The Pennsylvania State University

The Graduate School

Department of Energy and Mineral Engineering

ANALYSIS OF COMPOSITIONS OF FLOWBACK WATER FROM MARCELLUS SHALE WELLS BY UTILIZING DATAMINING TECHNIQUES

A Thesis in

Energy and Mineral Engineering

by

Yuewei Pan

© 2015 Yuewei Pan

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science

December 2015

The thesis of Yuewei Pan was reviewed and approved* by the following:

John Yilin Wang
Assistant Professor of Petroleum and Natural Gas Engineering
Thesis Advisor

Jeremy M. Gernand
Assistant Professor of Industrial Health and Safety

Eugene Morgan
Assistant Professor of Petroleum and Natural Gas Engineering

Luis F. Ayala H.
Associate Professor of Petroleum and Natural Gas Engineering
Associate Department Head for Graduate Education

*Signatures are on file in the Graduate School

ABSTRACT

Water that is produced from shale gas wells in three months after stimulation treatments and shut-in periods is called flowback water. Volume and salt concentrations of flowback water depend on geology, drilling and completions, stimulation and flowback operations. Recent studies include evaluations of geochemical origins based on the compostition concentrations, flowback sampling analysis and numerical studies. However, an in-depth understanding of chemical compositions as well as the changes of compositions is still needed.

In this work, we will firstly review the literature related to flowback in shale gas wells to fully understand the chemistry, geochemistry, and physics governing a fracture treatment, shut-in, and flowback. We will then gather all public and in-house flowback data, termed 3-week or 3-month flowback in this work, to build a flowback compositional database. After data screening, we will then analyze this compositional database which would potentially affect the wasterwater treatment design by using four different methods: geographical, changes over time, linear regression, clustering and multi-variable analysis. New understandings such as the magnitude and prevailing trends of concentrations for target constituents as well as the correlations among flowback compositions, the differentiation between early and late time flowback water were obtained and explained on the basis of geochemistry and physics.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost I would like to gratitude Dr. John Yilin Wang sincerely for his guidance and help not only for this thesis, but also throughout the entire two years of study in Penn State University as a student in petroleum engineering option. Working with Dr. Wang was an immeasurable learning experience for me, with continuous support and assistance along this path, I shall overcome the obstacles and prevail. During this impressed research experiences, I have gained imponderable abilities to handle technical problems not only are refrained in this small area, but when face to the unknown domain also shall I have no fear. Furthermore, Dr. Wang's personality also deeply affects me and keep changing my mentality about the living life as well as the way I reckon on science. Other than that, I also would like to thank to Mr. Arthur Rose for sharing the important information about valuable datasets and Dr. Li Li for her advice on my methodologies. I am also thankful for Dr. Gernand and Dr. Morgan attending to my final thesis defense and sharing their advice.

Finally, I would like to be devoutly thankful to the Penn State University EME Department as a whole. The dedication for each professor on every deliberate course have great impact on me, without gaining knowledge from those curriculums, I would never achieve such great improvement. The encouragements and efforts each professor sheds into us is valued and appreciated.

Chapter 1 Introduction

The flowback volume is regarded as a vital parameter to estimate the effective fracture volume in order to appraise the production performance and to plan field operations and water management. Nevertheless, flowback water compositions are seldom analyzed along the hydrocarbon production. In this study, investigations by utilizing different datamining approaches on the flowback compositions are the main tasks. Prior to the execution of these analysis, a complete database including all sources of datasets are included. Preliminary treatments for this database are carried out to evaluate the variable importance as well as corrleations among compositions and then several datamining models are developed. The assessments of each model is applied by using different dataset according to available informations.

In Marcellus shale gas wells, 10-100 thousands of barrels water are injected with 4-5 different chemical additives, but only about 10% of this volume flows back in 3 weeks and will contain inorganic solutes and organic components. An increasing or prominently decreasing concentrations of inorganic constituents could effectively influence the flowback attributes. Various possible mechanisms of different chemical processes occurred underground accounting for these concentration variations could be interpreted evidently, for example, dissolution of salts from the source rocks, mixing between flowback and formation water and reactions among active ions. The correlations among the compositions are still a mystery since the complexity of the chemical reactions and the impacts of those reactions on flowback deliverability which might dominate over each other along different periods. Multifarious datamining models are performed including geological change in concentrations, variations in concentrations along with chronological sequence, linear regression and multivariate approaches. These methods are adopted to capture the correlations among compositions. By using these models, a relative integral

understanding based on the flowback compositions would be generated which might be valuable for post-water treatment and gas production prediction.

Flowback water after a fracture treatment may contain high concentrations of chemicals which are affect to the environment as well as the human health. Being acquainted with flowback water compositions not only helps alleviate the environmental impacts as well as diminish the risks affecting the human health, but also a better understanding on flowback compositions would aid to optimize the operation in gas production industry.

Chapter 2 Literature Review

2.1 Marcellus Shale Formation

The extensive Marcellus shale formation, deposited over 350 million years ago which underlies in a shallow inland sea located in the eastern United States. Marcellus shale covers the most of northern Appalachian Basin (J.Soeder & M.Kappel, 2009). The Marcellus Shale formation stretches from southern New York across Pennsylvania, sweeps through part of western Maryland and most of West Virginia and attaches eastern Ohio as well (J.Soeder & M.Kappel, 2009). Stratigraphically, Marcellus shale was formed in the lowest or the basal part of a thick sequences of Devonian age, the stratigraphy of Devonian shales and rocks other than in the northern part of the Applachian Basin is shown in Fig.2-1 (Martin, 2008). Although the organic rich black shale dominates the lithology where the influx of the majority of the younger Devonian sediments buried above them. This event was triggered by sea level variation during the deposition progress almost 400 million years ago which black shale was precipitated in relatively deep water devoid of oxygen (Kostelnik, Gold, Doden, & Harper, 2003). Organic matter deposited in an anaerobic circumstance with a series of geologic events in the Marcellus Shale. The black shales contain iron ore which were used to play an important role in early development of the region, and some of the radioactive compounds such as uranium and pyrite which are now still environmentally hazardous (McBride, 2004). The fissile shales are easily to be crushed and eroded, indicating big challenges for civil and environmental engineering (P.Werne, Sageman, W.Lyons, & Hollander, 2002).

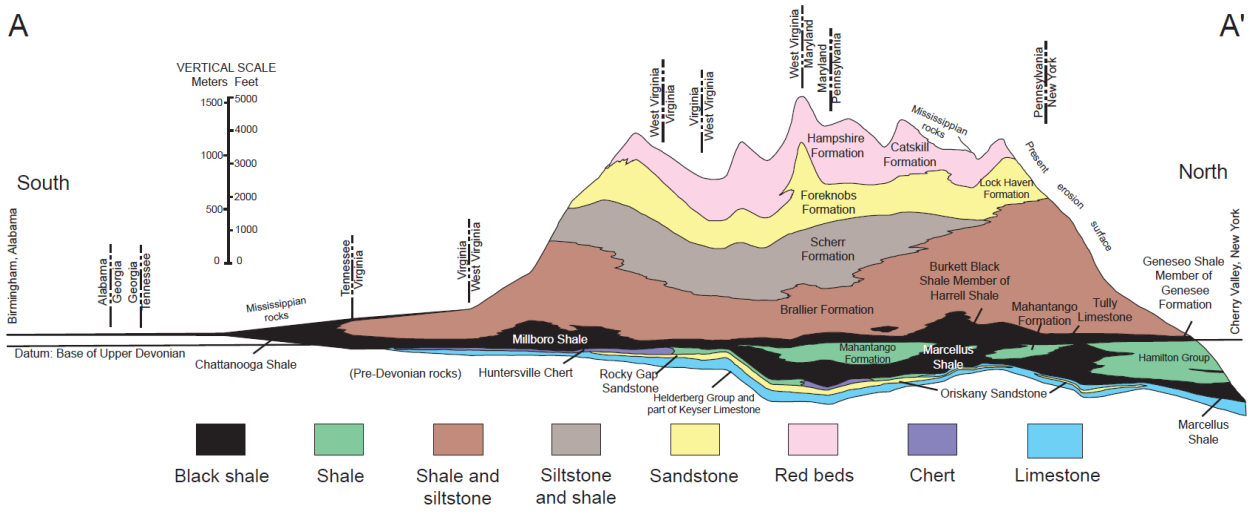
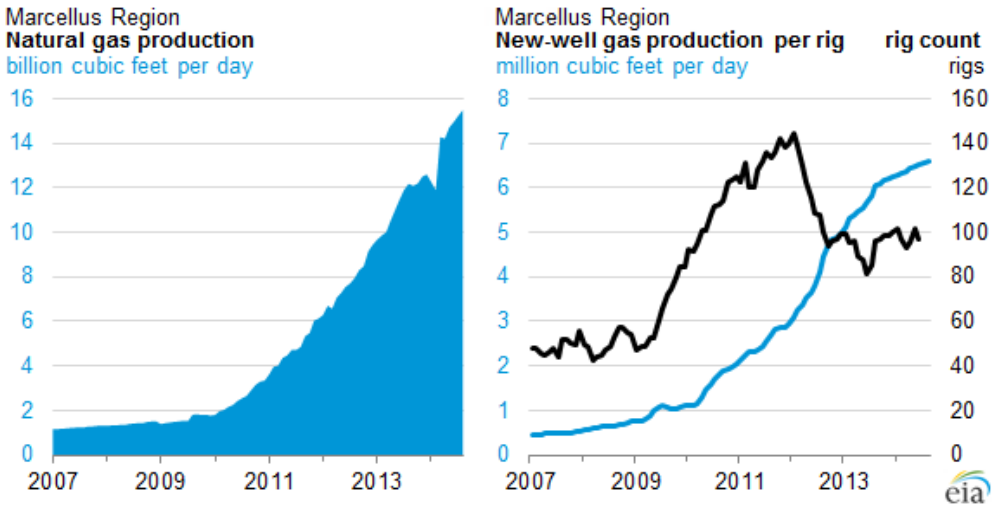


Figure 2-1. A geological cross section of upper to middle Devonian strata (Boughton & McCoy, 2006).

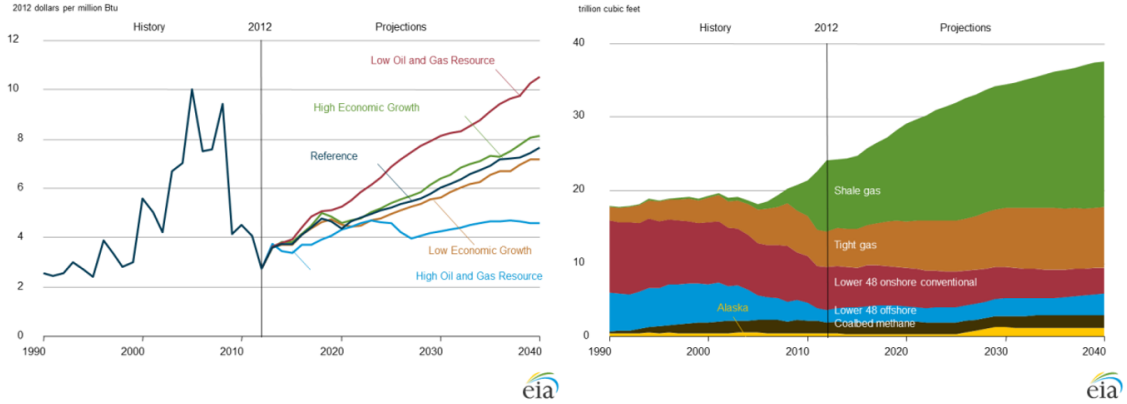
The Macellus Shale formation mainly consists of black shale in which the limestone might be interbeded and concentrations of iron pyrite (FeS_2) and siderite ($FeCO_3$) may also be noticeable (U.S Environmental Protection Agency, 2008). Fragments are formed after the exposure and the reactions between pyrite to air as well as the pyrite and limestone particles generating tiny gypsum ($CaSO_4 \cdot 2H_2O$) crystals. Moreover, the Marcellus also contains uranium, and isotope of the radioactive decay uranium-238 (^{238}U) which a source rock for radioactive radon gas is abundant. The black shale was formed from flysch, a fine mud precipitated in the deep water which buried the underlying Onodaga limestone beds. Since the deepening sea cut-out the supply of carbonates that form limestone interbeded (Hand & Banikowski, 2008). The organic content, which might be previously settled to the bottom with an anaerobic decay process, thereby was preserved through epochs. The radioactive compound such as uranium was also incorporated in these organic muds simultaneously. What is noteworthy, recent investigation reveals that hydrocarbons are founded including natural gas due to the compression and heat during geologic time. The gas presents in fractures, pore spaces and is absorbed onto organic matters within the shale.

2.2 Flowback Water From Marcellus Shale

Shale gas plays a critical role in the future development of natural gas industry. According to an annual energy outlook from 2014 in Fig.2-2, a prediction of total gas will be enlarged to 31.6 Tcf based on the natural gas consumption in 2012. The natural gas price rises with the economic growth (Fig.2-2 (c)) as well as the unexpected increase in production costs (Fig.2-2 (a)). As a consequence, 56% increase in total natural gas production from 2012 to 2040 in the AEO 2014 Reference Case are consist of the increment in shale gas, tight gas, and offshore natural gas resources (Fig.2-2 (d)). What is worthy to mention is that shale gas production dominates the contribution of gas production growth. The rest are declined more or less instantaneously and keep relatively stable for a long term. A paradigm could be exemplified, Pennsylvania State covers the most part of the Marcellus shale and 85% of state's shale gas production is from just 6 out of 67 counties (Fig.2-3). Up to the latest published report, natural gas production in the Marcellus Region surpassed 15 billion cubic feet per day (Bcf/day) through July 2014. Broadly, the West Virginia and Pennsylvania are the largest producing shale gas basin in the Marcellus Region, accounting for 40% of U.S. shale gas production, has increased dramatically over the past four years shown in Fig.2-2(a),(b) which benefit from recent advanced technologies (Administration, 2014).



(a) (b)



(c) (d)

Figure 2-2 a)b) Drilling productivity growth annual report, c)Annual average henry Hub spot natural gas prices in different cases, d)Natural gas production from different sources (Administration, 2014).

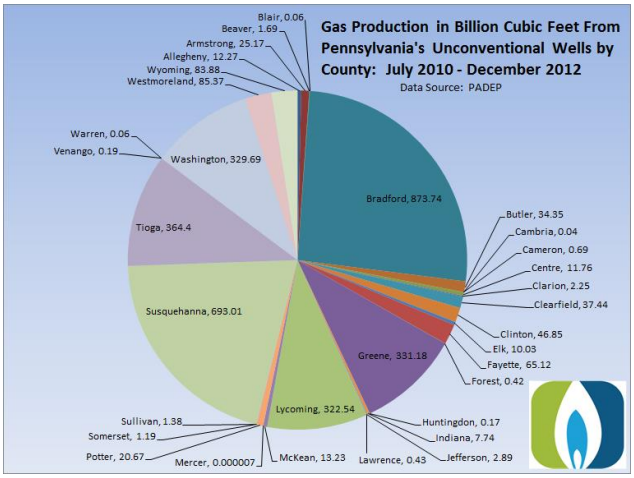


Figure 2-3 Unconventional gas production in different counties of Pennsylvania (PADEP, 2013)

In order to extract the large amount of natural gas reserves from such fine-grained rock in an economical and commercial level, higher permeable flow paths must be created in the formation by commonly using a combination technique of horizontal drilling and hydraulic (J.Soeder & M.Kappel, 2009). Within recent applications, hydraulic fracturing composition is one of the

emphasized issues and an illustration is presented in Fig.2-4. Million gallons of water in total, mixed with various additives, are pumped under high pressure fissuring the rock. The rock fractures are then propped open by sequential stages of fracturing water carrying the sand or other high-pressure resistance materials so as to propped a ‘Highway’ in the fissures for gas and other fluids to move out along the fracture towards the wellbore. The productivity of the well is changed after the treatment, low productivity of the reservoir is transformed into high productivity. Lower pressure drawdown is established within the stimulated zone. More details shown in Fig.2-5 annotate an integral process of hydraulic fracturing process. Hydraulic fracturing in Marcellus occurs after a cased well perforated within the target zones containing oil or gas. The fracturing fluid is injected flowing through the perforation, cracking the formation under the high pressure. Typically, fracturing fluids is a mixture of water, proppants and different purposes of chemicals (Table.2-1), the fraction of each constituent is elucidated in Fig.2-4, the pie chart assesses the weight percentages of each function group that might be used in a fracture treatment, yet not all of these chemicals are utilized in each phase of the treatment. Moreover, according to the chart, water and sands make the greatest contribution other than the minor chemicals, nevertheless, since the magnitude of the total amount of fluid is enormous, the absolute number of each functional group are still considerable though the proportion tends to be small.

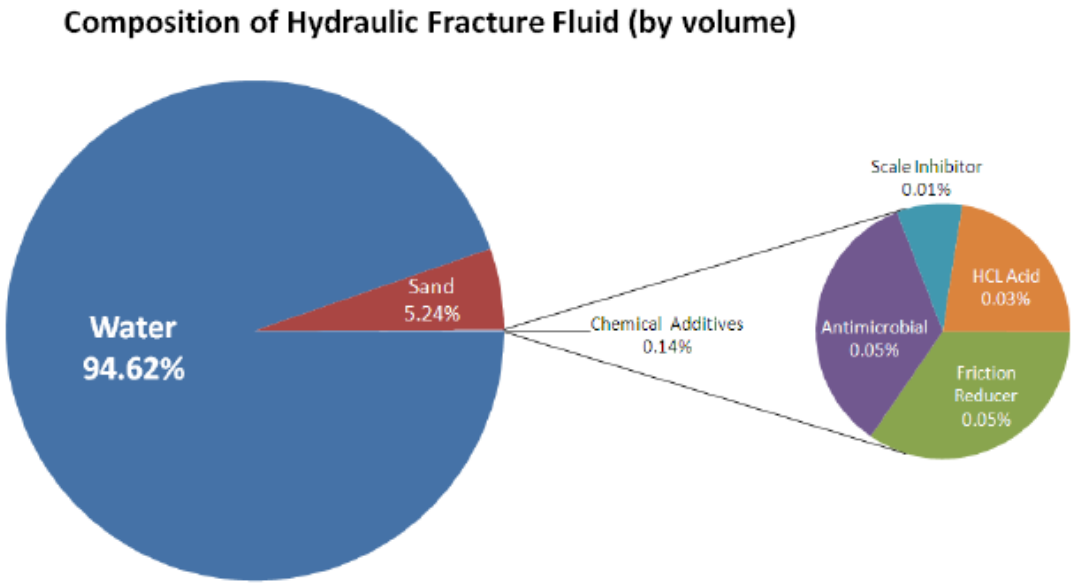


Figure 2-4 Volumetric Compositions in frac-fluids (Range Releases Frac Fact Sheet, 2010).

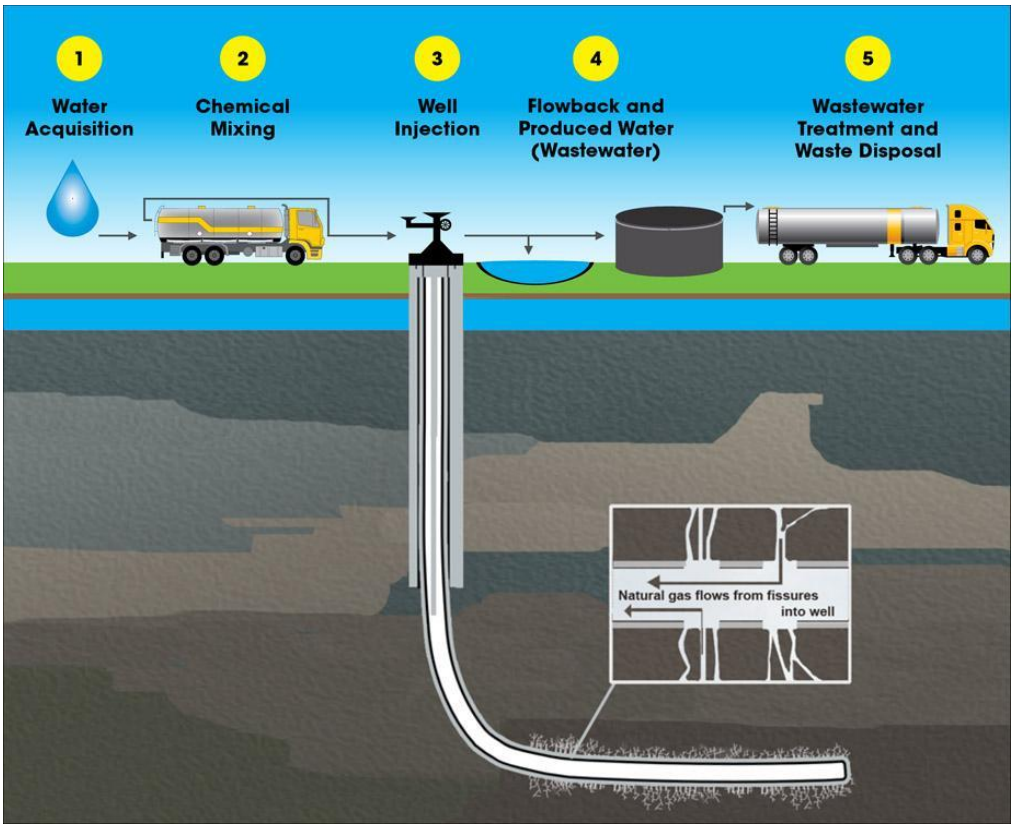


Figure 2-5 Interpretation of the integral process of hydraulic fracturing (The Hydraulic Fracturing Water Cycle, 2012)

Table 2-1 Basic chemical used in hydraulic fracturing (SGEIS, 2011).

Additive Type	Description of Purpose	Examples of Chemicals[47]
Proppant	"Props" open fractures and allows gas / fluids to flow more freely to the well bore.	Sand [Sintered bauxite; zirconium oxide; ceramic beads]
Acid	Removes cement and drilling mud from casing perforations prior to fracturing fluid injection, and provides accessible path to formation.	Hydrochloric acid (HCl, 3% to 28%) or muriatic acid
Breaker	Reduces the viscosity of the fluid in order to release proppant into fractures and enhance the recovery of the fracturing fluid.	Peroxydisulfates
Bactericide / Biocide / Antibacterial Agent	Inhibits growth of organisms that could produce gases (particularly hydrogen sulfide) that could contaminate methane gas. Also prevents the growth of bacteria which can reduce the ability of the fluid to carry proppant into the fractures.	Gluteraldehyde; 2,2-dibromo-3-nitrilopropionamide
Buffer / pH Adjusting Agent	Adjusts and controls the pH of the fluid in order to maximize the effectiveness of other additives such as crosslinkers	Sodium or potassium carbonate; acetic acid
Clay Stabilizer / Control /KCl	Prevents swelling and migration of formation clays which could block pore spaces thereby reducing permeability.	Salts (e.g., tetramethyl ammonium chloride Potassium chloride (KCl)
Corrosion Inhibitor (including Oxygen Scavengers)	Reduces rust formation on steel tubing, well casings, tools, and tanks (used only in fracturing fluids that contain acid).	Methanol; ammonium bisulfate for Oxygen Scavengers
Crosslinker	Increases fluid viscosity using phosphate esters combined with metals. The metals are referred to as crosslinking agents. The increased fracturing fluid viscosity allows the fluid to carry more proppant into the fractures.	Potassium hydroxide; borate salts
Friction Reducer	Allows fracture fluids to be injected at optimum rates and pressures by minimizing friction.	Sodium acrylate-acrylamide copolymer; polyacrylamide (PAM); petroleum distillates
Gelling Agent	Increases fracturing fluid viscosity, allowing the fluid to carry more proppant into the fractures.	Guar gum; petroleum distillates
Iron Control	Prevents the precipitation of metal oxides which could plug off the formation.	Citric acid;
Scale Inhibitor	Prevents the precipitation of carbonates and sulfates (calcium carbonate, calcium sulfate, barium sulfate) which could plug off the formation.	Ammonium chloride; ethylene glycol;
Solvent	Additive which is soluble in oil, water & acid-based treatment fluids which is used to control the wettability of contact surfaces or to prevent or break emulsions	Various aromatic hydrocarbons
Surfactant	Reduces fracturing fluid surface tension thereby aiding fluid recovery.	Methanol; isopropanol; ethoxylated alcohol

With this recent advancement of technology, Marcellus Shale gas play has been so far predicted that 489 Tcf of technically recoverable gas are existed, it is adequate to satisfy US demands for roughly two decades (Shale Energy Business Briefing, 2015). Therefore, an outburst of hydraulic fracturing would be implemented through the entire gas play, however, a concern of water quality and water consumption has brought over publics (Me.E.Blauch, R.R.Myers, T.R.Moore, & B.A.Lipinski, 2009). Technically, after hydraulic fracturing, water flowing back with contaminants resulting from the contact and reaction with formation rocks and naturally occurrence of sedimentary basin brines as well, might be unlikely to be disposes readily

(O.Vazquez, et al., 2014). The term ‘flowback’ is often used to describe this early stage of water production. This early periods shown in Fig.2-6.

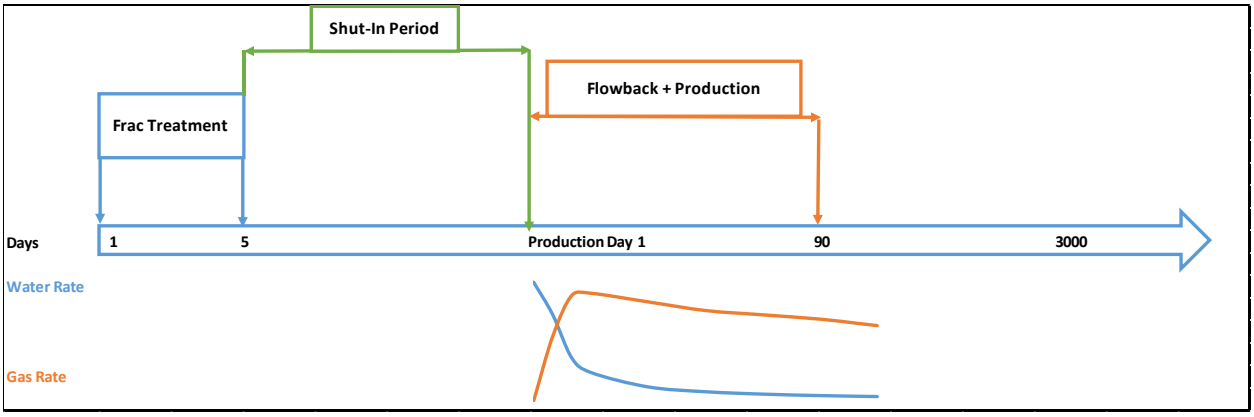


Figure 2-6 The flow sequence of a typical shale well with a frac-job followed by a flowback period.

In 2014, Wattenbarger and Alkouh concluded utilizing gas-water two phase modeling compiled with early flowback data could effectively estimate the fracture volume. The observation of bilinear regression has been identified after the early water flowback rates are coupled. Yet none of them consider about the correlation within the flowback compositions or between flowback compositions and production profile. The fracture flowback fluids or post-frac water contains massive quantity and various kinds of organic and inorganic compounds. The quality of the flowback fluids would be diversified due to the flowback condition as well as a variety of different exterior attributes. For example, the retention time for flowback contacting with formation rocks within the reservoir is critical to have the tendency of a relatively high proportion of TDS (total dissolved solids) for which it is necessary to be reduced to the regulatory levels prior to disposal. Several possible assumptions have been made to interpret how such high concentrations of chemicals happened. One possible theory refers to the dissolution of minerals which has been deposited within the formation, for instance, the evaporation of halite (Dresel & Rose, 2010). However, an indication of log(Br) vs. log(Cl) shows that the conventional brines in Pennsylvania

were derived from highly evaporated seawater that had been diluted by freshwater. Yet the halite dissolution is still considered to be a minor possible way for such high salinity phenomenon (Haluszczak, Rose, & Kump, 2012). Furthermore, the encroachment of basinal brine is another hypothesis that the formation fluid and fracture fluid are directly mixed with each other. A clear example from the Barnett shale play which a deep saline aquifer in Ellenburger formation was breached during hydraulic fracturing (Hayden & Pursell, 2005). Besides the deep aquifer, the mobilization of hypersaline connate water is another issue that immobile water trapped in the pores by capillary pressure would be diluted by fracturing water. The connate water trapped in the vicinity of the fracture would tend to have ion exchange due to the concentration difference driving force.

Comprehensive understanding the chemical compositions and different types of constituents existed in the flowback water is critical to establish a fundamental knowledge for proper water treatment process which benefits to alleviate health and environmental impacts or to effectively develop a datamining model to predict the gas production. The fracture flowback volume are varied from well to well, Hayes (2009) claims in Gas Technology Institute (GTI) Mission Report that 19 wells are recorded within 90 days, the accumulative weighted percentage of water flowback volume after 90 days is at the range 10% - 88% (Hayes, 2009). Another Spotts well Chief Oil & Gas in Lycoming County revealed the concentration of each ion at different weighted percentage of flowback that has been collected (Haluszczak L. , 2011). Dresel (1985) and Rose only claimed the compositions of oilfield brines at random one-time sample as well as the Bureau of Oil and Gas Management (BOGM) dataset. None of these two includes a consecutive flowback time or the flowback volume. The flowback water compositions consist of high concentrations of diversified compounds, such as arsenic, sodium, potassium, chloride, barium,

strontium, magnesium, sulfates and radionuclides. The range for each constituents are widely spread, with sodium (50-40,000ppm), chloride (5,000-80,000ppm), and barium (50-9,000ppm), and total dissolved solid (1,000-150,000ppm) which are particularly abundant in water (Hayes, 2009). The diversity of each composition depends on different locations and initial chemicals used in the injection fluids, in which different chemicals function properly within the formation. A robust case executed by Blauch et al. (2009) that over 100 flowback samples were collected through 18 months from the lower southwestern and upper northeastern regions of the Marcellus shale play. The basic idea initiated by Blauch was to address the public attention of the high salinity from flowback water. The concentration of chloride reaching 100,000mg/L at the highest were investigated which the retention time for flowback in the reservoir firstly came to the big picture. Other studies such as Haluszczak et al.(2012) also conducted an integral research about the flowback compositions from various wells and the results turned out to be identical. In Haluszczak et al (2012) analysis, the concentrations of Cl ranged from 1,070 to 151,000mg/L as other constituents which are specifically dominants in the flowback water including Ra-227(73-6,540pCi/L), Ba(76-13,600mg/L), Mg(22-1,800mg/L), K(8-1,010mg/L), and Ca(204-14,800mg/L).

Chapter 3 Problem Statement

The goal for this research is to understand the change of salt compositions in flowback water. The in-situ brines are taken as a reference to reasonably identify the characteristics of early-time and late-time flowback water by compositional analysis using different datamining approaches.

Several datasets used to build an integral database. In this study, robust datamining techniques such as PCA are implemented to better understand the flowback compositions' behavior. The steps to achieve the research goal is conveyed as below:

1. Different datasets are gathered after screening, the consistency is ascertained prior to feed in the datamining models (Advanced imputation, List-wise deletion, etc).
2. The time-series data and oilfield brines are sieved from the database and then plotted on the map by showing the variations geographically in concentrations of different compositions (Cl, Na).
3. The volume-series and time-series data are singled out to perform inorganic compositions' changes with time and volume such monovalent ions like sodium, chloride etc and divalent ions calcium, magnesium, sulfate etc.
4. The time-series data and oilfield brines from step 2 will be applied into different linear regression models (Cl vs. Br, Ca vs. Br, etc).
5. Only the major contribution constituents (13 different compositions) from time-series in step 2 will be feed in Principal Component Analysis (PCA) model which is introduced to minimize the complexity of the correlations and visualize the specified variances readily.

6. Time-series data for certain compositions (Na+K, Ca, Mg, Total Alkalinity, Cl, SO_4) are transformed into proportions and then plugged into Piper trilinear diagram which yields the coarse characteristics of different flowback water samples.

Chapter 4 Database Development Based on Field Data

Flowback data are available in different open sources and internal reports. This chapter focuses on backgrounds of each datasets and the development of an integrated database based on the characteristics and consistency of available datasets. Three volume-series and time-series datasets including GTI (2009), Blauch (2009) and Chief O&G Spotts Well (2011), as well as two one-time sampling flowback water observations including BOGM (2011) and Pritz & Kirby (2010) datasets are gathered and numbered in sequences. The futile data are screened manually while gathering the datasets. The inconsistency among datasets cannot be avoided so that each dataset are performed individually. The oilfield brine compositions originated by Dresel & Rose (2010) are incorporated in this research so as to have the comparison between flowback samples and in-situ brines.

4.1 GTI Dataset (Hayes, 2009)

The need for the compositional analysis of flowback water propelled 17 member companies of the Marcellus Shale Coalition (MSC) providing 19 sampling locations where hydraulic fracture had been performed. The GTI dataset is generated based on detections on 19 locations during a consistent 0, 1, 5, 14 or 15, 90-day sampling period. It comprises of almost all the inorganic and organic compositions under the PADEP sampling protocol.

Literally, 3 out of 19 well sites located upper north in West Virginia including two in Lewis County and one in Taylor County which are shown in and Fig.4-2. The rest are located in Pennsylvania, distributed on southwestern and northeastern part which are shown in Fig.4-1. At each of 19 well sites, samples of the influent prior to the injection was taken as the Day 0 indicated. A particular time series has been scheduled as the sampling criteria for each well sites so that the flowback water were sampled properly and in consistent. The compounds were determined and

divided into three groups, general chemistry or inorganic compounds, organic compounds, and metals. Preliminary results from this dedication turns out that magnitude values for pH, alkalinity, total dissolved solids (TDS), total organic carbon(TOC), oils and greases and other parameters from general chemical category are within the acceptable ranges for reported conventional produced water from USGS. In this report, sodium, calcium and potassium took part in nearly 90% of the TDS dominating the cations in flowback water. Metals are the second importance that make up to 5-9% of the TDS from the flowback, such as iron, calcium, barium, magnesium, and boron. Heavy metals are at low levels which might be a considerable concern in urban industrial wastewaters and POTW sludges. Furthermore, as for organic compounds, almost 96% of volatile organic constituents (VOCs) were under non-detectable levels and less than 0.5% were detected above 1 ppm. The similar events were also identified for semi-volatile organic constituents and pesticide (Hayes, 2009).

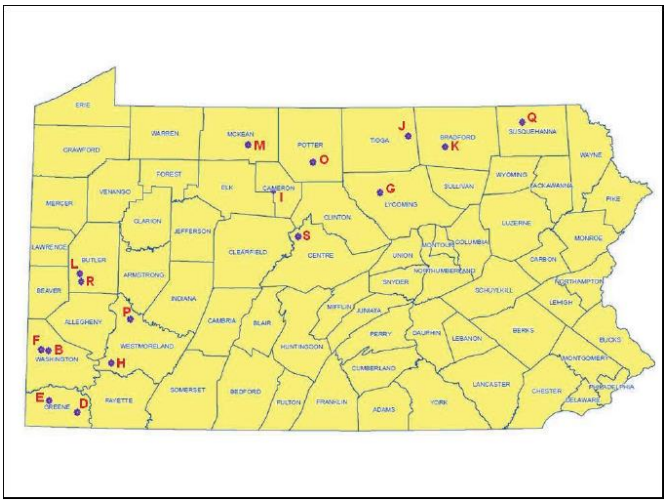


Figure 4-1 Sampling locations in Pennsylvania from GTI dataset (Hayes, 2009).

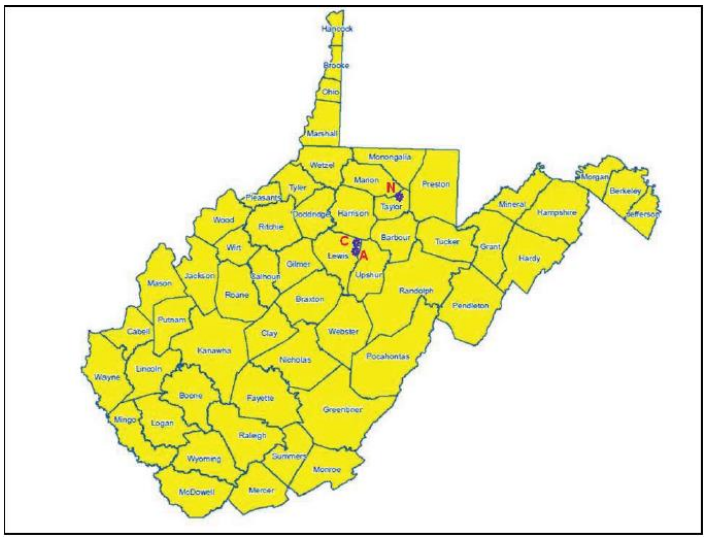


Figure 4-2 Sampling locations in West Virginia from GTI dataset (Hayes, 2009).

