EDGE
ENHANCING DIGITAL GOVERNMENT AND ECONOMY

জ্ঞানই শক্তি
বরিশাল বিশ্ববিদ্যালয়

Analysis of Aromatic Hydrocarbons (PAHs) Dataset

Prepared by:

Ibrahim Khalil Fahim

Data Science with Python Batch 03

Prepared for:

Dr. Rashid Al Asif

Assistant Professor

Department of Computer Science and Engineering

University of Barishal

Submission date: 15th December, 2024

Contents Page No

1. Abstract …………………………………………………………………….....2
2. Introduction …………………………………………………………………... 3
3. Objective …………………………………………………………………...... 3
4. Literature Review on Polycyclic Aromatic Hydrocarbons (PAHs)……………...... 4
5. Dataset Overview…………………………………………………………….... 5
6. Methodology for Analyzing the Dataset ……..………..………………………… ..5
7. Data Describe ...……………………………………………………………….5
8. **Data Visualization and Discussion:**

1.1: Geographic Distribution of Samples ……………………………………… 6
1.2: Sampling Frequency Over Time …………………………………………... 7
1.3: Distribution of Naph Concentrations ………………………………………. 8
1.4: Correlation Heatmap of Chemical Measurements …………………………. 9

9. Conclusion 10
10. Reference 11

Abstract:

PAHs (Polycyclic Aromatic Hydrocarbons) are a group of organic compounds formed from the incomplete combustion of fossil fuels and organic matter. They can contaminate air, water, and soil, posing significant risks to human health and wildlife by being carcinogenic, mutagenic, and toxic.

Here a dataset provides an in-depth analysis of polycyclic aromatic hydrocarbons (PAHs) collected from California's water, sediment, and tissue samples between 2011 and 2023. Using systematic methods, including data preprocessing, exploratory data analysis (EDA), and correlation analysis, the study establishes baseline PAH levels and examines spatial and temporal trends in pollution. Missing values were addressed through imputation techniques, and key variables like sampling dates and PAH concentrations were standardized for consistency.

Correlation analysis through heatmaps identified significant relationships between compounds like naphthalene and phenanthrene, indicating shared pollution sources. Diagnostic ratios further distinguished natural PAHs from anthropogenic contributions, aiding in source attribution and environmental risk assessment.

By combining robust data analysis techniques with comprehensive visualization, this dataset provides critical insights into PAH contamination patterns, high-risk areas, and ecological impacts. These findings support oil spill preparedness, pollution mitigation, and sustainable resource management efforts.

Keywords:

Polycyclic Aromatic Hydrocarbons (PAHs), Naphthalene (Naph), Phenanthrene (Phen), Dibenzothiophene (DBT), and Pyrene (Pyr). Correlation analysis, Environmental Monitoring, Oil Spills, Anthropogenic Pollution, Sediment Sampling, Water Quality Analysis, California Ecosystems.

Introduction

Polycyclic aromatic hydrocarbons (PAHs) are a class of organic pollutants commonly found in the environment due to both natural and anthropogenic activities. They are by-products of incomplete combustion of organic materials, oil spills, industrial processes, and natural phenomena such as forest fires and oil seeps. PAHs are persistent in the environment and have been widely studied due to their toxic, mutagenic, and carcinogenic properties, making them significant pollutants of concern in aquatic and terrestrial ecosystems.

Monitoring ambient levels of PAHs is critical for assessing baseline environmental conditions, particularly in regions prone to oil spills or industrial discharges. Baseline data help distinguish pre-existing contamination from new pollution events, enabling effective natural resource damage assessments (NRDA) and guiding remediation efforts. Understanding the spatial and temporal distribution of PAHs is also essential for identifying pollution hotspots, determining pollution sources, and evaluating potential ecological and human health risks.

This dataset contains comprehensive PAH measurements collected between 2011 and 2023 from various coastal and inland regions of California, including Humboldt Bay, Monterey Bay, Elkhorn Slough, and the Feather River Canyon. Samples were obtained from water, sediment, and tissue, reflecting the diverse environmental compartments where PAHs accumulate. These data were collected through collaborative efforts involving the California Department of Fish and Wildlife’s Office of Spill Prevention and Response (CDFW-OSPR), Chevron, and other trustee agencies to establish pre-spill ambient conditions and prepare for potential environmental emergencies.

