Spring 2022

Estimation of Radium-226 Concentrations in Produced Water from Shale Gas, Tight Gas and Conventional Hydrocarbon Wells

Richard Frank Haaker

Richard Haaker

Candidate

Mathematics and Statistics

Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

Prof James Degnan, PhD, Chairperson

Prof Yan Lu, PhD, Committee Member

Prof Laura Crossey, PhD, Committee Member

Estimation of Radium-226 Concentrations in Produced Water from Shale Gas, Tight Gas and Conventional Hydrocarbon Wells

BY

Richard Haaker

B.S. Biochemistry, Texas A&M University, 1975
M.S. Chemistry, Texas A&M University, 1978

Thesis

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Statistics

The University of New Mexico
Albuquerque, New Mexico

May, 2022

Acknowledgement

I would like to thank my wife, Cheryl, for her patience and support of me returning to graduate school after many years. Also, thanks to my committee chair, Professor James Degnan for his patience and input to this study. His support and willingness to answer questions was greatly appreciated. The efforts of committee member Professor Yan Lu are greatly appreciated. Her multiple imputation course introduced me to the problem of missing data. From her I learned that dealing with missing data is often a statistics problem hidden within a statistics problem. Committee member Professor Laura Crossey's efforts are appreciated in reviewing this document and for teaching the Geochemistry of Natural Waters class that I attended. Unfortunately, there was so much missing data that I was unable to provide more elegant models. Finally, I want to thank my classmates and the Math and Statistics Department faculty with whom I greatly enjoyed interacting these last several years.

Estimation of Radium-226 Concentrations in Produced Water from Shale Gas, Tight Gas and Conventional Hydrocarbon Wells

by

Richard Haaker

B.S. Biochemistry, Texas A&M University, 1975

M.S. Chemistry, Texas A&M University, 1978

M. S. Statistics, University of New Mexico, 2022

Abstract

This study examined data from the United States Geological Survey Produced Water database, version 2.3 (USGS DB) and built models to estimate the concentration of radium-226 in produced water given the values of other predictor variables. The dataset had only about 254 observations that were useable. Although the USGS DB had up to 190 possible attributes, it also had extreme rates of missingness, and many of the candidate variables were highly correlated. Multiple imputation techniques were employed using the **Mice**, **Hmisc**, and **RMS** packages for the R language to deal with the missing data. A multiple linear regression and two logistic regression main effects models were fitted to the data. The bootstrap was used as a means of internal validation of models. The models concluded that $\log_{10}$(total dissolved solids) and $\log_{10}$(barium) appear to be significant predictors of $\log_{10}$(radium-226) and radium exceedance probabilities.

Contents

List of Figures

List of Tables

Chapter 1: Introduction

The present study assesses the potential of produced water (PW) from specific oil and gas wells to be contaminated with elevated concentrations of naturally occurring radium-226, given information in their respective produced water (PW) chemistry reports and the United States Geological Survey PW database (USGS DB). The study design should lead to a set of figures or simple screening calculations that an environmental scientist or environmental attorney can use to estimate the potential of PW to have elevated radium-226 concentrations, based on data in Reports and general knowledge concerning the source geologic formation. This document uses many abbreviations, acronyms and symbols, and these are listed in Table 1.

Introduction to Radium

Radium is a radioactive element that was discovered by Marie Curie, PhD in 1898 *(Curie et al*., 1898). Radium-226 has a half-life of 1,600 years and an average life of 2,308 years. It is an indirect radioactive decay product of uranium-238, a long-lived naturally occurring radioactive isotope. Radium-226 atoms spontaneously transform to another radioactive isotope, radon-222, which is regarded as the leading cause of lung cancer among non-smokers (US EPA, 2014). Each radon-222 atom undergoes several more radioactive transformations, the final one being the radioactive decay of polonium-210 to stable lead-206. After its discovery, radium-226 was marketed as something wonderful. Eventually it was incorporated into a wide variety of commercial and industrial products, including luminescent dials, prescription medicines, bread, chocolate, jewelry, health products, cutlery, shampoo, fishing gear and condoms (Eriksson & O'Hagan, 2021). Radium-226 and its radioactive progeny were known to be a human health hazard since the 1920s.

Table 1. Acronyms and symbols used.

Term	Symbol
activity coefficient for species X	γ_X
activity of substance X	a_X
American Petroleum Institute	API
bootstrap	BS
coal bed methane	CBM
Code of Federal Regulations	CFR
coefficient	Coef.
Conventional hydrocarbon	CHC
database	DB
discharge limit	DL
false positive	FP
False negative	FN
ionic strength	IS
L	liter
$\log_{10}$(Barium, mg/L)	LBa
$\log_{10}$(Bicarbonate, mg/L)	LHCO
$\log_{10}$(Carbonate, mg/L)	LCO
$\log_{10}$(Calcium, mg/L)	LCa
$\log_{10}$(Chloride, mg/L)	LCl
$\log_{10}$(Total Iron, mg/L)	LFeT
$-\log_{10}$(Hydrogen ion activity, moles/L)	pH
$\log_{10}$(Potassium, mg/L)	LK
$\log_{10}$(Magnesium, mg/L)	LMg
$\log_{10}$(Sodium, mg/L)	LNa
$\log_{10}$(Sulfate, mg/L)	LSO
$\log_{10}$(Hydrogen sulfide, mg/L)	LH2S
$\log_{10}$(Bisulfide, mg/L)	LHS
$\log_{10}$(Radium-226, pCi/L)	LRa
$\log_{10}$(Total dissolved solids, mg/L)	LTDS
milligram	mg
missing at random	MAR
missing not at random	MNAR
multiple imputation	MI
naturally occurring radioactive material	NORM
not available	NA
pico-Curie	pCi
predictive mean matching	PMM
Preliminary Remediation Goal	PRG
probability	Pr
produced water	PW
quality control	QC
radium	Ra
radium-226	Ra-226
shale gas	SG
Simple random sample	SRS
tight gas	TG
total dissolved solids	TDS
true negative	TN
true positive	TP
United States Energy Information Administration	USEIA
United States Environmental Protection Agency	USEPA
United States Geological Survey	USGS
United States Nuclear Regulatory Commission	NRC
versus	vs.
well type	W.Type
within	wi
without	wo

Radium-226 is regulated in the United States in a piecemeal fashion. The United States Nuclear Regulatory Commission (USNRC) and certain states only regulate radium that qualifies as “discrete sources” or “by-product material.” The USNRC’s and states’ ability to regulate some radium flows down from the Atomic Energy Act of 1954, as amended. This authority includes radium that was (1) produced by the processing of ores for their uranium or thorium “source material” content and (2) large discrete sources that could be attractive to terrorists. The NRC regulations (10 CFR 20 Appendix B Table 2) restrict the concentrations of radium-226 that may be released to the environment in liquid effluent, such as runoff, to 60pCi/L from licensed operations, unless the effluent is discharged to a sanitary sewer. It also allows licensees to discharge an average of 600pCi/L of radium-226 to a sanitary sewer.

The United States Environmental Protection Agency (USEPA) also regulates the concentrations of radium-226 + radium-228 in drinking water to 5pCi/L pursuant to the Safe Drinking Water Act (*Radionuclides Rule: A Quick Reference Guide, EPA 816-F-01-003*, 2001). The agency has an online Preliminary Remediation Goal (PRG) Calculator that provides risk guidelines for radionuclides. The calculator estimates that there is an excess 10^{-4} lifetime incidence cancer risk associated with a 26 year duration of exposure to tap water containing only 2.84pCi/L radium-226, assuming no progeny are present. PRGs are not directly enforceable, but sometimes feed into decisions about remediation levels in cleanups that occur pursuant to the Comprehensive Environmental Response, Compensation and Liability Act.

Several states, such as Louisiana, Texas, and New Mexico regulate naturally occurring radioactive material (NORM) or technologically enhanced NORM in solid form, such as mineral scale in pipes and equipment, but not dissolved radium in produced water (PW) from oil and gas operations (Blackwell *et al*., 2021).