4.2 Dresel & Rose Dataset (Dresel & Rose, 2010)

With the growing interest on oilfield brines as well as the flowback water from Marcellus Shale, a dataset was generated originally in 1985 and then renewed and published in 2010 is named as Dresel & Rose dataset. 40 samples were taken and analyzed mostly one year or more after production initiated. The dataset mainly contains the inorganic compositions on which this study primarily concentrates based on different flowback datasets. The comparisons between each flowback datasets and oilfield brines is done in the following analysis.

Basically, despite the desired oil, gas and gas condensate production, vast wells produce saline solution or brine (Collins, 1975). High salinity has been found in previous analyses (C.W.Poth, 1962) and the major components are sodium and calcium chlorides which are along with other solutes at relatively low concentrations. The report was primarily to dedicate to better understand the geographic and geologic distribution of salinities and the pH, oxidation state, and chemistry of the brine depending on the depth. Minor concerns such as the origin of those solutes, history, and present hydrology of the brines as well as the content and trace elements discoveries

related to geography might be also involved in this effort. The terminology of “brine” is well defined that more than 35,000 mg/L (about 3.5%) total dissolved solids (Hem, 1985). But in this report, the terms “oil-field brine,” “saline formation water,” “oil-and-gas-field brine,” “sedimentary-basin brine,” are used to describe the chloride-rich waters discovered in sedimentary environments and brought to the atmospheric condition accompanied with oil and gas production. Therefore, several cases which a less than 35,000 mg/L TDS has been incorporated and under investigated. The structure for the dataset gathering is cultivated with 40 new analyses completed by Dresel (1985) and a relatively tiny amount of previously published analysis including Barb (1931) dedicated in some major components analysis on oilfield brines sampled from 45 wells, Poth (1962)’s major and trace-element data analysis from Cambrian through Pennsylvania aquifers.

The histories for different analysis on these datasets were variated until Dresel summarized and remediated the gaps in the previous data. The attempt for sampling was made for both oil wells and gas wells from Lower Silurian to Upper Devonian. The sample areas were subjected to avoid water flooding and in nearly all wells, the samples were taken after the production for a year or more. Thus, it indeed reduce the possibility of the contamination by drilling and completion fluids. Sampling dates were ranging between July and December 1982. The well locations are illustrated in Fig.4-3. Fourteen oil wells and twenty six gas wells were sample in which 11 of the gas wells and all the oil wells are sampled from the Upper Devonian. Other than that, 10 of the gas well samples were taken from the Medina Group, one is from the Tuscarora, the rest were sampled from the Ridgeley. Since the difficulties to collect adequate amount of samples for freshly produced Medina samples, only sample ED-82-27 is relatively completed. Sample ED-82-30 is only missing an alkalinity titration. Samples ED-82-28, 29 and 31 were lack of information to complete integral analysis. Previously, different methods including both field and laboratory

determinations are performed for this dataset. In this study, a variety of approaches are implemented for Devonian samples which would be interpreted in the following chapter. (Dresel & Rose, 2010).

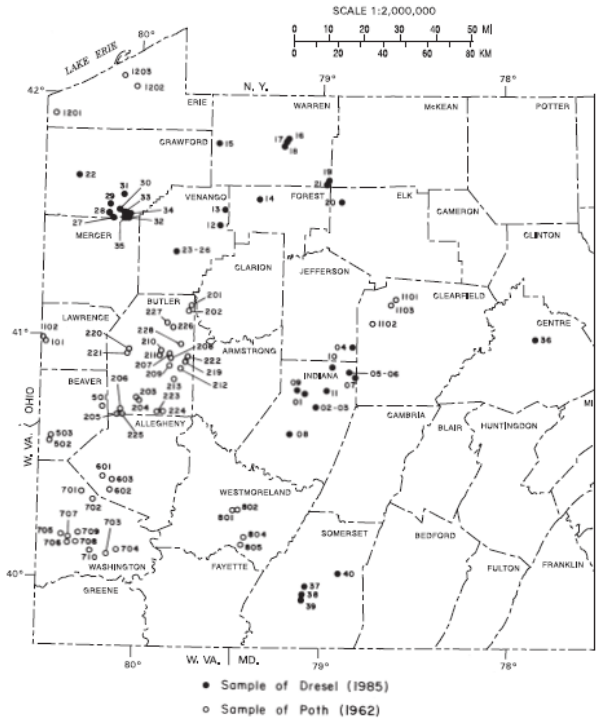


Figure 4-3 Well locations of samples by Dresel (1985) and Poth (1962)

4.3 BOGM Dataset (Haluszczak, Rose, & Kump, 2012)

The Marcellus flowback data acquired by the PA Department of Environmental Protection Bureau of Oil and Gas Management (BOGM) includes 40 flowback water samples distributed in Pennsylvania. These data are collected and obtained from James Fuller of PA DEP Bureau of Oil and Gas Management in Harrisburg. The date for only a couple of samples are available in 18 wells out of 40 indicated in the following interal database table.4-1. The sampling days may be ranged from 1 day to 51 days after fracturing. Unfortunately, time-series flowback analysis is unlikely to be performed because most of samples are taken only for one time and the inconsistency

of the sampling time. However, most of the observations are enclosed with well permit number so that the production prediction analysis could be performed.

According to the report, all the data were collected from southwestern Pennsylvania stretching along the orthogonal line towards the northeastern Pennsylvania as shown in Fig.4-4. The TDS concentration is similar as the previous GTI dataset. The upper and lower boundaries are approximately 264,000 mg/L and 5,090 mg/L. The monovalent and divalent cations such as sodium, calcium, potassium and magnesium are of high concentration dominating the TDS constituents. The possible coupling anions are chloride which are abundant supremely. Relatively low concentration of sulfate and bromide might be still considerable since specific chemical reactions may be triggered by these two anions. However, the sensitivity of the probe for determining sulfate is unable to detect the exact value but a range could still be identified (Protection, 2011).

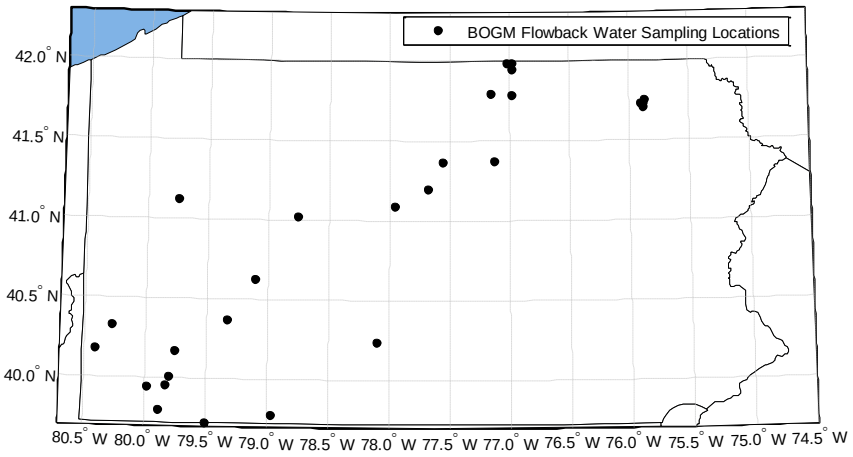


Figure 4-4 Well locations of BOGM dataset

4.4 Pritz & Kirby Dataset (Pritz, 2010)

Pritz and Dr. Kirby unveiled a collection of dataset which are also included in this research. These data are gathered under the permission of PA DEP (2009). However, in this dataset, most of the flowback water samples are without the sampling time. Therefore, most of the datamining tactics cannot applied to this dataset, nonetheless, this dataset is set to be a testing dataset for potential model of gas production prediction model which will be further discussed later.

4.5 Blauch (2009) and Chief Oil & Gas Spotts Well (2011) Dataset

Other than the major dataset gathering above, two relative small datasets are extracted and rebuilt. One of those is taken from a stimulated Marcellus Shale well belongs to Chief Oil and Gas Company in western Lycoming County presented in Haluszczak (2011) Bachelor thesis. An indication of increment in chloride in flowback water has been observed as the samples were taken via a two-week period. The concentration for each constituents are recorded as the volume of flowback returned, at the beginning the third, second third sampled and final third sampled of the flowback. The dataset itself interpreted a rising on ionic concentrations in the flowback flowed out of the Spotts Well (Haluszczak L. , 2011). The other one is coming from the Blauch et al. (2009) dataset. An analytical dataset of flowback waters from the Marcellus Shale well A located in southwestern Pennsylvania was collected as nearly 40% of flowback recovery. The similar trend for both sodium and calcium is identified. The major components such as divalent ions are analyzed in the following chapter 6. But other relative low level of components are remain disguised, therefore, the analysis based on this dataset is within limitation.

4.6 Overview of Database

Fundamentally, all the datasets are included in the database and numbered sequentially as illustrated in Table.4-1. The availability of the time and volume-series as well as the well locations

are also presented. Totally, flowback and oilfield brine samples from 125 Marcellus wells are included.

Table 4-1 Overview of database

Well #	Sample Collected on Flowback Days					Note	Well-Permit	Lat	Long
	0	1	5	14/15	90				
1	x	x	x	x	x	GTI		39.1	-80.36
2	x	x	x	x	x			40.27	-80.26
3	x	x	x	x	x			39.03	-80.34
4	x	x	x		x			39.81	-80.02
5	x	x	x	x	x			39.94	-80.29
6	x	x	x	x	x			40.26	-80.39
7	x	x	x	x	x			41.43	-77.28
8	x	x	x	x	x			40.19	-79.78
9	x	x	x	x	x			41.41	-78.23
10	x	x	x	x				41.88	-76.97
11	x	x	x	x				41.73	-76.62
12	x	x	x	x				40.8	-80.07
13	x	x*	x*	x				41.74	-78.51
14	x	x	x					39.41	-79.97
15	x	x	x	x				41.66	-77.9
16	x	x	x	x	x			40.5	-79.56
17	x	x	x	x	x			41.89	-75.88
18	x	x	x	x	x			40.74	-80.06
19	x	x	x	x				41.13	-78.05
20	Unknown					BOGM	115-20095	41.72631	-75.903102
21	Unknown						115-20171	41.72638	-75.902975
22	Unknown						035-21155	41.36923	-77.559436
23	Unknown						117-20268	41.78531	-76.987285
24	Unknown						117-20325	41.94702	-76.977267
25	Unknown						117-20295	41.98925	-76.979561
26	Unknown						063-36435	40.63046	-79.110816
27	Unknown						129-27856	40.37158	-79.33216
28	Unknown						115-20147	41.74964	-75.874509
29	Unknown						115-20148	41.74951	-75.874432
30	Day20						035-21162	41.19671	-77.682991
31	Unknown						051-24098	39.72282	-79.506643
32	Unknown						051-24209	39.79976	-79.891333
33	Unknown						117-20293	41.79178	-77.155424
34		x					051-24206	40.00652	-79.805407
35		x					059-25137	39.94537	-79.980466
36	Day2						111-20268	39.77665	-78.971834
37		x	x	x			033-26585	41.01973	-78.761283
38	Unknown						027-21478	41.09236	-77.955459
39	Unknown						027-21479	41.09233	-77.955481
40	Unknown						115-20045	41.70379	-75.881335
41	Day8						115-20036	41.72821	-75.87909
42	Unknown						027-21478	41.09236	-77.955459
43	Unknown						027-21479	41.09233	-77.955481
44	Unknown								
45	Day11						061-20004	40.23425	-78.106021
46	Day37						019-21476	41.1228	-79.753513
47	Unknown								
48	Day7						125-23048	40.17897	-80.410946
49	Day51						117-20197	41.98413	-77.026332
50	Day12						059-24394		
51	Day2						129-27331	40.17138	-79.75723
52	Day16						019-21476	41.1228	-79.753513
53		x					051-23926	39.95862	-79.83427
54	Day44						117-20197	41.98413	-77.026332
55	Day19						125-23023	40.33011	-80.281747
56	Unknown								
57	Unknown								
58		x					081-20063	41.37467	-77.1345
59		x					081-20063	41.37467	-77.1345

60	Unknown				Pritz&Kirby	035-21161	41.19501	-77.68769203
61	Unknown					035-21156	41.36937	-77.55943694
62	Unknown					015-20116	41.73669	-76.6032285
63	Unknown					081-20149	41.28121	-76.631075
64	Unknown					081-20109	41.27089	-76.65969119
65	Unknown					027-21476	41.13093	-78.04724025
66	Unknown					053-29019	41.55034	-79.16475833
67	Unknown					105-21484	41.84775	-78.15733778
68	Unknown					105-21298	41.61478	-77.87637681
69	Unknown					117-20620	41.68941	-77.56073333
70	Unknown					117-20198	41.67784	-77.58123061
71	Unknown					033-26680	41.17009	-78.44976722
72	Unknown					129-27866	40.27728	-79.64681389
73	Unknown					129-27880	40.47877	-79.56526944
74	Unknown					129-27882	40.49144	-79.56915278
75	Unknown					129-27881	40.49563	-79.58927778
76	Unknown					129-27879	40.49738	-79.56966389
77	Unknown					063-36255	40.75149	-79.04436389
78	Unknown					129-27886	40.50399	-79.57778611
79	Unknown					129-27883	40.4995	-79.54335278
80	Unknown					129-28022	40.49249	-79.57472778
81	Unknown					129-28097	40.50397	-79.560975
82	Unknown					129-28031	40.50906	-79.54695278
83	Unknown					129-27865	40.2905	-79.63706389
84	33% Flowback Vol. Returned		66% Flowback Vol. Returned		100% Flowback Vol. Returned	Chief O&G LLC	081-20060	41.2747	-77.2657
85	12000bbls	13000bbls	14000bbls	15000bbls		Blauch Well A			
86	Sampling after production for a year or more					Dresel & Rose (Oilfield Brines)	ED–82–01	40.79	-79.07
87							–02	40.74	-79.05
88							–03	40.74	-79.05
89							–04	40.99	-78.83
90							–05	40.86	-78.85
91							–06	40.86	-78.85
92							–07	40.84	-78.81
93							–08	40.6	-79.16
94							–09	40.78	-79.14
95							–10	40.89	-78.95
96							–11	40.78	-78.98
97							–12	41.46	-79.57
98							–13	41.54	-79.53
99							–14	41.57	-79.36
100							–15	41.69	-79.56
101							–16	41.82	-79.2
102							–17	41.82	-79.21
103							–18	41.81	-79.21
104							–19	41.64	-78.96
105							–20	41.57	-78.9
106							–21	41.63	-78.97
107							–22	41.66	-80.33
108							–23	41.36	-79.8
109							–24	41.36	-79.8
110							–25	41.36	-79.8
111							–26	41.36	-79.8
112							–27	41.49	-80.14
113							–28	41.51	-80.14
114							–29	41.55	-80.16
115							–30	41.53	-80.11
116							–31	41.59	-80.07
117							–32	41.49	-80.06
118							–33	41.51	-80.06
119							–34	41.5	-80.05
120							–35	41.49	-80.06
121							–36	41	-77.84
122							–37	39.97	-79.09
123							–38	39.94	-79.12
124							–39	39.93	-79.11
125							–40	40.04	-78.92

x* Some of compositions are missing on that day

Chapter 5 Data Analysis Methods

A series of tactics for data analysis are deployed in this study. The path to achieve the discovery of valuable information depends on the sequence from simplicity to complexity of the datamining approaches. By using colors and marks to represent different levels of compositions in different sampling groups, compositions changing geographically are able to be visualized. Chronological change in compositions is then generated so as to ascertain the hypothetical chemical reaction occurred in reservoir condition by differentiating the magnitude of increment and decrement on specific constituents. Singled out from the vast database, the major compositions are applied with linear-regression method. The distinction and similarity between flowback water and oilfield brines will be further discussed according to the results from linear-regression models. Other than that, the K-means clustering is adopted to concurrently demonstrate the phenomenon observed in linear-regression models. A more comprehensive approach nominated as Principal Component Analysis (PCA) is employed to magnify the variances among a various compositions. The discovered information based on PCA could be utilized in post waste water treatment industry. Coincidently, the Piper trilinear diagram is also an auxiliary for engineers on decision making on examination of waste water treatments. Ideally, a high-potential prospective datamining tool is exemplified by utilizing Artificial Neural Network (ANN), predicting the initial or average gas production based on flowback compositions from which the benefits would be amplified considerably.

5.1 Flowback Compositions Change Geographically

Given a geographical map, a view of the integral database location is acquired. By visualizing the locations, the universality of the samples is convincing from which each model is built for quantitative and qualitative analysis. The well locations are pinned on maps which are

generated from built-in function of matlab mapping tool box. Furthermore, the concentrations of each constituent are marked with distinguished shapes and each color indicates the origin of each sample. Based on these locations with different marks and colors, the tendency of each composition in a variety of concentration could be identified geologically. More importantly, the statistical ratio could be also spotted from those maps to analyze the data trends chronologically.

5.2 Flowback Compositions Change with Volume/Time

The compositions always variate along the time line, sometime the flowback volume are identical to the time line which can also record composition change with regarding to the flowback procedure. Therefore, each data point linked to the corresponding time are connected by straight lines as time point sorted in chronological sequence. The cumulative flowback volumes are recorded chronologically as well. The plots for flowback compositions change along with volume or time aim to observe the variation in various compositions at different wells. Those variation may introduce diverse possible hypothesis behind the piles of numbers. A momentous discovery from these chronologically created graphs is the reaction which might be occurred in reservoir conditions resulting in such magnitude of increment or decrement in specific compositions. The chemical reaction equations are then formulated to further explore how far the reaction goes by quantitative analysis on chemical equilibrium.

5.3 Linear-Regression

Regression analysis is a statistical methodology utilized frequently in distinguishing relations between two or more quantitative variables. The corresponding outcomes between variables form an organized and predictive tool so that approximate outputs could be generated by different entries of input without knowing the output through this tool (H.Kutner, Nachtsheim, Neter, & Li, 2005).

A linear regression model is said to be a simple where there is only one predictor variable and one control parameters, linear relation happens both in the parameter and in the predictor variable. In this study, the field data is plot in scatter and the linear fit model is built up based on these scatters. In order to find a relative accurate linear fit and the best of estimators of parameters β_0 and β_1, the method of least squares is employed and performed appropriately.

For each observation set (X_i, Y_i), the least squares is defined as the deviation from Y_i (Kutner, Nachtsheim, Neter, & Li, 2005)

$$[Y_i - (\beta_0 + \beta_1 X_1)] \qquad \textit{Equation 5-1}$$

$$\beta_0 = intercept$$

$$\beta_1 = slope$$

Particularly, an n number of data points would require a consideration of the sum of the n squared deviations.

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \qquad \textit{Equation 5-2}$$

The ultimate aim is to find out the best estimator β_0 and β_1 which are denoted as b_0 and b_1 to minimized the total deviation Q. By utilizing the analytical approach,

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})} \qquad \textit{Equation 5-3}$$

$$b_0 = \frac{1}{n}(\sum Y_i - b_1 \sum X_i) = \bar{Y} - b_1 \bar{X} \qquad \textit{Equation 5-4}$$

Where $\bar{X}, \bar{Y}$ are the means of the X_i, Y_i observations. Once the estimators are derived, the model is then evaluated with its accuracy.The estimated function is $\hat{Y} = b_0 + b_1 X$ where $\hat{Y}$ is the

value of the estimated regression function at the level X of the predictor variable.The mean square value is the sum of squares divided by the degrees of freedom.

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} \qquad \textit{Equation 5-5}$$

The root mean squared error can be denoted as $s = \sqrt{MSE}$ (Kutner, Nachtsheim, Neter, & Li, Chapter 1 Linear Regression With One Predictor Variable, 2005)

The confidence bounds on coefficients is well defined as followed:

$$C = b \pm t\sqrt{S} \qquad \textit{Equation 5-6}$$

Where b are the coefficients derived previously $[b_0, b_1]$, t depends on the confidence level and is computed using the inverse of Student's t cumulative distribution function. And S is a vector of the diagonal elements estimated from covariance matrix of the coefficient estimators, in linear fit, X is the design matrix which could be denoted as $[X_i\ X_{i+1} \ldots X_n]$

$$S = (X^T X)^{-1} s^2 \qquad \textit{Equation 5-7}$$

The confidence bounds are set as 95% in default but could be also specified with any certainty such as 90%, 99.9% and so on. It would perhaps take a 5% chance to take risk in being incorrect about predicting your outcome. Then, a 95% prediction interval would be selected and that interval indicates a 95% chance that new observation is actually enveloped within the lower and upper prediction bounds (Confidence Bound, 2015).

Sometimes, the anomalies would distort the authenticity of the correlations, those anomalies may be caused by system errors or random errors. In order to have a relative general model to describe the discovery, the anomalies must be segregated out of the observations. Cook's

distance is useful for identifying outliers or data out of trend in the X matrix or arrays (observations for predictor variables). Once the outlier is distinguished, the model would be reconstructed excluding the outliers. Cook's distance is the scaled change in fitted values which the original fitting model must be existed prior. It is defined as:

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{p\ MSE} \qquad \textit{Equation 5-8}$$

where

- $\hat{y}_j$ is the jth fitted response value.
- $\hat{y}_{j(i)}$ is the jth fitted response value, where the fit exclude observation i.
- MSE is the mean squared error.
- p is the number of coefficients in the regression model.

In the program, an observation with cook's distance larger than 3 times the mean Cook's distance might be considered as an outlier in default shown in Fig.5-1 (Cook's Distance, 2015).

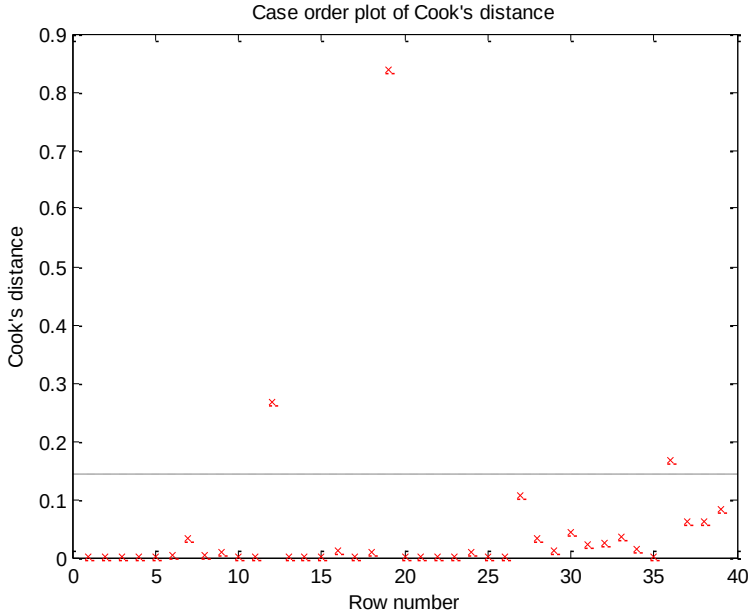


Figure 5-1 Case order plot of Cook's distance of one example in linear regression model, the threshold of Cook's distance to identify outliers in default is set as larger than 3 times the mean Cook's distance.

Cook's distance is always functional when the anomalies take a little part of the whole data pool, however, the outliers in some occasions may occupy up to 15% of the dataset which deteriorate the regression fit. Only small portion of the anomalies could be eliminated by cook's distance, therefore, the rest would still devastate the accuracy of the fitting process. Therefore, the combination in the application of cook's distance and residual check is introduced. It is critical to minimize the error in fitting but also to control the authenticity of the original data, thus by controlling the number of outliers eliminated increasing the accuracy of the regression model is the key in this treatment.

Residual check are useful to detecting the hypothetical fitting model with respect to the error term. Three different types of residuals is calculated in the program, each method is presented as below:

Raw Residuals

$$r_i = y_i - \hat{y}_i \qquad \textit{Equation 5-9}$$

Where r_i is the residual of the ith term, $y_i \; and \; \hat{y}_i$ are the corresponding observed and fitted values.

Pearson Residuals

$$pr_i = \frac{r_i}{\sqrt{MSE}} \qquad \textit{Equation 5-10}$$

Standardized Residuals

$$sr_i = \frac{r_i}{\sqrt{MSE_i(1-h_{ii})}} \qquad \textit{Equation 5-11}$$

Where MSE_i is the mean squared error of the regression fit calculated by removing observation i and h_{ii} is the leverage value for observation i. The leverage value is the provided by hat matrix

$$H = X(X^T X)^{-1} X^T \qquad \textit{Equation 5-12}$$

Where the diagonal element of H, h_{ii} satisfy $0 \leq h_{ii} \leq 1, \sum_{i=1}^{n} h_{ii} = p$; p is the number of coefficients and n is the number of observations (row of X) in the regression model (Residuals, 2015).

However, only raw residuals is adopted to perform the evaluation in this study. The histogram shown in Fig.5-2 would give an indication of how good the regression model is fitted. The more residual frequency histogram is approaching to the normal distribution, the one could expect the errors to be independently distributed in which the regression model would acquire more possibilities to account for this dataset. Huge departures usually allude any structures contained in the residuals that is not sufficiently interpreted by the regression model.

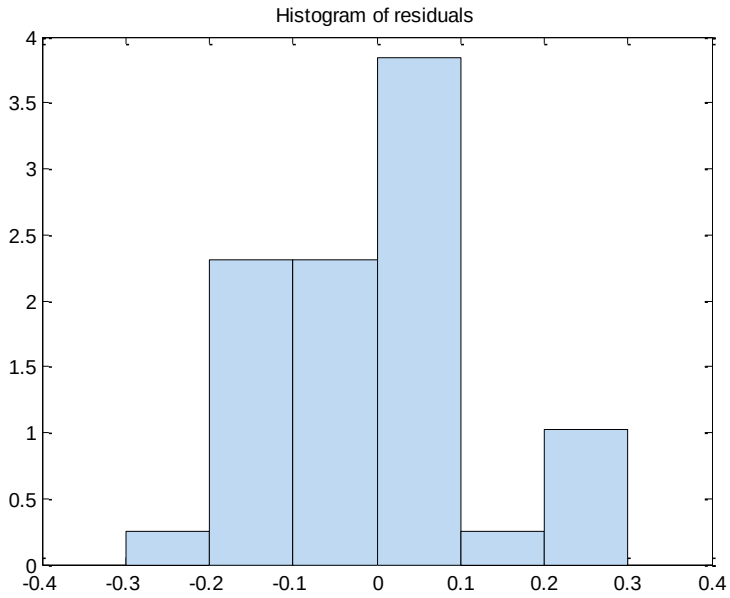


Figure 5-2 Histogram of residuals of one example in fitting linear regression model.

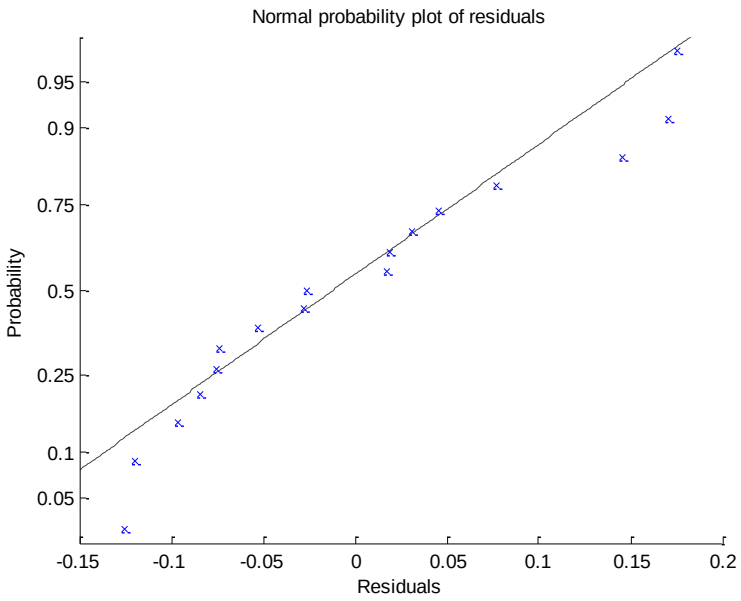


Figure 5-3 Normal probability plots of residuals as one example in a linear regression model

Another plot for examining the residuals is the normal probability plots Fig.5-3 in which it is useful when sample sizes of residuals are relatively small. By sorting the residuals into ascending order and then the cumulative probability of each residual is calculated through the equation:

$$P_i = \frac{i}{N+1} \qquad \textit{Equation 5-13}$$

With P_i on behalf of cumulative probability at the ith ascending order. And N is the total number of entries in the list. The scatter points should be lying on or in the vicinity of an approximately straight line only if the points possess a normal distribution property (Normal Probability of Residuals, 2015).

5.4 Clustering and Multivariate Approach

The representative-based clustering are popular in data mining which is regarded as a partitioning method that the given n number of observations in d-dimensional space $(x_1, x_2, x_3 \dots x_n)$ are aimed to be divided into k clusters. Each cluster is assigned as $C =$

$\{C_1, C_2, \dots, C_k\}$ and the mean or so-called centroid μ_i of all points in each cluster is specified as the representative to summarize the cluser (Mitchell, 1997):

$$\mu_i = \frac{1}{n_i}\sum_{x_j \in C_i} x_j \qquad \textit{Equation 5-14}$$

Where n_i is the number of points in cluster C_i. One approach for representative-based clustering is introduced as K-mean described as followed. In order to evaluate the quality of the given clustering $C = \{C_1, C_2, \dots, C_k\}$, a sum of squared errors scoring function shown below is in necessity .

$$SSE(C) = \sum_{i=1}^{k}\sum_{x_j \in C_i}\left\|x_j - \mu_i\right\|^2 \qquad \textit{Equation 5-15}$$

The goal is to find the clustering that SSE score is minimized:

$$C^* = arg\min_{c}\{SSE(C)\} \qquad \textit{Equation 5-16}$$

K-means randomly initialize k centroids in the data space. This is typically done by generating a value uniformly at random within the range of each dimension. Considering the traditional iteration includes two steps in K-means algorithm: (1) cluster assignment, and (2) centroid update. As k cluster means given as mentioned previously, simultaneously, each point $x_j \in D$ is allocated to the relatively closer mean point of which the certain cluster consist, with each cluster C_i enclsoing points that are closest to μ_i. That is, each point x_j is assigned to cluster C_j^*, where

$$j^* = arg\min^{k}\left\{\left\|x_j - \mu_i\right\|^2\right\} \qquad \textit{Equation 5-17}$$

Secondly in the centroid update step, clusters C_i, $i = 1, \dots, k$ with new mean values calculated for each cluster from the points in C_i updating along with the cluster assignment iteratively until the local minima is achieved. More specifically in our various cases, the number of k is known precedently, thus the centroids are the main focuses. The distances between centroids would be changed dramatically in different scenarios (Rokach & Maimon, 2015).

In statistics, principal component analysis (PCA) is probably the oldest and best known of the techniques of multivariate analysis. The main idea of PCA is to reduce the dimensionality of a dataset in which the interrelated variables are identified intrinsically by orthogonal transformation. Converting the possibly correlated variables in the observations into a new set of variables which is defined as the principal components that are uncorrelated while retaining as much as possible of the variation existed in the raw dataset. The principal components are yielded in order so that the first and second principal component has the top largest possible variance in which most of the variability in the raw dataset is described, sometimes the third one is also coupled with the first two (Jackson, 2003).

Consider a 95×13 matrix t, the row number 95 indicates that 19 wells consist of 5 time step for each well, Day 0, Day 1, Day 5, Day 14, Day90. The column number indicates 13 constituents of each well examined along the 5 differet times the 13 constituents are {Na, K, Ca, Cl, SO_4, Ba, B, Fe, Li, Sr, Mg, Br, Alk} and the units for each constituent are the same.

The procedure in conducting PCA is well exhibited as below:

The size of the matrix is recorded as $n = 95; m = 13$ where n represents the number of row and m represents the number of column. In order to summarize the raw dataset, the sample mean vector and sample standard deviation vector is determined by following equations.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} t_i \qquad \textit{Equation 5-18}$$

Where N is equivalent to n here which the mean of each column is computed and formulated as a vector. And then, based on this vector, the standard deviation can be calculated.

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}|t_i - \mu|^2} \qquad \textit{Equation 5-19}$$

Most often, the raw data should be standardized initially. Standardization or normalization is a process to alleviate the variance caused by rescaling, for example, $\mu g/L$ is transformed into mg/L, it may have impact on our principal components, the primary principal component may dominate over others as well. Standardize the variables by subtracting its mean from that variable and dividing it by its standard deviation.

$$Z_{ij} = \frac{t_{ij} - \bar{\mu}_j}{s_j} \qquad \textit{Equation 5-20}$$

Where t_{ij} is the variable in cell (i, j), $\bar{\mu}_j$ is the sample mean at jth coloumn and s_j is the sample standard deviation at jth coloumn (Jackson, 2003).

In computational terminology, the principal components are done by determining the eigenvectors or eigenvalues of the sample covariance matrix which is equivalent to finding the axis system where the covariance matrix is diagonal. The eigenvector comprising first two or three greatest eigenvalue are the direction of greatest variation that are also the axes orthogonal to each other. By assuming T is the normalization matrix of t, the eigenvalues of covariance matrix can be defined as:

$$determinant(cov(T) - \lambda I) = |(cov(T) - \lambda I)| = 0 \qquad \textit{Equation 5-21}$$

Where I is the identity matrix, in order to have this equation solved, the covariance matrix should be solved first.

$$Cov(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}(A_i - \mu_A) \times (B_i - \mu_B) \qquad \textit{Equation 5-22}$$

Where μ_A, μ_B are the meana of A and B respectively. Therefore in this study, the $Cov(T)$ would result in a diagonal matrix.

$$Cov(T) = \begin{bmatrix} cov(T_1,T_1) & \cdots & cov(T_1,T_{13}) \\ \vdots & \ddots & \vdots \\ cov(T_{13},T_1) & \cdots & cov(T_{13},T_{13}) \end{bmatrix} \qquad \textit{Equation 5-23}$$

For an instance, in our program, the command $[V,D] = eig(cov(T))$ returns two matrix V and D, the matrix V encompass the coefficients for the principal component but what is worthy to notice is that the order from this command turns out to be ascending which indicates that the largest three variances happen in our dataset is stored in the last three column of the V matrix. Matrix D's diagonal element storing the variance of the respective principal components that is also in an ascending order. The loadings indicated by arrays on the plot would be the last three column of the V matrix emanating from the origin.

Another graphical method presented in this study is the Piper diagram, it is widely used in the interpretation of water chemistry. The basic chemistry of flowback water samples were taken at the surface condition and examined with respect to presence of key inorganic constituents, for example, primary cations Na, K, Ca, Mg and chief anions $HCO_3^-, CO_3^{2-}, Cl^-$and SO_4^{2-} were basic element to construct the Piper trilinear diagrams. Two bottom ternary diagrams, the combination of elements on bottom triangles projected to a diamond-shaped quadrilateral at the top, comprise the Piper diagram. It has great advantage over other graphical plots like Stiff and star diagrams

that each observation on each time-series is shown as only one point in different color and shape. Therefore, the similarities and differences as well as the tendency between and along with tremendous amount of observations is readily to be identified with piper diagram. The relative abundant constituents are plotted in percentages which are transformed from mg/L (Me.E.Blauch, R.R.Myers, T.R.Moore, & B.A.Lipinski, 2009) (Helsel & Hirsch, 2002).

