychaudhuri, Patrick D Sutphin, Jeffrey T Chang, and Russ B Altman. Basic microarray analysis: grouping and feature reduction. *TRENDS in Biotechnology*, 19(5):189–193, 2001.

[43] Roman W Świniarski. Rough sets methods in feature reduction and classification. *International Journal of Applied Mathematics and Computer Science*, 11(3):565–582, 2001.

[44] Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Norwell, MA, USA, 1991.

[45] A. Gersho and B. Ramamurthi. Image coding using vector quantization. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82.*, volume 7, pages 428–431, May 1982.

[46] R Dony. Karhunen-loeve transform. *The transform and data compression handbook. CRC Press, Boca Raton, London, New York, Washington, DC*, 2001.

[47] BR Kowalski and CF Bender. Pattern recognition. powerful approach to inter-

preting chemical data. *Journal of the American Chemical Society*, 94(16):5632–5639, 1972.

[48] Lars Kryger. Interpretation of analytical chemical information by pattern recognition methodsa survey. *Talanta*, 28(12):871–887, 1981.

[49] ChotiratAnn Ratanamahatana, Jessica Lin, Dimitrios Gunopulos, Eamonn Keogh, Michail Vlachos, and Gautam Das. Mining time series data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 1049–1077. Springer US, 2010.

[50] John A Hartigan. *Clustering algorithms*. Wiley series in probability and mathematical statistics. Wiley, New York, NY, 1975.

[51] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 2:241–254, 1967.

[52] Hamidreza Ghasemi Damavandi, Ananya Sen Gupta, Chris Reddy, and Robert Nelson. Compound-cognizant feature compression of gas chromatographic data to facilitate environmental forensics. Data Compression Conference, April 2015.

[53] Hamidreza Ghasemi Damavandi, Ananya Sen Gupta, Robert Nelson, and Chris Reddy. Compressed forensic source image using source pattern map. Data Compression Conference, April 2016.

[54] Ananya Sen Gupta, Chris M.Reddy, and Robert Nelson. Systems and methods for topographic analysis. *U.S. Patent*, 2014.

[55] C. M.Reedy R K.Nelson H. Ghasemi Damavandi, A.Sen Gupta. Oil-spill forensics using two-dimensional gas chromatography: Differentiating highly correlated petroleum sources using peak manifold clusters. Proceedings of Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, Nov. 2015.

[56] Hamidreza Ghasemi Damavandi, Ananya Sen Gupta, Chris M.Reddy, and Robert Nelson. Interpreting comprehensive two-dimensional gas chromatography using peak topography maps with application to petroleum forensics. *under review in Chemistry Central*, 2015.

[57] Ian Jolliffe. *Principal component analysis*. 2002.

[58] Zhendi Wang and Merv Fingas. Differentiation of the source of spilled oil and monitoring of the oil weathering process using gas chromatography-mass spectrometry. *Journal of Chromatography A*, 712(2):321–343, 1995.

[59] Woodfin V Ligon and Ralph J May. Target compound analysis by two-dimensional gas chromatographymass spectrometry. *Journal of Chromatography A*, 294:77–86, 1984.

[60] Kenneth E Peters and J Michael Moldowan. *The biomarker guide: interpreting molecular fossils in petroleum and ancient sediments*. Englewood Cliffs, Prentice Hall, New Jersey (United States), 1993.

[61] Josep M Bayona, Carmen Domínguez, and Joan Albaigés. Analytical developments for oil spill fingerprinting. *Trends in Environmental Analytical Chemistry*, 5:26–34, 2015.

[62] Richard B Gaines, Glenn S Frysinger, Martha S Hendrick-Smith, and James D Stuart. Oil spill source identification by comprehensive two-dimensional gas chromatography. *Environmental science & technology*, 33(12):2106–2112, 1999.

[63] Scott A Stout and Zhendi Wang. Diagnostic compounds for fingerprinting petroleum in the environment. *Environmental Forensics*, 26:54, 2008.

[64] Standard practice for oil spill source identification by gas chromatography and positive ion electron impact low resolution mass spectrometry. *ASTM D5739-06, developed by ASTM International, DOI: 10.1520/D5739-06*.

[65] EPA8270D semivolatile organic compounds by gas chromatography/mass spectrometry (gc/ms). *developed by United States Environmental Protection Agency, http://www.epa.gov/osw/hazard/testmethods/sw846/pdfs/8270d.pdf*.

[66] Modified EPA8270. *Developed by United States Environmental Protection Agency*.

[67] Christoph Aeppli, Catherine A Carmichael, Robert K Nelson, Karin L Lemkau, William M Graham, Molly C Redmond, David L Valentine, and Christopher M Reddy. Oil weathering after the deepwater horizon disaster led to the formation of oxygenated residues, 2012.

[68] Mohammad R Nezami Ranjbar, Cristina D Poto, Yue Wang, and Habtom W Ressom. Simat: Gc-sim-ms data analysis tool. *BMC bioinformatics*, 16(1):259, 2015.

[69] K McAdam, Arif Faizi, Harriet Kimpton, Andrew Porter, and Brad Rodu. Polycyclic aromatic hydrocarbons in us and swedish smokeless tobacco products. *Chemistry Central*, 7:151, 2013.

[70] Jixiang Guo, Jia Fang, and Jingjing Cao. Characteristics of petroleum contaminants and their distribution in lake taihu, china. *Chemistry Central Journal*, 6(1):92, 2012.