During the 1930s, naturally occurring radioactivity came to the attention of the American Petroleum Institute (API) and they were aware of an association of radioactivity and petroleum in sedimentary rocks (Bell *et al*., 1940). In 1940, the API funded its “Project 43c” to study the possible role that natural radioactivity played in the formation of petroleum in sedimentary rocks from biological residue (Breger & Whitehead, 1951).

In 1951, the United States Geological Survey published observations of radioactive scale precipitating from PW at oil and gas fields in Kansas (Gott & Hill, 1951). They concluded that the radioactive scale was intimately associated with sulfate minerals. The focus of that study was on the potential for PW or PW solids to contain recoverable amounts of uranium, although the authors recognized it was radium (not uranium) that was being precipitated during scale formation. In 1984 the United Kingdom’s National Radiation Protection Board issued a report on radiation protection problems caused by radium in PW and PW solids during development of offshore North Sea oil and gas fields (Escott, 1984).

Several papers have been published concerning the radium content of PW. Kraemer and Reid studied the relationship between total dissolved solids (TDS) and radium-226 concentrations in PW from geothermal and oil & gas wells situated along the gulf coast of the United States (Kraemer & Reid, 1984) and found a generally monotonically increasing relationship between Log(TDS) and Log(radium-226 concentrations). A few years later the USGS published data on the radium and TDS content of PW from oil and gas fields off the coast of Mississippi (T. F. Kraemer, 1987). Taylor published a study of PW discharges in Texas coastal water (Taylor, 1993). The International Atomic Energy Agency published a number of monographs on the behavior of radium in the environment, including groundwater (IAEA, 1990a, 1990b, 2016).

In 1986, an oil and gas pipe descaling service operated by the Street family in Laurel, Mississippi was ordered to cease operations by the Mississippi State Department of Health Division of Radiological Health because of high radiation levels arising from radium-226 contamination on the property and in the dirty pipe inventory. Litigation against major oil companies ensued with the Street *et al*. v. Chevron *et al*. case, which was concerned with personal injury and property damage resulting from negligence (i.e. issues such as failure to inspect and warn the Streets, of the hazards of de-scaling oil and gas casing and pipe that were contaminated with radium) (Smith, 2015). In 1992, the Street case settled during trial (Smith, 2015) and subsequently a large number of law suits were filed against oil companies alleging personal injury and/or property damage from oil field NORM.

In December 1990 the New York Times published two articles that alerted the public about the association of oil and gas operations with naturally occurring radioactivity and the potential radiation hazard (Keith Schneider, 1990a, 1990b).

Recently, the U.S. Geological Survey published the National Produced Waters Geochemical Database, version 2.3 (USGS DB). It is a large database of publicly available information on

PW characteristics It is the data source for this study, and is described in the section of this document entitled “USGS Produced Water Dataset.”

Oil and gas field operators are naturally concerned with the characteristics of the PW their wells are producing. Pertinent water quality information for oil and gas wells is summarized in water chemistry reports (Reports), and an example Report is provided in Appendix A. Oil and gas producers use the information in Reports to develop control strategies to manage corrosion and scale accumulation problems. Historically, such Reports did not include information on radium-226 concentrations.

Geochemistry Background

This section provides background on chemical concepts that are important in understanding radium behavior in PW.

Total Dissolved Solids (TDS) and Ionic Strength (IS)

TDS is a bulk property of a solution that is simply the mass concentration of dissolved solids, usually in milligrams (mg) in one liter (L) of solution. It may be measured by evaporating a known volume of solution to dryness and weighing the resulting solids. Detailed analyses are not needed to measure TDS. IS may be thought of as a measure of the effective quantity of electric charge contained in a known volume of solution. It is calculated from the known concentrations of all the ionic species in solution and their electric charges. Calculation of IS requires considerably more information than does calculation of TDS.

Specifically, IS a function of the concentrations of the ions in solution. Calculating IS with a reasonable of precision often requires knowledge of the following ion concentrations as well as the other charged species that the elements form: hydrogen (H^+), sodium (Na^+), potassium (K^+), calcium (Ca^{+2}), magnesium (Mg^{+2}), hydroxide (OH^-), bicarbonate (HCO_3^-), carbonate (CO_3^{-2}), sulfate (SO_4^{-2}), chloride (Cl^-). The calculation is iterative and tedious to do by hand. Software such as Geochemist's Workbench or PHREEQC is normally used to do the calculation (GWB, 2021; Unknown, 2021).

As IS increases, the solubility of ionic compounds also tends to increase in a smooth but non-linear fashion. For natural waters, TDS and IS are positively correlated variables. TDS being the more easily obtained value, is sometimes used as an explanatory variable when IS would be more appropriately used.

Concentration, Activity and Activity Coefficients

At infinite dilution in pure water, the activity of an ion of interest and its concentration would be equal. As the IS of a solution increases, while holding the concentration in solution of the ion of interest constant, the corresponding activity of the ion is decreased. The activity coefficient is a function of IS and provides a means of correction for the non-ideal behavior of real solutions that have non-zero IS. PW tends to have high TDS and IS, which makes them very non-ideal.

Acidity, Basicity and pH

pH is an expression of hydrogen ion activity in water. It is defined as the negative of the base 10 log of the H^+ activity. As the amount of H^+ in solution increases, the acidity of the solution increases and the pH decreases. A basic solution is one where the concentration of OH^- exceeds the concentration of hydrogen ions. At a given temperature and IS, the product of the activities of H^+ and OH^- is constant, known as the dissociation constant of water, K_W. In pure water the concentrations of H^+ and OH^- are equal and the pH is neither acidic nor basic.

Carbon Dioxide, Bicarbonate, and Carbonate Equilibria

Carbon dioxide (CO_2) is an acid gas that is ubiquitous in the atmosphere and in gases associated with PW from oil and gas production. It readily dissolves in water to produce carbonic acid (H_2CO_3). Pure water that is in equilibrium with the atmosphere is slightly acidic; at 25C such water has a pH of about 5.6. Carbonic acid undergoes dissociation in a pH dependent fashion to produce bicarbonate and carbonate. The proportions of carbonic acid, bicarbonate and carbonate in solution are a function of pH and temperature. At 25C and a pH of 6.35, the activities of carbonic acid and bicarbonate are equal, while the activities of bicarbonate and carbonate are equal at a pH of about 10.33 (Drever, 1997). The concentrations and relative proportions of these species in PW can limit the concentrations of magnesium and calcium in solution and determine whether PW is saturated with respect to carbonate minerals.

Radium Chemistry

Radium-226 is a decay product in the uranium-238 decay chain. It is an alkaline earth element, which behaves very much like barium and strontium. Atoms of both barium and strontium are millions of times more abundant than those of radium-226 in the environment and in PW. Natural waters do not contain enough radium to be saturated with respect to a radium mineral, such as $RaSO_4$ (which would be the radium analogue to barite). Instead, it tends to co-precipitate in minerals where it occasionally takes the place of barium or strontium atoms (Langmuir & Riese, 1985). The most common barium mineral that precipitates from PW is barite $(Ba, Sr, Ra)SO_4$, which is insoluble over a wide range of pH values. The solubility of barite decreases with decreasing temperature (GWB, 2021). As PW travels up a well it cools and barite can precipitate as scale on the tubular surfaces, which decreases the concentration of Ra in PW that exits the well. Under strongly reducing conditions sulfate may no longer be the predominant form of sulfur in solution, and this can result in enhanced solubility of radium, barium, and strontium.

At higher pH values, mineral surfaces tend to be negatively charged and thus are capable of attracting and adsorbing positively charged ions, such as Ra^{2+}, while at lower pH values mineral surfaces tend to be positively charged (Drever, 1997). At higher pH values, radium is increasingly adsorbed onto fine grained clays, metal hydroxide particles, organic matter and mineral surfaces (IAEA, 2016).

Landis *et al*. (2018) studied the behavior of radium in the Marcellus Shale. They concluded that the presence of radium-226 in PW is largely attributable to desorption from organic shale constituents (Landis, Sharma, & Renock, 2018; Landis, Sharma, Renock, *et al.*, 2018). Landis, Sharma, & Renock, (2018) also reported that high TDS and high calcium concentrations were correlated with increased concentrations of radium-226 in PW .