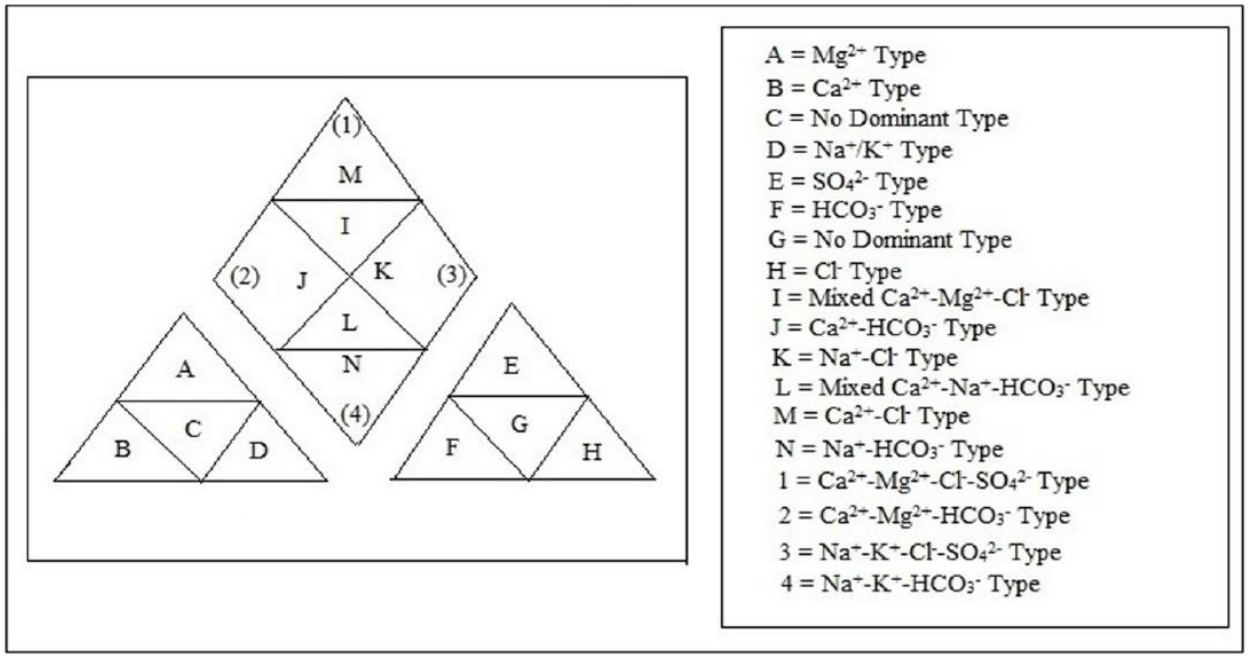


Figure 5-4 Piper trilinear diagram modified version (Manoj, Ghosh, & Padhy, 2013).

Chapter 6 Knowledge Discovery

In this chapter, geographical analysis on flowback compositions are firstly applied so as to determine whether geographical trend are existed. By comparing the late flowback water samples with oilfield brines in statistical manner, the similarity and imparity could be identified simultaneously. Simple plots among major compositions versus flowback sampling time or flowback volume are done, the variances occurred at different time are visualized to initiate different reaction mechanisms. Combined with each simple plots, diverse reaction mechanisms are examined thoroughly. Linear-Regression models compiled with clustering are adopted to proceed further investigation on distinguishing early-time flowback water and late-time flowback water. After this has been done, the multivariate approaches including Principal Component Analysis and Piper trilinear diagram are performed to better understand the variable importance and reduce the complexity of the correlations between different variables. Yet the Artificial Neural Network model is not sufficiently applicable to maximize the value in gas production industry, but it is still an extensively functional dataming tool and the development in gas production industry by using this tool shall be pervasive.

6.1 Geographical Analysis of Flowback Compositions

In order to survey sampled flowback waters distributed in the specific field that the comparison could be made, an integral map for each dataset are constructed. All 111 sampling

well locations including 40 samples of oilfield brines in our database are plotted on the Fig.4-5 which refers to the Table.4-1 in the Chapter 4 numerically displaying the location information of the wells. However, 6 wells, (#44, #47, #50, #56, #57 and #85) from BOGM dataset are excluded because the locations are not provided. Due to the difficulties to reconcile different sampling methods, the variance for each performance of the data still exists though field and laboratory analyses protocols are similar under the PA, DEP guidance.

Figure 6-1 a) All 111 sampling well locations in Pennsylvania.

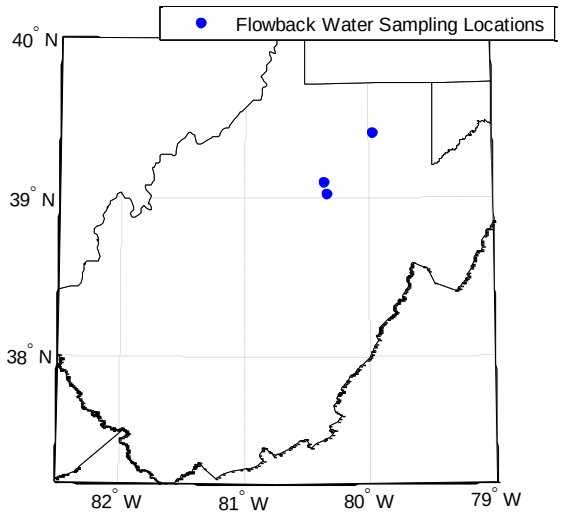


Figure 6-1 b) All 3 sampling well locations in West Virginia.

As Fig.6-1a) illustrated, most wells are located in upper northwestern part of Pennsylvania. Besides only three wells are located in West Virginia (Fig.6-1 b)). The indications of high salinity

in flowback water are commonly presented in all datasets respectively no matter the flowback is taken at any time or at any location. Additionally, neither significant reactions would cause a major variance on salinity concentration nor the concentration of influent would have such great impact on the flowback, thus the magnitude of high salinity could be an allusion on geological interpretation. Several figures are created to compare which are distinguished by time series.

Figure 6-2 a) Sodium concentration in 14 or 15-day flowback water for 15 wells vs. in in-situ brine for 33 wells.

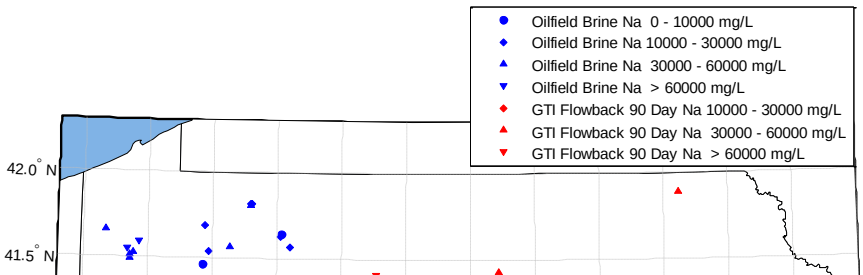


Figure 6-2 b) Sodium concentration in 90-day flowback water for 10 wells vs. in in-situ brine for 33 wells.

The Fig.6-2 a) and b) shown above delivers a critical information that the concentration of sodium from late time flowback are similar to the oilfiled brines. However, the initial point when the flowback become identical to the oilfield brines is still under investigation. And it is hard to conclude or to acquire the exact time line when two types of fluid diverse. The mixing procedure depends on different variables which are still complicated to explain. However, the results above illustrate that 14 or 15 days might be a possible entry the water produced from the formation becomes dominant. But there is still variance among distinct wells. Similar graphs are done for chloride as shown below in Fig.6-3 a) and b).

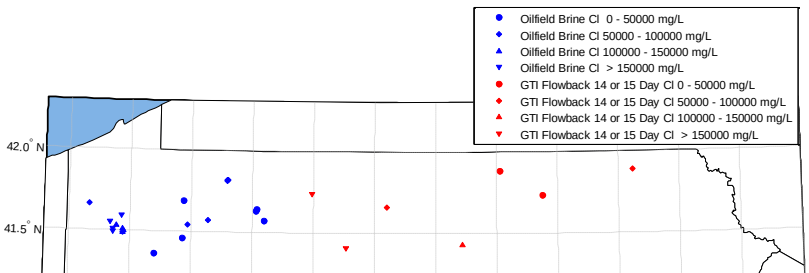


Figure 6-3 a) Chloride concentration in 14 or 15-day flowback water for 15 wells vs. in in-situ brine for 33 wells.

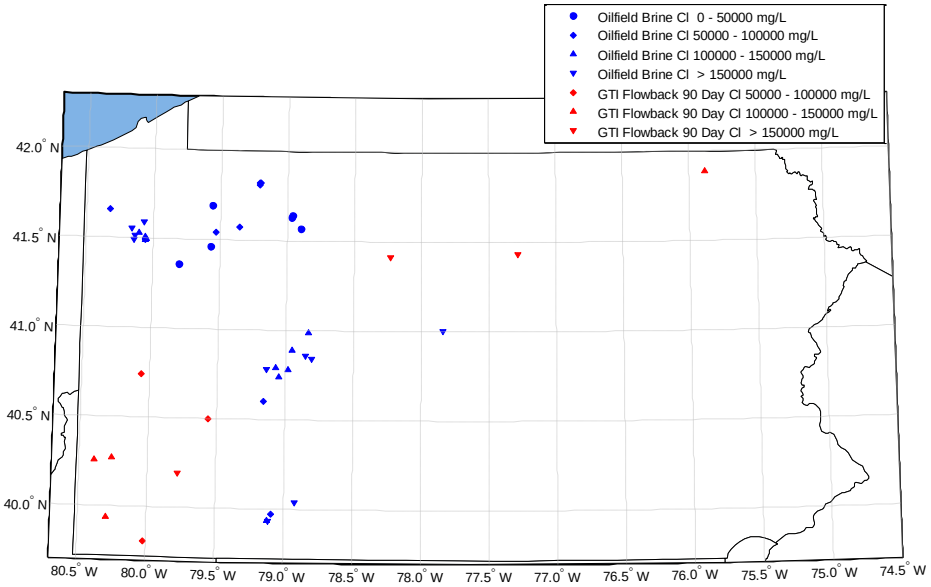


Figure 6-3 b) Chloride concentration in 90-day flowback water for 10 wells vs. in in-situ brine for 33 wells.

As the Fig.6-2 and Fig.6-3 shown, no prominent tendency based on geological graphs could be observed. However, one noteworthy variation along the time-series is identical for both sodium concentration and chloride concentration that the lower concentration accounted for the lowest proportion of total during the transient time from day 14 to day 90. This can be concluded via Fig.6-4 indicated by filled circle (for $\mathrm{Na} < 10^4\ ppm$ and for $\mathrm{Cl} < 5 \cdot 10^4\ ppm$). Another observation is the ratio of high concentration (shown as triangle and inversed triangle) over all are the dominant shown in Fig.6-4 which are also identical to oilfield brines. Similarly, in Fig.6-5, a

transferring tendency from low concentration of chloride to high level of chloride could be illustrated by comparing the ratios.

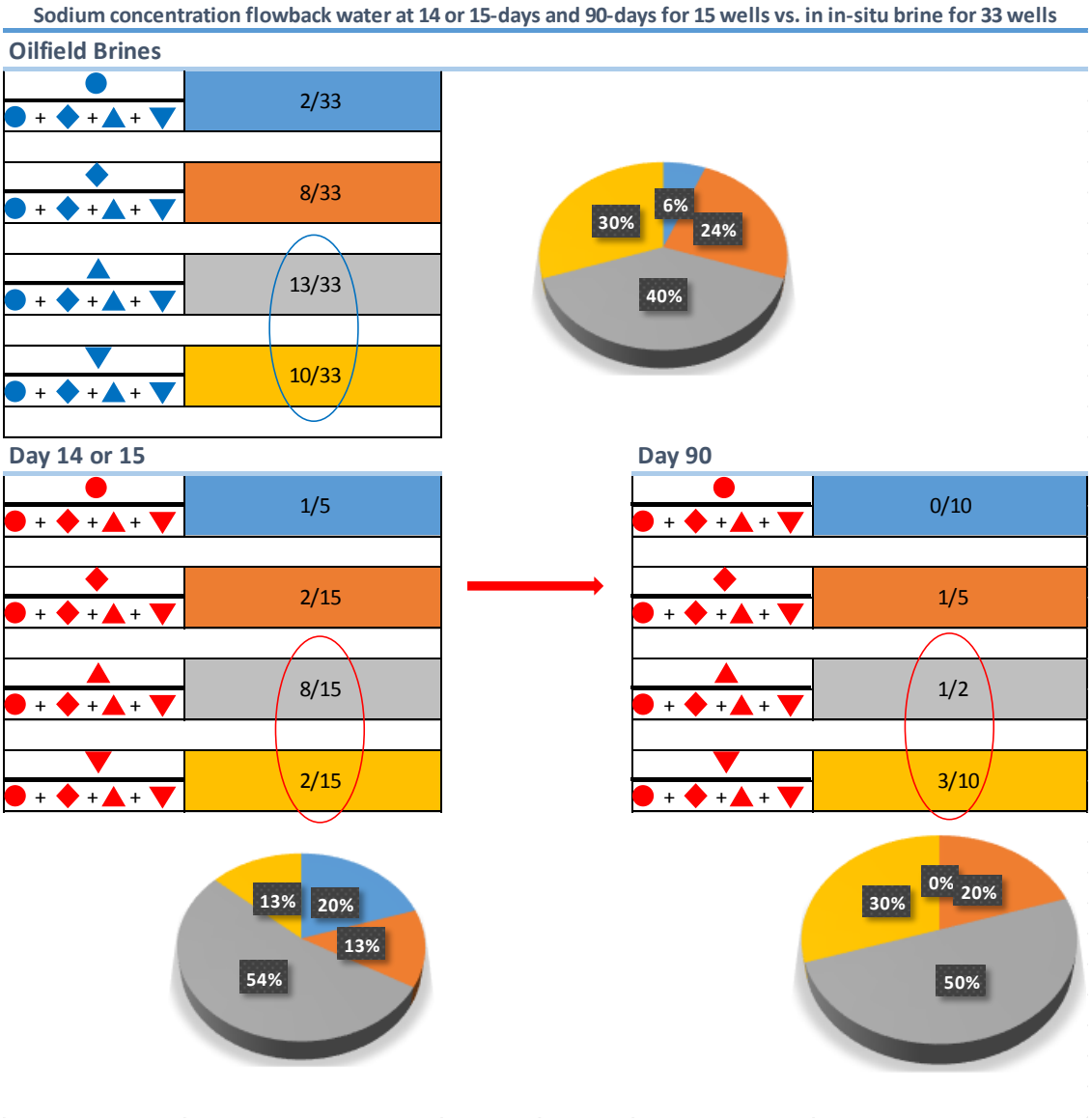


Figure 6-4 Pie chart indicating the proportion of each level of sodium concentration for 14 or 15-day flowback samples as well as 90-day flowback samples vs. the in-situ brines.

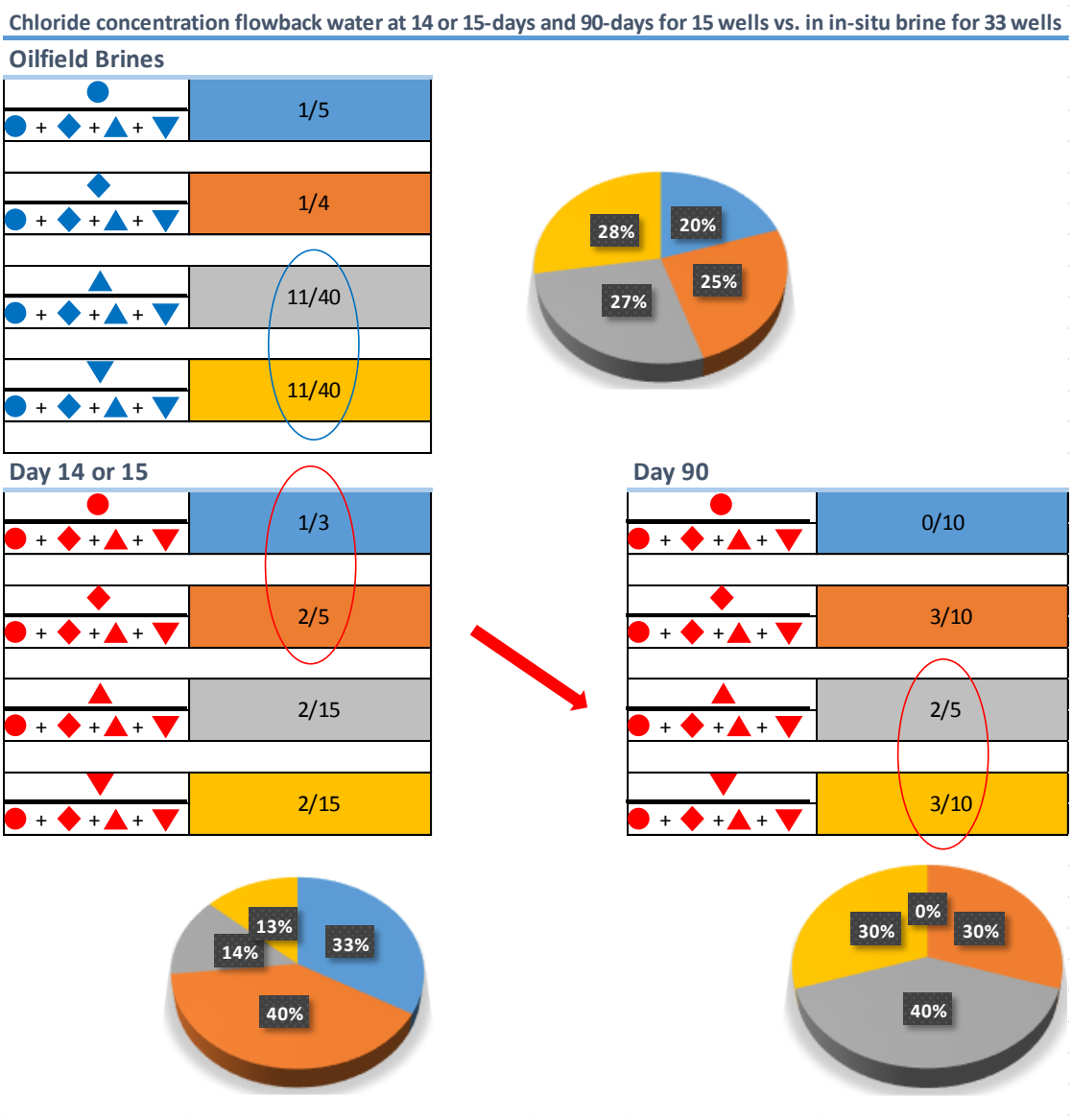


Figure 6-5 Pie chart indicating the proportion of each level of chloride concentration for 14 or 15-day flowback samples as well as 90-day flowback samples vs. the in-situ brines.

6.2 Analysis of Flowback Water Compositions Change with Volume/Time

Time-series dataset are valuable since the concentrations of each constituent are monitored with time. Without time, the discoveries of critical change, equilibrium or other significant information would remain unveiled.

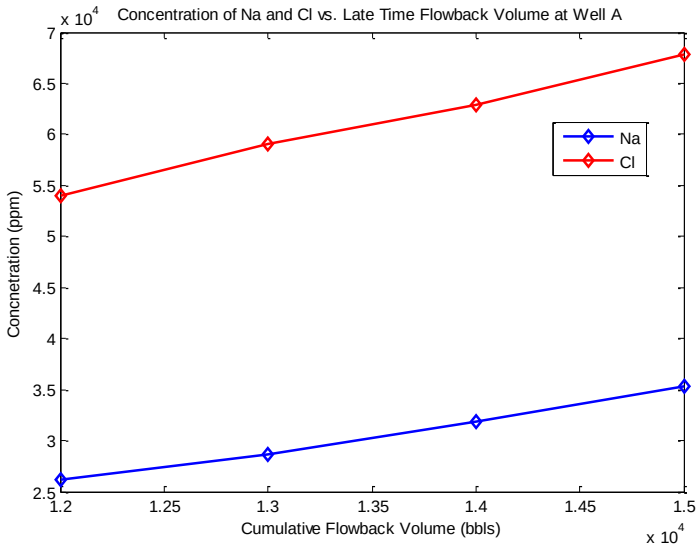


Figure 6-6 Concentration of sodium and chloride with the cumulative flowback volume for well #85 at the late time flowback period.

Fig.6-6 shows the Na and Cl concentrations in flowback water in southwestern Pennsylvania. Each point represents concentration at different cumulative flowback volume. The concentrations increase as flowback continues.

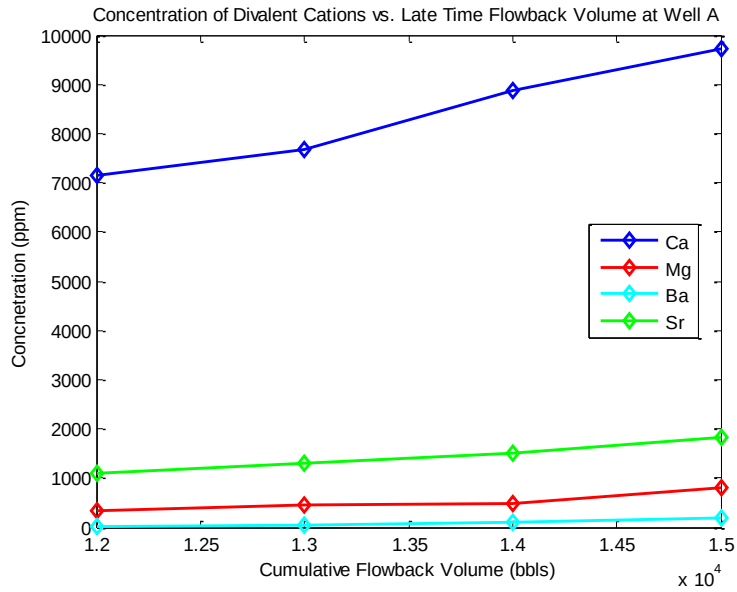


Figure 6-7 The variation of different divalent cations level along with cumulative flowback volume for well #85 at late time flowback period.

Simultaneously, the relatively high level divalent cations are screened out of the data pool and a similar graph is generated as shown in Fig.6-7. Each point indicates the concentration corresponding to distinguish cumulative flowback volume. The rising of concentrations as flowback cumulative collected could be observed. Furthermore, the calcium is the dominant divalent cation superior to others such as magnesium, barium and strontium. Literally, based the reactivity series for divalent element, barium performs more affiliation of metal. On the contrary, calcium behaves the weakest affiliation of metal. In this case, the high concentration of calcium gives an indication of high hardness of the water as the related carbonate radical are hypothetically not supposed to be abundant. Besides, the barium might be reacted and precipitated with sulfate, details of interpretation would be discussed in the following data pretreatments.

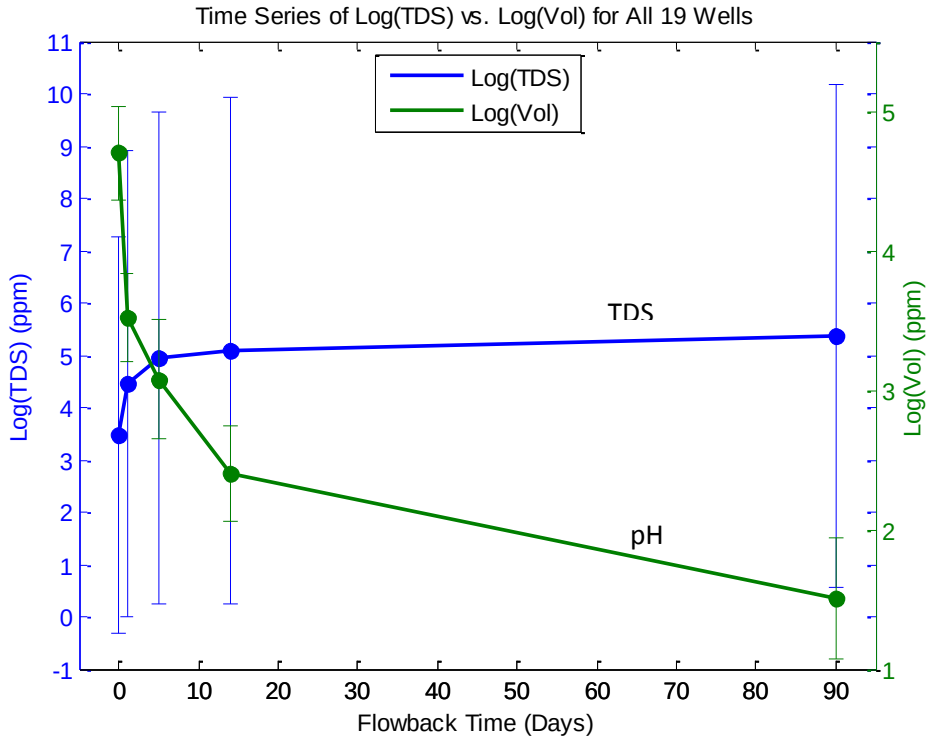


Figure 6-8 Average TDS and Volume showing maximum and minimum at each flowback time from well #1 to well #19 vs. time.

Undoubtedly, the flowback water rate is decreasing progressively and the total dissolved solid increasing speed reduces at the mean time shown in Fig. 6-8. The variation for both TDS and volume of flowback are quite unpredictable, but as all the value are averaged, the key tendencies are destined to have such comparison. One possible interpretation for such phenomenon is that with less residual flowback water remaining in the formation, the maximum capacity for dissolving the ions is gradually approached, therefore, at initial flowback days, the increment of TDS is prominent and then periodically lowering the dissolution speed.

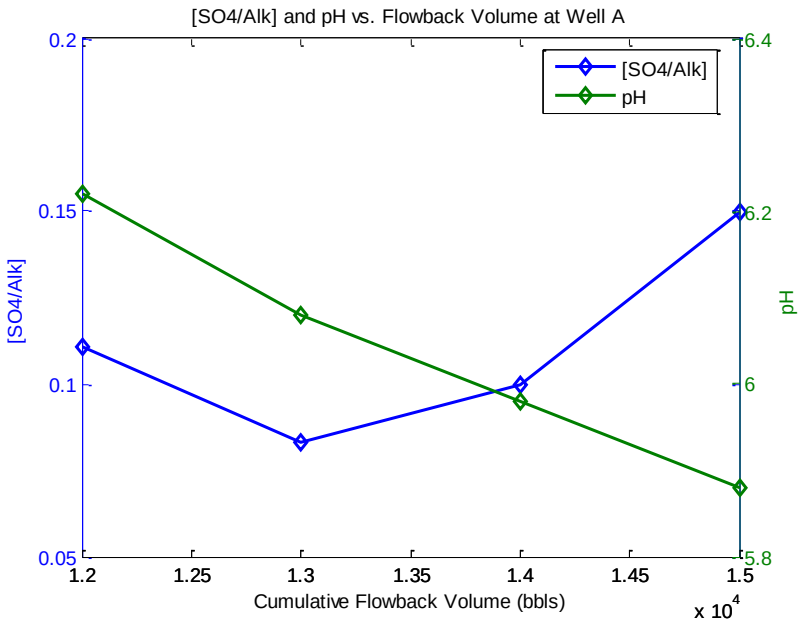


Figure 6-9 Ratio of sulfate over alkalinity and pH value with the cumulative flowback volume for well #85 at the late time flowback period.

Fig.6-9 shows a comparison between $\left[\frac{SO_4}{Alk}\right]$ and pH value vs. flowback volume. The ratio $\left[\frac{SO_4}{Alk}\right]$ at late time flowback is decreasing and then increasing gradually. On the other hand, the pH

value is decreasing all along with time passing by in the late time flowback. Note that the decline of ratio $\left[\frac{SO_4}{Alk}\right]$ at the beginning would be possibly triggered by a precipitation procedure, with the flowback water becoming more acid, the bicarbonate radical would possibly decomposed and as a consequence, the ratio $\left[\frac{SO_4}{Alk}\right]$ increased progressively.

The total dissolved solids (TDS) is missing so that the proportion of each ion possesses is unknown.

A similar graph is drawn below in Fig.6-9 to show how much chloride occupied in TDS. With cumulative flowback consecutively returned in proportion, the concentrations are ascending. The graph also shows chloride concentration takes up to over 50% of TDS concentration.

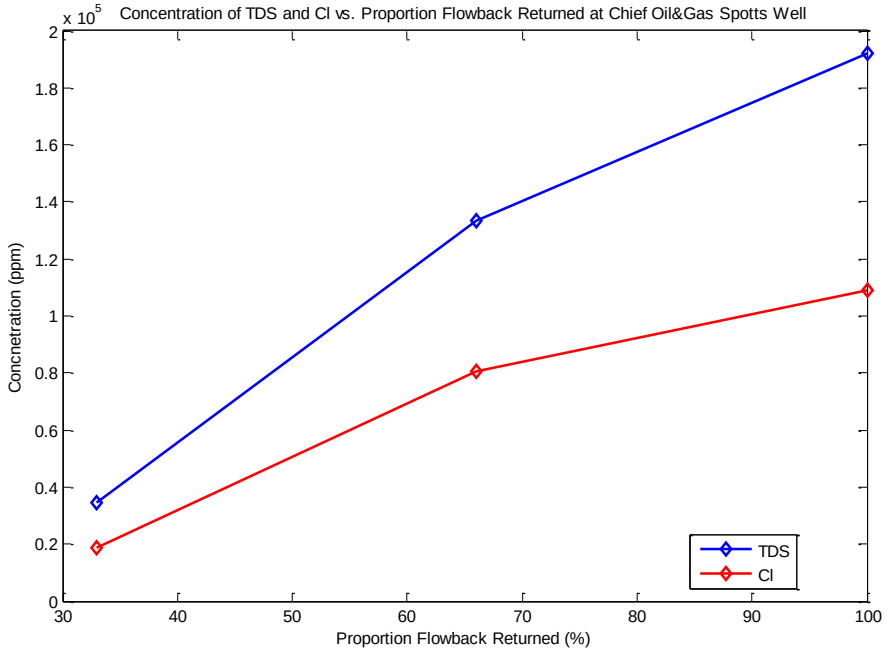


Figure 6-10 Correlation between TDS and the most abundant ion Cl with the cumulative percentage of flowback water returned to the surface for well #84.

Wells #1 to #19 have water samples at 0, 1, 5, 14 or 15, 90 days after a fracture treatment. The samples were collected by an independent contractor under the protocol enacted by state regulators and analyzed by a single laboratory for 45 inorganic constituents and about 200 organic constituents. However, in this study, only inorganic constituents are investigated and discussed.

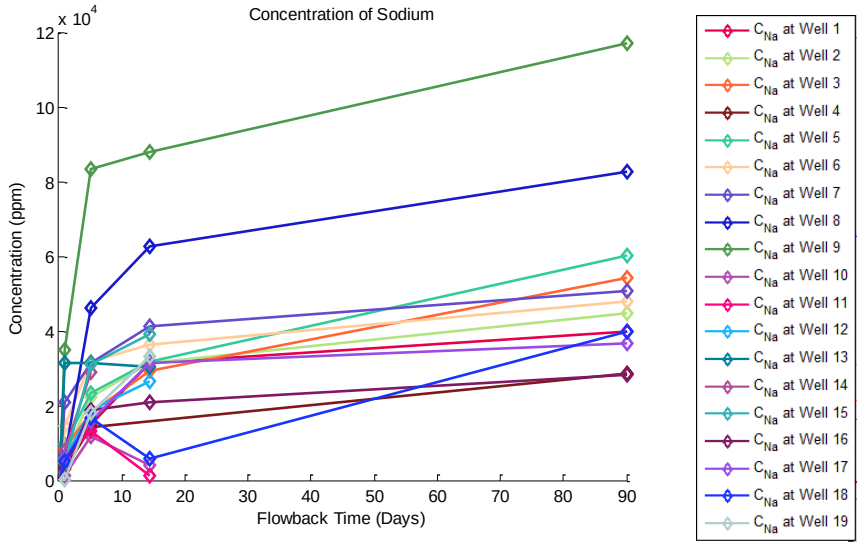


Figure 6-11 Sodium concentration changes along 90 flowback days from well #1 to #19.

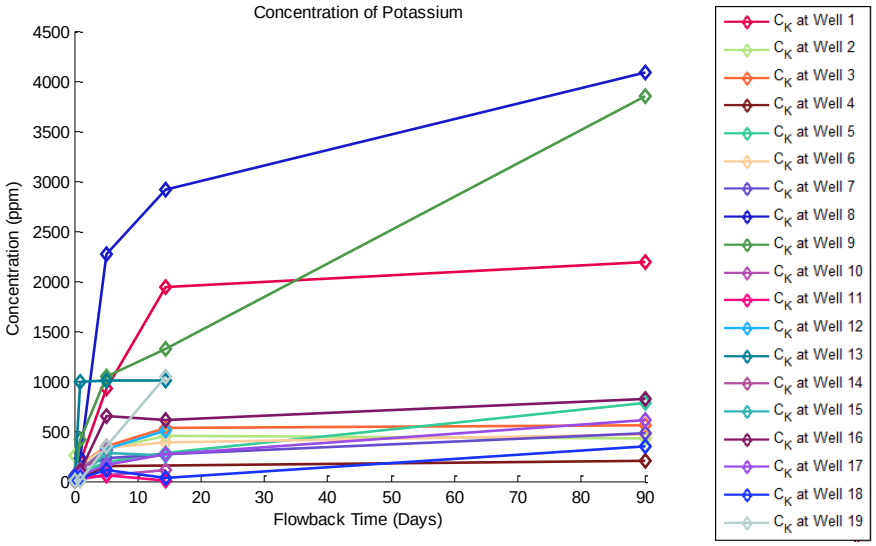


Figure 6-12 Potassium concentration changes along 90 flowback days from well #1 to #19.

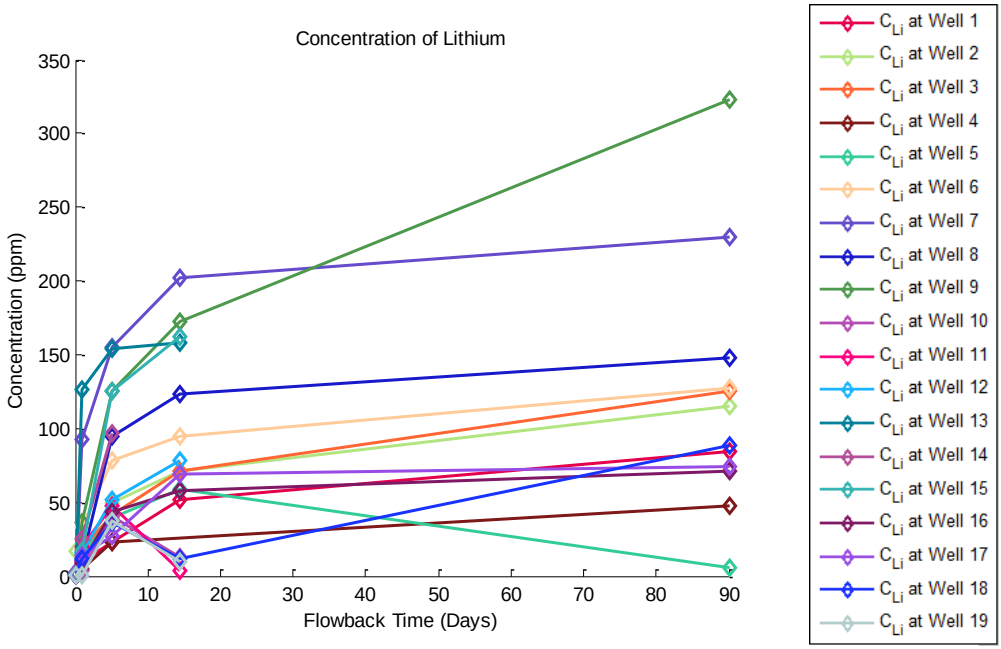


Figure 6-13 Lithium concentration changes along 90 flowback days from well #1 to #19.

According the Fig.6-11 – Fig.6-13, the concentration of sodium, potassium and lithium are plotted along with the sampling date in Cartesian coordinates. Samples collected from different wells are illustrated by different colors and the color for each well is consistent for the following graphs. The major monovalent anion "sodium" could be obviously distinguished from the magnitude of y – axis. Tables of the raw data for each graphs including following graphs will be presented in the appendix. Fig.6-11 exhibits the Na content of flowback water from 19 wells, most of the concentration of Na tends to elevate intensively at the beginning and then with time extending, the rate of the increment mitigate gradually. Nonetheless, 4 wells out of 19 wells are spotted behaves that they encounter a relatively small decrease from the day 5 to the day 14 or 15. The explanation for these specific phenomenon is sophisticated, the mild decrease may correlate to the permeability and the formation heterogeneity or errors in sampling. But all in all, there must be a low concentration source contacting with the high concentration flowback so that the reduction of the concentration could be triggered by dilution process. Same events could be

identified on lithium and potassium in Fig.6-12 and Fig.6-13. The potassium chloride used to add in as a constituent in slick water pumped into the well cracking the formation, it is functional to control the clay swelling and avoid plugging the pores. Thus, the initial concentration of potassium chloride is crucial for itself in the flowback water.