Objective:

The objective of this analysis is to explore the dataset to establish baseline PAH levels, analyze spatial and temporal trends, identify pollution sources, and provide actionable insights for environmental monitoring and policy-making. By leveraging this dataset, researchers and policymakers can better understand the environmental distribution of PAHs, assess ecological risks, and improve preparedness for oil spill response and resource protection efforts.

Literature Review on Polycyclic Aromatic Hydrocarbons (PAHs)

1. Environmental Presence and Sources of PAHs

Polycyclic aromatic hydrocarbons (PAHs) are persistent organic pollutants found in various environmental compartments such as water, soil, sediment, and biota. They originate from both natural processes, such as wildfires and volcanic activity, and anthropogenic sources, including fossil fuel combustion, industrial discharges, oil spills, and vehicular emissions (Wang et al., 2017). Coastal and inland ecosystems are particularly vulnerable to PAH contamination due to the prevalence of both natural oil seeps and human activities like shipping and petroleum refining (Neff, 2002).

2. PAHs in Sediments and Water

Sediments act as reservoirs for hydrophobic pollutants like PAHs, making them crucial for long-term monitoring of contamination levels (Yang et al., 2020). Studies have shown that sediment PAH concentrations are highly influenced by proximity to urban or industrial areas and natural events such as floods that mobilize pollutants (Achten & Hofmann, 2009). In water, PAHs are subject to degradation through photolysis and microbial activity, but their persistence varies depending on environmental conditions, including temperature and light (Beyer et al., 2010).

3. Biological Impacts of PAHs

PAHs are known to be toxic, mutagenic, and carcinogenic, posing significant risks to aquatic organisms and human health. Chronic exposure to PAHs can lead to bioaccumulation in organisms such as fish and crustaceans, disrupting ecological balance (Baumard et al., 1998). Biota sampling, such as tissue analysis in crayfish, helps in understanding bioavailability and trophic transfer of PAHs (Mai et al., 2002).

4. Baseline Monitoring and Oil Spill Response

Establishing baseline levels of PAHs in sediments, water, and biota is critical for assessing the impacts of oil spills and other contamination events. Pre-spill data provide a reference for distinguishing background levels from spill-related pollution, aiding in natural resource damage assessments (NRDA) (Michel & Rutherford, 2014).

Dataset Overview:

This dataset contains measurements of polycyclic aromatic hydrocarbons (PAHs**)** collected from environmental samples in California between 2011 and 2023**.** The data aims to establish baseline PAH levels for assessing environmental conditions before and after potential contamination events, such as oil spills.

This dataset provides comprehensive measurements of polycyclic aromatic hydrocarbons (PAHs) collected from various environmental samples in California between 2011 and 2023. It consists of 387 samples (rows) analyzed across 119 variables (Columns)**,** capturing both chemical concentrations and metadata related to sample collection.

Methodology for Analyzing the Dataset:

Steps in Analysis

1. Data Loading: Load the dataset into a pandas DataFrame.
2. Exploratory Data Analysis (EDA): Understand the dataset's structure, distributions, and relationships.
3. Data Cleaning: Handle missing or inconsistent data.
4. Correlation Analysis: Heatmaps are generated.
5. Visualization: Seaborn and Matplotlib are used.

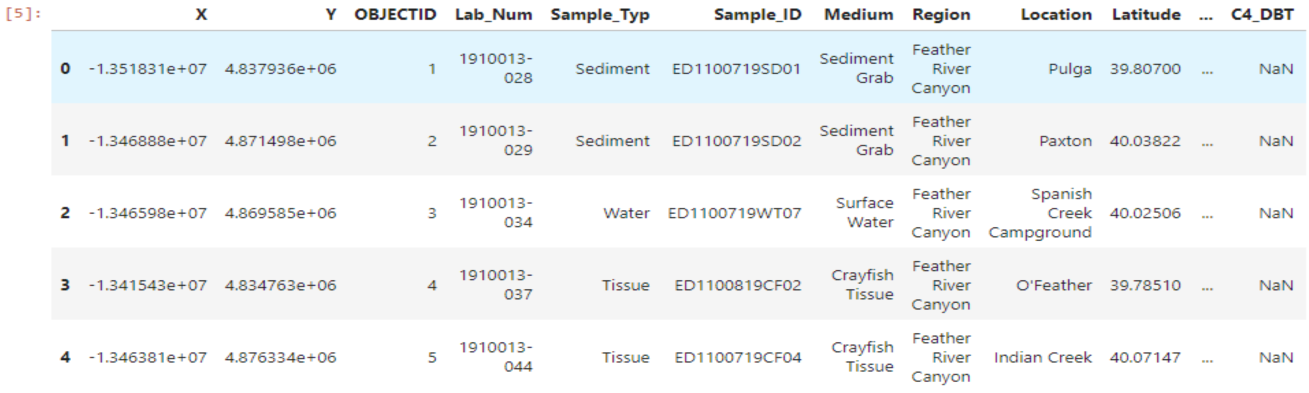
Data Describe:

Rows: 387, representing individual samples and **Columns**: 119 including:

Numeric Columns: PAH concentrations, recovery percentages, and geographic coordinates.

Categorical Columns: Sample types, regions, and site names.

Datetime Columns: Dates for sampling, extraction, and analysis.



[5]:

	X	Y	OBJECTID	Lab_Num	Sample_Typ	Sample_ID	Medium	Region	Location	Latitude	...	C4_DBT
0	-1.351831e+07	4.837936e+06	1	1910013-028	Sediment	ED1100719SD01	Sediment Grab	Feather River Canyon	Pulga	39.80700	...	NaN
1	-1.346888e+07	4.871498e+06	2	1910013-029	Sediment	ED1100719SD02	Sediment Grab	Feather River Canyon	Paxton	40.03822	...	NaN
2	-1.346598e+07	4.869585e+06	3	1910013-034	Water	ED1100719WT07	Surface Water	Feather River Canyon	Spanish Creek Campground	40.02506	...	NaN
3	-1.341543e+07	4.834763e+06	4	1910013-037	Tissue	ED1100819CF02	Crayfish Tissue	Feather River Canyon	O'Feather	39.78510	...	NaN
4	-1.346381e+07	4.876334e+06	5	1910013-044	Tissue	ED1100719CF04	Crayfish Tissue	Feather River Canyon	Indian Creek	40.07147	...	NaN

Data Visualization and Discussion:

1.1: Geographic Distribution of Samples

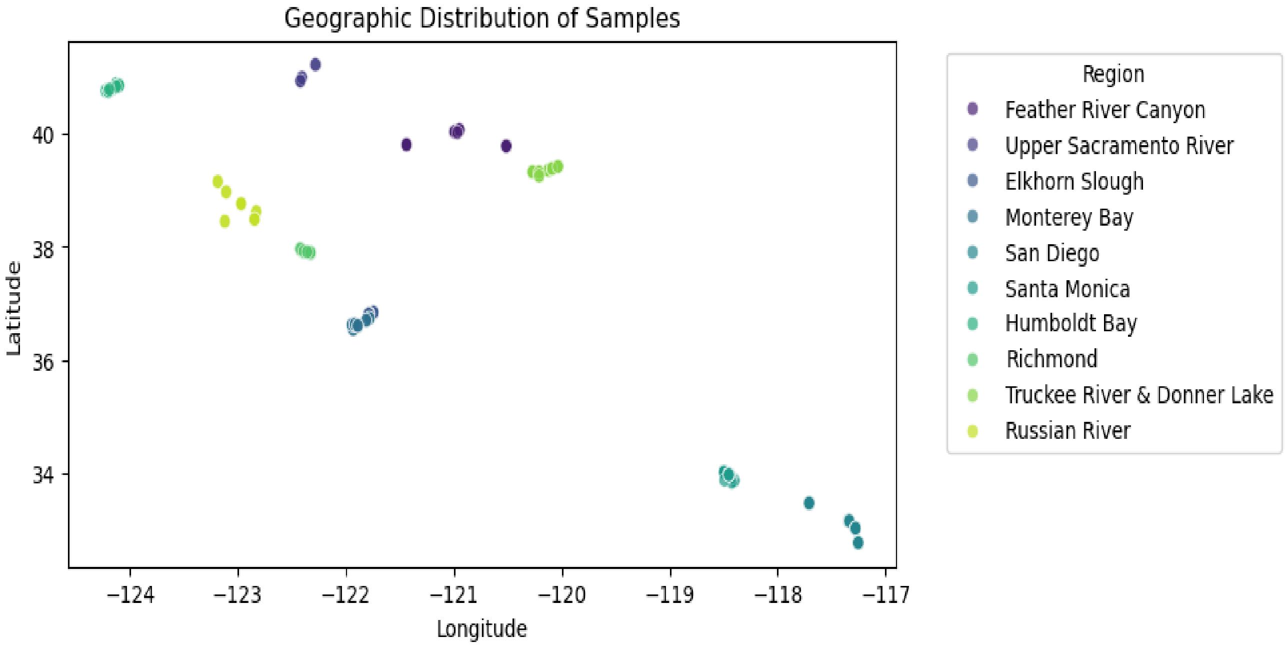


Figure: 1.1

- **Description**: The scatterplot shows the geographic distribution of samples across various Latitude and Longitude coordinates. Different colors represent distinct regions (e.g., "Feather River Canyon").
- **Insights**: Samples appear clustered in specific areas, suggesting focused sampling efforts. Sparse regions might indicate areas of lesser environmental focus or logistical challenges in sampling.
- **Implication**: There may be geographic sampling gaps in underrepresented regions. This might necessitate additional sampling for comprehensive environmental monitoring.

1.2: Sampling Frequency Over Time

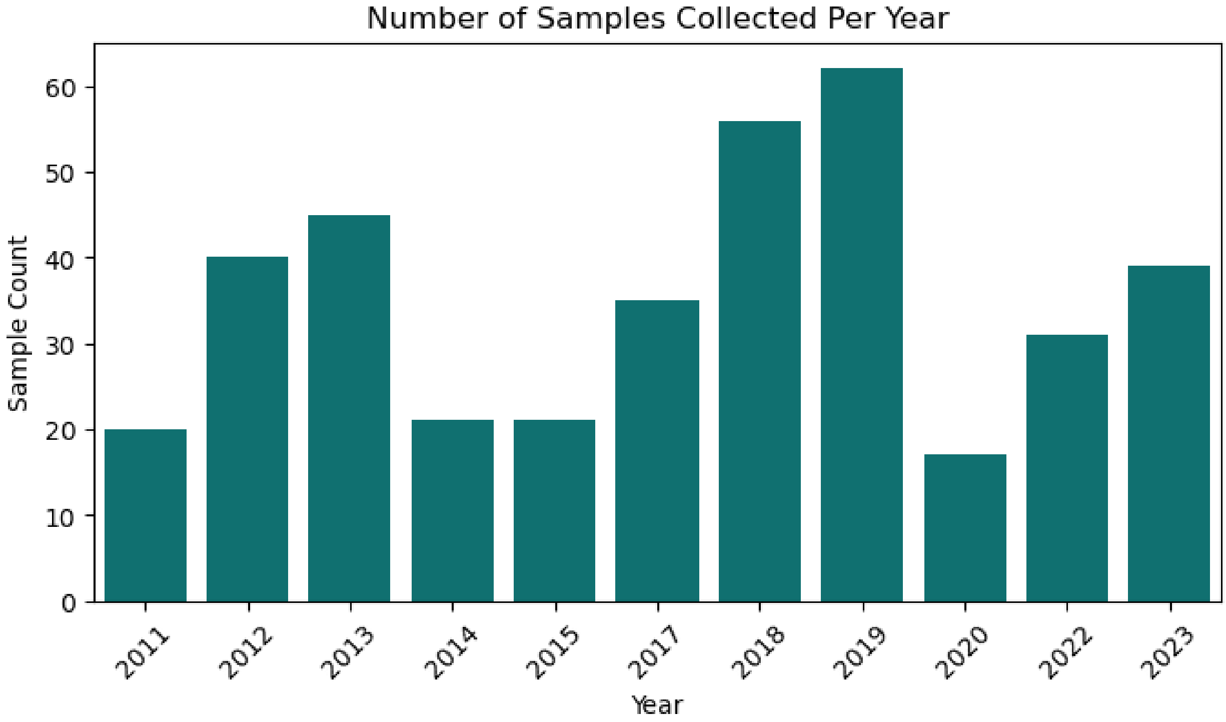


Figure: 1.2

- **Description**: The bar chart illustrates the number of samples collected annually, with the x-axis representing years and the y-axis the sample count.
- **Insights**: The plot highlights fluctuations in sampling activity. Years with high counts might reflect targeted studies or significant environmental events, while low counts could indicate operational constraints or data gaps.
- **Implication**: The uneven sampling trends could be influenced by factors like funding availability, environmental events, or research priorities. Consistent sampling is essential for long-term trend analysis.