Overall, radium-226 concentrations in PW are expected to be related to many factors, including the mineral phases present in the formation and their respective radium-226 concentrations, IS, temperature, pH, the amounts of SO_4^{-2}, inorganic carbon (carbonic acid + bicarbonate + carbonate) in the formation, the presence of complexing agents, and the ability of solids in the formation with high surface areas to adsorb radium-226 from solution. One expects that (1) high concentrations of dissolved solids favor higher concentrations of radium-226 and that (2) those conditions that favor high dissolved barium concentrations will also favor higher radium-226 concentrations.

Overview of Major Oil and Gas Well Types

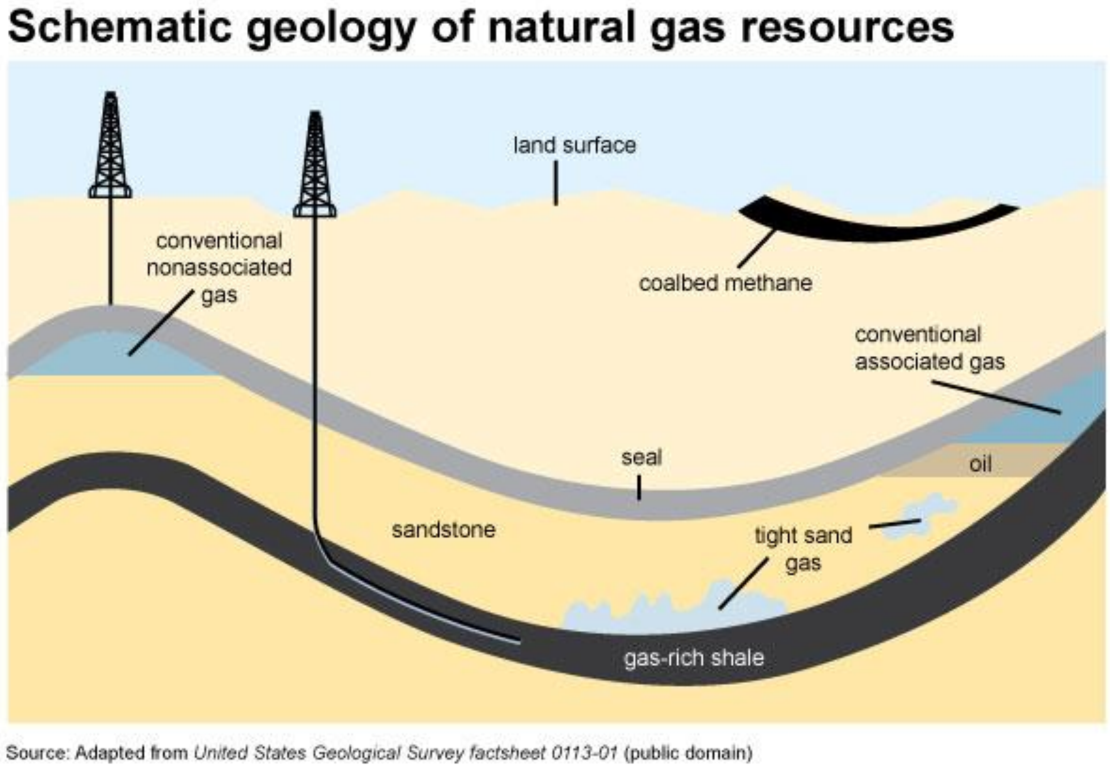
Oil and gas wells may be broadly categorized as conventional or unconventional. Each category is described in this section. Figure 1 is a depiction of major types of oil and gas wells.

Conventional Hydrocarbon

Conventional hydrocarbon resources are those that are trapped in permeable and porous rock formations, where they are confined in dome-like structures or folds. Conventional hydrocarbon wells are an older technology and most that have been in production did not require extensive hydraulic fracturing or horizonal drilling techniques (BC, undated).

Nonconventional Hydrocarbon Wells

Tight gas and shale gas are the nonconventional well types considered in this analysis. Coal bed methane wells are a third type of non-conventional well that are included in the USGS DB. Modern wells drilled into tight formations and shale gas formations typically employ hydraulic fracturing and or horizontal drilling techniques to increase production rates of hydrocarbons.



Schematic geology of natural gas resources

Source: Adapted from *United States Geological Survey factsheet 0113-01* (public domain)

Figure 1. Major types of gas wells (USEIA, 2020).

Tight Gas Formations

Tight natural gas formations are a legal category of natural gas resource created by the Natural Gas Policy Act of 1978. They are typically low-permeability sandstones and carbonate formations (USEIA, 2021). Low permeability sandstones often have intergranular pores that are largely occluded by cements, such as silica or carbonate minerals. Usually most of the hydrocarbons in formations described as “tight” are not believed to have formed *in situ.* The gas in tight gas formations is mostly in pore space and fractures.

Shale Gas Formations

Technically a shale gas formation is a type of tight gas formation and shale gas is natural gas that comes from shale formations. Shale is a fine-grained sedimentary rock that is largely composed of silt and clays, often with a significant proportion of organic material. The hydrocarbons in shale gas formations are generally considered to have formed *in situ* and are sorbed into some of the shale’s constituents. Shale formations often require hydraulic fracturing and horizontal drilling techniques to stimulate production of natural gas in practical quantities (USEIA, 2021).

USGS Produced Water Dataset

The USGS DB, data dictionary and metadata are available on the USGS Produced Water website (Engle *et al*., 2019). It provides an extensive compilation of publicly available data on PW. The data dictionary is included in Appendix B, Table B.1. The dataset was downloaded in “.Rdata” file format for this investigation. It includes 114,943 observations and 190 variables. Of these, there are only 720 observations that have analytical results for radium-226. This study is largely concerned with evaluating the radium-226 data in this dataset.

Chapter 2: Statistical Analysis Techniques and Concepts

Regression Bootstrapping

Non-parametric bootstrapping is a statistical inference technique where an existing sample of size *n* is treated as a population and from it a set of simple random samples (resamples), also of size n, are drawn with replacement (Efron, 1979). Since the resampling occurs with replacement, some observations will be drawn more than once, while others may not be drawn at all in a particular sample. Each of the resulting resamples is analyzed independently to obtain an approximately normally distributed set of estimates of the statistics of interest. In this study the statistics of primary interest are expected value of the response variable given the values of the predictor variable, its confidence interval and prediction interval.

In this study a series of linear models are fitted to bootstrapped resamples to obtain information of the distribution of the response variables: E(LRa|predictors), E(log(odds(Ra-226 > 60|predictors))) and E(log(odds(Ra-226 > 600|predictors))). Bootstrapping was implemented in various ways. These include:

1. The *validate()* function in **RMS** performs a bootstrap, given a linear model, and provides estimates of the measures of fit and optimism (see "Optimism and Overfitting"), but does not provide a table of regression coefficients for each resample. The function *validate()* can accept a simple linear model object that was produced directly by the functions *ols()* or *lrm()*. Or *validate()* can accept an MI linear model object produced by *fit.mult.impute()* after Rubin's rules have been applied;

Those rules are described in the section entitled "Rubin's Rules". The function *validate()* in the **RMS** package has attractive features: It will produce a linear model object that can be used by the function *predict()* to produce graphs of confidence intervals and prediction intervals. It also will execute a fast backward elimination for each of the *n* bootstrap resamples linear models. This allows the tally of how often the various coefficients in a preliminary model were judged significant (Harrell Jr., 2021b).

2. For a complete case analysis, the bootstrap can be performed within a loop by a series of R commands that repeatedly:
    a. create a simple random resample of size n with replacement,
    b. fit the linear models with *ols()* or *lrm()* without MI,
    c. extract and save both the regression coefficients and the performance measures for later analysis.

    Upon exiting the loop, regression coefficients and performance measures are averaged, and standard deviations calculated.

3. The bootstrap was also used in a slightly more complicated manner. It was sometimes performed within a loop by a series of R commands that repeatedly:
    a. created a simple random resample with replacement,
    b. executed *mice(),* a function in the **MICE** package to perform multiple imputation for each bootstrap resample,

c. fit the linear models with *ols()* or *lrm()* arguments inside of the function *fit.mult.impute()*, which also applied Rubin's rules (Harrell Jr., 2021a, 2021b),

d. extracted and saved both the resulting regression coefficients and the performance measures for later analysis.

e. Upon exiting the loop, regression coefficients and performance measures were averaged, and standard deviations calculated.