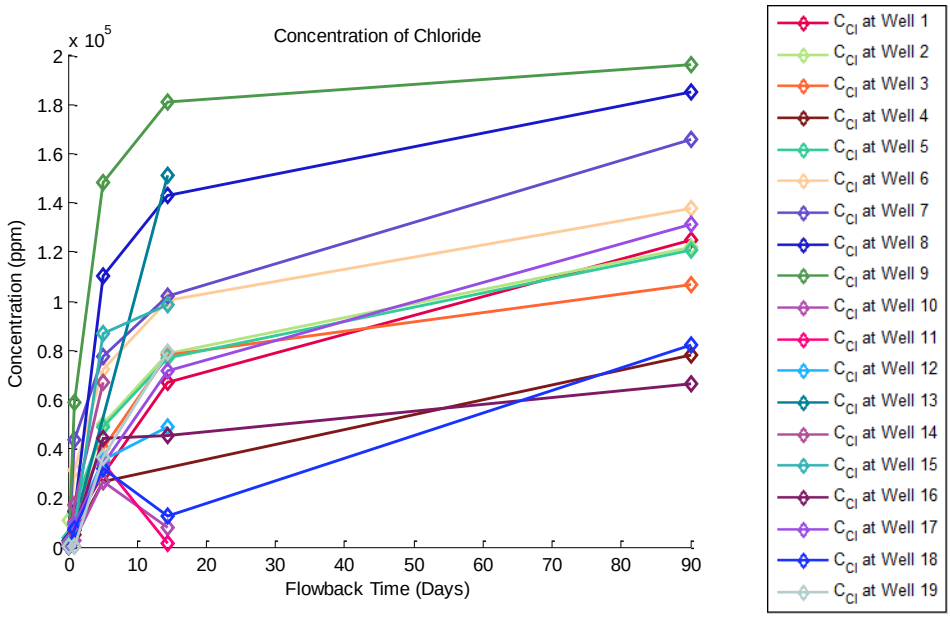


Figure 6-14 Chloride concentration changes along 90 flowback days from well #1 to #19.

As the statement declared above, Fig.6-14 is done to represent the Cl changes in time so that the origin of chloride could be differentiated. It is affirmative that the Cl take over the most occupancy in anions to balance the cations. Similarly, the concentration increases steeply at the beginning and then gently afterwards and several wells indicate a drop in contents. The level of Cl is 10 times to Na, therefore, the divalent cations hereby is considerable.

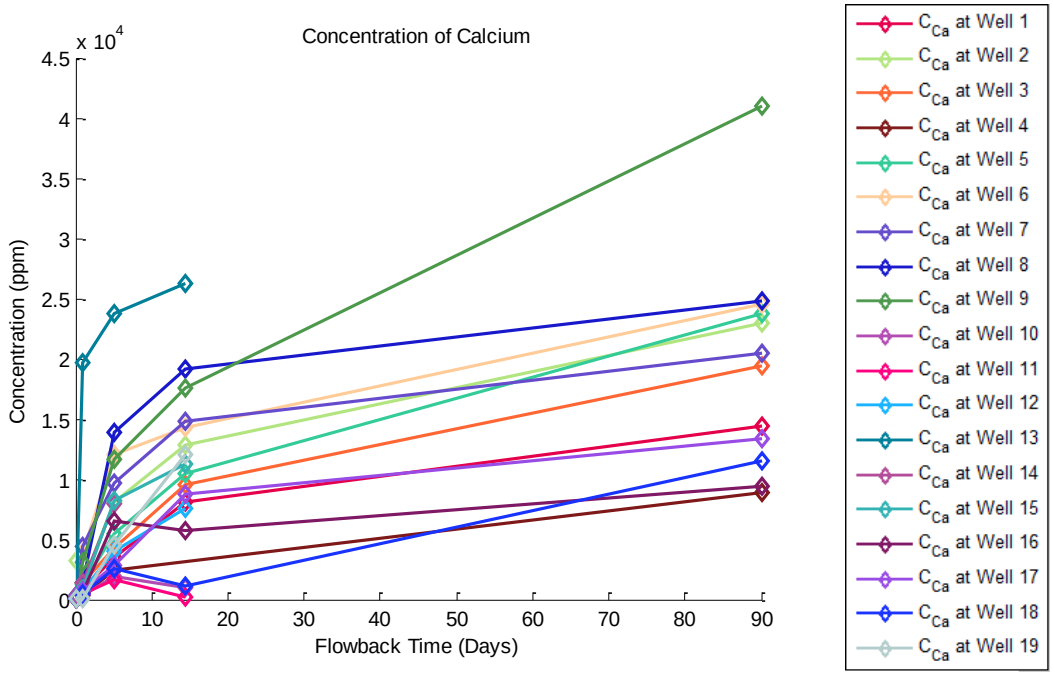


Figure 6-15 Calcium concentration changes along 90 flowback days from well #1 to #19.

The calcium （Fig.6-15） dominates the divalent cations content which the magnitude of concentration reaches 10^4 ppm. The barium (Fig.6-16) and strontium (Fig.6-17) as well as magnesium (Fig.6-18) achieve the equivalent concentration despite only one well does the barium reach the unusual high level. The colors correlated with well numbers help to cross out the most frequent event of a decrement from day 5 to day 14 or 15 including well #11, #17 and #18. The consequence for this phenomenon might be caused by systematic errors or random errors. It is unlikely to conclude any significant reactions for these divalent cations, nevertheless, the possibility of precipitation that sulfate bonded with barium separated out of the solution through crystallization with hydration is highly considered. The concentration of sulfate is then evaluated.

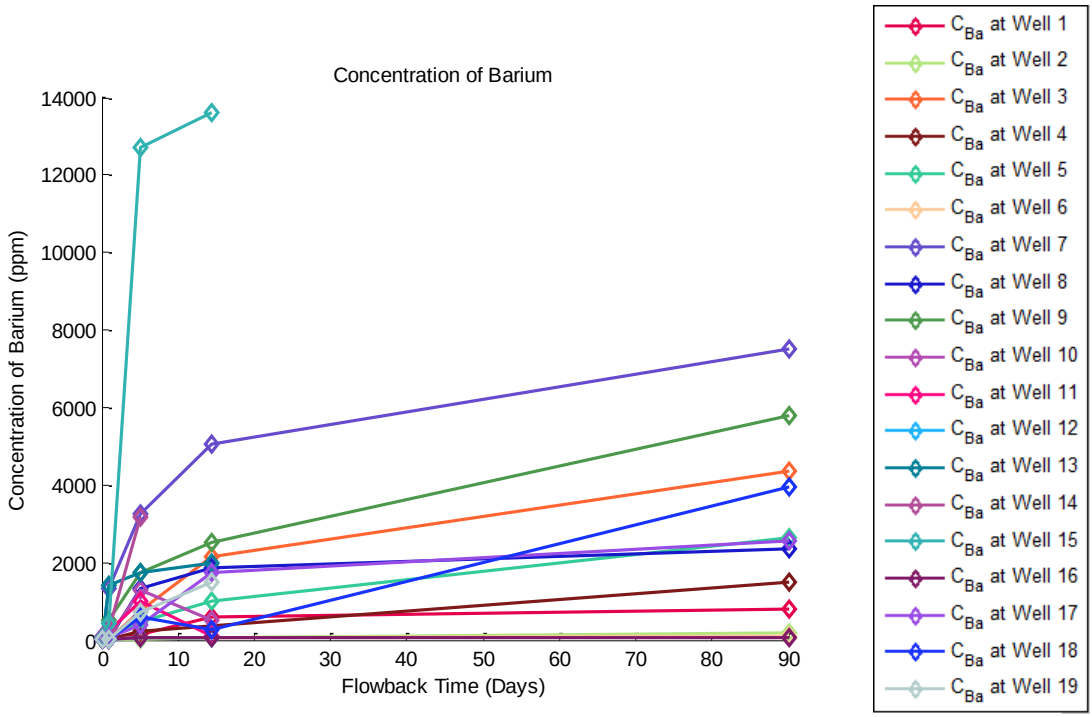


Figure 6-16 Barium concentration changes along 90 flowback days from well #1 to #19.

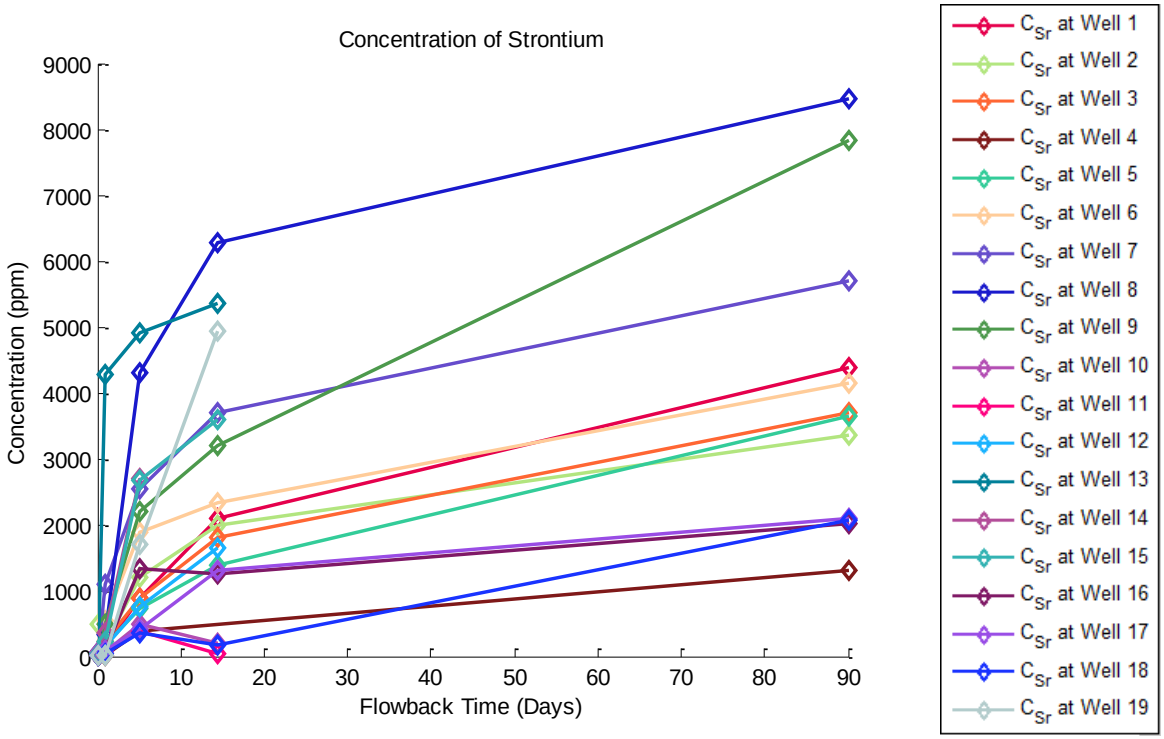


Figure 6-17 Strontium concentration changes along 90 flowback days from well #1 to #19.

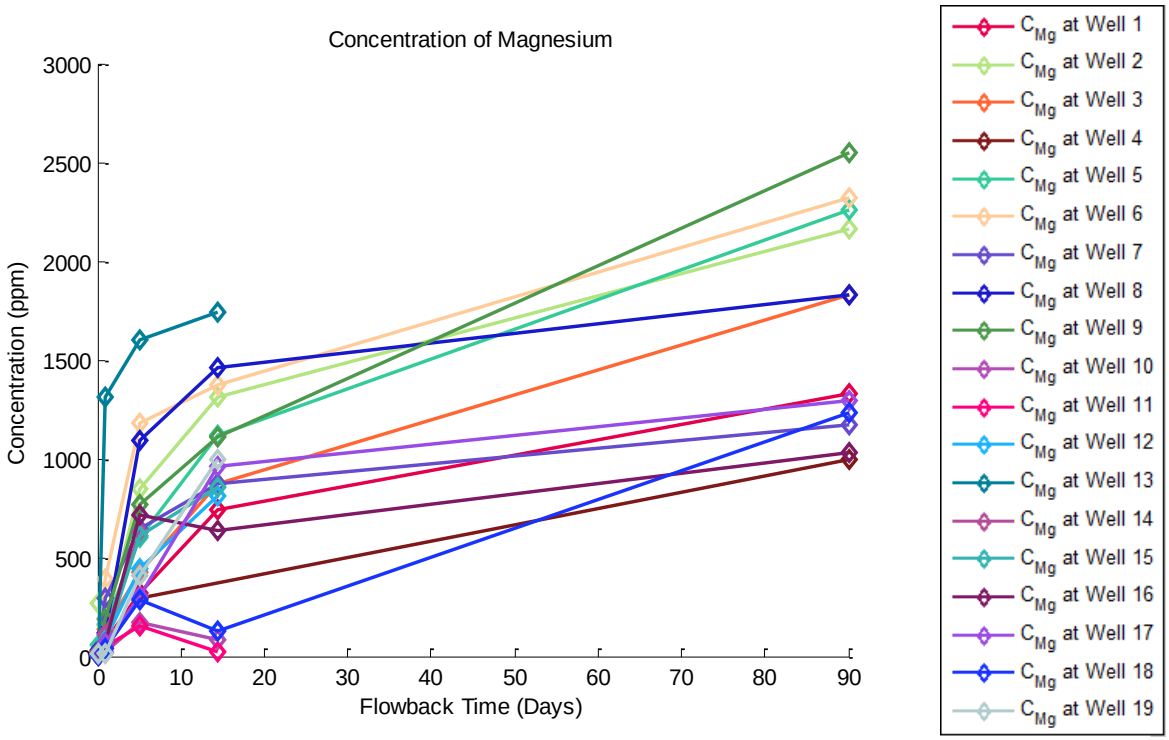


Figure 6-18 Magnesium concentration changes along 90 flowback days from well #1 to #19.

The evaluation of sulfate is done in Fig.6-19 by using semi-log plot, the variation can be clearly observed. The initial content of sulfate in the injection frac water at each well sites is highly variable. However, the concentration of sulfate seems to be not stable for which the dissolution process and reaction process may be incorporated simultaneously. A couple of possible reactions are enumerated based on different hypothesis. The lack of other information such as the exposure to the air or the temperature of sample and so on impose the difficulties on ascertaining the hypothesis.

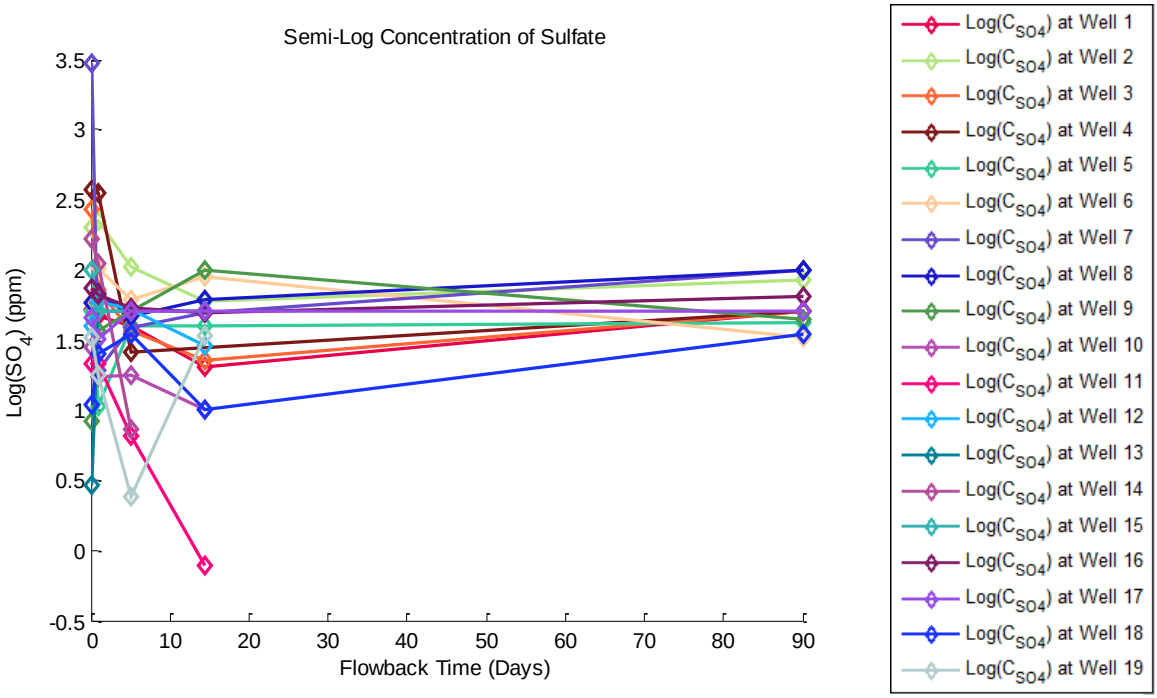


Figure 6-19 Sulfate concentration changes along 90 flowback days from well #1 to #19.

The possible dissolution and ionization reactions:

$$MgSO_4 \rightleftharpoons Mg^{2+} + SO_4^{2-}$$

$$CaSO_4 \rightleftharpoons Ca^{2+} + SO_4^{2-}$$

$$Ba^{2+} + SO_4^{2-} \rightleftharpoons BaSO_4 \downarrow$$

According to the low level of sulfate, the precipitation could be barely observed from the original database.Other reactions are also critical to be considered, the carbonate might reduce the concentration of carbon dioxide as the divalent cations are abundant and form the carbonate precipitates. Nevertheless, the hydrogen ion existed in the flowback would result in a re-dissolution of the precipitates back to the flowback water which may cause the concentration of both divalent cations and carbonate enriched. The origin of the hydrogen ion were likely coming from the sour gas CO_2 absorbed into the flowback water.

$$CaCO_3 + CO_2 + H_2O \rightleftharpoons Ca^{2+} + 2HCO_3^-$$

$$CO_2 + H_2O \rightleftharpoons H^+ + HCO_3^-$$

$$HCO_3^- \rightleftharpoons H^+ + CO_3^{2-}$$

The second ionization might not be significant as the first stage.

Therefore, the pH value which the H^+ concentration can be represented are evaluated along with the ratio of $[\frac{SO_4}{Alk}]$. The Fig.6-20 is shown below, the ratio $[\frac{SO_4}{Alk}]$ is taken as a log scale so that the variance could be easily identified, the highest and lowest value based on 19 wells are illustrated with a bar at the top and bottom therefore all the other values are within the bar range. Obviously, the pH value is increasing at the abrupt flowback between day 0 and day 1 and then decreasing gradually. Another result could be appraised accurately as the flowback remains its acidity after day 1 by decreasing the pH value. On the contrary, the ratio $[\frac{SO_4}{Alk}]$ at the initial flowback from day 0 and day 1 is decreasing instantaneously and then increasing progressively afterwards.

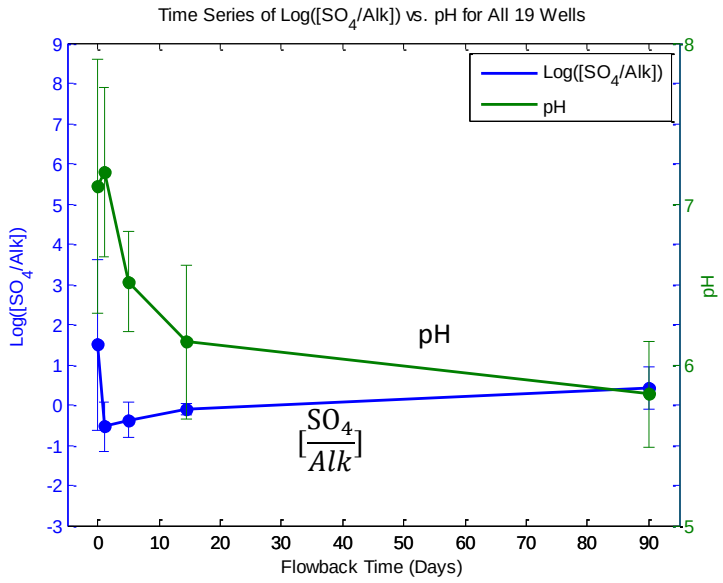


Figure 6-20 Average ratio of sulfate over alkalinity and pH value coupling with maximum and minimum at each sampling time vs. the fowback time for well #1 to #19.

After all these data pretreatments done, the TDS for 19 wells is constructed as Fig.6-21 illustrated. Most wells show an increasing tendency but several wells especially well 18 and well 11 are the frequently to be identified a small abatement on concentration of any constituents. Therefore, these two wells can be well defined as the low contents source had been breached and mixed with the flowback water to dilute the concentration. The possible sources could be the formation water or connate water.

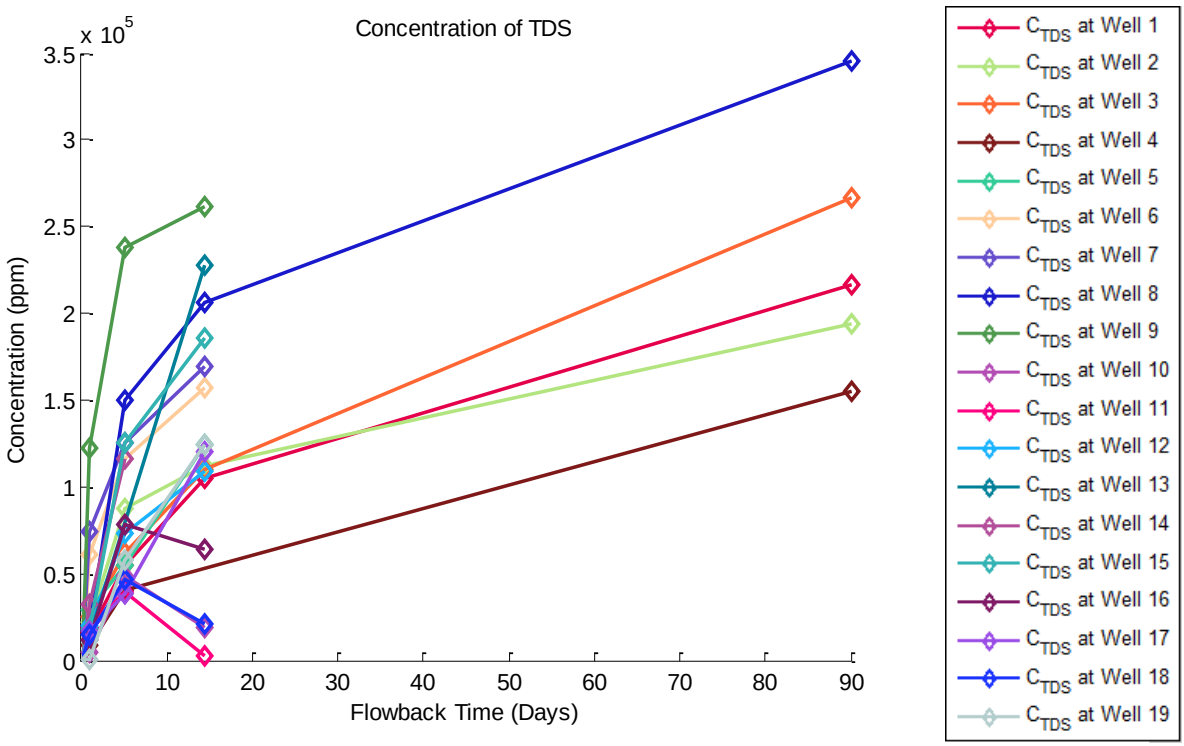


Figure 6-21 TDS concentration changes along 90 flowback days from well #1-19.

6.3 Analysis Using Linear Regression

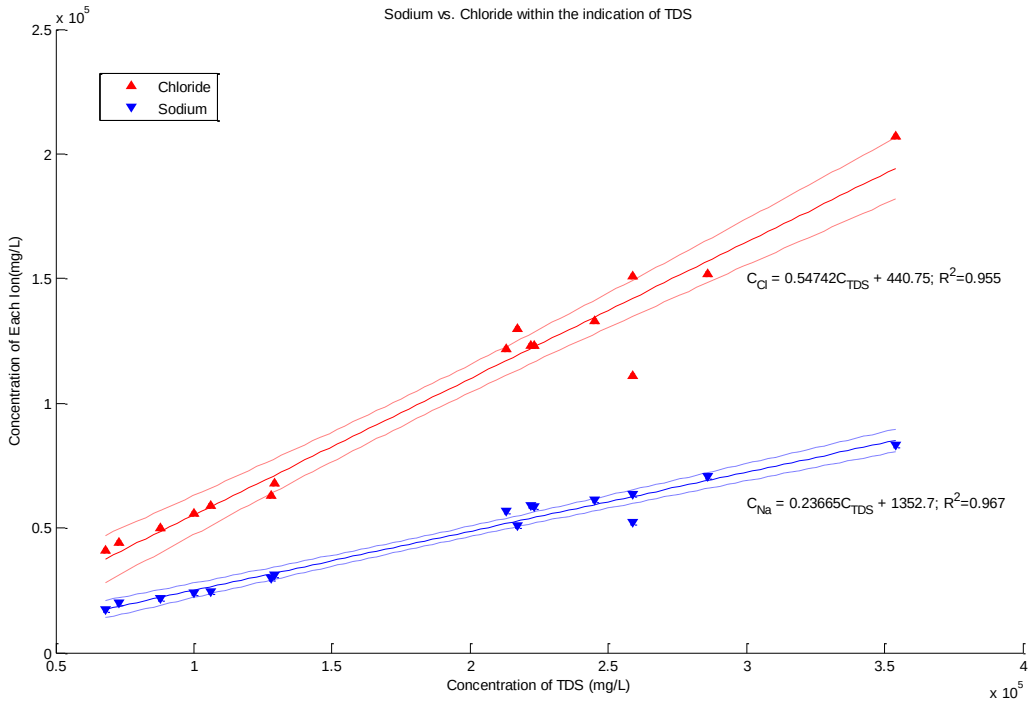


Figure 6-22 Linear-regression model for sodium chloride from oilfield brines for well #86 to #125.

According to the linear-regression model for chloride and sodium vs. TDS (Fig.6-22), the dissolution and dilution as well as the crystallization of both ions are linear correlated. In this oilfield brines, the retention time of formation water contacted with rock in the vicinity is assumed to be infinite long, and the saline source rock is conceived to be an infinite large reservoir of saline, the sodium is more linear correlated with TDS because the confidence boundary is more convergent than the chloride and the R-square is closer to approach to one.

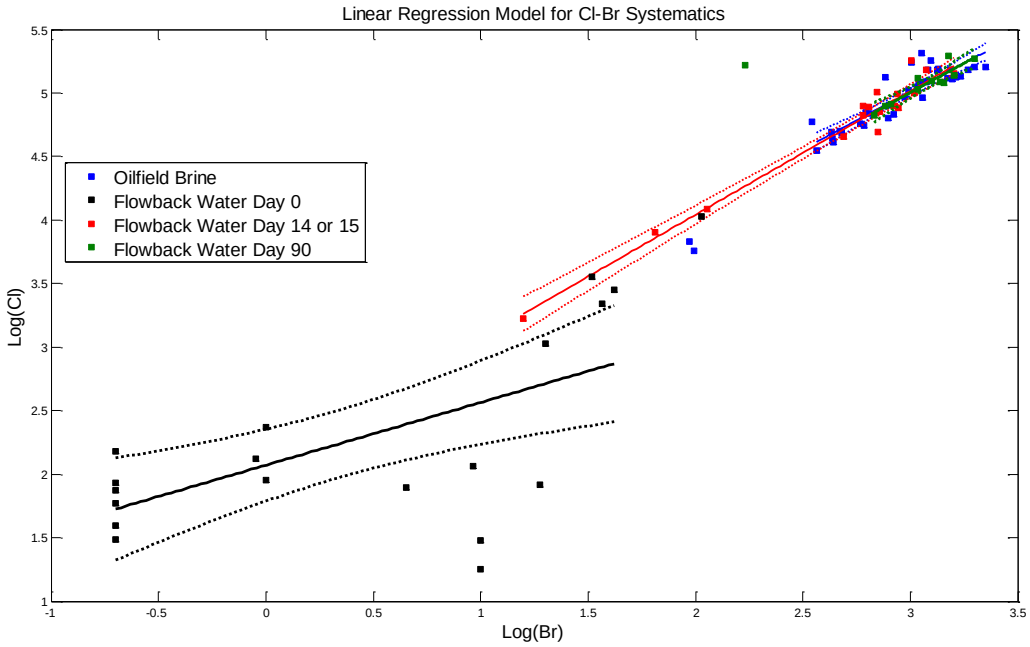


Figure 6-23 Linear-regression models for Cl-Br systematics on oilfield brines for well #86 to #125, injection fluid samples from day 0 and late time flowback water samples from well #1 to #19.

The late time flowback water containing Cl, Br are plotted along with Oilfield Brines. As illustrated in Fig.6-23, each color represent one species. The outliers are filtered and discarded by Cook's distance so that the linear regression model could be applied without huge errors generated. The results turn out that the Cl-Br systematics could be one representative indicating the late-time flowback compositions acquiring affiliation to the oilfield brines. A transition in linear regression from injection to late time flowback could be observed. However, other proofs should be discovered and discussed to backup this hypothesis.

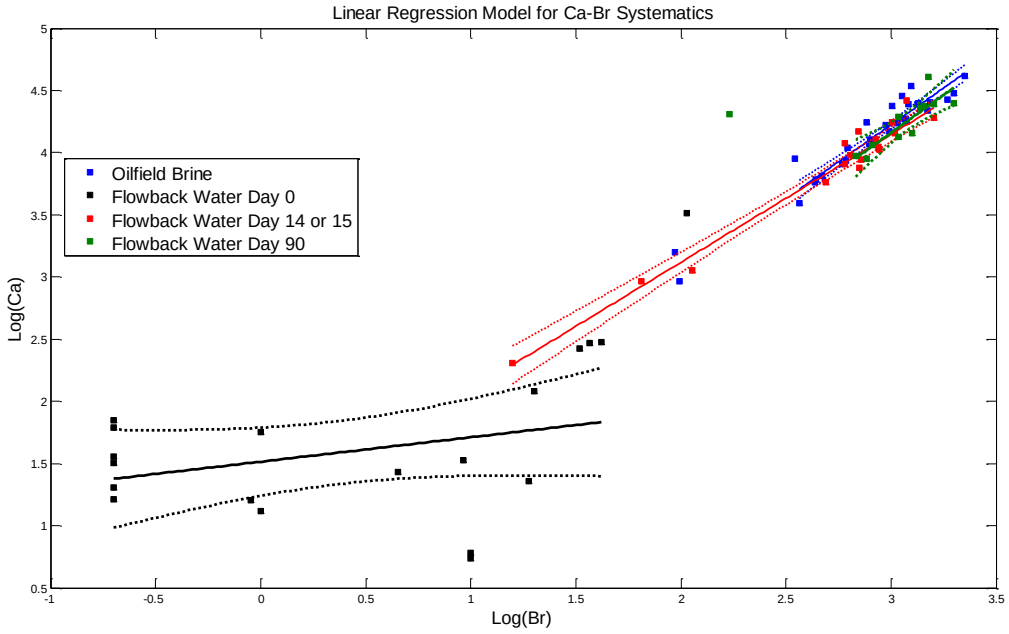


Figure 6-24 Linear-regression models for Ca-Br systematics on oilfield brines for well #86 to #125, injection fluid samples from day 0 late time flowback water samples from well #1 to #19.

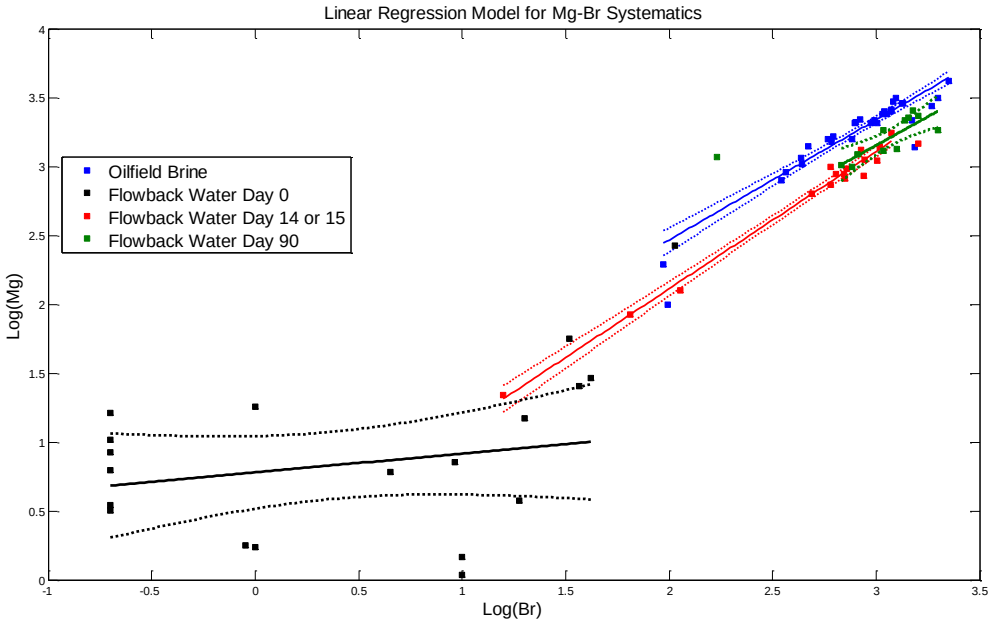


Figure 6-25 Linear-regression models for Mg-Br systematics on oilfield brines for well #86 to #125, injection fluid samples from day 0 late time flowback water samples from well #1 to #19.

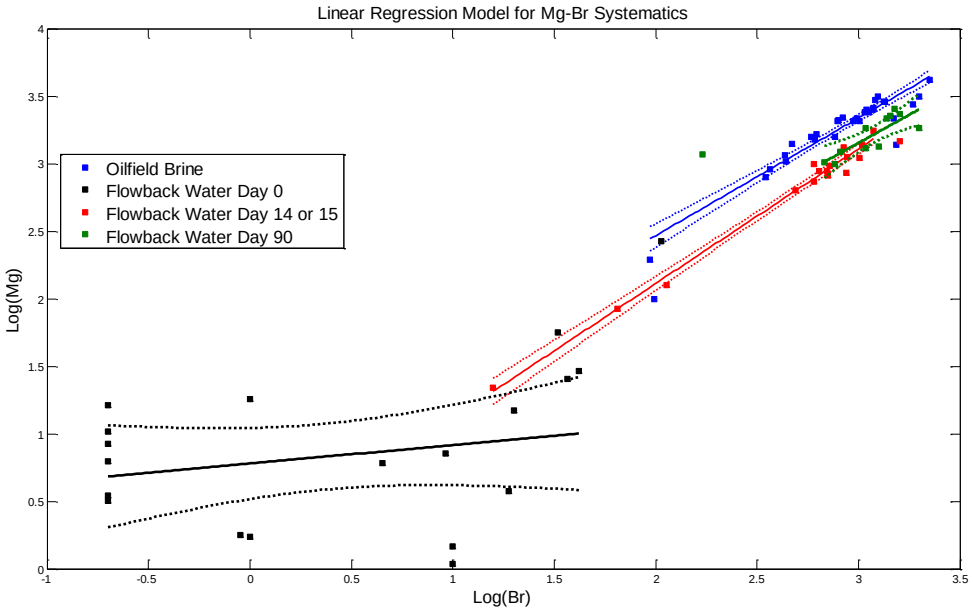


Figure 6-26 Linear-regression models for Na-Br systematics on oilfield brines for well #86 to #125 and late time flowback water samples #1 to #19.

Fig.6-24 – Fig.6-26 show various systematics including Ca-Br, Mg-Br, Na-Br. The categories are distinguished by colors and the linear regression models are created in the same pattern as it is done for Cl-Br systematics, the outliers are filtered by Cook's distance. According to the discrete data on oilfield brines, the data out of training are outcropped by residual check. As for the Ca-Br systematics, the late flowback waters behave a strong affiliation with the oilfield brines no matter at low or high concentration of Br. Furthermore, as for the Mg-Br systematics, the late flowback performs a relatively small affiliation at low concentration of Br, however, it is not sufficient for us to draw this consequence due to the scarcity of the data points obtained from the original report. On the other hand, the Na-Br systematics show the affiliation in the opposite way, the late flowbacks indicate affiliation at high Br contents.

The time-series dataset gives good indications that correlations between constituents could be identified by utilizing linear regression method. Nevertheless, the dataset without time dimension could still be analyzed under the guidance of these indications. The BOGM and Pritz&Kirby datasets are such datasets evaluated by linear regression method.

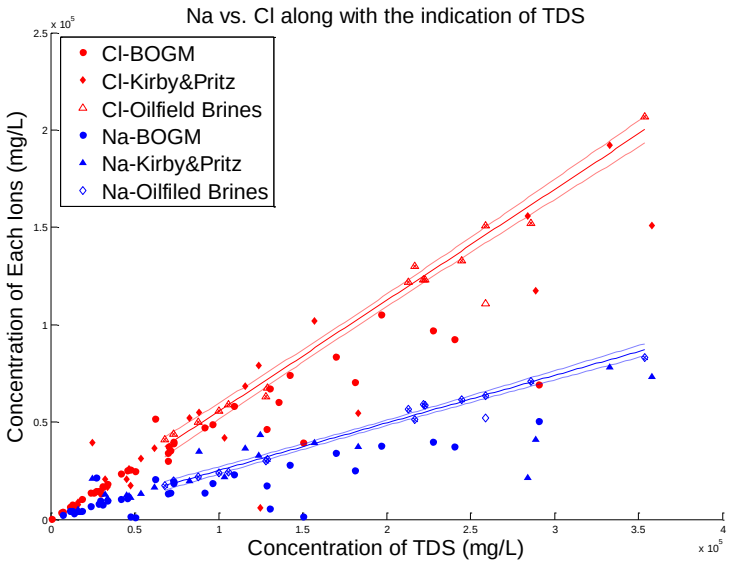


Figure 6-27 Comparison on Na-Cl systematics between oilfield brines from well #86 to #125 and flowback samples from well #20 to #83.