[71] G Todd Ventura, Gregory J Hall, Robert K Nelson, Glenn S Frysinger, Bhavani Raghuraman, Andrew E Pomerantz, Oliver C Mullins, and Christopher M Reddy. Analysis of petroleum compositional similarity using multiway principal components analysis (mpca) with comprehensive two-dimensional gas chromatographic data. *Journal of Chromatography A*, 1218(18):2584–2592, 2011.

[72] Diako Ebrahimi, Jianfeng Li, and David Brynn Hibbert. Classification of weathered petroleum oils by multi-way analysis of gas chromatography–mass spectrometry data using parafac2 parallel factor analysis. *Journal of Chromatography A*, 1166(1):163–170, 2007.

[73] Jan H Christensen, Giorgio Tomasi, and Asger B Hansen. Chemical fingerprinting of petroleum biomarkers using time warping and pca. *Environmental science & technology*, 39(1):255–260, 2005.

[74] Richard B Gaines, Gregory J Hall, Glenn S Frysinger, Wayne R Gronlund, and Kristy L Juaire. Chemometric determination of target compounds used to fingerprint unweathered diesel fuels. *Environmental Forensics*, 7(1):77–87, 2006.

[75] Ayhan Demiriz, Kristin P Bennett, Curt M Breneman, and Mark J Embrechts. Support vector machine regression in chemometrics. In *In Computing Science and Statistics: Proceedings of the 33rd Symposium on the Interface*. Citeseer, 2001.

[76] Tom Howley, Michael G Madden, Marie-Louise OConnell, and Alan G Ryder. The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. *Knowledge-Based Systems*, 19(5):363–370, 2006.

[77] Barry K Lavine, D Brzozowski, Anthony J Moores, CE Davidson, and Howard T Mayfield. Genetic algorithm for fuel spill identification. *Analytica Chimica Acta*, 437(2):233–246, 2001.

[78] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

[79] J Samuel Arey, Robert K Nelson, Li Xu, and Christopher M Reddy. Using comprehensive two-dimensional gas chromatography retention indices to estimate environmental partitioning properties for a complete set of diesel fuel hydrocarbons. *Analytical chemistry*, 77(22):7172–7182, 2005.

[80] Christopher M Reddy and James G Quinn. Gc-ms analysis of total petroleum hydrocarbons and polycyclic aromatic hydrocarbons in seawater samples after the north cape oil spill. *Marine Pollution Bulletin*, 38(2):126–135, 1999.

[81] Christoph Aeppli, Robert K Nelson, Jagos R Radovic, Catherine A Carmichael, David L Valentine, and Christopher M Reddy. Recalcitrance and degradation of petroleum biomarkers upon abiotic and biotic natural weathering of deepwater horizon oil. *Environmental science & technology*, 48(12):6726–6734, 2014.

[82] J Samuel Arey, Robert K Nelson, and Christopher M Reddy. Disentangling oil weathering using gc× gc. 1. chromatogram analysis. *Environmental science & technology*, 41(16):5738–5746, 2007.

[83] J Samuel Arey, Robert K Nelson, Desiree L Plata, and Christopher M Reddy. Disentangling oil weathering using gc× gc. 2. mass transfer calculations. *Environmental science & technology*, 41(16):5747–5755, 2007.

[84] Christopher M Reddy, J Samuel Arey, Jeffrey S Seewald, Sean P Sylva, Karin L Lemkau, Robert K Nelson, Catherine A Carmichael, Cameron P McIntyre, Judith Fenwick, G Todd Ventura, et al. Composition and fate of gas and oil released to the water column during the deepwater horizon oil spill. *Proceedings of the National Academy of Sciences*, 109(50):20229–20234, 2012.

[85] George D Wardlaw, J Samuel Arey, Christopher M Reddy, Robert K Nelson, G Todd Ventura, and David L Valentine. Disentangling oil weathering at a marine seep using gc× gc: Broad metabolic specificity accompanies subsurface petroleum biodegradation. *Environmental science & technology*, 42(19):7166–7173, 2008.

[86] Richard Camilli, Christopher M Reddy, Dana R Yoerger, Benjamin AS Van Mooy, Michael V Jakuba, James C Kinsey, Cameron P McIntyre, Sean P Sylva, and James V Maloney. Tracking hydrocarbon plume transport and biodegradation at deepwater horizon. *Science*, 330(6001):201–204, 2010.

[87] Glenn S Frysinger, Gregory J Hall, Ariana L Pourmonir, Heather N Bischel, Emily E Peacock, Robert N Nelson, and Christopher M Reddy. Tracking and modeling the degradation of a 30 year old fuel oil spill with comprehensive two-dimensional gas chromatography. In *International Oil Spill Conference Proceedings (IOSC)*, volume 2011, page abs428. American Petroleum Institute, 2011.

[88] Harald Martens and Tormod Naes. *Multivariate calibration*. John Wiley & Sons, New Jersey, 1992.

[89] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

[90] Paul McA Harvey and Robert A Shellie. Data reduction in comprehensive two-dimensional gas chromatography for rapid and repeatable automated data analysis. *Analytical chemistry*, 84(15):6501–6507, 2012.

[91] Richard Camilli, Daniela Di Iorio, Andrew Bowen, Christopher M Reddy, Alexandra H Techet, Dana R Yoerger, Louis L Whitcomb, Jeffrey S Seewald, Sean P Sylva, and Judith Fenwick. Acoustic measurement of the deepwater horizon macondo well flow rate. *Proceedings of the National Academy of Sciences*, 109(50):20235–20239, 2012.