Optimism and Overfitting

Ordinarily a regression model is fitted to a set of data, and typically the model will fit that dataset better than it will fit new data. For example, assume that a regression model is fitted to a dataset of size n. Then the dataset is treated as a population, and several more SRS of size *n* (i.e. resamples) are drawn from it, each with replacement. If the original model coefficients are used with the resamples to calculate apparent R^2 values (R^2_{app}), they should be lower than the R^2value of the original fit (R^2_{orig}). The optimism in the R^2_{orig} measure of fit is the difference between R^2_{app} and the average value of R^2_{app}. The same concept may be applied to assess the optimism other measures of fit such as the concordance (C) or Somers' D. Thoya *et al*. provides a detailed summary of methods for assessing optimism (Thoya *et al.*, 2018).

Having correlated predictor variables in a regression model can cause overfitting. It may result in a model having a marginal improvement in measures of fit, such as the coefficient of determination, R^2, but it also causes the variances of the coefficients of the correlated predictors to increase substantially. The variance inflation factor (vif) is a convenient measure of the severity of the collinearity of predictor variables; a value near 1 is an indication that severe correlation is absent. However, there is no general agreement of what values of vif indicate a serious collinearity issue exists, and values such as 2.5, 5, and 10 (O'Brien, 2007).

Missingness Pattern

Missing data can be of three general types. Missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) and each of these is defined by a missing data mechanism. MCAR means that each datum has the same probability of being missing. Data that is MCAR is benign in the respect that it should not introduce bias into an analysis if only complete cases are considered. Other missingness patterns are not benign and analyzing only complete cases may introduce bias into an analysis (van Buuren, 2018).

One type of MNAR data is data that is missing in a manner that depends on its value; there is one datum in our dataset that clearly is MNAR, it is a negative concentration of radium-226, and of course concentrations are constrained to be positive numbers. Since the concentrations will be log transformed, it was changed to "NA" and treated with caution as MAR.

The assumption in this study has been made that the overall missingness pattern is MAR and verifying this assumption in depth is not part of the scope of this investigation.

Predictive Mean Matching (PMM) Algorithm

Multiple imputation is based on the idea that variables that are not missing contain information about those that are. PMM is a variation on an older method of imputation known as hot deck imputation. Hot deck imputation methods assign to each missing value the value of an observed response from a similar unit (Andridge & Little, 2010). PMM is a nonparametric technique for MI, and the *mice()* function by default uses Type 1 PMM (van Buuren, 2018). PMM is an iterative technique that assumes that the set of all possible values for a variable are contained in its set not-missing observations(van Buuren, 2018). PMM calculations that were performed in this study all used a “reverse monotone” imputation order, except for one sensitivity case where the imputation order was “monotone.” In reverse monotone order, imputation proceeds from the variables with the highest missingness to the lowest missingness. In the main part of this study, PMM imputed values in the order: LFeT (47.2% missing), LBa (35.8%), LTDS (30.3%), and then LRa (0.3%); each of these variable names are defined in Table 1. W.Type was included in the imputation process as a dichotomous variable but was never missing.

Those wishing to know exactly how *mice()* does Type 1 PMM are referred to the mice() source code (van Buuren & *et al*., 2022). A concise but dense description is also found the Algorithm 3.1 and Algorithm 3.3 boxes in van Buuren’s book (van Buuren, 2018).

Generally, mice() begins its first iteration by noting the values of LFeT in the complete cases and the values of other predictor variables. For each unit with missing LFeT, it makes a short list of the values of LFeT in the complete cases that have the most similar set of values of the other predictor variables. For each unit with missing LFeT, it draws a random value of LFeT from its list of similar units. Then it proceeds to do the same for LBa, LTDS and LRa.

At the beginning of the second iteration mice() begins by taking each unit that originally had a missing value of LFeT; generally, each of these now has estimates for all other variables. For each of these units, it makes a short list of all units that are most similar in their values of LBa, LTDS and LRa, and randomly draws a new value for LFeT for each. Then it proceeds to LBa. For each unit that originally had missing LBa, it constructs a list of the units that are most similar, taking in the account the newly assigned values of LFeT, and the previously obtained values of LTDS and LR and W.Type. Then for each, it draws a random value of LBa from their respective short lists. Next it updates its picks for LTDS and then LRa in a similar fashion. This procedure continues until the prescribed number of iterations is performed. The values for the last iteration are retained and these constitute one multiply imputed dataset. For this study, the maximum number of iterations was set to 10.

This process is then repeated for the number of multiply imputed data sets that were desired, in this study 20 were always created.

Rubin's Rules

The set of 20 multiply imputed datasets created by *mice()* each have linear models fitted to them by running lrm() or *ols()* within *fit.mult.impute().* That function applies Rubin's rules (Rubin, 1986) to obtain a single linear model from the 20 linear model fits. Obtaining point estimates of the linear coefficients is straight forward, since the point estimate of the value of a coefficient is just the average of the 20 estimates. Estimating the variances is more complicated because they must take into account the variation within and between each of the 20 estimates. Rubin's rule (van Buuren, 2018) for calculating the total variance for each of the coefficients is

(Eq 1.) $$T = U_{bar} + (B + \frac{B}{m}),$$

Where:

- *T* is the total variance,
- U_{bar} is the average of the individual complete data variances,
- *B* is the standard unbiased estimates of the variance between the coefficient and estimates for the *m*=20 estimates. It represents the variance introduced by having missing values in the sample.

Chapter 3: Characteristics of the USGS Dataset, Inclusion and Exclusion Criteria

Of the four well types considered in the USGS DB, CBM wells are the most problematic. PW from CBM wells tends to have the low radium-226 concentrations, low TDS, and high pH. While interesting, the CBM data has very high rates of left-censored (MNAR) radium-226 observations that would interfere with MI of predictor and response variables using standard and well-known MI software packages available in R. For this reason, CBM wells were deemed infeasible to include in the analysis.

Table B-2 provides the overall missingness rates for the 254 observations that had radium-226 measurements on the CHC, TG and SG well types addressed in this study.

Table 2 provides a summary of the frequency of well type by geographic area. The totals take into account that two CBM wells had been misclassified as some other well type and the USGS QC flagged one observation due to a suspect (extreme) pH value.

Table 2. Frequency of well type by basin in the dataset.

Basin	Conventional Hydrocarbon (N=142)	Shale Gas (N=106)	Tight Gas (N=6)	Overall (N=254)
Appalachian	91 (64.1%)	101 (95.3%)	0 (0%)	192 (75.6%)
Arkoma	0 (0%)	5 (4.7%)	0 (0%)	5 (2.0%)
Big Horn	5 (3.5%)	0 (0%)	0 (0%)	5 (2.0%)
Green River	15 (10.6%)	0 (0%)	1 (16.7%)	16 (6.3%)
Gulf Coast	17 (12.0%)	0 (0%)	0 (0%)	17 (6.7%)
Hanna	0 (0%)	0 (0%)	4 (66.7%)	4 (1.6%)
Powder River	13 (9.2%)	0 (0%)	0 (0%)	13 (5.1%)
Wind River	1 (0.7%)	0 (0%)	1 (16.7%)	2 (0.8%)

A further breakdown of the frequency of observations by formation is provided for the Appalachian Basin in Table B-3.

Table 3 provides summary statistics of candidate predictor variables that possibly could be useful in developing linear models describing Ra-226 levels in PW. To actually be useful, a candidate variable must: (1) have some association or role in explaining the behavior of radium-226, (2) have a sufficiently low rate of missingness that it can reasonably be included in multiple imputation of missing data procedures and (3) not have strong correlations with other variables that have lower rates of missingness.

Several potentially important predictor variables were excluded from initial consideration based having extremely high missingness rates (66.9% to 100% missing). These included the attributes: temperature, pressure, sulfide, bisulfide, iron(II), iron(III), carbonate, bicarbonate and sulfate. pH was not immediately excluded from consideration despite having a missingness rate of 69.6% based on my professional judgement as a chemist.