1.3: Distribution of Naph (Naphthalene) Concentrations

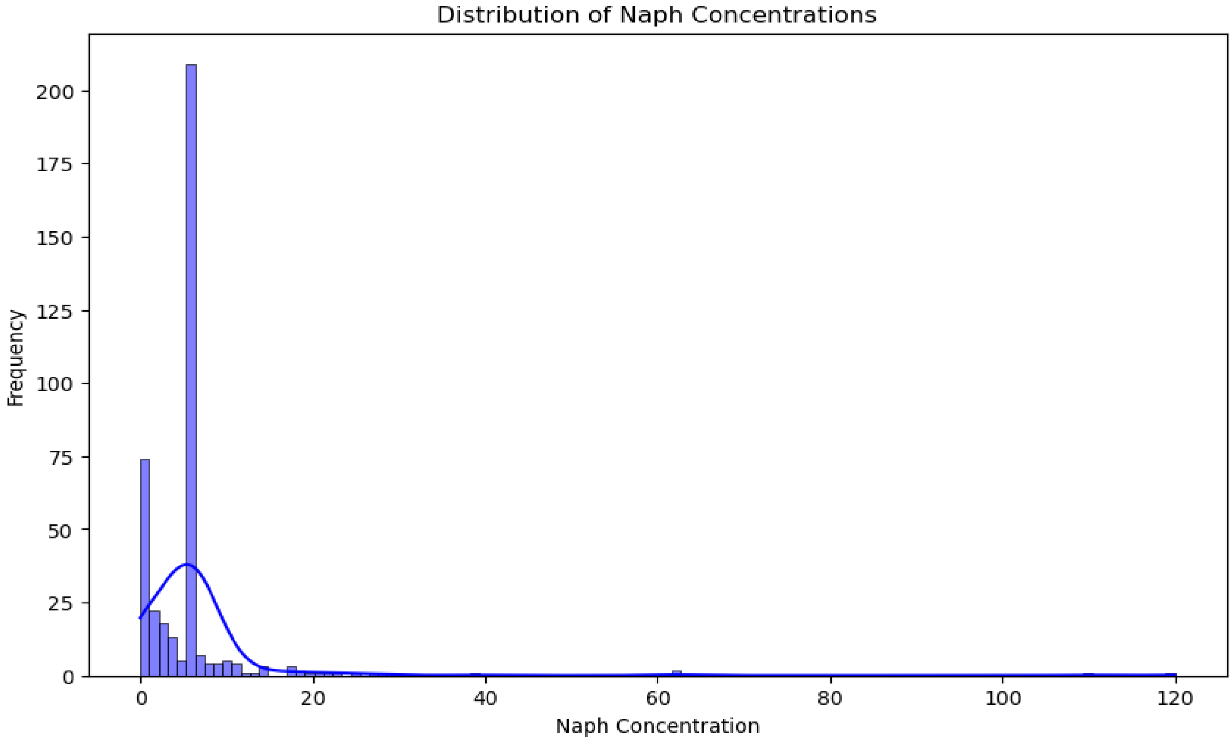


Figure: 1.3

- **Description**: The histogram shows the frequency distribution of Naph concentrations, overlaid with a smooth density curve.
- **Insights**: The data distribution appears skewed, with a concentration peak at lower values. This suggests that most samples have low levels of Naph, while a few may represent elevated or anomalous concentrations.
- **Implication**: These outliers with elevated concentrations might indicate localized contamination events or hotspots. Further investigation into these anomalies is recommended to identify sources of contamination.