Based on the Fig.6-27 shown above, two linear regression models set up previously in Fig.6-22, are reproduced here and extended with the adjusted linear regression equations. The adjusted linear regression equations are re-written below with the help of Cook's distance.

$$C_{Cl} = 0.57012C_{TDS} - 1579.4, R^2 = 0.988$$

$$C_{Na} = 0.24432C_{TDS} + 670.4, R^2 = 0.987$$

Comparing with the linear equation obtained before, the adjusted R^2 of the fitting models are closer approaching to one. By extending the linear fitting regression line, the flowback water compositions are well fitted at low concentration of TDS in which a possible speculation that low TDS flowback water performs more likely like Oilfield brines however this conjecture should be reevaluated with other constituents. Correspondingly, at high TDS concentration, the flowback water data in BOGM and Kirby&Pritz datasets are mostly out of trend (out of the 95% confidence boundaries). In this case, the flowback water are defined as "under-saturated", the hierarchy of saturation depends on the retention time. Oilfield brines have a relatively long time contacting with the formation which ion exchanges reach the equilibrium. The "under-saturated" flowbacks could be regarded as unequilibrium status therefore the engaging time is not enough resulting in insufficient ion exchanges.

6.4 Variable Importance Discovery by Cluster and Multivariate Approach

The theorem of K-means clustering has been explained in chapter 5, the algorithm is applied to well #1 to #19.

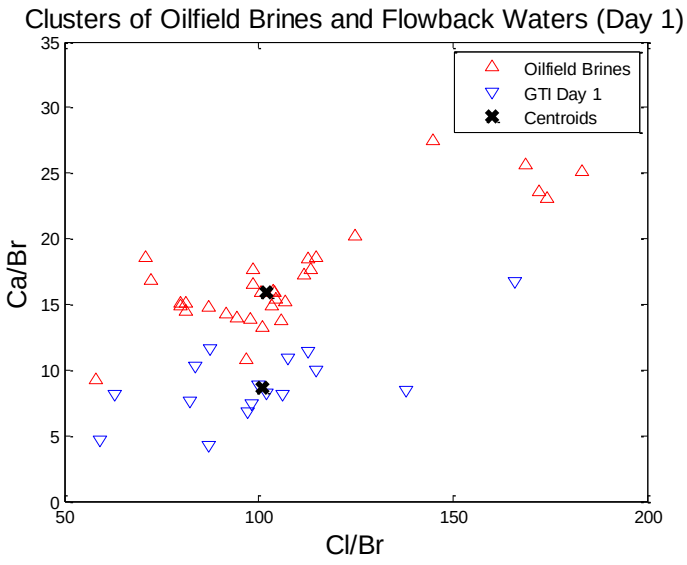


Figure 6-28 K-means clusterings in data analysis of $[\frac{Ca}{Br}]$ *over* $\left[\frac{Cl}{Br}\right]$ *between oilfield brines for well #86 to #125 and early flowback samples at Day 1 from well #1 to #19.*

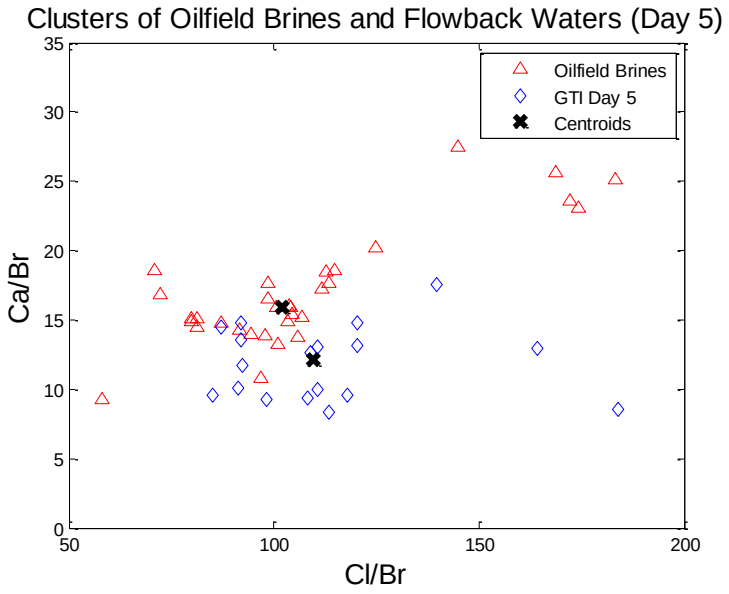


Figure 6-29 K-means clusterings in data analysis of $[\frac{Ca}{Br}]$ *over* $\left[\frac{Cl}{Br}\right]$ *between oilfield brines for well #86 to #125 and early flowback samples at Day 5 from well #1 to #19.*

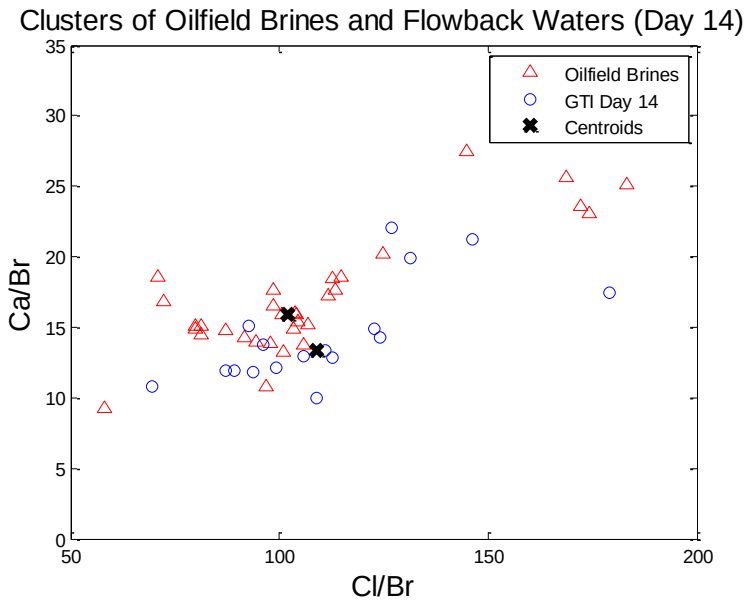


Figure 6-30 K-means clusterings in data analysis of $[\frac{Ca}{Br}]$ *over* $\left[\frac{Cl}{Br}\right]$ *between oilfield brines for well #86 to #125 and early flowback samples at Day 14 from well #1 to #19.*

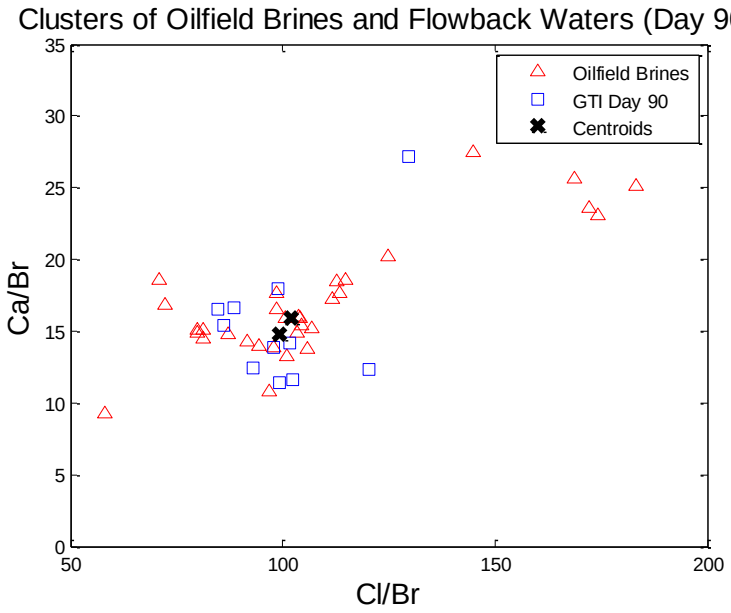


Figure 6-31 K-means clusterings in data analysis of $[\frac{Ca}{Br}]$ *over* $\left[\frac{Cl}{Br}\right]$ *between oilfield brines for well #86 to #125 and early flowback samples at Day 90 from GTI dataset from well #1 to #19.*

Fig.6-28 – Fig.6-31 as illustrated above are ascribed to the time-series GTI dataset, the indication of the flowback water attributes could be identified by K-mean clusters, the distance between oilfield brines and flowback waters centroids significantly changes as the flowbacks varies from time to time. Four time records are plotted by consistent color with consistent shapes in each sample collected from the same source. The ratio Cl/Br and Ca/Br are chosen to relegate the variances in magnitude among different wells. Yet the abnormal magnitude are neglected to display on the plots. Different trials are executed, the abnormalities have little impacts on the centroids due to the high volume of datasets. As the flowback water contents altered from early time towards the late time, the centroids is moving from far apart to adjacent. These events combined with linear regression model somehow prove that the late time flowback tends to behave affiliated in chemistry with the oilfield brines. The reason why bromide is chosen is that bromide

is always the target toxic element the water treatment plant would like to eliminate. Different trials are applied.

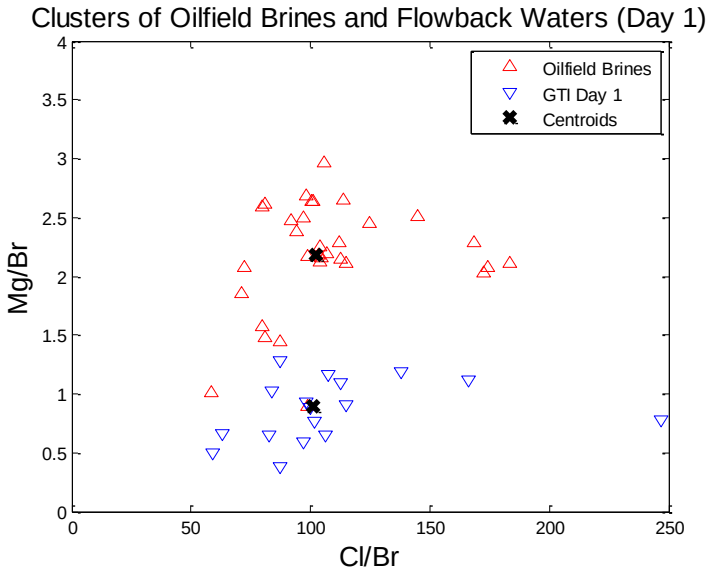


Figure 6-32 K-means clusterings in data analysis of $[\frac{Mg}{Br}]$ *over* $\left[\frac{Cl}{Br}\right]$ *between oilfield brines for well #86 to #125 and early flowback samples at Day 1 from well #1 to #19.*

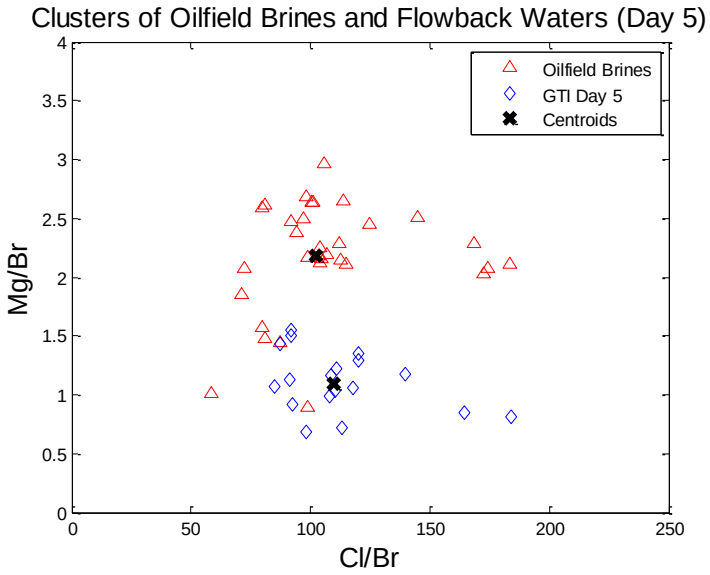


Figure 6-33 K-means clusterings in data analysis of $[\frac{Mg}{Br}]$ *over* $\left[\frac{Cl}{Br}\right]$ *between oilfield brines for well #86 to #125 and early flowback samples at Day 5 from well #1 to #19..*

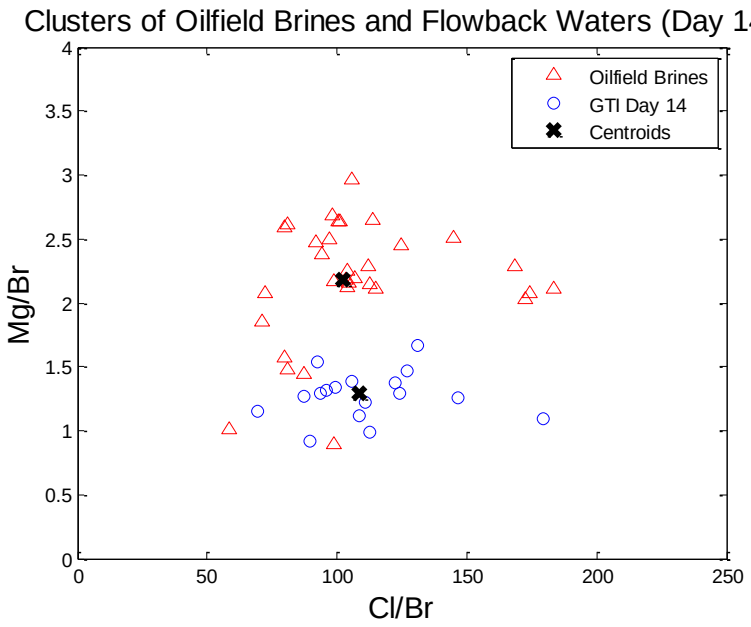


Figure 6-34 K-means clusterings in data analysis of $[\frac{Mg}{Br}]$ *over* $\left[\frac{Cl}{Br}\right]$ *between oilfield brines for well #86 to #125 and early flowback samples at Day 14 from well #1 to #19.*

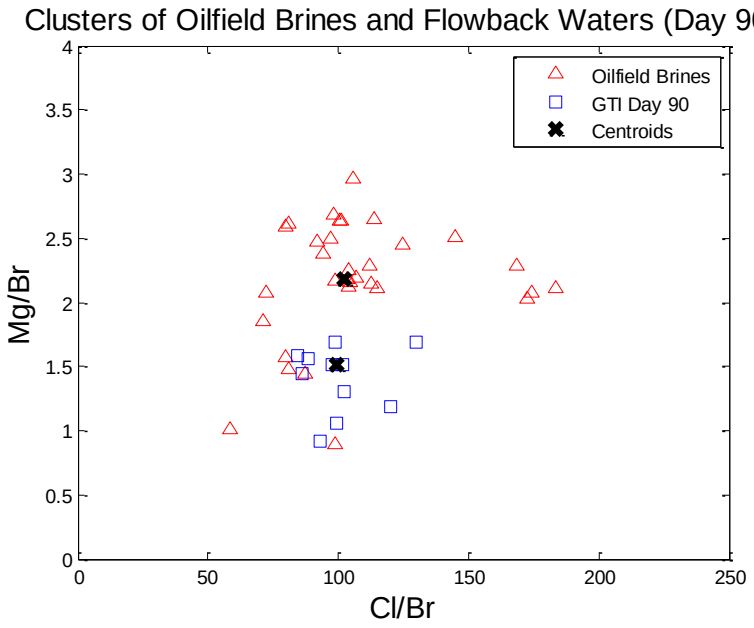


Figure 6-35 K-means clusterings in data analysis of $[\frac{Mg}{Br}]$ *over* $\left[\frac{Cl}{Br}\right]$ *between oilfield brines for well #86 to #125 and early flowback samples at Day 90 from well #1 to #19.*

Similarly, the Fig.6-32 – Fig.6-35 is done with the coordinates in ratio Mg/Br and Cl/Br. The change of the distance between centroids is not as prominent as the former features but still could be distinguished. Numerically, the results of distance are shown in table.

Table 6-1 Distance between centroids as flowback time proceeding

	Day 1	Day 5	Day 14 or 15	Day 90
Ca/Br vs. Cl/Br	7.37	8.39	7.18	3.20
Mg/Br vs. Cl/Br	1.74	7.58	6.77	3.07

Before the principal component analysis (PCA) is applied, a simple interpretation is performed on GTI dataset by utilizing the boxplots. In this case, 13 constituents are chosen to be exhibited including monovalent ions (Na, Cl, Li, K, Br) and divalent ions (Mg, Ca, Sr, Ba, SO_4) and also other abundant parameters (B, Alk, Fe).

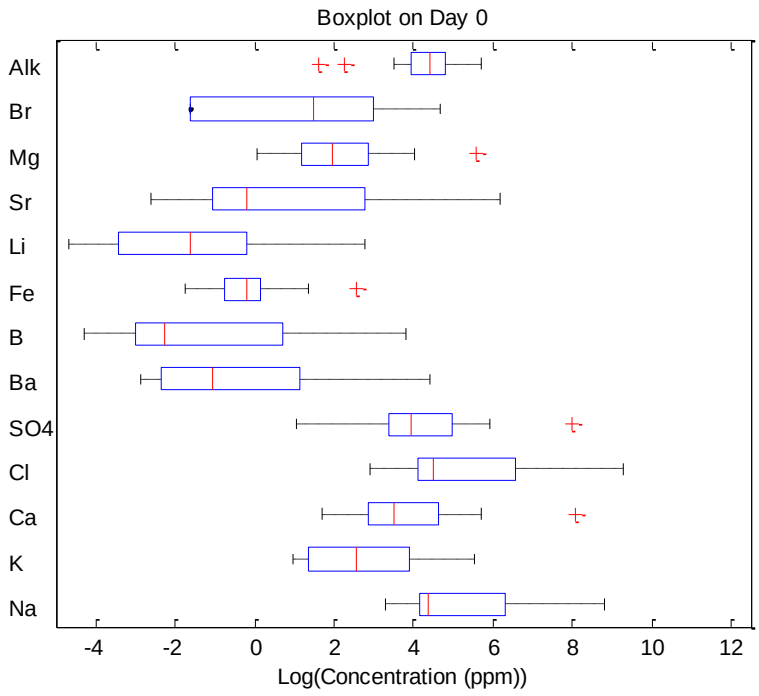


Figure 6-36 Boxplot for 13 different compositions in flowback at Day 0 from well #1 to #19.

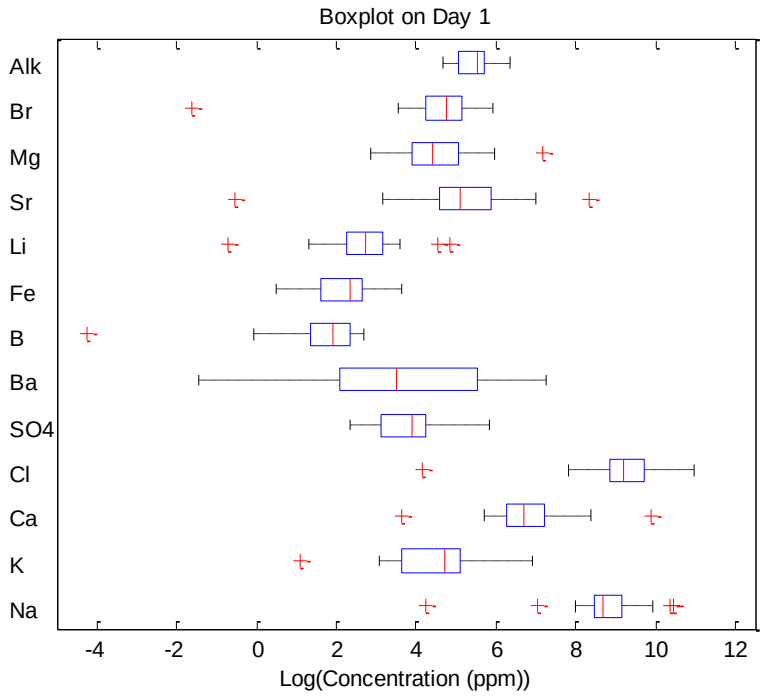


Figure 6-37 Boxplot for 13 different compositions in flowback at Day 1 from well #1 to #19.

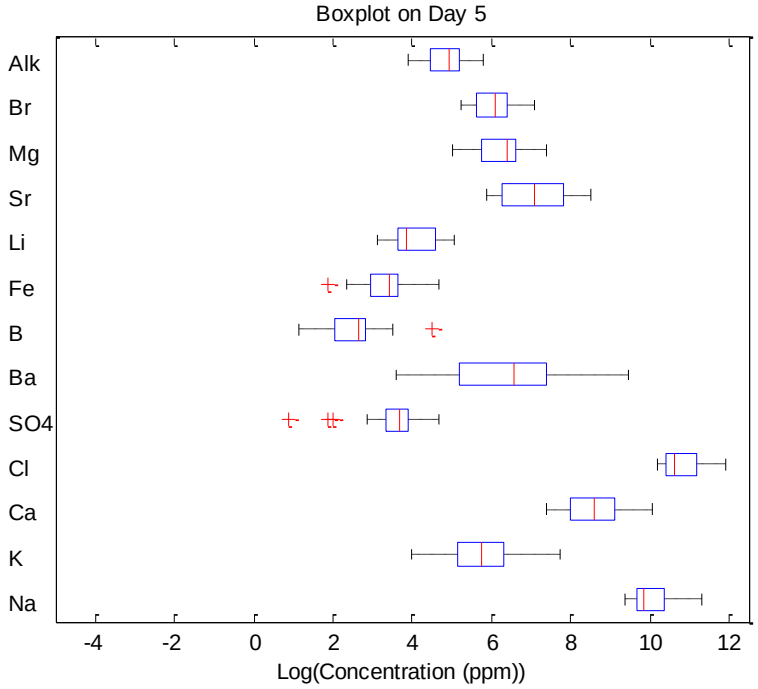


Figure 6-38 Boxplot for 13 different compositions in flowback at Day 5 from well #1 to #19.

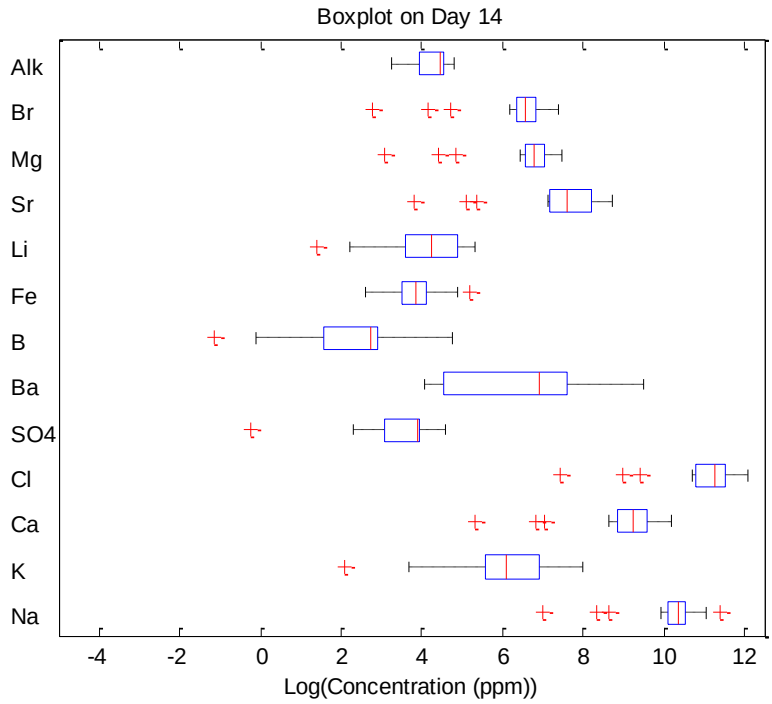


Figure 6-39 Boxplot for 13 different compositions in flowback at Day 14 from well #1 to #19.

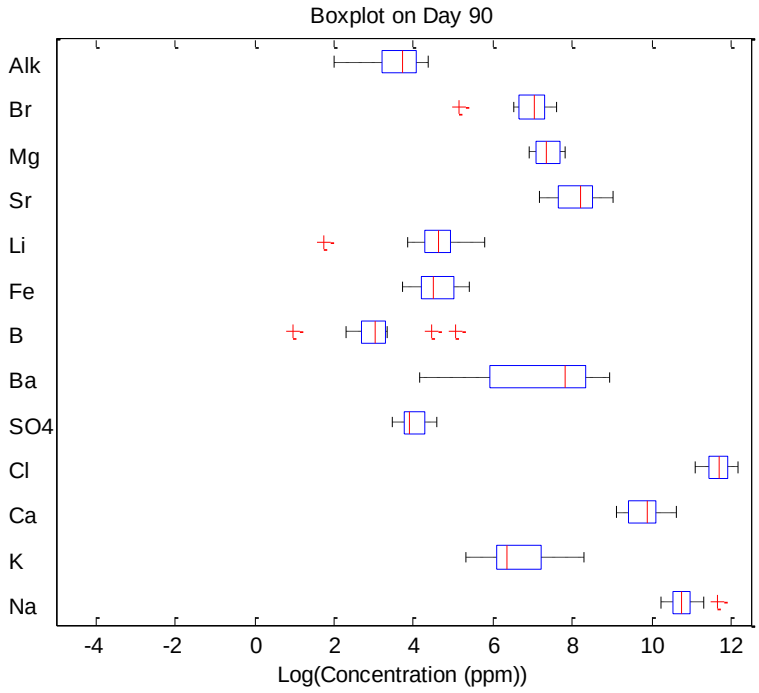


Figure 6-40 Boxplot for 13 different compositions in flowback at Day 90 from well #1 to #19.

By plotting Fig.6-36 – Fig.6-40, the data distribution could be visualized more primitively. The outliers are varied randomly. Despite the Day 0 dataset, the significant shifts for specific boxes could be pointed out, for instance, the divalent ions and the alkalinity, while the divalent ions tend to increase, simultaneously the alkalinity decreases. As it is mentioned before, the outlier effects are minimized by the adoption of dimensionless ratio, other researchers also use dimensional ratio which is functional as well. It is of certain that dataset are divided into 5 groups to represent each flowback days during to the data variation during multivariate approach.

In this study, the dataset contains random 'NaN' value which errors would be definitely caused. Thus, the alternating least squares (ALS) algorithm (PCA Using ALS for Missing Data, 2015) is utilized to remediate the missing data. The relative importance of each principal component is determined after the original matrix is reduced to lower dimension, and the result is shown below:

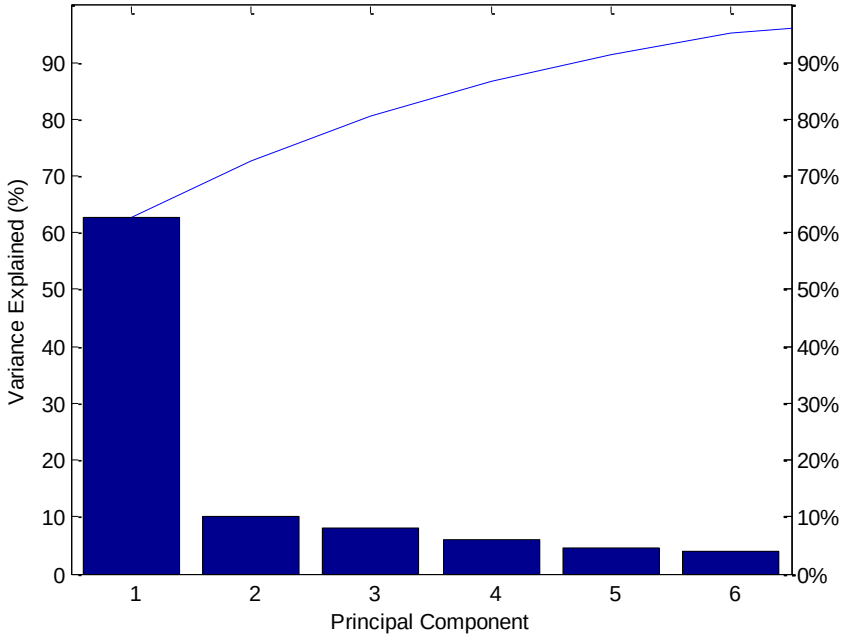


Figure 6-41 Scree plot shows major first 6 components that account for over 95% of the total variance.

In this Fig.6-41, the first 2 principal component take up to over 70% variance of the total variance. However, the first 3 considerably make up to over 80% in which 3 components is outweigh to be selected to perform the PCA.

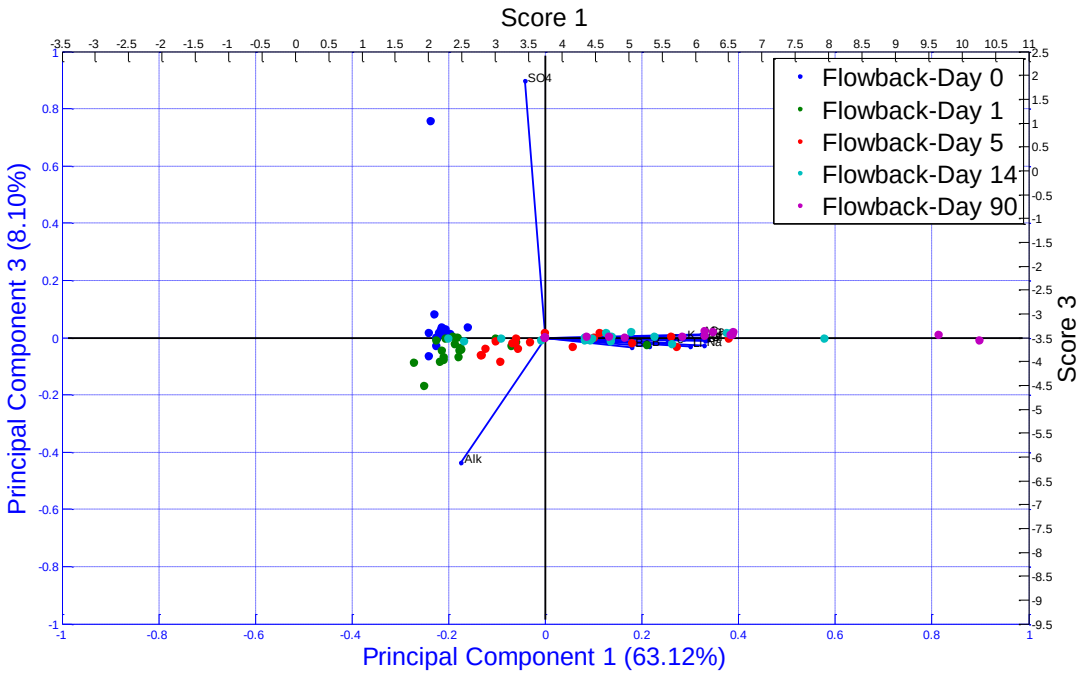


Figure 6-42 XZ view of PCA model in 3-D (3 principal components) is built up based on 13 independent flowback compositions from well #1 to #19.

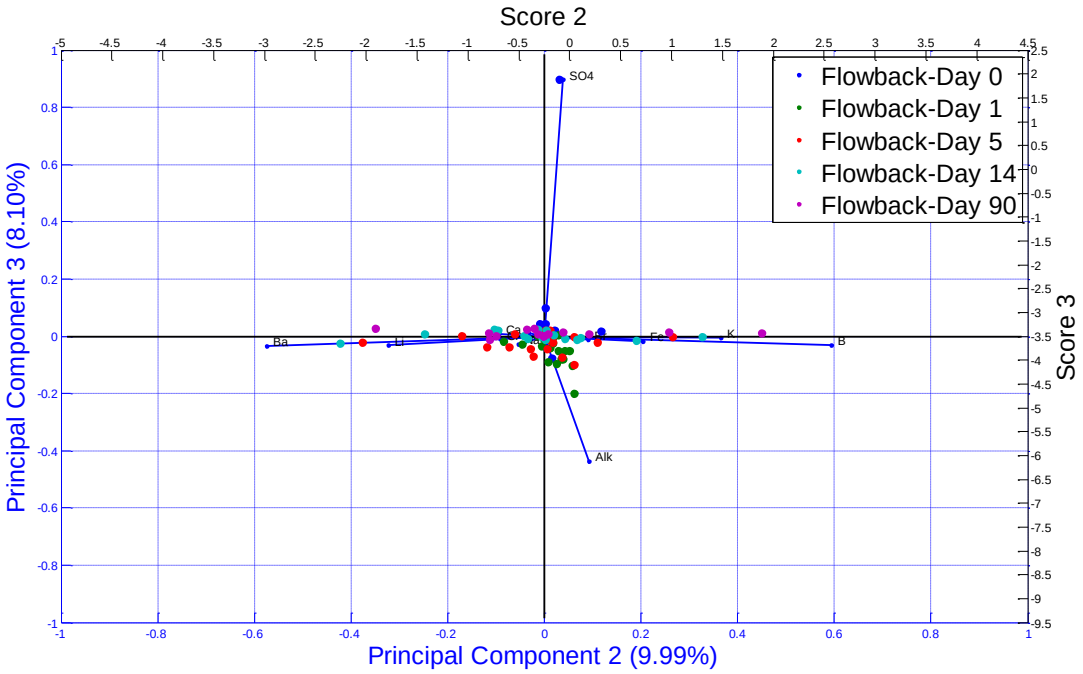


Figure 6-43 YZ view of PCA model in 3-D (3 principal components) is built up based on 13 independent flowback compositions from well #1 to #19.

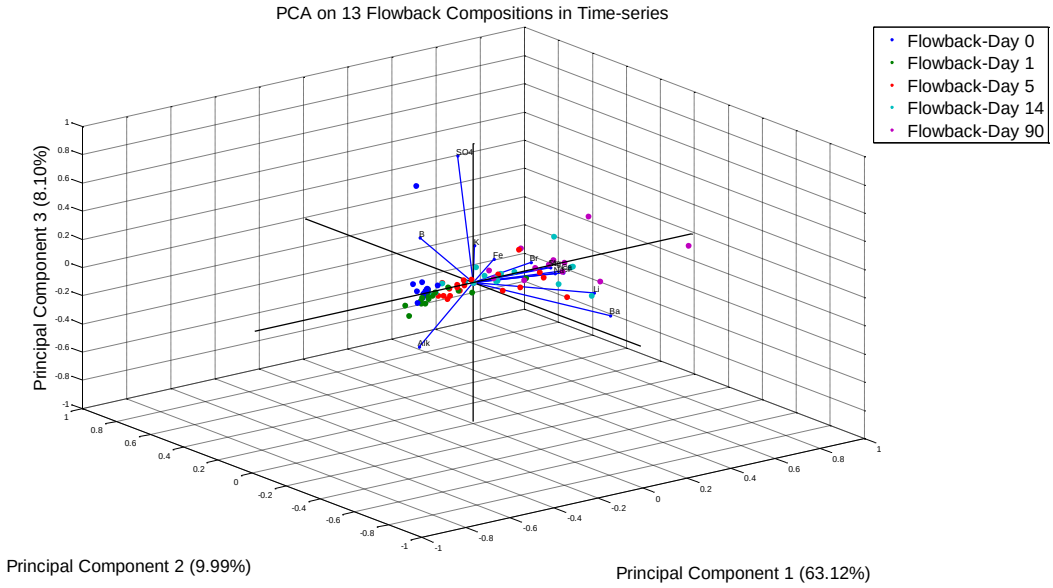


Figure 6-44 Over view of PCA model in 3-D (3 principal components) is built up based on 13 independent flowback compositions from well #1 to #19.