There is not consistent guidance on how much missing data is too much. Madley-Dowd, *et al.* advise against using the proportion of missing data as a criteria for excluding variables (Madley-Dowd *et al*., 2019). They provide evidence that in the case of MAR and MCAR data, the fraction of missing information is more important that the proportion of missing data. Their result is based on the premise that there is sufficient auxiliary information, however. Van Buuren is more cautious and notes in section 6.2 of his book that the risk of introducing more bias into regression coefficients increases as the missingness rates increase if the

Table 3. Summary statistics for potentially useful variables.

Analyte	Conventional Hydrocarbon (N=142)	Shale Gas (N=106)	Tight Gas (N=6)	Overall (N=254)
LRa				
Mean (SD)	2.12 (1.11)	2.78 (0.966)	0.358 (0.584)	2.36 (1.12)
Median [Min, Max]	2.49 [-1.30, 3.72]	3.10 [-0.788, 4.23]	0.322 [-0.215, 1.13]	2.63 [-1.30, 4.23]
Missing	0 (0%)	0 (0%)	1 (16.7%)	1 (0.4%)
LTDS				
Mean (SD)	4.65 (0.875)	4.97 (0.408)	3.59 (0.423)	4.77 (0.719)
Median [Min, Max]	5.12 [2.98, 5.60]	5.09 [3.93, 5.52]	3.54 [3.02, 4.13]	5.09 [2.98, 5.60]
Missing	54 (38.0%)	13 (12.3%)	0 (0%)	67 (26.4%)
pH				
Mean (SD)	7.03 (1.37)	6.67 (0.583)	8.20 (0.383)	7.04 (1.25)
Median [Min, Max]	6.80 [4.73, 10.4]	6.80 [5.50, 7.59]	8.23 [7.68, 8.72]	6.93 [4.73, 10.4]
Missing	87 (61.3%)	90 (84.9%)	0 (0%)	177 (69.7%)
LBa				
Mean (SD)	1.20 (1.23)	2.88 (1.05)	0.541 (1.35)	2.04 (1.43)
Median [Min, Max]	1.48 [-1.15, 3.64]	3.20 [-0.155, 4.18]	0.312 [-0.670, 2.72]	2.25 [-1.15, 4.18]
Missing	67 (47.2%)	23 (21.7%)	1 (16.7%)	91 (35.8%)
LCa				
Mean (SD)	3.16 (1.47)	3.81 (0.547)	0.844 (0.607)	3.39 (1.24)
Median [Min, Max]	3.92 [0, 4.69]	4.06 [2.45, 4.64]	0.854 [0, 1.45]	3.93 [0, 4.69]
Missing	57 (40.1%)	23 (21.7%)	0 (0%)	80 (31.5%)
LNa				
Mean (SD)	3.85 (0.884)	4.35 (0.368)	3.20 (0.388)	4.11 (0.683)
Median [Min, Max]	3.78 [2.49, 4.90]	4.49 [3.44, 4.91]	3.17 [2.68, 3.74]	4.37 [2.49, 4.91]
Missing	89 (62.7%)	24 (22.6%)	0 (0%)	113 (44.5%)

Table 3. (continued).

Analyte	Conventional Hydrocarbon (N=142)	Shale Gas (N=106)	Tight Gas (N=6)	Overall (N=254)
LCl				
Mean (SD)	4.20 (1.25)	4.71 (0.407)	2.88 (0.806)	4.41 (0.986)
Median [Min, Max]	4.85 [0.934, 5.30]	4.85 [3.69, 5.28]	3.20 [2.02, 3.82]	4.84 [0.934, 5.30]
Missing	56 (39.4%)	18 (17.0%)	1 (16.7%)	75 (29.5%)
LFeT				
Mean (SD)	0.971 (1.23)	1.76 (0.520)	-0.249 (0.469)	1.45 (0.938)
Median [Min, Max]	0.778 [-1.30, 2.67]	1.88 [0, 2.64]	-0.500 [-0.538, 0.292]	1.82 [-1.30, 2.67]
Missing	97 (68.3%)	20 (18.9%)	3 (50.0%)	120 (47.2%)

missingness pattern is partly MNAR (van Buuren, 2018). It seems clear that one can contrive ideal datasets and then introduce a MAR missingness pattern with a high proportion of missing values and still obtain relatively unbiased models. However, in the case of the USGS DB, we don’t have a way to know whether missingness is truly MAR or partly MNAR.

Table 4 provides an updated summary of potentially useful variables that excludes some attributes based on extreme missingness after merging the CHC and TG well types into one group, CHC/TG. The decision process for pooling CHC and TG wells is described in sections entitled “Dichotomizing the W.Type variable.” Observation ID 26278 had a negative reported radium-226 concentration, and it was reset to NA.

Table 4. Summary of observations after pooling CHC and TG well types.

Analyte	CHC/TG (N=148)	SG (N=106)	Overall (N=254)
LRa			
Mean (SD)	2.06 (1.14)	2.78 (0.966)	2.36 (1.12)
Median [Min, Max]	2.46 [-1.30, 3.72]	3.10 [-0.788, 4.23]	2.63 [-1.30, 4.23]
Missing	1 (0.7%)	0 (0%)	1 (0.4%)
LTDS			
Mean (SD)	4.58 (0.891)	4.97 (0.408)	4.77 (0.719)
Median [Min, Max]	5.07 [2.98, 5.60]	5.09 [3.93, 5.52]	5.09 [2.98, 5.60]
Missing	54 (36.5%)	13 (12.3%)	67 (26.4%)
pH			
Mean (SD)	7.14 (1.35)	6.67 (0.583)	7.04 (1.25)
Median [Min, Max]	7.31 [4.73, 10.4]	6.80 [5.50, 7.59]	6.93 [4.73, 10.4]
Missing	87 (58.8%)	90 (84.9%)	177 (69.7%)
LBa			
Mean (SD)	1.16 (1.24)	2.88 (1.05)	2.04 (1.43)
Median [Min, Max]	1.41 [-1.15, 3.64]	3.20 [-0.155, 4.18]	2.25 [-1.15, 4.18]
Missing	68 (45.9%)	23 (21.7%)	91 (35.8%)
LCa			
Mean (SD)	3.01 (1.54)	3.81 (0.547)	3.39 (1.24)
Median [Min, Max]	3.85 [0, 4.69]	4.06 [2.45, 4.64]	3.93 [0, 4.69]
Missing	57 (38.5%)	23 (21.7%)	80 (31.5%)

Table 4. (Continued)

Analyte	CHC/TG (N=148)	SG (N=106)	Overall (N=254)
LNa			
Mean (SD)	3.79 (0.867)	4.35 (0.368)	4.11 (0.683)
Median [Min, Max]	3.63 [2.49, 4.90]	4.49 [3.44, 4.91]	4.37 [2.49, 4.91]
Missing	89 (60.1%)	24 (22.6%)	113 (44.5%)
LCl			
Mean (SD)	4.12 (1.26)	4.71 (0.407)	4.41 (0.986)
Median [Min, Max]	4.78 [0.934, 5.30]	4.85 [3.69, 5.28]	4.84 [0.934, 5.30]
Missing	57 (38.5%)	18 (17.0%)	75 (29.5%)
LFeT			
Mean (SD)	0.895 (1.23)	1.76 (0.520)	1.45 (0.938)
Median [Min, Max]	0.563 [-1.30, 2.67]	1.88 [0, 2.64]	1.82 [-1.30, 2.67]
Missing	100 (67.6%)	20 (18.9%)	120 (47.2%)

Figure 2 provides a scatterplot matrix of selected analytes. Figures 3, 4 and 5 provide unstacked histograms of analyte concentrations by well type.

.