1.4: Correlation Heatmap of Chemical Measurements

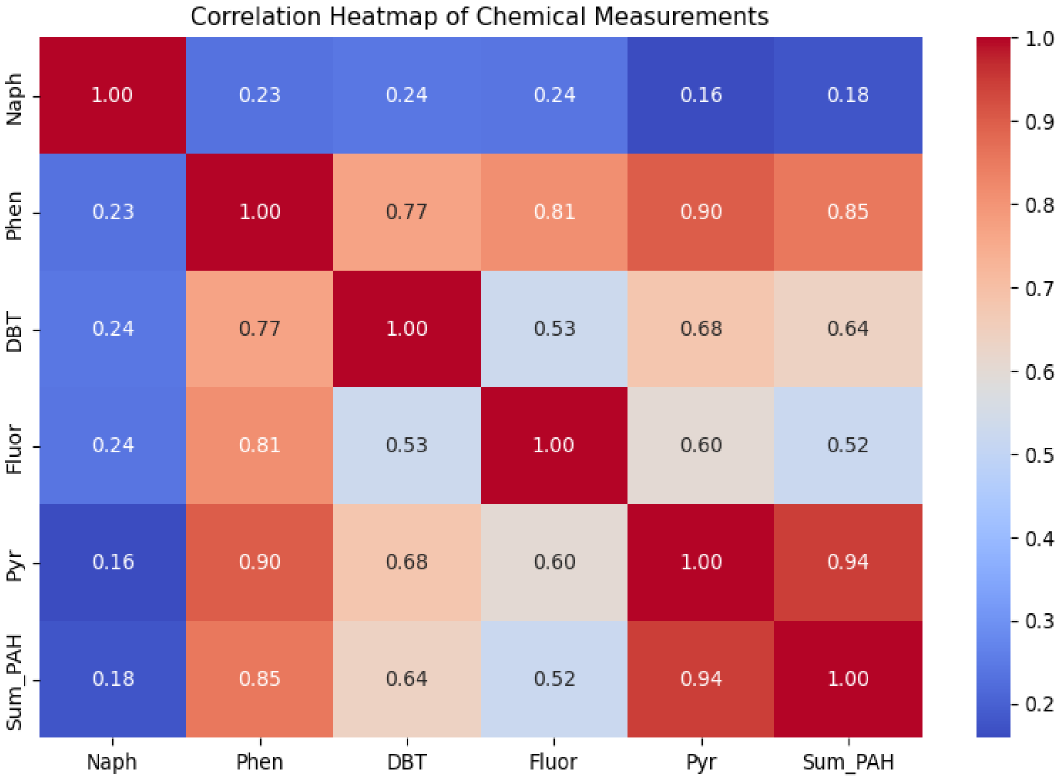


Figure: 1.4

- **Description**: The heatmap visualizes the correlation between selected chemical measurements (e.g., Naph, Phen, DBT). Values range from -1 (strong negative correlation) to +1 (strong positive correlation).
- **Insights**: Strong positive correlations (e.g., Naph and Phen) might indicate shared sources or co-occurrence in environmental matrices. Weak or negative correlations could highlight independent sources or distinct behavior in the environment.
- **Implication**: Such relationships can help identify chemical groups for monitoring and aid in source apportionment studies.

Conclusion:

The analysis of the dataset reveals important insights into environmental sampling and chemical distribution patterns. The geographic distribution of samples indicates a focused sampling effort in specific regions, such as Feather River Canyon, while other areas appear underrepresented, suggesting potential gaps in data coverage. This uneven distribution could limit the overall understanding of environmental conditions across the broader landscape. Similarly, temporal trends show fluctuations in sampling activity, with some years exhibiting high levels of data collection, likely driven by specific events or research priorities, while others show a marked decline, which may hinder consistent monitoring and long-term assessments.

The chemical concentration analysis, particularly for Naph, highlights a predominantly low concentration across samples, with a few notable outliers suggesting localized contamination events or hotspots. These anomalies are critical to investigate further, as they may signify environmental risks or pollution sources that require targeted mitigation efforts. The observed strong correlations between certain chemicals, such as Naph and Phen, point to shared sources or co-occurrence mechanisms, offering valuable clues about potential pollutant origins and transport pathways.

In summary, while the dataset provides significant insights into environmental conditions, addressing geographic and temporal gaps in sampling, investigating contamination hotspots, and leveraging multivariate analyses will enhance the comprehensiveness and impact of future environmental assessments.

Reference:

[1] https://data.gov/

[2] https://catalog.data.gov/dataset/polycyclic-aromatic-hydrocarbon-samples-2011-to
2021-ospr-ds714-1e70b

[3] https://www.sciencedirect.com/science/article/pii/S0146638016303060

[4] https://www.sciencedirect.com/science/article/pii/S0048969708012710

[5] https://www.sciencedirect.com/science/article/pii/S0025326X98000885

[6]https://books.google.com/books?hl=en&lr=&id=sZpPDQAAQBAJ&oi=fnd&pg=PP1&dq=baseline+Monitoring+and+Oil+Spill+Response+Michel+%26+Rutherford+(2014)&ots=JBcEPyej-T&sig=OnKP30rNEvrgoOi8YJvQ32lkEHI