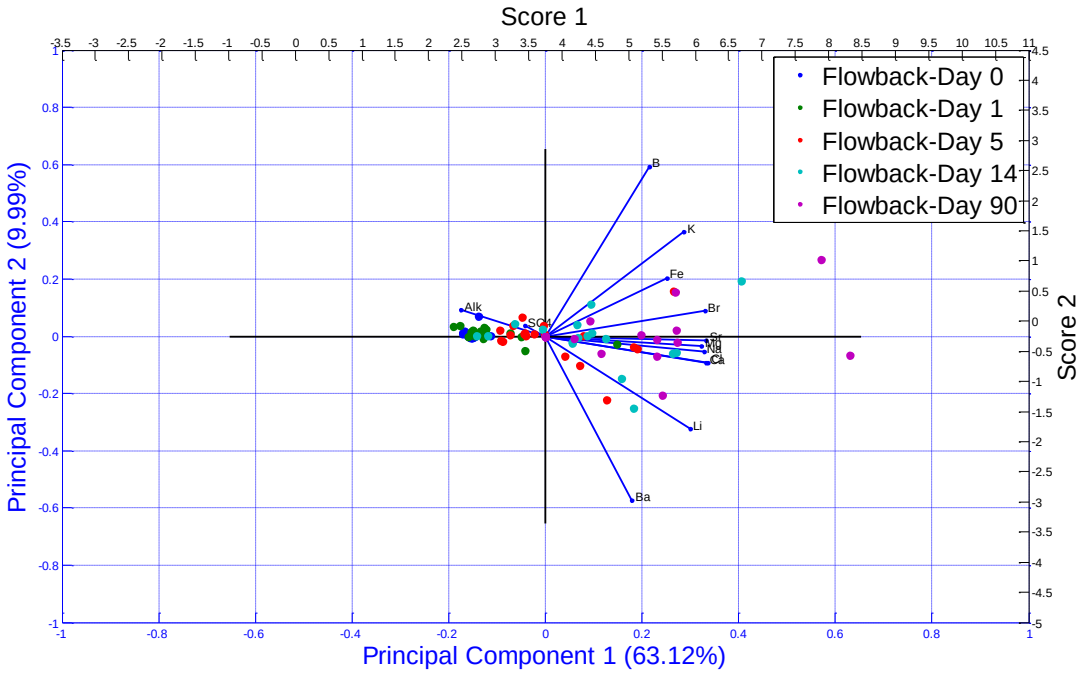


Figure 6-45 PCA model in 2-D (2 principal components) or 3-D XY view is built up based on 13 independent flowback compositions in from well #1 to #19.

A more comprehensive picture of how relative importance of all the 13 constituents is constructed. Each flowback water sample is also differentiated by the correspondence through the PCA biplots individually. The principal components loadings are plotted as blue rays from the geometric center (origin point). More importantly, the robust variance (proportional to the length of the ray) of the constituents are captured and recorded as shown in Fig.6-42 – Fig.6-45, the first three components contribute the most variance capturing all 13 constituents which includes 81.2748% of the total robust variance. The SO_4, barium and boron exhibit the somehow longest rays in 3-D, and therefore contain the greatest variance. Another biplots in 2-D with the contribution of first 2 components is done as illustrated in Fig.6-45. In 2-D case, the rays are mostly in the same direction except the alkalinity and SO_4. It is reasonable to conclude constituents whose rays are close to one other may be proportional, therefore SO_4 is proportional to alkalinity but the difference could be obviously identified in 3-D.Additionally in 2-D figure, the score 1 along the

first principal component successfully distinguishes the prevailing trend in time dependence. PC 2 also somehow discriminates a discrete trend in time series.

Possible interpretations of the PCA biplots could be proposed that most constituents tend to be dominant in late stage flowback waters except the alkalinity and SO_4. Rationally to be concluded in the late change on the concentration of alkalinity and SO_4, complicated reaction under the ground and on the surface might be the issues. However, the early stage flowback water might be influenced by the injection fluid compositions in which a transferring tendency appear on flowback day 5, a cluster of Day-0 points moved slowly from cluster Day-1 and then transfer to Day-5, discretizing during and after Day-5. When looking at Fig.6-43, SO_4, alkalinity and barium are most likely to be controlled by PC 3. Others are controlled by PC 1 and PC 2 including barium.

By using piper diagram to distinguish each observation at different flowback days, the back square indicates the flowback day 1 and each single one represents one observation. Similarly, the red circle, green triangle and blue inverse triangle illustrate different observation at different time respectively. As shown in Fig.6-46, one anomaly happens on day 1 which might be deteriorated by the injection frac-water compositions and thus it is reasonable to be classified as an outlier. In the left bottom cations triangle, with retention time gradually increased on the flowback samples, the proportion of monovalent ion combination $Na + K$ decreases consorting with the significant increment of proportion on Ca and a slightly augment on Mg. On the contrary, due to the significant comparison of proportion between Cl and the rest of anions, it seems that low proportion of anions are invisible at the right corner of the right bottom triangle. However, as time passing by, part of the Cl proportion is taken over by other anions so that the increasing of other ions at late flowback period is more considerable than Cl. When the projections of cations and anions are done in the upper diamond, the flowback water samples induced by the arrow shows

the early flowback time towards the late flowback. In this transferring period, the divalent cations tend to increase as the monovalent cations decrease, coincidently, the $Cl + SO_4$ incline to be stabilized after reaching the equilibrium with total alkalinity mildly decreases. Based on the illustration of Fig.6-46, the major cations in the flowback are at domain D where $Na - K$ type represents. Cl type dominates in anions and when the combination of cations and anions is done, it is conclusive that the transfer period of early flowback towards the late flowback is accompanied with the transferring from $Na - Cl$ type to $Mixed\ Ca - Mg - Cl$ type, however, most cases would remain in $Na - Cl$ type.

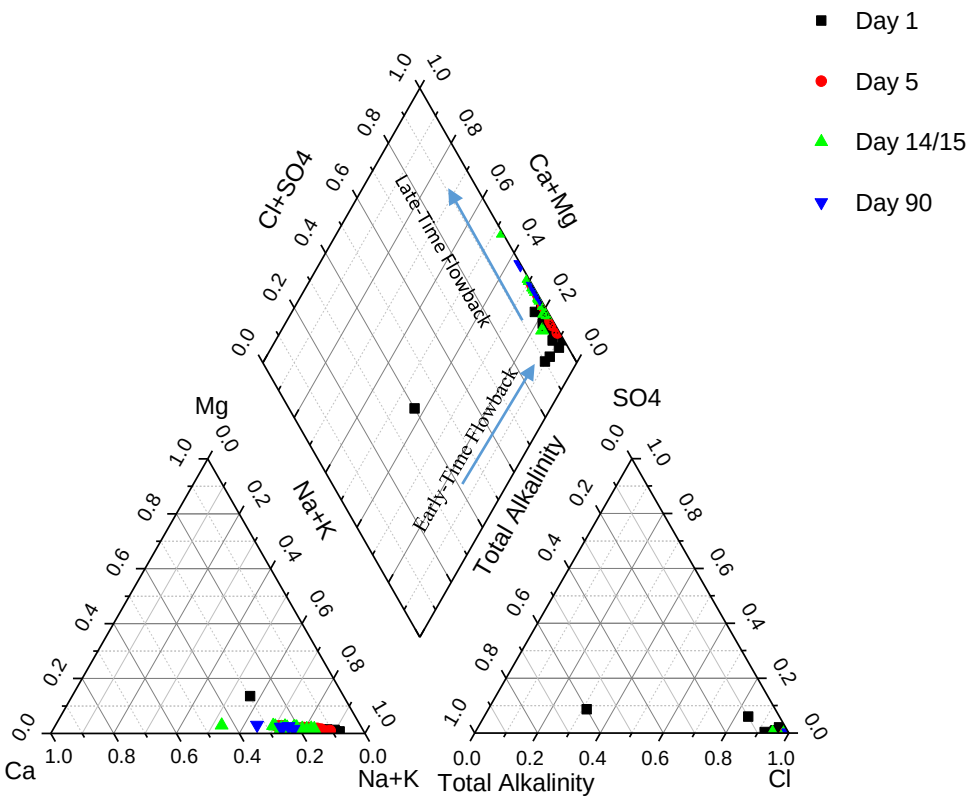


Figure 6-46 Piper trilinear diagram for entire flowback time series from well #1 to #19.

Chapter 7 Conclusions

In this work, an integrated flowback water database was built, including chemical compositions in water samples from 125 Marcellus wells and compositions in in-situ brines from 40 wells. By performing four different datamining techniques on the geographic change in concentrations, and the variation in concentrations across time, with linear-regression as well as multivaraiate approaches, a broad view and changes of chemicals in flowback have been generated and understood based on geochemistry and physics. Main conclusions are included as following:

- Geographic analysis indicates most wells have sodium concentration similar to in-situ brine after 14 days, while chloride concentration takes more than 14 days up to 90 days to reach the level of in-situ brine.
- Most non-H_2O constituents show a consistent increase in concentration over time during entire flowback period, however, there is no consistent change of sulfate concentration during flowback period.
- pH values in flowback water decreases with time and are less than 7 indicating in a weak acid. These phenomenon could be possibly triggered by dissolution of carbon dioxide and the first order of deionization of carbonic acid.
- Late time flowback water samples tends to have an affiliation with oilfield brines and the traditional treatment methods in coping with oilfield brines could also be feasible for late time flowback as it is limited in inorganic consitutents.
- An examination on water treatments based on PCA models is developed so as to better understand the features of side products along with hydrocarbon production, any implementation based on the existed PCA model could be performed for current treatment affairs.

A preliminary-Artificial Neural Network (ANN) datamining model was built up upon 5 compositions and it seems that reasonable prediction salt concentrations could be made. It is worth the efforts in future research.

Many fundamental questions about the geochemical processes and hydrodynamic influences are still veiled. Despite the behavior of early flowback are capable to be one possible recycle source of fracture fluids to alleviate the abuse of surface freshwater, the late flowback compositions especially after 3 months are yet to be fully understood.

- Water samples during the treatment
- One water sample per day during the the first 5 days
- One water sample per week week for the following three weeks.
- After this, one water sample per month for the flowback period.
- Then one water sample per year if see fit by the operator
- In summary, at least 14 water samples should be taken at day 0, 1, 2, 3, 4, 5, 12, 19, 26, 30, 60, 90, 120, 365.

References

U.S Environmental Protection Agency. (2008). *HRS Documentation Record, Safety Light Corporation, EPA.* United States Environmental Protection Agency.

Administration, U. E. (2014). *Annual Energy Outlook 2014 with projections to 2040.* Washington, DC: U.S Energy Information Administration.

Alkouh, A., McKetta, S., & Wattenbarger, R. A. (2014). *Estimation of Effecive-Fracture Volume Using Water-Flowback and Production Data for Shale-Gas Wells.* Journal of Canadian Petroleum Technology.

Alkouh, A., McKetta, S., & Wattenbarger, R. A. (2014). Estimation of Effective-Fracture Volume Using Water-Flowback and Production Data for Shale-Gas Wells. *Journal of Canadian Petroleum Technology*.

Boughton, C. J., & McCoy, K. J. (2006). *Hydrogeology, Aquifer Geochemistry, and Ground-Water Quality in Morgan County, West Virginia.* U.S. Geological Survey.

C.W.Poth. (1962). *Occurrence of brine in wester Pennsylvania: Pennsylvania geological Survey.* Mineral Resource Report.

Coleman, J. L. (2011). *Assessment of Undiscovered Oil and Gas Resources of the Devonian Marcellus Shaleof the Appalachian Basin Province.* U.S Geological Survey.

Collins, A. (1975). *Geochemistry of oilfield waters: New York.* Elsevier.

Confidence Bound. (2015). Retrieved from Mathwork Documentation: http://www.mathworks.com/help/curvefit/confidence-and-prediction-bounds.html

Cook's Distance. (2015). Retrieved from Mathwork Documentation: http://www.mathworks.com/help/stats/cooks-distance.html

Dresel, P., & Rose, A. W. (2010). *Chemistry and Origin of Oil and Gas Well Brines in Western Pennsylvania .*

H.Kutner, M., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models.* Sigapore: McGRAW HILL .

Haluszczak, L. (2011). *Geochemical analysis of flow back and production waters from oil and gas wells in Pennsylvania.* State College.

Haluszczak, L. O., Rose, A. W., & Kump, L. R. (2012). *Geochemical evaluation of flowback brine form Marcellus gas wells in Pennsylvania.* Applied Geochemistry.

Hand, B. M., & Banikowski, J. E. (2008). Radon in Onondaga County, New York: Paleohydrogeology and redistribution of uranium in Paleozoic sedimentary rocks. *Geology*, 775-778.

Hayden, J., & Pursell, D. (2005). Visitors Guide to the Hottest Gas Play in the US. In *The Barnett Shale.* Pickering.

Hayes, T. (2009). *Sampling and Analysis of Water Streams Associated with the Development of Marcellus Shale Gas.* Marcellus Shale Coalition .

Helsel, D., & Hirsch, R. (2002). Chapter 2 Graphical Data Analysis . In *Statistical Methods in Water Resources.* U.S. Geological Survey.

Hem, J. (1985). *Study and interpretation of the chemical characteristics of natural water.* U.S. Geological Survey Water-Supply.

J.Soeder, D., & M.Kappel, W. (2009). *Water Resources and Natural Gas Production from the Marcellus Shale.* Baltimore: U.S.Geol.Surv. Fact Sheet 2009-3032.

Jackson, J. (2003). Weighted PCA. In *AUser's Guide to Principal Components* (pp. 75-77). John Wiley & Sons,Inc.

Kostelnik, J., Gold, D. P., Doden, A. G., & Harper, J. A. (2003). Trenton and Black River Carbonates in the Union Furnace Area of Blair and Huntingdon Counties, Pennsylvania. In *Field Trip Guidebook for the Eastern Section AAPG Annual Meeting.* Pennsylvania Geological Survey.

Kutner, M. H., Nachtsheim, C. j., Neter, J., & Li, W. (2005). Applied Linear Statistical Models. McGRAW HILL.

Kutner, M. H., Nachtsheim, C. j., Neter, J., & Li, W. (2005). Chapter 1 Linear Regression With One Predictor Variable. McGRAW HILL.

Manoj, K., Ghosh, S., & Padhy, P. (2013). Characterization and Classification of Hydrochemistry using Multivariate Graphical and Hydrostatistical Techniques. *Research Journal of Chemical Sciences*, 32-42.

Martin, J. P. (2008). *The Middle Devonian Hamilton Group Shales in the Northern Appalachian Basin: Production and Potential.* New York State Energy Research and Development Authority.

McBride, P. S. (2004). *FACIES ANALYSIS OF THE DEVONIAN GORDON STRAY SANDSTONE IN WEST VIRGINIA.* Morgantown West Virginia .

Me.E.Blauch, R.R.Myers, T.R.Moore, & B.A.Lipinski. (2009). Marcellus Shale Post-Frac Flowback Waters – Where is All the Salt Coming From and What are the Implications ? *SPE Eastern Regional Metting .* Charleston: SPE.

Mitchell, T. (1997). Derivation of the k Means Algorithm. In *Machine Learning* (pp. 207-211). McCraw Hill.

Normal Probability of Residuals. (2015). Retrieved from Engineering Statistics Handbook: http://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm

O.Vazquez, R.Mehta, E.Mackay, S.Linares-Samaniego, M.Jordan, & J.Fidoe. (2014). Post-frac Flowback Water Chemistry Matching in a Shale Development. *SPE International Oilfield Scale Conference and Exhibition .* Aberdeen : SPE.

P.Werne, J., Sageman, B. B., W.Lyons, T., & Hollander, D. J. (2002). An integrated assessment of a “type euxinic” deposit: Evidence for multiple controls on black shale deposition in the middle Devonian Oatka Creek formation. *American Journal of Science*, 110-143.

PADEP. (2013). Retrieved from PA DEP Oil & Gas Reporting Website: https://www.paoilandgasreporting.state.pa.us/publicreports/Modules/Welcome/Welcome.aspx

PCA Using ALS for Missing Data. (2015). Retrieved from Mathwork Documentation: http://www.mathworks.com/help/stats/pca.html

Pritz, M. E. (2010). *Geochemical Modeling and Analysis of the Frac Water Used in the Hydraulic Fracturing of the Marcellus Formation, Pennsylvania.*

Protection, P. D. (2011). *Chemcial Characterization of Marcellus Shale Fracturing Flowback and Drill Cuttings.* PA Department of Environmental Protection.

Range Releases Frac Fact Sheet. (2010). Retrieved from Drilling Info: http://info.drillinginfo.com/range-releases-frac-fact-sheet-say-that-3-times-fast/

Residuals. (2015). Retrieved from Mathwork Documentation: http://www.mathworks.com/help/stats/residuals.html

Rokach, L., & Maimon, O. (2015). Chapter 15 Clustering Methods.

SGEIS. (2011). *Natural Gas Development Activities & High-Volume Hydraulic Fracturing.* Supplemental Generic Environmental Impact Statement.

Shale Energy Business Briefing. (2015). Retrieved from http://shaleenergybusinessbriefing.com/

The Hydraulic Fracturing Water Cycle. (2012). Retrieved from EPA: http://www2.epa.gov/hfstudy/hydraulic-fracturing-water-cycle

U.S. Energy Information Administration. (2012). Retrieved from http://www.eia.gov/analysis/studies/usshalegas/