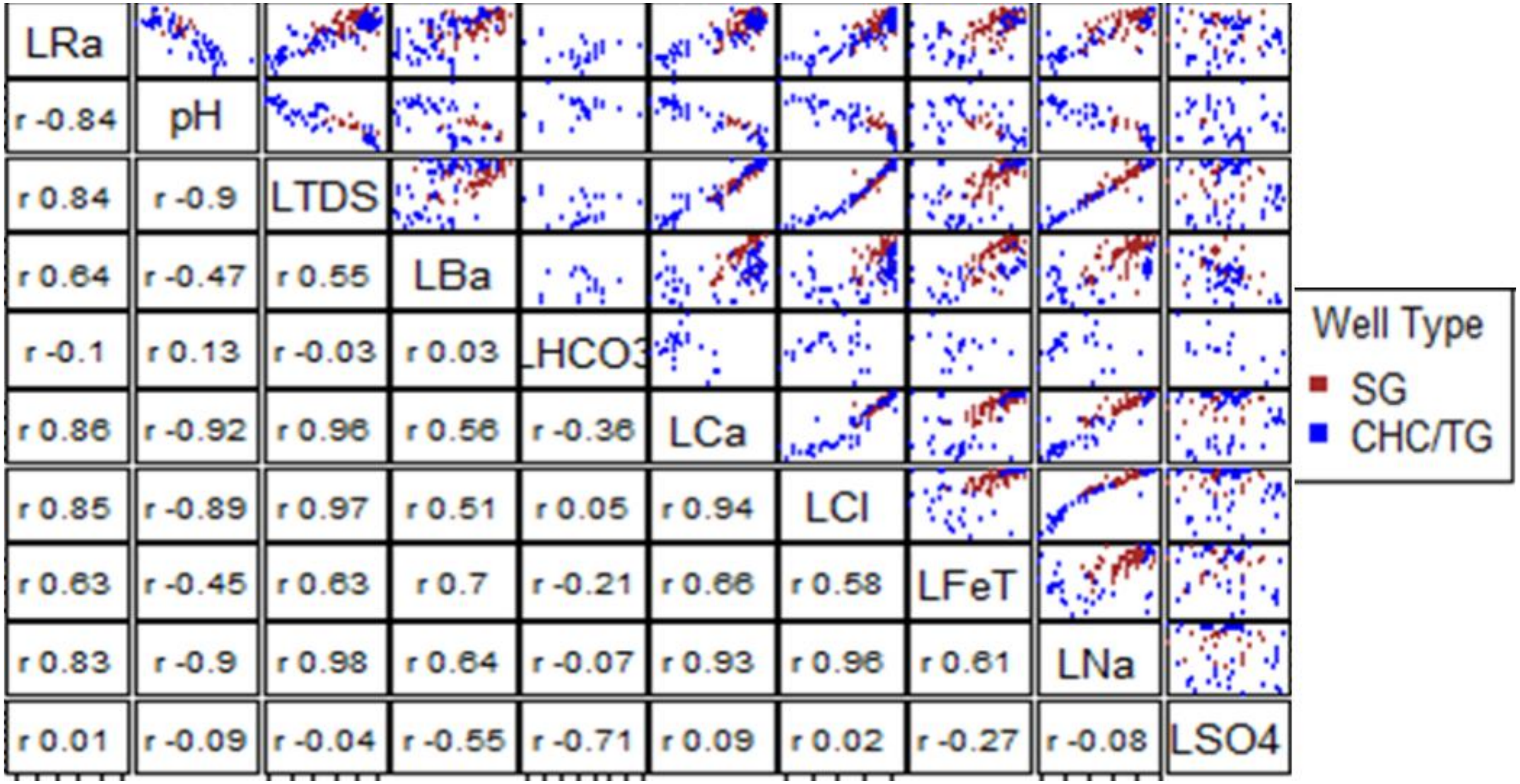


Figure 2. Scatterplot matrix of analytes initially of interest.

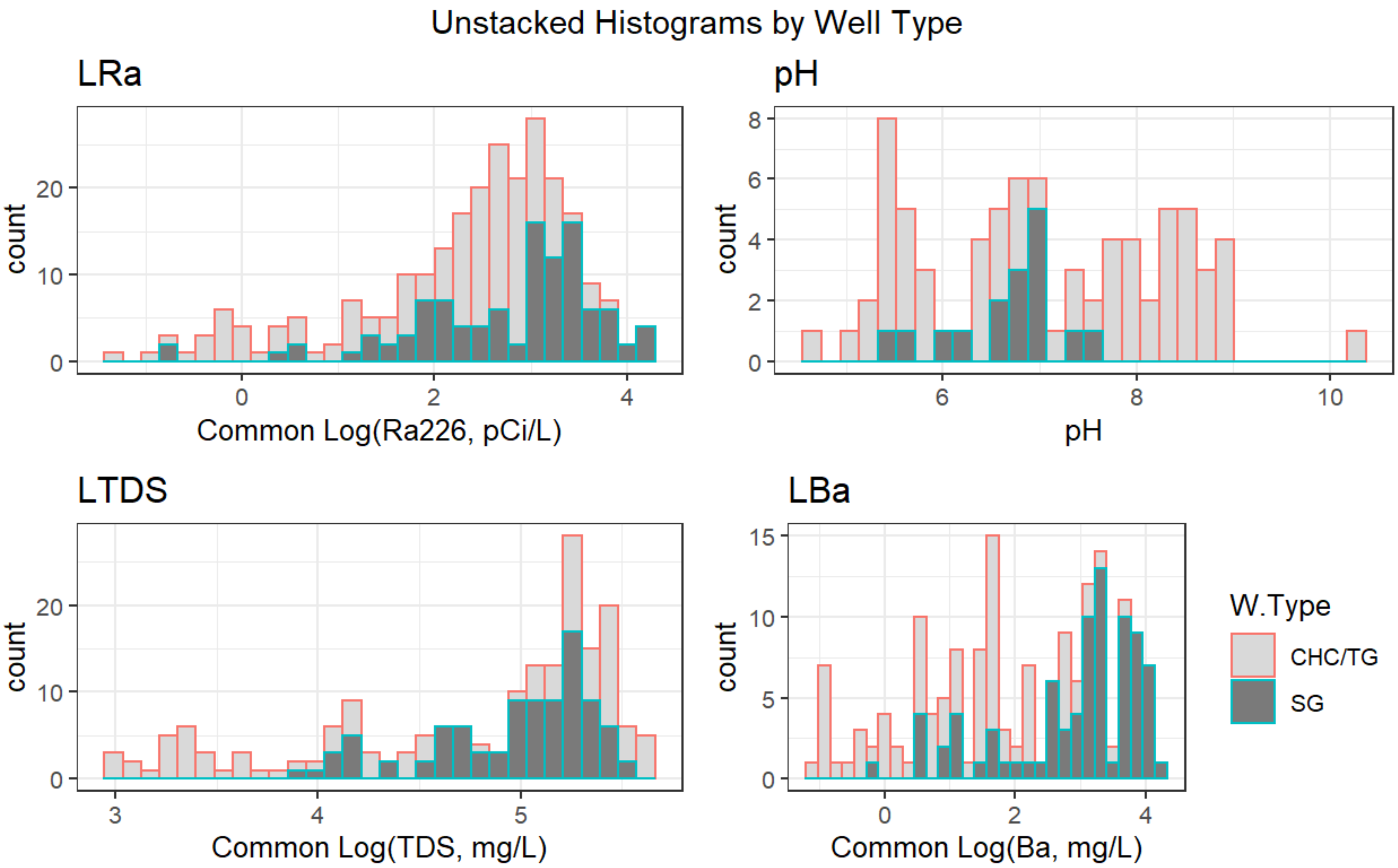


Figure 3. Histograms of LRa, pH, LTDS and LBa by Well Type.

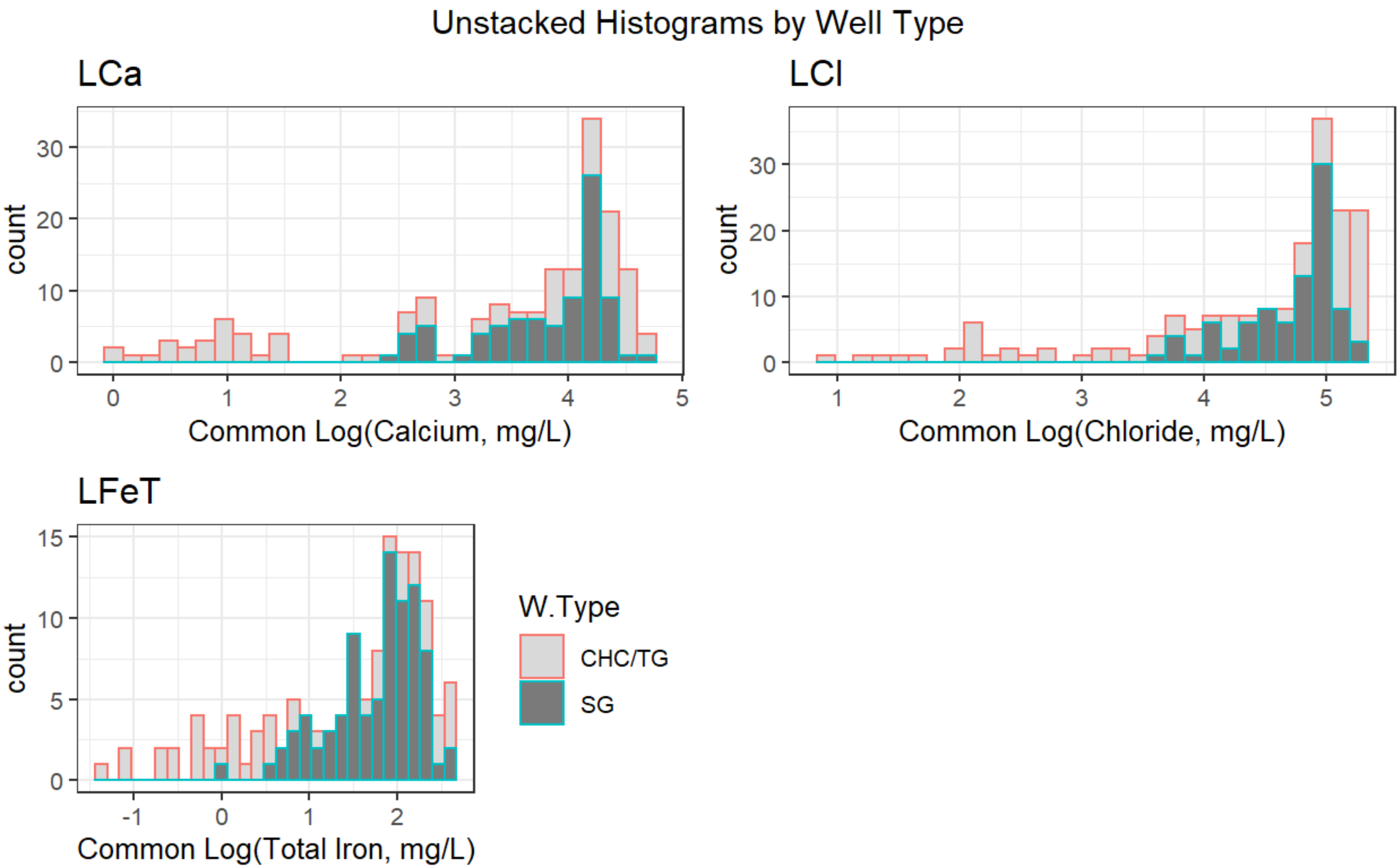


Figure 4. Histograms of LCa, LCl and LFeT by well type.

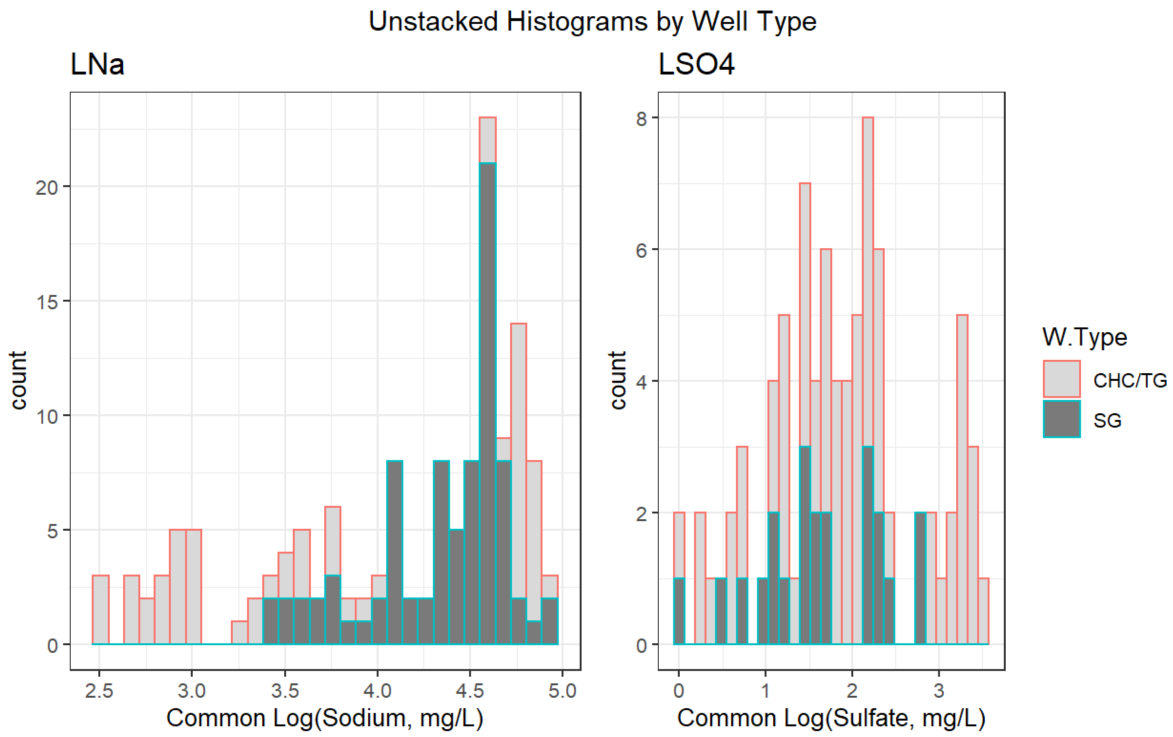


Figure 5. Histograms of LNa and LSO4 by well type.

Chapter 4: Method of Analysis Overview

This section describes the major steps in analyzing the dataset. The overall workflow is depicted in Figure 6. It includes: (1) treating the single left-censored (MNAR) radium-226 concentration and reducing the three-factor variable W.Type to a dichotomous one, (2) performing MI on the original dataset and on the bootstrap resampled datasets to address the missingness of predictor and response variables using the predictive mean matching technique, (3) performing regressions (multiple linear regression and logistic) on the MI within BS datasets with backward elimination after applying Rubin's rules to identify the "form of the full model," (4) fitting the full model form to the MI within BS datasets and to the original MI datasets and (5) comparing and assessing results.