Appendix Program

Program Code

```
Hayes (2010)

clc;
clear all;
clf;
close all;
%%
Ha=xlsread('Data(Hayes2009)_PAN','TDS');
save ('Ha')

%%
load ('Ha');

%% TDS vs. Flowback Volume

i=19 %Note that i~ 1 to 19 since we just have 19 wells.
t=[0 1 5 14 90];% if we do not consider neither the influent concetration or
the influent volume
                 % delete '0'


if i==4;
    i_TDS=[log10(Ha(i,1:3)) log10(Ha(i,5))];
    t1=[0 1 5 90];% if we do not consider neither the influent concetration
or the influent volume
                  % delete '0'
i_Vol=[log10(Ha(i,6)) log10(Ha(i,7)) log10((Ha(i,8)-Ha(i,7))/4)
log10((Ha(i,9)-Ha(i,8))/9) log10((Ha(i,10)-Ha(i,9))/76)];
                  % if we do not consider neither the influent concetration
or the influent volume
                  % delete Ha(i,6)
[ax,p1,p2] = plotyy(t1,i_TDS,t,i_Vol,'plot');
elseif i==13;
    i_TDS=[log10(Ha(i,1)) log10(Ha(i,4))];
    t13=[0 14];
    i_Vol=[log10(Ha(i,6)) log10(Ha(i,7)) log10((Ha(i,8)-Ha(i,7))/4)
log10((Ha(i,9)-Ha(i,8))/9) log10((Ha(i,10)-Ha(i,9))/76)];
    [ax,p1,p2] = plotyy(t13,i_TDS,t,i_Vol,'plot');
else
    i_TDS=log10(Ha(i,1:5));
    i_Vol=[log10(Ha(i,6)) log10(Ha(i,7)) log10((Ha(i,8)-Ha(i,7))/4)
log10((Ha(i,9)-Ha(i,8))/9) log10((Ha(i,10)-Ha(i,9))/76)];
                  % if we do not consider neither the influent concetration
or the influent volume
                  % delete Ha(i,6)
    [ax,p1,p2] = plotyy(t,i_TDS,t,i_Vol,'plot');
end
set(p1,'LineStyle','-','Marker','d','LineWidth',2)
set(p2,'LineStyle','-','Marker','d','LineWidth',2)
title(sprintf('Log(C_T_D_S) vs. Log(V_F_L_B) at Well %i', i));
```

```
ylabel(ax(1),'Log(C_T_D_S)') % left y-axis
ylabel(ax(2),'Ave Flowback Log(V_F_L_B) in the Interval') % right y-axis
xlabel('Flowback Time (Days)')
legend('Log(C_T_D_S)','Ave Flowback Log(V_F_L_B) in the
Interval','location','best')

%%
t=[0 1 5 14 90];
TDS=Ha(:,1:5);
Vol=[log10(Ha(:,6)) log10(Ha(:,7)) log10((Ha(:,8)-Ha(:,7))/4) log10((Ha(:,9)-
Ha(:,8))/9) log10((Ha(:,10)-Ha(:,9))/76)]
for j=1:1:5;
for i=1:1:19;
   K2(i,j)=isnan(TDS(i,j));
   row=find(K2(:,j)==0);
   TDS_mean(j)=mean(TDS(row,j),1);
end
end

for j=1:1:5;
for i=1:1:19;
   K3(i,j)=isnan(TDS(i,j));
   row=find(K2(:,j)==0);
   TDS_std(j)=std(TDS(row,j),1);
end
end

for j=1:1:5;
for i=1:1:19;
   K4(i,j)=isnan(Vol(i,j));
   row=find(K4(:,j)==0);
   Vol_mean(j)=mean(Vol(row,j),1);
end
end

for j=1:1:5;
for i=1:1:19;
   K5(i,j)=isnan(Vol(i,j));
   row=find(K5(:,j)==0);
   Vol_std(j)=std(Vol(row,j),1);
end
end

figure()

[ax,p1,p2] = plotyy(t,log10(TDS_mean),t,Vol_mean,'plot');
hold(ax(1))
hh1=errorbar(ax(1),t,log10(TDS_mean),log10(TDS_std),'bo');
hold(ax(2))
hh2=errorbar(ax(2),t,Vol_mean,Vol_std,'o');
set(hh2,'Color',[0,0.5 0]);
set(ax,'xlim',[-5,95]);
set(ax(1),'ylim',[-1,11]);
set(ax(1),'YTick',-1:1:11);
set(ax(2),'ylim',[1,5.5]);
set(ax(2),'YTick',1:1:5.5)
```

```
set(p1,'LineStyle','-','LineWidth',2)
set(p2,'LineStyle','-','LineWidth',2)
set(hh1,'MarkerFace','b')
set(hh2,'MarkerFace',[0,0.5,0])
title('Time Series of Log(TDS) vs. Log(Vol) for All 19 Wells');
ylabel(ax(1),'Log(TDS) (ppm)') % left y-axis
ylabel(ax(2),'Log(Vol) (ppm)') % right y-axis
xlabel('Flowback Time (Days)')
legend([p1;p2],'Log(TDS)','Log(Vol)','location','best');
%% Not useful

i=1   %Note that i~ 1 to 19 since we just have 19 wells.
t=[1 5 14 90];% if we do not consider neither the influent concetration or
the influent volume
Int=[1:2 2:5 5:14 ];                % delete '0'


if i==4;
    i_TDS=[Ha(i,2:3) Ha(i,5)];
    t1=[1 5 90];% if we do not consider neither the influent concetration or
the influent volume
                % delete '0'
i_Vol=[Ha(i,7) (Ha(i,8)-Ha(i,7))/4 (Ha(i,9)-Ha(i,8))/9 (Ha(i,10)-
Ha(i,9))/76];
                % if we do not consider neither the influent concetration
or the influent volume
                % delete Ha(i,6)
[ax,p1,p2] = plotyy(t1,i_TDS,Int,i_Vol,'plot','area');
else
    i_TDS=Ha(i,2:5);
    i_Vol=[ones(1,2).*Ha(i,7) ones(1,4).*((Ha(i,8)-Ha(i,7))/4)
ones(1,10).*((Ha(i,9)-Ha(i,8))/9) ];
                % if we do not consider neither the influent concetration
or the influent volume
                % delete Ha(i,6)
    [ax,p1,p2] = plotyy(Int,i_Vol,t,i_TDS,'area','plot');
end
set(p1,'LineWidth',2)
xlim([0 90])
title(sprintf('Concentration of TDS vs. Water Volume at Well %i', i));
ylabel(ax(2),'TDS (mg/L)') % left y-axis
ylabel(ax(1),'Ave Flowback Vol in the Interval (bbls/Day)') % right y-axis
xlabel('Flowback Time (Days)')



%% Concentration for each ion from 19 wells

Ha_Ions=xlsread('Data(Hayes2009)_PAN','sheet1');
t=[0 1 5 14.5 90]; % t=0 is the influent of fracing water

Na=Ha_Ions(1:11:199,1:5);
K=Ha_Ions(2:11:200,1:5);
Ca=Ha_Ions(3:11:201,1:5);
```

```
Cl=Ha_Ions(4:11:202,1:5);
SO4=Ha_Ions(5:11:203,1:5);
Ba=Ha_Ions(6:11:204,1:5);
B=Ha_Ions(7:11:205,1:5);
Fe=Ha_Ions(8:11:206,1:5);
Li=Ha_Ions(9:11:207,1:5);
Sr=Ha_Ions(10:11:208,1:5);
Mg=Ha_Ions(11:11:209,1:5);
TDS=Ha(:,1:5);
rng('default'); % For reproducibility
for i=1:19;
XX(i,:)=(randi([0,10],1,3))./10; %Fixed the index color for each well
end

load XX

figure()
for i=1:1:19;
hold on
 I(i,:)=isnan(Na(i,:));
 col=find(I(i,:)==0);
h=plot(t(col),Na(i,col)./1000,'marker','d');
set(h,'color',XX(i,:),'linewidth',1.5);
hold on
legendInfo{i} = ['C_N_a at Well ' num2str(i)];
end
legend(legendInfo)
xlabel('Flowback Time (Days)');
ylabel('Concentration (ppm)');
title('Concentration of Sodium')

figure()
for i=1:1:19;
hold on
 I(i,:)=isnan(K(i,:));
 col=find(I(i,:)==0);
h=plot(t(col),K(i,col)./1000,'marker','d');
set(h,'color',XX(i,:),'linewidth',1.5);
hold on
legendInfo{i} = ['C_K at Well ' num2str(i)];
end
legend(legendInfo)
xlabel('Flowback Time (Days)');
ylabel('Concentration (ppm)');
title('Concentration of Potassium')

for i=1:1:19;
hold on
 I(i,:)=isnan(Ca(i,:));
 col=find(I(i,:)==0);
h=plot(t(col),Ca(i,col)./1000,'marker','d');
set(h,'color',XX(i,:),'linewidth',1.5);
hold on
legendInfo{i} = ['C_C_a at Well ' num2str(i)];
end
legend(legendInfo)
```

```
xlabel('Flowback Time (Days)');
ylabel('Concentration (ppm)');
title('Concentration of Calcium')

for i=1:1:19;
hold on
I(i,:)=isnan(Cl(i,:));
 col=find(I(i,:)==0);
h=plot(t(col),Cl(i,col),'marker','d');
set(h,'color',XX(i,:),'linewidth',1.5);
hold on
legendInfo{i} = ['C_C_l at Well ' num2str(i)];
end
legend(legendInfo)
xlabel('Flowback Time (Days)');
ylabel('Concentration (ppm)');
title('Concentration of Chloride')

figure()
for i=1:1:19;
hold on
I(i,:)=isnan(SO4(i,:));
 col=find(I(i,:)==0);
h=plot(t(col),log10(SO4(i,col)),'marker','d');
set(h,'color',XX(i,:),'linewidth',1.5);
hold on
legendInfo{i} = ['Log(C_S_O_4) at Well ' num2str(i)];
end
legend(legendInfo)
xlabel('Flowback Time (Days)');
ylabel('Log(SO_4) (ppm)');
title('Semi-Log Concentration of Sulfate')


for i=1:1:19;
hold on
I(i,:)=isnan(Ba(i,:));
 col=find(I(i,:)==0);
h=plot(t(col),Ba(i,col)./1000,'marker','d');
set(h,'color',XX(i,:),'linewidth',1.5);
hold on
legendInfo{i} = ['C_B_a at Well ' num2str(i)];
end
legend(legendInfo)
xlabel('Flowback Time (Days)');
ylabel('Concentration of Barium (ppm)');
title('Concentration of Barium')

for i=1:1:19;
hold on
I(i,:)=isnan(B(i,:));
 col=find(I(i,:)==0);
h=plot(t(col),B(i,col)./1000,'marker','d');
set(h,'color',XX(i,:),'linewidth',1.5);
hold on
legendInfo{i} = ['C_B at Well ' num2str(i)];
```

```
end
legend(legendInfo)
xlabel('Flowback Time (Days)');
ylabel('Concentration (ppm)');
title('Concentration of Boron')

for i=1:1:19;
hold on
I(i,:)=isnan(Fe(i,:));
 col=find(I(i,:)==0);
h=plot(t(col),Fe(i,col)./1000,'marker','d');
set(h,'color',XX(i,:),'linewidth',1.5);
hold on
legendInfo{i} = ['C_F_e at Well ' num2str(i)];
end
legend(legendInfo)
xlabel('Flowback Time (Days)');
ylabel('Concentration (ppm)');
title('Concentration of Iron')

for i=1:1:19;
hold on
I(i,:)=isnan(Li(i,:));
 col=find(I(i,:)==0);
h=plot(t(col),Li(i,col)./1000,'marker','d');
set(h,'color',XX(i,:),'linewidth',1.5);
hold on
legendInfo{i} = ['C_L_i at Well ' num2str(i)];
end
legend(legendInfo)
xlabel('Flowback Time (Days)');
ylabel('Concentration (ppm)');
title('Concentration of Lithium')

for i=1:1:19;
hold on
I(i,:)=isnan(Sr(i,:));
 col=find(I(i,:)==0);
h=plot(t(col),Sr(i,col)./1000,'marker','d');
set(h,'color',XX(i,:),'linewidth',1.5);
hold on
legendInfo{i} = ['C_S_r at Well ' num2str(i)];
end
legend(legendInfo)
xlabel('Flowback Time (Days)');
ylabel('Concentration (ppm)');
title('Concentration of Strontium')


for i=1:1:19;
hold on
I(i,:)=isnan(Mg(i,:));
 col=find(I(i,:)==0);
h=plot(t(col),Mg(i,col)./1000,'marker','d');
set(h,'color',XX(i,:),'linewidth',1.5);
hold on
```

```
legendInfo{i} = ['C_M_g at Well ' num2str(i)];
end
legend(legendInfo)
xlabel('Flowback Time (Days)');
ylabel('Concentration (ppm)');
title('Concentration of Magnesium')

for i=1:1:19;
hold on
I(i,:)=isnan(TDS(i,:));
 col=find(I(i,:)==0);
h=plot(t(col),TDS(i,col),'marker','d');
set(h,'color',XX(i,:),'linewidth',1.5);
hold on
legendInfo{i} = ['C_T_D_S at Well ' num2str(i)];
end
legend(legendInfo)
xlabel('Flowback Time (Days)');
ylabel('Concentration (ppm)');
title('Concentration of TDS')

Well_A=Ha_Ions(1:11,1:5);
for i=1:1:11;
hold on
if i==4;
    h=plot(t,Well_A(i,:));
elseif i==5;
     h=plot(t,Well_A(i,:));
     else h=plot(t,Well_A(i,:)./1000);
end
set(h,'color',rand(1,3),'linewidth',1.5);
hold on
legend('Sodium','Potassium','Calcium','Chloride','Sulfate','Barium','Boron','
Iron','Lithium','Strontium','Magnesium')
end

xlabel('Flowback Time (Days)');
ylabel('Concentration (ppm)');


%% Same thing as above but more detail
% Change the TDS vs. Ions. Details of each well and each ions
i=3 % i index stands for each well
t=[0 1 5 14 90];

I1(i,:)=isnan(TDS(i,:));
 col1=find(I1(i,:)==0);
I2(i,:)=isnan(Ba(i,:)); % The Ions can be changed
 col2=find(I2(i,:)==0);

        h1 =plot(t(col1),log(TDS(i,col1)),'d-');
    hold on
        h2=plot(t(col2),log(Ba(i,col2)./1000),'d-'); % The Ions can be
changed
```

```
set(h1,'color','blue','LineWidth',2)
set(h2,'color',[0 0.5 0],'LineWidth',2)
title(sprintf('Semi-Log Plot for Concentration of TDS vs. Ion at Well %i',
i));
ylabel('Log(Concentration)')
xlabel('Flowback Time (Days)')
legend('TDS','Ba') % Don't forget to change the name of ions


%% Time Series Divalent Ions
i=1;
t=[0 1 5 14 90];
I3(i,:)=isnan(Ba(i,:));
 col3=find(I3(i,:)==0);
I4(i,:)=isnan(Sr(i,:));
 col4=find(I4(i,:)==0);
I5(i,:)=isnan(Ca(i,:));
 col5=find(I5(i,:)==0);
I6(i,:)=isnan(Mg(i,:));
 col6=find(I6(i,:)==0);
I7(i,:)=isnan(SO4(i,:));
 col7=find(I7(i,:)==0);
h1=plot(t(col3),log(Ba(i,col3)/1000),'d-');
hold on
h2=plot(t(col4),log(Sr(i,col4)/1000),'d-');
h3=plot(t(col5),log(Ca(i,col5)/1000),'d-');
h4=plot(t(col6),log(Mg(i,col6)/1000),'d-');
h5=plot(t(col7),log(SO4(i,col7)),'d-');

set(h1,'color','red','LineWidth',2);
set(h2,'color','blue','LineWidth',2);
set(h3,'color','green','LineWidth',2);
set(h4,'color','cyan','LineWidth',2);
set(h5,'color','magenta','LineWidth',2);
xlabel('Flowback Time (Days)')
ylabel('Log(Concentration)')
legend('Ba','Sr','Ca','Mg','SO_4','Location','best')
title(sprintf('Time Series Divalent Ions at Well %i', i));

%% Bad Results
i=5
h1=plot(log(SO4(:,i)),log(Ba(:,i)/1000),'r.');
hold on
h2=plot(log(SO4(:,i)),log(Sr(:,i)/1000),'b.');
h3=plot(log(SO4(:,i)),log(Ca(:,i)/1000),'m.');
h4=plot(log(SO4(:,i)),log(Mg(:,i)/1000),'c.');


set(h1,'MarkerSize',20);
set(h2,'MarkerSize',20);
set(h3,'MarkerSize',20);
set(h4,'MarkerSize',20);

%% Principal Component Results
Ha2=xlsread('Data(Hayes2009)_PAN','Sheet4');
```

```
Br=Ha2(1:4:73,1:5);
S=Ha2(2:4:74,1:5);
Alk=Ha2(3:4:75,1:5);
pH=Ha2(4:4:76,1:5);

names = cell(5,1);
t=[0 1 5 14 90]
for i=1:5;
    names{i}=['Day ' num2str(t(i))];
end


figure() % Boxplots for each ion for different days
boxplot(Na/1000,'orientation','horizontal','labels',names)
xlabel('Concentration of Sodium (ppm)')
title('Boxplot of Sodium')

% Boxplots for all ions on each flowback days
for i=1:5;
    figure(i)

day_i=[Na(:,i)/1000 K(:,i)/1000 Ca(:,i)/1000 Cl(:,i) SO4(:,i)
Ba(:,i)/1000 ...
    B(:,i)/1000 Fe(:,i)/1000 Li(:,i)/1000 Sr(:,i)/1000 Mg(:,i)/1000 Br(:,i)
Alk(:,i)];
Categories = cell(13,1)
Categories={'Na';'K';'Ca';'Cl';'SO4';'Ba';'B';'Fe';'Li';'Sr';'Mg';'Br';'Alk'}
boxplot(log(day_i),'plotstyle','traditional','orientation','horizontal','labe
ls',Categories)
xlabel('Log(Concentration (ppm))')
title(sprintf('Boxplot on Day %i',t(i)))
xlim([-5 12.5]);
end

% PCA Datapool
i=1;
Day_0=[Na(:,i)/1000 K(:,i)/1000 Ca(:,i)/1000 Cl(:,i) SO4(:,i)
Ba(:,i)/1000 ...
    B(:,i)/1000 Fe(:,i)/1000 Li(:,i)/1000 Sr(:,i)/1000 Mg(:,i)/1000 Br(:,i)
Alk(:,i)];
i=2;
Day_1=[Na(:,i)/1000 K(:,i)/1000 Ca(:,i)/1000 Cl(:,i) SO4(:,i)
Ba(:,i)/1000 ...
    B(:,i)/1000 Fe(:,i)/1000 Li(:,i)/1000 Sr(:,i)/1000 Mg(:,i)/1000 Br(:,i)
Alk(:,i)];
i=2;
i=3;
Day_5=[Na(:,i)/1000 K(:,i)/1000 Ca(:,i)/1000 Cl(:,i) SO4(:,i)
Ba(:,i)/1000 ...
    B(:,i)/1000 Fe(:,i)/1000 Li(:,i)/1000 Sr(:,i)/1000 Mg(:,i)/1000 Br(:,i)
Alk(:,i)];
i=4;
Day_14=[Na(:,i)/1000 K(:,i)/1000 Ca(:,i)/1000 Cl(:,i) SO4(:,i)
Ba(:,i)/1000 ...
    B(:,i)/1000 Fe(:,i)/1000 Li(:,i)/1000 Sr(:,i)/1000 Mg(:,i)/1000 Br(:,i)
Alk(:,i)];
```

```
i=5;
Day_90=[Na(:,i)/1000 K(:,i)/1000 Ca(:,i)/1000 Cl(:,i) SO4(:,i)
Ba(:,i)/1000 ...
    B(:,i)/1000 Fe(:,i)/1000 Li(:,i)/1000 Sr(:,i)/1000 Mg(:,i)/1000 Br(:,i)
Alk(:,i)];

Datapool=[Day_0 Day_1 Day_5 Day_14 Day_90];

for i=1:1:19;
for    j=1:1:65;
   K0(i,j)=isnan(Datapool(i,j));

end
end
for j=1:65;
[row,col]=find(K0(:,j)==0);
for i=1:size(row,1);
Data_New(i,j)=Datapool(row(i),j);
end
end


Datapool_Categories=[Day_0; Day_1; Day_5; Day_14; Day_90];

for i=1:1:95;
for    j=1:1:13;
   K1(i,j)=isnan(Datapool_Categories(i,j));

end
end
for j=1:13;
[row,col]=find(K1(:,j)==0);
for i=1:size(row,1);
Data_New1(i,j)=Datapool_Categories(row(i),j);
end
end

Datapool_Categories(isnan(Datapool_Categories))=0;
Data_New2=Datapool_Categories

Categories_new = cell(11,1)
Categories_new={'Na';'K';'Ca';'Cl';'SO4';'Ba';'B';'Fe';'Li';'Sr';'Mg';'Br';'A
lk'}
w = 1./var(Data_New2);

[wcoeff,score,latent,tsquared,explained] = pca(Data_New2,...
'VariableWeights',w);

 coefforth = inv(diag(std(Data_New2)))*wcoeff;

 biplot(coefforth(:,1:3),'score',score(:,1:3),'varlabels',Categories_new);
 hold on

 sc_Day0=score(1:19,:);
```

```
 sc_Day0=sc_Day0(:);
 sc_Day1=score(20:38,:);
 sc_Day1=sc_Day1(:);
 sc_Day5=score(39:57,:);
 sc_Day5=sc_Day5(:);
 sc_Day14=score(58:76,:);
 sc_Day14=sc_Day14(:);
 sc_Day90=score(77:95,:);
 sc_Day90=sc_Day90(:);

 scatter3( sc_Day0,sc_Day1,sc_Day5,'r^')


%  hold on
%  [wcoeff,score,latent,tsquared,explained] = pca(Day_0,...
% 'VariableWeights',w);
% plot(score(:,1),score(:,2),'r^')
%  [wcoeff,score,latent,tsquared,explained] = pca(Day_1,...
% 'VariableWeights',w);
% plot(score(:,1),score(:,2),'bv')
%  [wcoeff,score,latent,tsquared,explained] = pca(Day_5,...
% 'VariableWeights',w);
% plot(score(:,1),score(:,2),'go')
%  [wcoeff,score,latent,tsquared,explained] = pca(Day_14,...
% 'VariableWeights',w);
% plot(score(:,1),score(:,2),'c*')
%  [wcoeff,score,latent,tsquared,explained] = pca(Day_90,...
% 'VariableWeights',w);
% plot(score(:,1),score(:,2),'md')
% xlabel('1st Principal Component')
% ylabel('2nd Principal Component')
% legend on
% xlim([-1.5,1.5])
% ylim([-1.5,1.5])
%% Log(Br) vs. Log(Cl)

i=4
% hh=plot(log(Br(:,i)),log(Cl(:,i)),'r.');
% hold on
% set(hh,'MarkerSize',15);
mdl = LinearModel.fit(log(Br(:,i)),log(Cl(:,i)))
h=plotAdded(mdl)
title('Log(Br) vs. Log(Cl)')

xlabel('Log(Br)')
ylabel('Log(Cl)')

%%
i=2
Rat=SO4./Alk
figure()
t=[0 1 5 14 90];% if we do not consider neither the influent concetration or
the influent volume
                % delete '0'
```

```
I(i,:)=isnan(Rat(i,:));
 col=find(I(i,:)==0);
 I0(i,:)=isnan(pH(i,:));
 col0=find(I0(i,:)==0);

    [ax,p1,p2] = plotyy(t(col),Rat(i,col),t(col0),pH(i,col0),'plot');


set(p1,'LineStyle','-','Marker','d','LineWidth',2)
set(p2,'LineStyle','-','Marker','d','LineWidth',2)
title(sprintf('[SO4/Alk] vs. pH at Well %i', i));
ylabel(ax(1),'[SO4/Alk]') % left y-axis
ylabel(ax(2),'pH ') % right y-axis
xlabel('Flowback Time (Days)')
legend('[SO4/Alk]','pH','location','best')

%%
t=[0 1 5 14.5 90];
for j=1:1:5;
for i=1:1:19;
   K2(i,j)=isnan(Rat(i,j));
   row=find(K2(:,j)==0);
   y_mean(j)=mean(Rat(row,j),1);
end
end

for j=1:1:5;
for i=1:1:19;
   K3(i,j)=isnan(Rat(i,j));
   row=find(K2(:,j)==0);
   s_std(j)=std(Rat(row,j),1);
end
end

for j=1:1:5;
for i=1:1:19;
   K4(i,j)=isnan(pH(i,j));
   row=find(K4(:,j)==0);
   pH_mean(j)=mean(pH(row,j),1);
end
end

for j=1:1:5;
for i=1:1:19;
   K5(i,j)=isnan(pH(i,j));
   row=find(K5(:,j)==0);
   pH_std(j)=std(pH(row,j),1);
end
end

figure()

[ax,p1,p2] = plotyy(t,log10(y_mean),t,pH_mean,'plot');
hold(ax(1))
hh1=errorbar(ax(1),t,log10(y_mean),log10(s_std),'bo');
```

```
hold(ax(2))
hh2=errorbar(ax(2),t,pH_mean,pH_std,'o');
set(hh2,'Color',[0,0.5 0]);
set(ax,'xlim',[-5,95]);
set(ax(1),'ylim',[-3,9]);
set(ax(1),'YTick',-3:1:9)

set(p1,'LineStyle','-','LineWidth',2)
set(p2,'LineStyle','-','LineWidth',2)
set(hh1,'MarkerFace','b')
set(hh2,'MarkerFace',[0,0.5,0])
title('Time Series of Log([SO_4/Alk]) vs. pH for All 19 Wells');
ylabel(ax(1),'Log([SO_4/Alk])') % left y-axis
ylabel(ax(2),'pH') % right y-axis
xlabel('Flowback Time (Days)')
legend([p1;p2],'Log([SO_4/Alk])','pH');

%% Prework for Piper Diagram
Ba_pre=[Ba(:,2);Ba(:,3);Ba(:,4);Ba(:,5)];
Ca_pre=[Ca(:,2);Ca(:,3);Ca(:,4);Ca(:,5)];
Mg_pre=[Mg(:,2);Mg(:,3);Mg(:,4);Mg(:,5)];
Na_pre=[Na(:,2);Na(:,3);Na(:,4);Na(:,5)];
K_pre=[K(:,2);K(:,3);K(:,4);K(:,5)];
Cl_pre=[Cl(:,2);Cl(:,3);Cl(:,4);Cl(:,5)];
SO4_pre=[SO4(:,2);SO4(:,3);SO4(:,4);SO4(:,5)];
Alk_pre=[Alk(:,2);Alk(:,3);Alk(:,4);Alk(:,5)];
TDS_pre=[TDS(:,2);TDS(:,3);TDS(:,4);TDS(:,5)];
Br_pre=[Br(:,2);Br(:,3);Br(:,4);Br(:,5)];
Piper=[(Ca_pre./1000)./TDS_pre (Mg_pre./1000)./TDS_pre...
    (Na_pre./1000+K_pre./1000)./TDS_pre Cl_pre./TDS_pre  ...
    SO4_pre./TDS_pre Alk_pre./TDS_pre TDS_pre];

Sample_ID = cell(5,1)
Sample_ID ={'Day 1';'Day 5';'Day 14/15';'Day 90'}


%% Prework for Ternary Diagram


Ter=[Ba_pre./1000 SO4_pre Cl_pre];

j=3;
for i=1:1:76;
   L2(i,j)=isnan(Ter(i,j));
   row=find(L2(:,j)==0);
end
Ter_new=[Ba_pre(row)./1000 SO4_pre(row) Cl_pre(row)]
alf= [1 2; 3 4];

aa=normr(Ter_new)

[Pn,ps] = mapminmax(Ter_new,0,1)
```

```
Mapping tool
clc;
clear all;
clf;
close all;

%% Integral Map
GTI_L=xlsread('Location','GTI');
Dresel_L=xlsread('Location','Dresel');
ChiefOG_L=xlsread('Location','ChiefO&G');
PritKir_L=xlsread('Location','PritzKirb');
BOGM_L=xlsread('Location','BOGM');

states = geoshape(shaperead('usastatehi', 'UseGeoCoords', true));
oceanColor = [.5 .7 .9];
latlim = [39.7 42.3];
lonlim = [-80.7 -74.5];

Lat_GTI_PA=[GTI_L(2,1);GTI_L(4:13,1);GTI_L(15:19,1)];
Lon_GTI_PA=[GTI_L(2,2);GTI_L(4:13,2);GTI_L(15:19,2)];
Lat_GTI_WV=[GTI_L(1,1);GTI_L(3,1);GTI_L(14,1)];
Lon_GTI_WV=[GTI_L(1,2);GTI_L(3,2);GTI_L(14,2)];
Lat_Dresel=Dresel_L(:,1);
Lon_Dresel=Dresel_L(:,2);
Lat_ChiefOG=ChiefOG_L(:,1);
Lon_ChiefOG=ChiefOG_L(:,2);
Lat_PritKir=PritKir_L(:,1);
Lon_PritKir=PritKir_L(:,2);
Lat_BOGM=BOGM_L(:,1);
Lon_BOGM=BOGM_L(:,2);

Lat_Flo=[Lat_GTI_PA;Lat_ChiefOG;Lat_PritKir;Lat_BOGM];
Lon_Flo=[Lon_GTI_PA;Lon_ChiefOG;Lon_PritKir;Lon_BOGM];

ax = usamap(latlim, lonlim);
setm(ax, 'FFaceColor', oceanColor);
geoshow(states,'DefaultFaceColor', 'white','DefaultEdgeColor', 'black');
[latlim, lonlim] = bufgeoquad(latlim, lonlim, .05, .05);
% h=geoshow(Lat,
Lon,'DisplayType','point','Color','r','Marker','.','MarkerSize',20);
% title({'Map of Well or Sample Locations Based on the Integral Datasets',
'Pennsylvania'});

% t1=scatterm(Lat1,Lon1,50,'r','d','filled');
t1= geoshow(Lat_Flo,Lon_Flo, 'DisplayType', 'Point', 'Marker', 'o',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
hold on
t2=geoshow(Lat_Dresel,Lon_Dresel, 'DisplayType', 'Point', 'Marker', 's',
'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r','MarkerSize',5);
% t3=geoshow(Lat_ChiefOG,Lon_ChiefOG, 'DisplayType', 'Point', 'Marker', '^',
'MarkerFaceColor', 'g', 'MarkerEdgeColor', 'g','MarkerSize',5);
% t4=geoshow(Lat_PritKir,Lon_PritKir, 'DisplayType', 'Point', 'Marker', 'd',
'MarkerFaceColor', 'c', 'MarkerEdgeColor', 'c','MarkerSize',5);
% t5=geoshow(Lat_BOGM,Lon_BOGM, 'DisplayType', 'Point', 'Marker', 'v',
'MarkerFaceColor', 'm', 'MarkerEdgeColor', 'c','MarkerSize',5);
legend([t1,t2],{'Flowback Water Sampling Locations',...
```

```
    'Oilfield Brines Sampling Locations'});


figure()
states = geoshape(shaperead('usastatehi', 'UseGeoCoords', true));
oceanColor = [.5 .7 .9];
latlim = [37.2 40];
lonlim = [-82.5 -79];

ax = usamap(latlim, lonlim);
setm(ax, 'FFaceColor', oceanColor);
geoshow(states,'DefaultFaceColor', 'white','DefaultEdgeColor', 'black');
[latlim, lonlim] = bufgeoquad(latlim, lonlim, .05, .05);
% h=geoshow(Lat,
Lon,'DisplayType','point','Color','r','Marker','.','MarkerSize',20);
% title({'Map of Well or Sample Locations Based on the Integral Datasets',
'West Virginia'});

% t1=scatterm(Lat1,Lon1,50,'r','d','filled');
t11= geoshow(Lat_GTI_WV,Lon_GTI_WV, 'DisplayType', 'Point', 'Marker', 'o',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
legend([t11],{'Flowback Water Sampling Locations'})

%% BOGM locations
states = geoshape(shaperead('usastatehi', 'UseGeoCoords', true));
oceanColor = [.5 .7 .9];
latlim = [39.7 42.3];
lonlim = [-80.7 -74.5];

ax = usamap(latlim, lonlim);
setm(ax, 'FFaceColor', oceanColor);
geoshow(states,'DefaultFaceColor', 'white','DefaultEdgeColor', 'black');
[latlim, lonlim] = bufgeoquad(latlim, lonlim, .05, .05);
t1= geoshow(Lat_BOGM,Lon_BOGM, 'DisplayType', 'Point', 'Marker', 'o',
'MarkerFaceColor', 'k', 'MarkerEdgeColor', 'k','MarkerSize',5);
legend([t1],{'BOGM Flowback Water Sampling Locations'});

%% Sodium Concentration in Geological Distribution

Na=xlsread('Location','Sodium');
Na_GTI=Na(1:19,:);
Na_GTI_1_PA=[Na_GTI(2,1);Na_GTI(4:13,1);Na_GTI(15:19,1)];
Na_GTI_5_PA=[Na_GTI(2,2);Na_GTI(4:13,2);Na_GTI(15:19,2)];
Na_GTI_14_PA=[Na_GTI(2,3);Na_GTI(4:13,3);Na_GTI(15:19,3)];
Na_GTI_90_PA=[Na_GTI(2,4);Na_GTI(4:13,4);Na_GTI(15:19,4)];
Na_GTI_1_WV=[Na_GTI(1,1);Na_GTI(3,1);Na_GTI(14,1)];
Na_GTI_5_WV=[Na_GTI(1,2);Na_GTI(3,2);Na_GTI(14,2)];
Na_GTI_14_WV=[Na_GTI(1,3);Na_GTI(3,3);Na_GTI(14,3)];
Na_GTI_90_WV=[Na_GTI(1,4);Na_GTI(3,4);Na_GTI(14,4)];
Na_Dresel=Na(20:59,1);

states = geoshape(shaperead('usastatehi', 'UseGeoCoords', true));
oceanColor = [.5 .7 .9];
latlim = [39.7 42.3];
lonlim = [-80.7 -74.5];
```

```
Lat_GTI_PA=[GTI_L(2,1);GTI_L(4:13,1);GTI_L(15:19,1)];
Lon_GTI_PA=[GTI_L(2,2);GTI_L(4:13,2);GTI_L(15:19,2)];
Lat_GTI_WV=[GTI_L(1,1);GTI_L(3,1);GTI_L(14,1)];
Lon_GTI_WV=[GTI_L(1,2);GTI_L(3,2);GTI_L(14,2)];
Lat_Dresel=Dresel_L(:,1);
Lon_Dresel=Dresel_L(:,2);


Lat1_PA=Lat_GTI_PA(find(isnan(Na_GTI_1_PA)==0));
Lon1_PA=Lon_GTI_PA(find(isnan(Na_GTI_1_PA)==0));
Lat5_PA=Lat_GTI_PA(find(isnan(Na_GTI_5_PA)==0));
Lon5_PA=Lon_GTI_PA(find(isnan(Na_GTI_5_PA)==0));
Lat14_PA=Lat_GTI_PA(find(isnan(Na_GTI_14_PA)==0));
Lon14_PA=Lon_GTI_PA(find(isnan(Na_GTI_14_PA)==0));
Lat90_PA=Lat_GTI_PA(find(isnan(Na_GTI_90_PA)==0));
Lon90_PA=Lon_GTI_PA(find(isnan(Na_GTI_90_PA)==0));
Lat0_Dresel=Lat_Dresel(find(isnan(Na_Dresel)==0));
Lon0_Dresel=Lon_Dresel(find(isnan(Na_Dresel)==0));


Na_Dresel=Na_Dresel(find(isnan(Na_Dresel)==0));
for i=1:length(Lat0_Dresel);
    if Na_Dresel(i)>0&&Na_Dresel(i)<=10000;
        i1(i)=i;
    elseif Na_Dresel(i)>10000&&Na_Dresel(i)<=30000;
        i2(i)=i;
    elseif Na_Dresel(i)>30000&&Na_Dresel(i)<=60000;
        i3(i)=i;
    elseif Na_Dresel(i)>60000;
        i4(i)=i;
    end
end

Lat1=Lat0_Dresel(i1(find(i1~=0)));
Lon1=Lon0_Dresel(i1(find(i1~=0)));
Lat2=Lat0_Dresel(i2(find(i2~=0)));
Lon2=Lon0_Dresel(i2(find(i2~=0)));
Lat3=Lat0_Dresel(i3(find(i3~=0)));
Lon3=Lon0_Dresel(i3(find(i3~=0)));
Lat4=Lat0_Dresel(i4(find(i4~=0)));
Lon4=Lon0_Dresel(i4(find(i4~=0)));

% 14 day
Na_GTI_14_PA=Na_GTI_14_PA(find(isnan(Na_GTI_14_PA)==0));
for i=1:length(Lat14_PA);
    if Na_GTI_14_PA(i)>0&&Na_GTI_14_PA(i)<=10000;
        i11(i)=i;
    elseif Na_GTI_14_PA(i)>10000&&Na_GTI_14_PA(i)<=30000;
        i22(i)=i;
    elseif Na_GTI_14_PA(i)>30000&&Na_GTI_14_PA(i)<=60000;
        i33(i)=i;
    elseif Na_GTI_14_PA(i)>60000;
        i44(i)=i;
    end
```

```
end
Lat11=Lat14_PA(i11(find(i11~=0)));
Lon11=Lon14_PA(i11(find(i11~=0)));
Lat22=Lat14_PA(i22(find(i22~=0)));
Lon22=Lon14_PA(i22(find(i22~=0)));
Lat33=Lat14_PA(i33(find(i33~=0)));
Lon33=Lon14_PA(i33(find(i33~=0)));
Lat44=Lat14_PA(i44(find(i44~=0)));
Lon44=Lon14_PA(i44(find(i44~=0)));
% 14 Day
ax = usamap(latlim, lonlim);
setm(ax, 'FFaceColor', oceanColor);
geoshow(states,'DefaultFaceColor', 'white','DefaultEdgeColor', 'black');
[latlim, lonlim] = bufgeoquad(latlim, lonlim, .05, .05);
% title({'Sodium Concentraiton of Oilfield Brine and GTI Flowback Sampled at
90 Days', 'Pennsylvania'});
t1= geoshow(Lat1,Lon1, 'DisplayType', 'Point', 'Marker', 'o',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
hold on
t2=geoshow(Lat2,Lon2, 'DisplayType', 'Point', 'Marker', 'd',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
t3=geoshow(Lat3,Lon3, 'DisplayType', 'Point', 'Marker', '^',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
t4=geoshow(Lat4,Lon4, 'DisplayType', 'Point', 'Marker', 'V',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
tt1=geoshow(Lat11,Lon11, 'DisplayType', 'Point', 'Marker', 'o',
'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r','MarkerSize',5);
tt2=geoshow(Lat22,Lon22, 'DisplayType', 'Point', 'Marker', 'd',
'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r','MarkerSize',5);
tt3=geoshow(Lat33,Lon33, 'DisplayType', 'Point', 'Marker', '^',
'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r','MarkerSize',5);
tt4=geoshow(Lat44,Lon44, 'DisplayType', 'Point', 'Marker', 'V',
'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r','MarkerSize',5);
legend([t1,t2,t3,t4,tt1,tt2,tt3,tt4],{'Oilfiled Brine Na  0 - 10000 mg/L',...
     'Oilfiled Brine Na 10000 - 30000 mg/L'...
    ,'Oilfiled Brine Na  30000 - 60000 mg/L'...
    ,'Oilfiled Brine Na  > 60000 mg/L' ...
    ,'GTI Flowback 14 or 15 Day Na 0 - 10000 mg/L'...
    ,'GTI Flowback 14 or 15 Day Na 10000 - 30000 mg/L'...
    ,'GTI Flowback 14 or 15 Day Na  30000 - 60000 mg/L'...
    ,'GTI Flowback 14 or 15 Day Na  > 60000 mg/L'});

% 90 days
Na_GTI_90_PA=Na_GTI_90_PA(find(isnan(Na_GTI_90_PA)==0));
for i=1:length(Lat90_PA);
    if Na_GTI_90_PA(i)>10000&&Na_GTI_90_PA(i)<=30000;
        i22(i)=i;
    elseif Na_GTI_90_PA(i)>30000&&Na_GTI_90_PA(i)<=60000;
        i33(i)=i;
    elseif Na_GTI_90_PA(i)>60000;
        i44(i)=i;
    end
end

Lat22=Lat90_PA(i22(find(i22~=0)));
Lon22=Lon90_PA(i22(find(i22~=0)));
```

```
Lat33=Lat90_PA(i33(find(i33~=0)));
Lon33=Lon90_PA(i33(find(i33~=0)));
Lat44=Lat90_PA(i44(find(i44~=0)));
Lon44=Lon90_PA(i44(find(i44~=0)));
% 90 day
ax = usamap(latlim, lonlim);
setm(ax, 'FFaceColor', oceanColor);
geoshow(states,'DefaultFaceColor', 'white','DefaultEdgeColor', 'black');
[latlim, lonlim] = bufgeoquad(latlim, lonlim, .05, .05);
% title({'Sodium Concentraiton of Oilfield Brine and GTI Flowback Sampled at
90 Days', 'Pennsylvania'});
t1= geoshow(Lat1,Lon1, 'DisplayType', 'Point', 'Marker', 'o',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
hold on
t2=geoshow(Lat2,Lon2, 'DisplayType', 'Point', 'Marker', 'd',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
t3=geoshow(Lat3,Lon3, 'DisplayType', 'Point', 'Marker', '^',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
t4=geoshow(Lat4,Lon4, 'DisplayType', 'Point', 'Marker', 'V',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);

tt2=geoshow(Lat22,Lon22, 'DisplayType', 'Point', 'Marker', 'd',
'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r','MarkerSize',5);
tt3=geoshow(Lat33,Lon33, 'DisplayType', 'Point', 'Marker', '^',
'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r','MarkerSize',5);
tt4=geoshow(Lat44,Lon44, 'DisplayType', 'Point', 'Marker', 'V',
'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r','MarkerSize',5);
legend([t1,t2,t3,t4,tt2,tt3,tt4],{'Oilfiled Brine Na  0 - 10000 mg/L',...
     'Oilfiled Brine Na 10000 - 30000 mg/L'...
    ,'Oilfiled Brine Na  30000 - 60000 mg/L'...
    ,'Oilfiled Brine Na  > 60000 mg/L' ...
     'GTI Flowback 90 Day Na 10000 - 30000 mg/L'...
    ,'GTI Flowback 90 Day Na  30000 - 60000 mg/L'...
    ,'GTI Flowback 90 Day Na  > 60000 mg/L'});


%% Chloride Concentration in Geological Distribution
Cl=xlsread('Location','Chloride');
Cl_GTI=Cl(1:19,:);
Cl_GTI_1_PA=[Cl_GTI(2,1);Cl_GTI(4:13,1);Cl_GTI(15:19,1)];
Cl_GTI_5_PA=[Cl_GTI(2,2);Cl_GTI(4:13,2);Cl_GTI(15:19,2)];
Cl_GTI_14_PA=[Cl_GTI(2,3);Cl_GTI(4:13,3);Cl_GTI(15:19,3)];
Cl_GTI_90_PA=[Cl_GTI(2,4);Cl_GTI(4:13,4);Cl_GTI(15:19,4)];
Cl_GTI_1_WV=[Cl_GTI(1,1);Cl_GTI(3,1);Cl_GTI(14,1)];
Cl_GTI_5_WV=[Cl_GTI(1,2);Cl_GTI(3,2);Cl_GTI(14,2)];
Cl_GTI_14_WV=[Cl_GTI(1,3);Cl_GTI(3,3);Cl_GTI(14,3)];
Cl_GTI_90_WV=[Cl_GTI(1,4);Cl_GTI(3,4);Cl_GTI(14,4)];
Cl_Dresel=Cl(20:59,1);

states = geoshape(shaperead('usastatehi', 'UseGeoCoords', true));
oceanColor = [.5 .7 .9];
latlim = [39.7 42.3];
lonlim = [-80.7 -74.5];

Lat_GTI_PA=[GTI_L(2,1);GTI_L(4:13,1);GTI_L(15:19,1)];
Lon_GTI_PA=[GTI_L(2,2);GTI_L(4:13,2);GTI_L(15:19,2)];
```

```
Lat_GTI_WV=[GTI_L(1,1);GTI_L(3,1);GTI_L(14,1)];
Lon_GTI_WV=[GTI_L(1,2);GTI_L(3,2);GTI_L(14,2)];
Lat_Dresel=Dresel_L(:,1);
Lon_Dresel=Dresel_L(:,2);


Lat1_PA=Lat_GTI_PA(find(isnan(Cl_GTI_1_PA)==0));
Lon1_PA=Lon_GTI_PA(find(isnan(Cl_GTI_1_PA)==0));
Lat5_PA=Lat_GTI_PA(find(isnan(Cl_GTI_5_PA)==0));
Lon5_PA=Lon_GTI_PA(find(isnan(Cl_GTI_5_PA)==0));
Lat14_PA=Lat_GTI_PA(find(isnan(Cl_GTI_14_PA)==0));
Lon14_PA=Lon_GTI_PA(find(isnan(Cl_GTI_14_PA)==0));
Lat90_PA=Lat_GTI_PA(find(isnan(Cl_GTI_90_PA)==0));
Lon90_PA=Lon_GTI_PA(find(isnan(Cl_GTI_90_PA)==0));
Lat0_Dresel=Lat_Dresel(find(isnan(Cl_Dresel)==0));
Lon0_Dresel=Lon_Dresel(find(isnan(Cl_Dresel)==0));


Cl_Dresel=Cl_Dresel(find(isnan(Cl_Dresel)==0));
for i=1:length(Lat0_Dresel);
    if Cl_Dresel(i)>0&&Cl_Dresel(i)<=50000;
        i1(i)=i;
    elseif Cl_Dresel(i)>50000&&Cl_Dresel(i)<=100000;
        i2(i)=i;
    elseif Cl_Dresel(i)>100000&&Cl_Dresel(i)<=150000;
        i3(i)=i;
    elseif Cl_Dresel(i)>150000;
        i4(i)=i;
    end
end

Lat1=Lat0_Dresel(i1(find(i1~=0)));
Lon1=Lon0_Dresel(i1(find(i1~=0)));
Lat2=Lat0_Dresel(i2(find(i2~=0)));
Lon2=Lon0_Dresel(i2(find(i2~=0)));
Lat3=Lat0_Dresel(i3(find(i3~=0)));
Lon3=Lon0_Dresel(i3(find(i3~=0)));
Lat4=Lat0_Dresel(i4(find(i4~=0)));
Lon4=Lon0_Dresel(i4(find(i4~=0)));

% 14
Cl_GTI_14_PA=Cl_GTI_14_PA(find(isnan(Cl_GTI_14_PA)==0));
for i=1:length(Lat14_PA);
    if Cl_GTI_14_PA(i)>0&&Cl_GTI_14_PA(i)<=50000;
        i11(i)=i;
    elseif Cl_GTI_14_PA(i)>50000&&Cl_GTI_14_PA(i)<=100000;
        i22(i)=i;
    elseif Cl_GTI_14_PA(i)>100000&&Cl_GTI_14_PA(i)<=150000;
        i33(i)=i;
    elseif Cl_GTI_14_PA(i)>150000;
        i44(i)=i;
    end
end
Lat11=Lat14_PA(i11(find(i11~=0)));
Lon11=Lon14_PA(i11(find(i11~=0)));
Lat22=Lat14_PA(i22(find(i22~=0)));
```

```
Lon22=Lon14_PA(i22(find(i22~=0)));
Lat33=Lat14_PA(i33(find(i33~=0)));
Lon33=Lon14_PA(i33(find(i33~=0)));
Lat44=Lat14_PA(i44(find(i44~=0)));
Lon44=Lon14_PA(i44(find(i44~=0)));

ax = usamap(latlim, lonlim);
setm(ax, 'FFaceColor', oceanColor);
geoshow(states,'DefaultFaceColor', 'white','DefaultEdgeColor', 'black');
[latlim, lonlim] = bufgeoquad(latlim, lonlim, .05, .05);
% title({'Chloride Concentraiton of Oilfield Brine and GTI Flowback Sampled
at 90 Days', 'Pennsylvania'});
t1= geoshow(Lat1,Lon1, 'DisplayType', 'Point', 'Marker', 'o',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
hold on
t2=geoshow(Lat2,Lon2, 'DisplayType', 'Point', 'Marker', 'd',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
t3=geoshow(Lat3,Lon3, 'DisplayType', 'Point', 'Marker', '^',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
t4=geoshow(Lat4,Lon4, 'DisplayType', 'Point', 'Marker', 'V',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
tt1=geoshow(Lat11,Lon11, 'DisplayType', 'Point', 'Marker', 'o',
'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r','MarkerSize',5);
tt2=geoshow(Lat22,Lon22, 'DisplayType', 'Point', 'Marker', 'd',
'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r','MarkerSize',5);
tt3=geoshow(Lat33,Lon33, 'DisplayType', 'Point', 'Marker', '^',
'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r','MarkerSize',5);
tt4=geoshow(Lat44,Lon44, 'DisplayType', 'Point', 'Marker', 'V',
'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r','MarkerSize',5);
legend([t1,t2,t3,t4,tt1,tt2,tt3,tt4],{'Oilfiled Brine Cl  0 - 50000 mg/L',...
     'Oilfiled Brine Cl 50000 - 100000 mg/L'...
    ,'Oilfiled Brine Cl 100000 - 150000 mg/L'...
    ,'Oilfiled Brine Cl  > 150000 mg/L' ...
    ,'GTI Flowback 14 or 15 Day Cl 0 - 50000 mg/L'...
    ,'GTI Flowback 14 or 15 Day Cl 50000 - 100000 mg/L'...
    ,'GTI Flowback 14 or 15 Day Cl 100000 - 150000 mg/L'...
    ,'GTI Flowback 14 or 15 Day Cl  > 150000 mg/L'});

% 90
Cl_GTI_90_PA=Cl_GTI_90_PA(find(isnan(Cl_GTI_90_PA)==0));
for i=1:length(Lat90_PA);
    if Cl_GTI_90_PA(i)>50000&&Cl_GTI_90_PA(i)<=100000;
        i22(i)=i;
    elseif Cl_GTI_90_PA(i)>100000&&Cl_GTI_90_PA(i)<=150000;
        i33(i)=i;
    elseif Cl_GTI_90_PA(i)>150000;
        i44(i)=i;
    end
end

Lat22=Lat90_PA(i22(find(i22~=0)));
Lon22=Lon90_PA(i22(find(i22~=0)));
Lat33=Lat90_PA(i33(find(i33~=0)));
Lon33=Lon90_PA(i33(find(i33~=0)));
Lat44=Lat90_PA(i44(find(i44~=0)));
Lon44=Lon90_PA(i44(find(i44~=0)));
```

```
ax = usamap(latlim, lonlim);
setm(ax, 'FFaceColor', oceanColor);
geoshow(states,'DefaultFaceColor', 'white','DefaultEdgeColor', 'black');
[latlim, lonlim] = bufgeoquad(latlim, lonlim, .05, .05);
% title({'Chloride Concentraiton of Oilfield Brine and GTI Flowback Sampled
at 90 Days', 'Pennsylvania'});
t1= geoshow(Lat1,Lon1, 'DisplayType', 'Point', 'Marker', 'o',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
hold on
t2=geoshow(Lat2,Lon2, 'DisplayType', 'Point', 'Marker', 'd',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
t3=geoshow(Lat3,Lon3, 'DisplayType', 'Point', 'Marker', '^',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
t4=geoshow(Lat4,Lon4, 'DisplayType', 'Point', 'Marker', 'V',
'MarkerFaceColor', 'b', 'MarkerEdgeColor', 'b','MarkerSize',5);
% tt1=geoshow(Lat11,Lon11, 'DisplayType', 'Point', 'Marker', 'o',
'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r','MarkerSize',5);
tt2=geoshow(Lat22,Lon22, 'DisplayType', 'Point', 'Marker', 'd',
'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r','MarkerSize',5);
tt3=geoshow(Lat33,Lon33, 'DisplayType', 'Point', 'Marker', '^',
'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r','MarkerSize',5);
tt4=geoshow(Lat44,Lon44, 'DisplayType', 'Point', 'Marker', 'V',
'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r','MarkerSize',5);
legend([t1,t2,t3,t4,tt2,tt3,tt4],{'Oilfiled Brine Cl  0 - 50000 mg/L',...
     'Oilfiled Brine Cl 50000 - 100000 mg/L'...
    ,'Oilfiled Brine Cl 100000 - 150000 mg/L'...
    ,'Oilfiled Brine Cl  > 150000 mg/L' ...
    'GTI Flowback 90 Day Cl 50000 - 100000 mg/L'...
    ,'GTI Flowback 90 Day Cl 100000 - 150000 mg/L'...
    ,'GTI Flowback 90 Day Cl  > 150000 mg/L'});

Dresel (2010)
clc;
clear all;
clf;
close all;
%%
Dresel_Raw1=xlsread('DreselData','Raw1');
save ('Dresel_Raw1')
Dresel_Raw2=xlsread('DreselData','Raw2');
save ('Dresel_Raw2')
%% Na-Cl-TDS systematics
load ('Dresel_Raw1')
load ('Dresel_Raw2')

TDS=Dresel_Raw1(:,3)*10^3;
Cl=Dresel_Raw2(:,10);
Na=Dresel_Raw1(:,9);

figure(1)
h=scatter(TDS,Cl,'filled','r^');
hold on
hh=scatter(TDS,Na,'filled','bv');
xlabel('Concentration of TDS (mg/L)')
ylabel('Concentration of Ions (mg/L)')
```

```
mdl1 = LinearModel.fit(TDS,Cl,'linear','RobustOpts','on');
hold on
h1=plot(mdl1);
set(h1,'color','red');
mdl2=LinearModel.fit(TDS,Na,'linear','RobustOpts','on');
h2=plot(mdl2);
set(h2,'color','blue');

xlabel('Concentration of TDS (mg/L)')
ylabel('Concentration of Each Ion(mg/L)')
title('Sodium vs. Chloride within the indication of TDS')
text(3*10^5,1.5*10^5,'C_{Cl} = 0.56995C_{TDS} -1460.1; R^2=0.98')
text(3*10^5,0.6*10^5,'C_{Na} = 0.24487C_{TDS} + 511.39; R^2=0.979')

legend([h,hh],{'Chloride','Sodium'},'location','best');

%% Cl-Br systematics Compiled with Flowback
Cl_Dre=Dresel_Raw2(:,10);
Br_Dre=Dresel_Raw2(:,11);
Br_Dr=Br_Dre(find(isnan(Br_Dre)==0))
Cl_Dr=Cl_Dre(find(isnan(Br_Dre)==0))
% Flowback
Ha2=xlsread('Data(Hayes2009)_PAN','Sheet4');
Br_Flo=Ha2(1:4:73,1:5);
Ha_Ions=xlsread('Data(Hayes2009)_PAN','sheet1');
Cl_Flo=Ha_Ions(4:11:202,1:5);

Br_Day14=Br_Flo(:,4);
Br_Day14=Br_Day14(find(isnan(Br_Day14)==0));
Br_Day90=Br_Flo(:,5);
Br_Day90=Br_Day90(find(isnan(Br_Day90)==0));
Cl_Day14=Cl_Flo(:,4);
Cl_Day14=Cl_Day14(find(isnan(Cl_Day14)==0));
Cl_Day90=Cl_Flo(:,5);
Cl_Day90=Cl_Day90(find(isnan(Cl_Day90)==0));

figure()
h3=plot(log10(Br_Dr),log10(Cl_Dr),'bs');
set(h3,'MarkerFaceColor','blue','MarkerSize',5)
mdl1=LinearModel.fit(log10(Br_Dr),log10(Cl_Dr));
hold on
% plotDiagnostics(mdl1,'cookd')
% plotResiduals(mdl1)
% plotResiduals(mdl1,'probability')
larg=find((mdl1.Diagnostics.CooksDistance)>3*mean(mdl1.Diagnostics.CooksDista
nce));
mdl21 = LinearModel.fit(log10(Br_Dr),log10(Cl_Dr),'Exclude',larg);
hold on
L1=plot(mdl21)
set(L1,'color','blue','LineWidth',1.5);


h4=plot(log10(Br_Day14),log10(Cl_Day14),'rs');
set(h4,'MarkerFaceColor','red','MarkerSize',5)
```

```
mdl2=LinearModel.fit(log10(Br_Day14),log10(Cl_Day14));
% plotDiagnostics(mdl2,'cookd')
larg=find((mdl2.Diagnostics.CooksDistance)>3*mean(mdl2.Diagnostics.CooksDista
nce));
mdl22 = LinearModel.fit(log10(Br_Day14),log10(Cl_Day14),'Exclude',larg);
hold on
L2=plot(mdl22)
set(L2,'color','red','LineWidth',2);

h5=plot(log10(Br_Day90),log10(Cl_Day90),'s');
set(h5,'MarkerEdgeColor',[0 0.5 0],'MarkerFaceColor',[0 0.5
0],'MarkerSize',5)
mdl3=LinearModel.fit(log10(Br_Day90),log10(Cl_Day90));
% plotDiagnostics(mdl3,'cookd')
larg=find((mdl3.Diagnostics.CooksDistance)>3*mean(mdl3.Diagnostics.CooksDista
nce));
mdl23 = LinearModel.fit(log10(Br_Day90),log10(Cl_Day90),'Exclude',larg);
hold on
L3=plot(mdl23)
set(L3,'color',[0 0.5 0],'LineWidth',2.5);

xlabel('Log(Br)','FontSize',15)
ylabel('Log(Cl)','FontSize',15)

legend([h3,h4,h5],{'Oilfield Brine','Flowback Water Day 14 or 15','Flowback
Water Day 90'},'location','best')

title('Linear Regression Model for Cl-Br Systematics','FontSize',15)

%% Other

Br_Dre=Dresel_Raw2(:,11);
Br_Dr=Br_Dre(find(isnan(Br_Dre)==0));
Ca_Dre=Dresel_Raw1(:,12);
Ca_Dr=Ca_Dre(find(isnan(Br_Dre)==0));
Ca_D=Ca_Dr(find(isnan(Ca_Dr)==0));
Br_D=Br_Dr(find(isnan(Ca_Dr)==0));

Ha_Ions=xlsread('Data(Hayes2009)_PAN','sheet1');
Na_Flo=Ha_Ions(1:11:199,1:5);
K_Flo=Ha_Ions(2:11:200,1:5);
Ca_Flo=Ha_Ions(3:11:201,1:5);
SO4_Flo=Ha_Ions(5:11:203,1:5);
Ba_Flo=Ha_Ions(6:11:204,1:5);
Mg_Flo=Ha_Ions(11:11:209,1:5);


% Flowback
Br_Day14=Br_Flo(:,4);
Br_Day14=Br_Day14(find(isnan(Br_Day14)==0));
Br_Day90=Br_Flo(:,5);
Br_Day90=Br_Day90(find(isnan(Br_Day90)==0));
Ca_Day14=Ca_Flo(:,4);
Ca_Day14=Ca_Day14(find(isnan(Ca_Day14)==0));
Ca_Day90=Ca_Flo(:,5);
```

```
Ca_Day90=Ca_Day90(find(isnan(Ca_Day90)==0));
Mg_Day14=Mg_Flo(:,4);
Mg_Day14=Mg_Day14(find(isnan(Mg_Day14)==0));
Mg_Day90=Mg_Flo(:,5);
Mg_Day90=Mg_Day90(find(isnan(Mg_Day90)==0));
Na_Day14=Na_Flo(:,4);
Na_Day14=Na_Day14(find(isnan(Na_Day14)==0));
Na_Day90=Na_Flo(:,5);
Na_Day90=Na_Day90(find(isnan(Na_Day90)==0));

% Ca-Br
figure()
hh3=plot(log10(Br_D),log10(Ca_D),'bs');
set(hh3,'MarkerFaceColor','blue','MarkerSize',5)
hold on
mdl1=LinearModel.fit(log10(Br_D),log10(Ca_D));
% plot(mdl1)
% plotDiagnostics(mdl1,'cookd')
% plotResiduals(mdl1)
% plotResiduals(mdl1,'probability')
larg=find((mdl1.Diagnostics.CooksDistance)>3*mean(mdl1.Diagnostics.CooksDista
nce));
mdl21 = LinearModel.fit(log10(Br_D),log10(Ca_D),'Exclude',larg);
hold on
L1=plot(mdl21)
set(L1,'color','blue','LineWidth',1.5);


hh4=plot(log10(Br_Day14),log10(Ca_Day14./1000),'rs');
set(hh4,'MarkerFaceColor','red','MarkerSize',5)
mdl2=LinearModel.fit(log10(Br_Day14),log10(Ca_Day14./1000));
% plotResiduals(mdl2,'probability')
% plotDiagnostics(mdl2,'cookd')
larg=find((mdl2.Diagnostics.CooksDistance)>3*mean(mdl2.Diagnostics.CooksDista
nce));
mdl22 =
LinearModel.fit(log10(Br_Day14),log10(Ca_Day14./1000),'Exclude',larg);
hold on
L2=plot(mdl22)
set(L2,'color','red','LineWidth',2);

hh5=plot(log10(Br_Day90),log10(Ca_Day90./1000),'s');
set(hh5,'MarkerEdgeColor',[0 0.5 0],'MarkerFaceColor',[0 0.5
0],'MarkerSize',5)
mdl3=LinearModel.fit(log10(Br_Day90),log10(Ca_Day90./1000));
% plotResiduals(mdl3,'probability')
% plotDiagnostics(mdl3,'cookd')
larg=find((mdl3.Diagnostics.CooksDistance)>3*mean(mdl3.Diagnostics.CooksDista
nce));
mdl23 =
LinearModel.fit(log10(Br_Day90),log10(Ca_Day90./1000),'Exclude',larg);
hold on
L3=plot(mdl23)
set(L3,'color',[0 0.5 0],'LineWidth',2.5);

xlabel('Log(Br)','FontSize',15)
```

```
ylabel('Log(Ca)','FontSize',15)

legend([hh3,hh4,hh5],{'Oilfield Brine','Flowback Water Day 14 or
15','Flowback Water Day 90'},'location','best','FontSize',15)

title('Linear Regression Model for Ca-Br Systematics','FontSize',15)

% Mg-Br
Mg_Dre=Dresel_Raw1(:,11);
Mg_Dr=Mg_Dre(find(isnan(Br_Dre)==0));
Mg_D=Mg_Dr(find(isnan(Mg_Dr)==0));
Br_D=Br_Dr(find(isnan(Mg_Dr)==0));

figure()
hh3=plot(log10(Br_D),log10(Mg_D),'bs');
set(hh3,'MarkerFaceColor','blue','MarkerSize',5)
hold on
mdl1=LinearModel.fit(log10(Br_D),log10(Mg_D));
% plotResiduals(mdl1,'probability')
% plotDiagnostics(mdl1,'cookd')
outl = find(mdl1.Residuals.Raw > 0.2 | mdl1.Residuals.Raw < -0.2)
mdl21 = LinearModel.fit(log10(Br_D),log10(Mg_D),'Exclude',outl);
hold on
L1=plot(mdl21)
set(L1,'color','blue','LineWidth',1.5);


hh4=plot(log10(Br_Day14),log10(Mg_Day14./1000),'rs');
set(hh4,'MarkerFaceColor','red','MarkerSize',5)
mdl2=LinearModel.fit(log10(Br_Day14),log10(Mg_Day14./1000));
% plotResiduals(mdl2,'probability')
% plotDiagnostics(mdl2,'cookd')
larg=find((mdl2.Diagnostics.CooksDistance)>3*mean(mdl2.Diagnostics.CooksDista
nce));
mdl22 =
LinearModel.fit(log10(Br_Day14),log10(Mg_Day14./1000),'Exclude',larg);
hold on
L2=plot(mdl22)
set(L2,'color','red','LineWidth',2);

hh5=plot(log10(Br_Day90),log10(Mg_Day90./1000),'s');
set(hh5,'MarkerEdgeColor',[0 0.5 0],'MarkerFaceColor',[0 0.5
0],'MarkerSize',5)
mdl3=LinearModel.fit(log10(Br_Day90),log10(Mg_Day90./1000));
% plotResiduals(mdl3,'probability')
% plotDiagnostics(mdl3,'cookd')
larg=find((mdl3.Diagnostics.CooksDistance)>3*mean(mdl3.Diagnostics.CooksDista
nce));
mdl23 =
LinearModel.fit(log10(Br_Day90),log10(Mg_Day90./1000),'Exclude',larg);
hold on
L3=plot(mdl23)
set(L3,'color',[0 0.5 0],'LineWidth',2.5);

xlabel('Log(Br)','FontSize',15)
```

```
ylabel('Log(Mg)','FontSize',15)

legend([hh3,hh4,hh5],{'Oilfield Brine','Flowback Water Day 14 or
15','Flowback Water Day 90'},'location','best','FontSize',15)

title('Linear Regression Model for Mg-Br Systematics','FontSize',15)

% Na-Br
Na_Dre=Dresel_Raw1(:,9);
Na_Dr=Na_Dre(find(isnan(Br_Dre)==0));
Na_D=Na_Dr(find(isnan(Na_Dr)==0));
Br_D=Br_Dr(find(isnan(Na_Dr)==0));

figure()
hh3=plot(log10(Br_D),log10(Na_D),'bs');
set(hh3,'MarkerFaceColor','blue','MarkerSize',5)
hold on
mdl1=LinearModel.fit(log10(Br_D),log10(Na_D));
% plotResiduals(mdl1,'probability')
% plotDiagnostics(mdl1,'cookd')
larg=find((mdl1.Diagnostics.CooksDistance)>3*mean(mdl1.Diagnostics.CooksDista
nce));
outl = find(mdl1.Residuals.Raw > 0.15|mdl1.Residuals.Raw < -0.3)
mdl21 = LinearModel.fit(log10(Br_D),log10(Na_D),'Exclude',outl);
hold on
L1=plot(mdl21)
set(L1,'color','blue','LineWidth',1.5);


hh4=plot(log10(Br_Day14),log10(Na_Day14./1000),'rs');
set(hh4,'MarkerFaceColor','red','MarkerSize',5)
mdl2=LinearModel.fit(log10(Br_Day14),log10(Na_Day14./1000))
% plotResiduals(mdl2,'probability')
% plotDiagnostics(mdl2,'cookd')
larg=find((mdl2.Diagnostics.CooksDistance)>3*mean(mdl2.Diagnostics.CooksDista
nce));
outl = find(mdl2.Residuals.Raw > 0.2)
mdl22 =
LinearModel.fit(log10(Br_Day14),log10(Na_Day14./1000),'Exclude',outl);
hold on
L2=plot(mdl22)
set(L2,'color','red','LineWidth',2);

hh5=plot(log10(Br_Day90),log10(Na_Day90./1000),'s');
set(hh5,'MarkerEdgeColor',[0 0.5 0],'MarkerFaceColor',[0 0.5
0],'MarkerSize',5)
mdl3=LinearModel.fit(log10(Br_Day90),log10(Na_Day90./1000));
% plotResiduals(mdl3,'probability')
% plotDiagnostics(mdl3,'cookd')
larg=find((mdl3.Diagnostics.CooksDistance)>3*mean(mdl3.Diagnostics.CooksDista
nce));
outl = find(mdl3.Residuals.Raw > 0.2)
mdl23 =
LinearModel.fit(log10(Br_Day90),log10(Na_Day90./1000),'Exclude',larg);
hold on
```

```
L3=plot(mdl23)
set(L3,'color',[0 0.5 0],'LineWidth',2.5);

xlabel('Log(Br)','FontSize',15)
ylabel('Log(Na)','FontSize',15)

legend([hh3,hh4,hh5],{'Oilfield Brine','Flowback Water Day 14 or
15','Flowback Water Day 90'},'location','best','FontSize',15)

title('Linear Regression Model for Na-Br Systematics','FontSize',15)
```

BOGM

```
clc;
clear all;
clf;
close all;
%% Linear-Regression [Sodim vs. Chloride with Increment of TDS]
BG_data1=xlsread('PreliminaryAnalysisBOGM.XLSX','MajorComp');
KP_Well=xlsread('KirbyPritz.XLSX','Sheet1');
save ('BG_data1')
save('KP_Well')
%% Na-Cl-TDS systematics
load ('BG_data1')
load ('Dresel_Raw1')
load ('Dresel_Raw2')
load ('KP_Well')

TDS_Dre=Dresel_Raw1(:,3)*10^3;
Cl_Dre=Dresel_Raw2(:,10);
Na_Dre=Dresel_Raw1(:,9);

TDS_BG=BG_data1(:,8);
Na_BG=BG_data1(:,1);
Cl_BG=BG_data1(:,4);

TDS_KP=KP_Well(:,5);
Cl_KP=KP_Well(:,18);
Na_KP=KP_Well(:,15);

TDS_Dr=TDS_Dre(find(isnan(TDS_Dre)==0));
Cl_Dr=Cl_Dre(find(isnan(TDS_Dre)==0));
Cl_D=Cl_Dr(find(isnan(Cl_Dr)==0));
TDS_D=TDS_Dr(find(isnan(Cl_Dr)==0));

TDS_B=TDS_BG(find(isnan(TDS_BG)==0));
Cl_B=Cl_BG(find(isnan(TDS_BG)==0));
Cl_B1=Cl_B(find(isnan(Cl_B)==0));
TDS_B1=TDS_B(find(isnan(Cl_B)==0));

TDS_K=TDS_KP(find(isnan(TDS_KP)==0));
Cl_K=Cl_KP(find(isnan(TDS_KP)==0));
Cl_K1=Cl_K(find(isnan(Cl_K)==0));
```

```
TDS_K1=TDS_K(find(isnan(Cl_K)==0));

figure()
l=scatter(TDS_B1,Cl_B1,'ro','filled');
hold on
m=scatter(TDS_K1,Cl_K1,'rd','filled')
h=scatter(TDS_Dre,Cl_Dre,'r^');
mdl12=LinearModel.fit(TDS_D,Cl_D);
[~,larg] = max(mdl12.Diagnostics.CooksDistance);
mdl2 = LinearModel.fit(TDS_D,Cl_D,'Exclude',larg);
L2=plot(mdl2)
set(L2,'color','red');



TDS_Dr=TDS_Dre(find(isnan(TDS_Dre)==0));
Na_Dr=Na_Dre(find(isnan(TDS_Dre)==0));
Na_D=Na_Dr(find(isnan(Na_Dr)==0));
TDS_D=TDS_Dr(find(isnan(Na_Dr)==0));

TDS_B=TDS_BG(find(isnan(TDS_BG)==0));
Na_B=Na_BG(find(isnan(TDS_BG)==0));
Na_B1=Na_B(find(isnan(Na_B)==0));
TDS_B1=TDS_B(find(isnan(Na_B)==0));

TDS_K=TDS_KP(find(isnan(TDS_KP)==0));
Na_K=Na_KP(find(isnan(TDS_KP)==0));
Na_K1=Na_K(find(isnan(Na_K)==0));
TDS_K1=TDS_K(find(isnan(Na_K)==0));



ll=scatter(TDS_B1,Na_B1,'bo','filled');
hh=scatter(TDS_D,Na_D,'bd');
mm=scatter(TDS_K1,Na_K1,'b^','filled');
mdl11=LinearModel.fit(TDS_D,Na_D);
[~,larg] = max(mdl11.Diagnostics.CooksDistance);
mdl1 = LinearModel.fit(TDS_D,Na_D,'Exclude',larg);
hold on
L1=plot(mdl1)
set(L1,'color','blue');

x1=0:0.4*10^5:3.8*10^5;
y1=0.57012.*x1-1579.4;
y2=0.24432.*x1+670.4;
plot(x1,y1,'r--')
plot(x1,y2,'b--')
ylim([0, 2.5*10^5]);
legend([l,m,h,ll,mm,hh],{'Cl-BOGM','Cl-Kirby&Pritz','Cl-Oilfield Brines','Na-
BOGM','Na-Kirby&Pritz','Na-Oilfiled
Brines'},'location','best','FontSize',20);
xlabel('Concentration of TDS (mg/L)','FontSize',20)
ylabel('Concentration of Each Ions (mg/L)','FontSize',20)
title('Na vs. Cl along with the indication of TDS','FontSize',20)
```

```
%% Br-Cl Br-Na Br-Ca Br-Mg Classification
Cl_Dre=Dresel_Raw2(:,10);
Br_Dre=Dresel_Raw2(:,11);
Br_Dr=Br_Dre(find(isnan(Br_Dre)==0))
Cl_Dr=Cl_Dre(find(isnan(Br_Dre)==0))

% Flowback
Br_BG=BG_data1(:,11);
Na_BG=BG_data1(:,1);
Cl_BG=BG_data1(:,4);
Ca_BG=BG_data1(:,3);
Mg_BG=BG_data1(:,9);

% Cl-Br
Br_B=Br_BG(find(isnan(Br_BG)==0));
Cl_B=Cl_BG(find(isnan(Br_BG)==0));
Cl_B1=Cl_B(find(isnan(Cl_B)==0));
Br_B1=Br_B(find(isnan(Cl_B)==0));

figure()
h3=plot(log10(Br_Dr),log10(Cl_Dr),'bs');
set(h3,'MarkerFaceColor','blue','MarkerSize',5)
mdl1=LinearModel.fit(log10(Br_Dr),log10(Cl_Dr));
hold on
[~,larg] = max(mdl1.Diagnostics.CooksDistance);
mdl21 = LinearModel.fit(log10(Br_Dr),log10(Cl_Dr),'Exclude',larg);
hold on
L1=plot(mdl21)
set(L1,'color','blue','LineWidth',1.5);


h4=plot(log10(Br_B1),log10(Cl_B1),'rs');
set(h4,'MarkerFaceColor','red','MarkerSize',5)
mdl2=LinearModel.fit(log10(Br_B1),log10(Cl_B1));
[~,larg] = max(mdl2.Diagnostics.CooksDistance);
mdl22 = LinearModel.fit(log10(Br_B1),log10(Cl_B1),'Exclude',larg);
hold on
L2=plot(mdl22)
set(L2,'color','red','LineWidth',2);

xlabel('Log(Br)','FontSize',15)
ylabel('Log(Cl)','FontSize',15)

legend([h3,h4],{'Oilfield Brine','Flowback Water Day 14 or 15','Flowback
Water Day 90'},'location','best')

title('Linear Regression Model for Cl-Br Systematics','FontSize',15)


PCA

clc;
clear all;
clf;
close all;
%%
```

```
load ('Ha');
load ('DOC');
Ha_Ions=xlsread('Data(Hayes2009)_PAN','sheet1');
Ha2=xlsread('Data(Hayes2009)_PAN','Sheet4');

Vol=[Ha(:,6) Ha(:,7) (Ha(:,8)-Ha(:,7))/4 (Ha(:,9)-Ha(:,8))/9 (Ha(:,10)-
Ha(:,9))/76];

Na=Ha_Ions(1:11:199,1:5);
K=Ha_Ions(2:11:200,1:5);
Ca=Ha_Ions(3:11:201,1:5);
Cl=Ha_Ions(4:11:202,1:5);
SO4=Ha_Ions(5:11:203,1:5);
Ba=Ha_Ions(6:11:204,1:5);
B=Ha_Ions(7:11:205,1:5);
Fe=Ha_Ions(8:11:206,1:5);
Li=Ha_Ions(9:11:207,1:5);
Sr=Ha_Ions(10:11:208,1:5);
Mg=Ha_Ions(11:11:209,1:5);
TDS=Ha(:,1:5);

Br=Ha2(1:4:73,1:5);
S=Ha2(2:4:74,1:5);
Alk=Ha2(3:4:75,1:5);
pH=Ha2(4:4:76,1:5);

% PCA Datapool
i=1;
Day_0=[Na(:,i)/1000 K(:,i)/1000 Ca(:,i)/1000 Cl(:,i) SO4(:,i)
Ba(:,i)/1000 ...
    B(:,i)/1000 Fe(:,i)/1000 Li(:,i)/1000 Sr(:,i)/1000 Mg(:,i)/1000 Br(:,i)
Alk(:,i)];
i=2;
Day_1=[Na(:,i)/1000 K(:,i)/1000 Ca(:,i)/1000 Cl(:,i) SO4(:,i)
Ba(:,i)/1000 ...
    B(:,i)/1000 Fe(:,i)/1000 Li(:,i)/1000 Sr(:,i)/1000 Mg(:,i)/1000 Br(:,i)
Alk(:,i)];
i=2;
i=3;
Day_5=[Na(:,i)/1000 K(:,i)/1000 Ca(:,i)/1000 Cl(:,i) SO4(:,i)
Ba(:,i)/1000 ...
    B(:,i)/1000 Fe(:,i)/1000 Li(:,i)/1000 Sr(:,i)/1000 Mg(:,i)/1000 Br(:,i)
Alk(:,i)];
i=4;
Day_14=[Na(:,i)/1000 K(:,i)/1000 Ca(:,i)/1000 Cl(:,i) SO4(:,i)
Ba(:,i)/1000 ...
    B(:,i)/1000 Fe(:,i)/1000 Li(:,i)/1000 Sr(:,i)/1000 Mg(:,i)/1000 Br(:,i)
Alk(:,i)];
i=5;
Day_90=[Na(:,i)/1000 K(:,i)/1000 Ca(:,i)/1000 Cl(:,i) SO4(:,i)
Ba(:,i)/1000 ...
    B(:,i)/1000 Fe(:,i)/1000 Li(:,i)/1000 Sr(:,i)/1000 Mg(:,i)/1000 Br(:,i)
Alk(:,i)];

% Datapool=[Day_0 Day_1 Day_5 Day_14 Day_90];
```

```
% for i=1:1:19;
% for    j=1:1:65;
%    K0(i,j)=isnan(Datapool(i,j));
%
% end
% end
% for j=1:65;
% [row,col]=find(K0(:,j)==0);
% for i=1:size(row,1);
% Data_New(i,j)=Datapool(row(i),j);
% end
% end


Datapool_Categories=[Day_0; Day_1; Day_5; Day_14; Day_90];

% for i=1:1:95;
% for    j=1:1:13;
%    K1(i,j)=isnan(Datapool_Categories(i,j));
%
% end
% end
% for j=1:13;
% [row,col]=find(K1(:,j)==0);
% for i=1:size(row,1);
% Data_New1(i,j)=Datapool_Categories(row(i),j);
% end
% end

% Datapool_Categories(isnan(Datapool_Categories))=0;

[coeff1,score1,latent1,tsquared1,explained1,mu1] =
pca(Datapool_Categories,...
'algorithm','als');
coeff1
mu1
t = score1*coeff1' + repmat(mu1,95,1);


Categories_new = cell(13,1);
Categories_new={'Na';'K';'Ca';'Cl';'SO4';'Ba';'B';'Fe';'Li';'Sr';'Mg';'Br';'A
lk'};


% [coeff2,score2,latent2,tsquared2,explained2,mu2] = pca(t,...
% 'algorithm','eig');
% [coeff3,score3,latent3,tsquare3,explained3,mu3] =
princomp(t,'algorithm','svd');
% cumsum(latent3)./sum(latent3)

% figure()
% biplot(coeff1(:,1:2),'Scores',score1(:,1:2),'VarLabels',...
%       Categories_new)
%
% figure()
```

```
% biplot(coeff2(:,1:2),'Scores',score2(:,1:2),'VarLabels',...
%       Categories_new)
%
%
%
% figure()
%     biplot(coeff3(:,1:2),'Scores',score3(:,1:2),'VarLabels',...
%       Categories_new)
%
%


w = 1./var(t);
[wcoeff,score,latent,tsquared,explained,mu] = pca(t,...
'VariableWeights',w);
 coefforth = inv(diag(std(t)))*wcoeff;
  coefforth2 = diag(sqrt(w))*wcoeff;
% figure()
% pareto(explained)
%  coefforth*coefforth'
%  cumsum(latent)./sum(latent)


%  figure()
% biplot(coefforth(:,1:2),'Scores',score(:,1:2),'varlabels',Categories_new);


figure()

hp=biplot(coefforth(:,1:3),'VarLabels',Categories_new);
set(hp,'LineWidth',2)
grid on
xlabel('Principal Component 1 (63.12%)','FontSize',20);
ylabel('Principal Component 2 (9.99%)','FontSize',20);
zlabel('Principal Component 3 (8.10%)','FontSize',20);
title('PCA on 13 Flowback Compositions in Time-series','FontSize',20)
xlim([-1,1])
ylim([-1,1])
zlim([-1,1])

hold on
xxx = coefforth (:,1:3);
yyy= score(:,1:3);
%Taken from biplot.m; This is alter the data the same way biplot alters data
- having the %data fit on grid axes no larger than 1.**
[n,d2] = size(yyy);
[p,d] = size(xxx); %7 by 3
[dum,maxind] = max(abs(xxx),[],1);
colsign = sign(xxx(maxind + (0:p:(d-1).*p)));
% xxx = xxx .* repmat(colsign, p, 1)
% yyy= (yyy ./ max(abs(yyy(:)))) .* repmat(colsign, 95, 1)

 maxCoefLen = sqrt(max(sum(xxx.^2,2)));
    scores_new = bsxfun(@times, maxCoefLen.*(yyy ./ max(abs(yyy(:)))),
colsign);
```

```
nans = NaN(n,1);
ptx = [scores_new(:,1) nans]';
pty = [scores_new(:,2) nans]';
ptz = [scores_new(:,3) nans]';
%I grouped the pt matrices for my benefit**
plotdataholder(:,1) = ptx(1,:);
plotdataholder(:,2) = pty(1,:);
plotdataholder(:,3) = ptz(1,:);
% %my original score matrix is 42x3 - wanted each 14x3 to be a different
color**
size=350;
sc0=scatter3(plotdataholder(1:19,1),plotdataholder(1:19,2),plotdataholder(1:1
9,3),size,'marker', '.');
%
scatter3(mean(plotdataholder(1:19,1)),mean(plotdataholder(1:19,2)),mean(plotd
ataholder(1:19,3)),'marker', 'o');
sc1=scatter3(plotdataholder(20:38,1),plotdataholder(20:38,2),plotdataholder(2
0:38,3),size,'marker', '.') ;
%
scatter3(mean(plotdataholder(20:38,1)),mean(plotdataholder(20:38,2)),mean(plo
tdataholder(20:38,3)),'marker', 'd') ;
sc5=scatter3(plotdataholder(39:57,1),plotdataholder(39:57,2),plotdataholder(3
9:57,3),size,'marker', '.') ;
%
scatter3(mean(plotdataholder(39:57,1)),mean(plotdataholder(39:57,2)),mean(plo
tdataholder(39:57,3)),'marker', '^') ;
sc14=scatter3(plotdataholder(58:76,1),plotdataholder(58:76,2),plotdataholder(
58:76,3),size,'marker', '.') ;
%
scatter3(mean(plotdataholder(58:76,1)),mean(plotdataholder(58:76,2)),mean(plo
tdataholder(58:76,3)),'marker', 's') ;
sc90=scatter3(plotdataholder(77:95,1),plotdataholder(77:95,2),plotdataholder(
77:95,3),size,'marker', '.') ;
%
scatter3(mean(ploataholder(77:95,1)),mean(plotdataholder(77:95,2)),mean(plotd
ataholder(77:95,3)),'marker', 'V') ;
legend([sc0,sc1,sc5,sc14,sc90],{'Flowback-Day 0','Flowback-Day 1','Flowback-
Day 5','Flowback-Day 14','Flowback-Day 90'},'FontSize',20);
```

ANN

```
clc;
clear all;
clf;
close all;
% %%
%
% ANN_Raw=xlsread('ANN_Data','sheet1');
%
% save('ANN_Raw');
% ANN_Predict=xlsread('ANN_Data','sheet2');
% save('ANN_Predict');
%%
load('ANN_Raw')
load('ANN_Predict')
```

```
Na=ANN_Raw(:,1);
K=ANN_Raw(:,2);
Ca=ANN_Raw(:,3);
Cl=ANN_Raw(:,4);
Sr=ANN_Raw(:,5);
Ba=ANN_Raw(:,6);
Fe=ANN_Raw(:,7);
TDS=ANN_Raw(:,8);
Mg=ANN_Raw(:,9);
Br=ANN_Raw(:,11);
Gas_Prod=ANN_Raw(:,12);
Prod_Day=ANN_Raw(:,13);

Na_KP=ANN_Predict(:,1);
K_KP=ANN_Predict(:,2);
Ca_KP=ANN_Predict(:,3);
Cl_KP=ANN_Predict(:,4);
Sr_KP=ANN_Predict(:,5);
Ba_KP=ANN_Predict(:,6);
Fe_KP=ANN_Predict(:,7);
TDS_KP=ANN_Predict(:,8);
Mg_KP=ANN_Predict(:,9);
Br_KP=ANN_Predict(:,11);
Gas_Prod_KP=ANN_Predict(:,12);
Prod_Day_KP=ANN_Predict(:,13);



%%
for i=1:1000;
Pn=[log10(Na./TDS) log10(Cl./Br) log10(Ca./Br) log10(Br)]';
Tn=[log10(Gas_Prod./Prod_Day)]';

[Pn,ps] = mapminmax(Pn,0,1); % gives all values between 0 & 1
[Tn,ts] = mapminmax(Tn,0,1); % gives all values between 0 & 1
[mi,ni] = size(Pn);
[mo,no] = size(Tn);

N_in = mi; % number of inputs in the network
N_out = mo; % number of outputs in the network
Tot_in = ni; %total no. of simulations
N_train = 699;
% N val = 50;
N_test = 80;
%seperating training, testing & validation data when random
%selection command is available through higher version dividing random
[Pn_train,Pn_val,Pn_test,trainInd,valInd,testInd] =
dividerand(Pn,0.7,0.15,0.15);
[Tn_train,Tn_val,Tn_test] = divideind(Tn,trainInd,valInd,testInd);


val.T = Tn_val;
val.P = Pn_val;
test.T = Tn_test;
test.P = Pn_test;
```

```
%Initiating network parameters
NNeu1 =28;
NNeu2 = 31;
NNeu3 = 27;
% NNeu5 = 69;NNeu6 = 50;
% NNeu3 = 50;NNeu3 = 12;
% creating the cascade backpropagation network
net = newff(Pn,Tn,[NNeu1,NNeu2,NNeu3]...
,{'logsig','logsig','logsig','purelin'},'trainscg','learngdm','msereg');
%setting training parameters for the network
 net.performFcn = 'msereg';
%net.performParam.ratio = 0.5;
net.trainParam.goal = 0.0001; %accuracy within this range
net.trainParam.epochs = 10000; % number of iteration sets
net.trainParam.show = 1;
net.trainParam.max_fail = 500;
NET.efficiency.memoryReduction = 60; %to reduce memory requirements

[net,tr] = train(net,Pn_train,Tn_train,[],[],test,val);
plotperf(tr)

Tn_train_ann = sim(net,Pn_train);
Tn_test_ann = sim(net,Pn_test);
%denormalizing the data sets obtained
%output reversal
T_train = mapminmax('reverse',Tn_train,ts);
T_test = mapminmax('reverse',Tn_test,ts);
T_train_ann = mapminmax('reverse',Tn_train_ann,ts);
T_test_ann = mapminmax('reverse',Tn_test_ann,ts);
% %input reversal
Pn_train = mapminmax('reverse',Pn_train,ps);
Pn_val = mapminmax('reverse',Pn_val,ps);
Pn_test = mapminmax('reverse',Pn_test,ps);

[row,col]=size(T_test);
n_funcLinks=0;
logorizeOutput=0;

if logorizeOutput==1
    Target=10.^T_test((1:row-n_funcLinks),:);
    Prediction= 10.^T_test_ann((1:row-n_funcLinks),:);
 elseif logorizeOutput==2
        Target=exp(T_test((1:row-n_funcLinks),:));
    Prediction=exp(T_test_ann((1:row-n_funcLinks),:));

     elseif logorizeOutput==3
        Target=[10.^(T_test((1:11),:));exp(T_test((12:23),:))];
    Prediction=[10.^(T_test_ann((1:11),:));exp(T_test_ann((12:23),:))];
else
    Target=T_test((1:row-n_funcLinks),:);
    Prediction=T_test_ann((1:row-n_funcLinks),:);
end

e=100*abs(Target-Prediction)./(Target);
```

```
Target_crossplot=reshape(Target,(row-n_funcLinks)*col,1);
Prediction_crossplot=reshape(Prediction,(row-n_funcLinks)*col,1);
%plot(Prediction_crossplot,Target_crossplot,'.');
MeanE=mean(mean(e));
fprintf('\n Overall Training Error is %.3f %s', MeanE,'%');

if MeanE<20;
    break;
end
end

%%
filename=sprintf('net_BOGM4')
save(filename,'net');
filename2=sprintf('tr_BOGM4')
save(filename2,'tr');

%% Pritz & Kirby
% load net_BOGM4
% load tr_BOGM4
%denormalizing the data sets obtained
%output reversal
Pn=[log10(Na./TDS) log10(Cl./Br) log10(Ca./Br) log10(Br)]';
Tn=[log10(Gas_Prod./Prod_Day)]';

[Pn,ps] = mapminmax(Pn,0,1); % gives all values between 0 & 1
[Tn,ts] = mapminmax(Tn,0,1); % gives all values between 0 & 1
% Pn_predict=[log10(Na_KP./TDS_KP) log10(Cl_KP./Br_KP) log10(Ca_KP./Br_KP)
log10(Br_KP)]';


%
Pn_predict=[-0.608542119    2.131932358 1.062899281 2.36361198
            -0.570880387    2.170214383 1.120096072 2.394451681
            -0.620314857    2.23949617  1.128554865 2.475671188
            -0.578873933    2.1180306   1.292285516 2.779596491
            -0.598204466    2.08020432  1.198708946 2.928395852
            -0.623361967    1.912798593 0.924230277 2.324282455
            -0.503356606    1.925935685 0.997850023 2.908485019]'



[Pn_predict,ps] = mapminmax(Pn_predict,0,1); % gives all values between 0 & 1
% view(net);
outputs=net(Pn_predict);
e=gsubtract(outputs,Pn_predict);
GAS_prod_ann= sim(net,Pn_predict);
GAS_prod_ann=mapminmax('reverse',GAS_prod_ann,ts);
GAS_prod_ann=10.^GAS_prod_ann;
%%
view(net)
plotperform(tr)
```

```
Inputs=[Na./TDS Cl./Br Ca./Br Br]';
Targets=[Gas_Prod./Prod_Day]';
[Inputs,ps] = mapminmax(Inputs,0,1); % gives all values between 0 & 1
[Targets,ts] = mapminmax(Targets,0,1); % gives all values between 0 & 1

Outputs=net(Inputs)
trOut =Outputs(tr.trainInd);
vOut = Outputs(tr.valInd);
tsOut = Outputs(tr.testInd);
trTarg = Targets(tr.trainInd);
vTarg = Targets(tr.valInd);
tsTarg = Targets(tr.testInd);
plotregression(trTarg,trOut,'Train',vTarg,vOut,'Validation',...
tsTarg,tsOut,'Testing')
%% Production Curve
figure()
hold on
plot(1:length(GAS_prod_ann),GAS_prod_ann,'r');
Tn_Predict=[4703.181818
5323.129032
1229.847909
50.68327402
90.17808219
100.8541096
42.38389041];
plot(1:length(Tn_Predict),Tn_Predict,'b');
% plot(1:length(Gas_Prod_KP./Prod_Day_KP),Gas_Prod_KP./Prod_Day_KP,'b');
legend('ANN Predicted Gas Production','Production Curve','Location','best')
ylabel('Gas Production (Mcf/Day)')
xlabel('Well Index')
%% Mapping
figure()
% Lat_KP=ANN_Predict(:,15);
Lat_KP=[41.36936633
41.28121389
41.27089244
41.13093381
41.61478442
40.50398611
40.50906111
];
% Lon_KP=ANN_Predict(:,16);
Lon_KP=[-77.55943694
-76.631075
-76.65969119
-78.04724025
-77.87637681
-79.57778611
-79.54695278
];

axesm sinusoid;view(3)
states = geoshape(shaperead('usastatehi', 'UseGeoCoords', true));
oceanColor = [.5 .7 .9];
latlim = [39.7 42.3];
```

```
lonlim = [-80.7 -74.5];

ax = usamap(latlim, lonlim);
setm(ax, 'FFaceColor', oceanColor);
geoshow(states,'DefaultFaceColor', 'white','DefaultEdgeColor', 'black');
[latlim, lonlim] = bufgeoquad(latlim, lonlim, .05, .05);


ptz = GAS_prod_ann';

for i=1:length(ptz);
    if ptz(i)<1000;
        i1(i)=i;
    elseif ptz(i)>1000 && ptz(i)<3000;
        i2(i)=i;
    elseif ptz(i)>3000;
        i3(i)=i;
end
end

i1=i1(find(i1~=0));
i2=i2(find(i2~=0));
i3=i3(find(i3~=0));

hh1=stem3m(Lat_KP(i1),Lon_KP(i1),50*ptz(i1),'r-','LineWidth', 3)
hold on
hh2=stem3m(Lat_KP(i2),Lon_KP(i2),50*ptz(i2),'b-','LineWidth', 3)
hh3=stem3m(Lat_KP(i3),Lon_KP(i3),50*ptz(i3),'g-','LineWidth', 3)

legend([hh1(1),hh2(1),hh3(1)],{'Gas Production Prediction 0-1000 Mcf/Day'...
    ,'Gas Production Prediction 1000-3000 Mcf/Day'...
    ,'Gas Production Prediction over 3000 Mcf/Day'},'FontSize',15);


title('ANN Gas Production Prediction for Kirby&Pritz Dataset','FontSize',15)
```