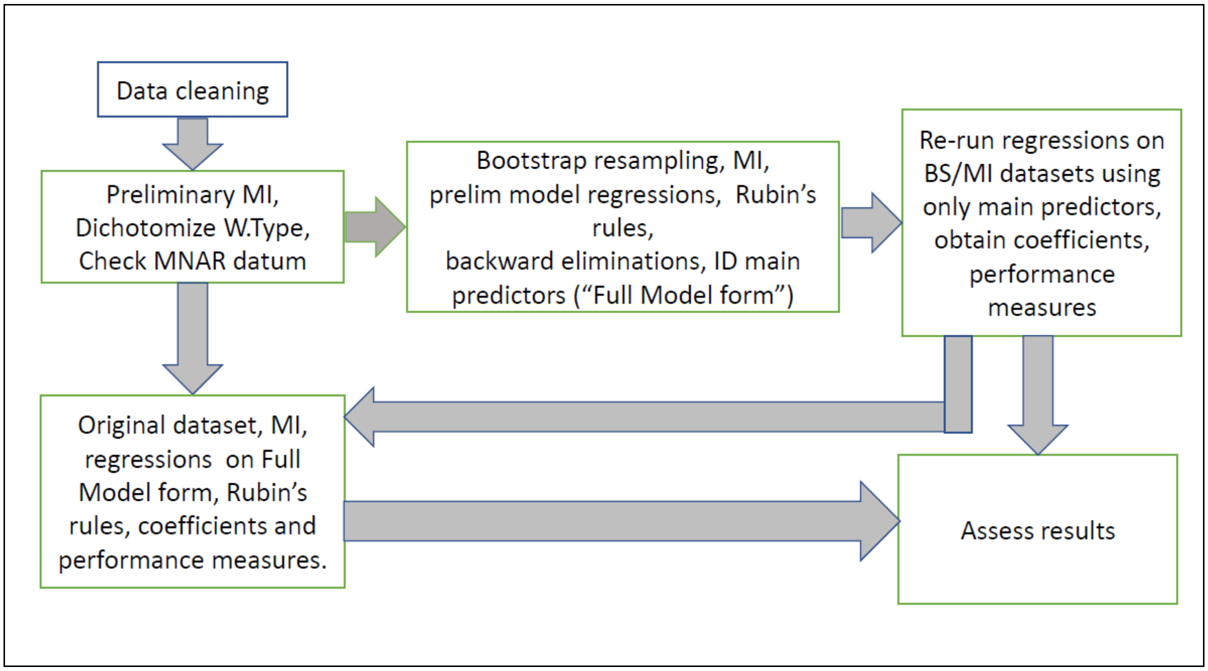

Data cleaning
Preliminary MI, Dichotomize W.Type, Check MNAR datum
Bootstrap resampling, MI, prelim model regressions, Rubin's rules, backward eliminations, ID main predictors ("Full Model form")
Re-run regressions on BS/MI datasets using only main predictors, obtain coefficients, performance measures
Original dataset, MI, regressions on Full Model form, Rubin's rules, coefficients and performance measures.
Assess results

Figure 6. Workflow of the study.

Data Cleanup

The 254 observations summarized in Tables 2 and 3 were all available observations that, at a minimum, had values for radium-226 concentration, basin and well type. It excludes two observations (ID # 26259 and 26330) because they were actually of CBM wells that were erroneously mis-classified as CHC or TG (WOGCC, 2021). The set of 254 observations also excludes observation ID # 1643 because it had been QC-flagged by USGS for an implausible pH (greater than 10).

Left-censored radium-226 observation and dichotomizing W.Type

Of the 254 observations, only ID # 26278 was reported as having a negative radium-226 concentration. In reality, concentrations of radium-226 will never be truly zero or negative as there should always be at least a few atoms of it present in any medium. When a negative concentration is reported for an analysis, it can be interpreted as having a signal that consisted of fewer counts in a time period than was expected from a "blank" that was prepared using the same ingredients as the unit, but with pure water substituted for the medium of interest (*i.e.*, produced water). A concentration that is reported as zero or less than zero is clearly a non-detect and can be interpreted as a left censored value. It also presents a problem because negative numbers cannot be log transformed. The options considered for this observation were to: (1) delete it, (2) formally treat it as MNAR, or (3) set to NA and allow the value to be chosen during multiple imputation treating the missing value to be MAR. Alternative (3) was chosen as the preferred option. In all cases, the function *mice()*, in the R package **Mice** was allowed to assign values during MI by PMM.

The minimum detectable concentration of Ra-226 in PW by standard analytical methods is of interest. For reference: (1) the standard method for radium-226 determination in water is EPA Method 903.1, and it provides a lower bound for minimal detectable concentration (MDC) of 0.1pCi/L (USEPA, EMSL, 1980); and (2) a real dataset of 41 radium-226 MDC observations from a brine-contaminated unconfined aquifer in the Erath gas field (Vermilion Parish, Louisiana) had a maximum MDC of 1.4pCi/L (Haaker, 2021). If treating the missing radium-226 value for observation 26278 as MAR yields imputed concentration estimates near the range of 0.1 to 1.4pCi/L for the original MI dataset, then the MAR assumption will be considered satisfactory.

Dichotomizing the W.Type variable

There are only 6 observations for the category W.Type = TG, which is an insufficient number to include it as its own category. The options for this attribute include (1) deletion of the TG observations altogether or (2) reset the W.Type to NA for these observations and then allow the R function *mice()* to multiply impute W.Type as either CHC or SG by logistic regression imputation during the preliminary MI step. The latter option was chosen since it does not waste data. The final decision rule will be: “All TG wells will be assigned one W.Type or other based on which W.Type receives the highest proportion of the 6 TG observations during the preliminary MI step using the original dataset.”

Multiple Imputation

Multiple imputation of missing data will be accomplished using R Studio with the *mice()* function from **Mice** package (van Buuren, 2018; van Buuren & Groothuis-Oudshoorn, 2021). This will be accomplished in several steps as described below:

1. Identify redundant variables using the *redun()* function from the **Hmisc** R package (Harrell Jr., 2021a) and eliminate them so that MI does not rely on highly correlated predictor variables. This is a necessary step because PMM does not work, or does not work well, if strong correlations among predictor variables exist. Based on the appearance of the scatterplots in Figure 2, it can be reasonably anticipated that several potential predictor variables will be dropped due to collinearity issues. In practice, predictor variables that can be predicted from other variables with a linear correlation coefficient greater than 0.9 will be dropped. When a group of highly correlated predictor variables is detected, the preference will be to retain only the variable with the lowest missingness. Using principal component analysis was considered as a way to pool the information from several highly correlated variables but was rejected for reasons that will be discussed later.
2. Use the *quickpred()* and *mice()* functions from the **Mice** package. This will create an initial prediction matrix considering only main effects that defines the dependencies among variables and provides an initial list of the imputation methods to be used for each variable. A single prediction matrix and list of imputation methods for the variables will be defined based on the original dataset and these will also be used

with the MI within BS part of the study. The matrix and list will be reviewed and edited as necessary. It is anticipated that two dichotomous response variables, GT60 and GT600, will be created from the continuous variable LRa by passive imputation. In practice, this means that the GT60 and GT600 variables will be calculated based on a deterministic formula. These are indicators of whether the radium-226 concentrations exceed NRC effluent discharge limits. A value of 1 will be assigned in instances where the respective discharge limit is exceeded and 0 in instances where it is not. These two dichotomous variables will not be used to impute other variables.

3. After the prediction matrix and methods list have been edited, run the *mice()* function to generate the MI datasets based on the original dataset. Then graphs of the object created by *mice()* will be produced to view the behavior of means and of standard deviations of variables as imputations proceed, Figure B-1. It will be impractical to produce similar graphs for the resampled datasets within the bootstrap.
4. Perform bootstrap resampling of the original dataset and then use *mice()* to create the series of MI datasets for each of the bootstrap resamples.

Multiple Linear and Logistic Regression

Informal preliminary modelling efforts to identify important interactions did not identify any significant ones. It was judged not worthwhile to attempt to include interaction terms in an empirical model where there was good reason to believe that significant main effects terms were missing due to extreme rates of missingness for some analytes.

The entire set of predictor variables that *redun()* identifies as not highly correlated and have relatively lower rates of missingness will be used for the preliminary multiple regression and logistic regression modeling. The MI within BS datasets will be fitted to the preliminary models, and Rubin's rules applied to obtain one regression result for each model for each bootstrap resample. Backward elimination will be conducted on each of these models and a tally kept of the frequency of each coefficient being significant. It is anticipated that the forms of the final main effects models will include those coefficients that are significant in approximately 70% or more of the MI within BS models. Once the final form is determined, the regressions will be rerun using that form, and for each of the three models (MLR, GT60 logistic, and GT600 logistic) the coefficients will be tabulated and averaged to obtain the three final models.

Chapter 5: Results

Left-censored radium-226 observation

Observation ID 26278 had the left censored MNAR value that was set to NA. The function ***mice()*** was allowed to impute values for it using the PMM technique and the original dataset. MI provided 20 estimates for radium-226 concentrations as depicted in Figure 7. The set of 20 imputed values ranged from 0.16 to 1.70pCi/L, with a median of 0.666 and mean of 0.616pCi/L. Consequently, ID 26278 was retained in the dataset since the range of imputed values of Ra-226 are similar to those observed in the Erath Field, Vermilion Parish, Louisiana dataset.

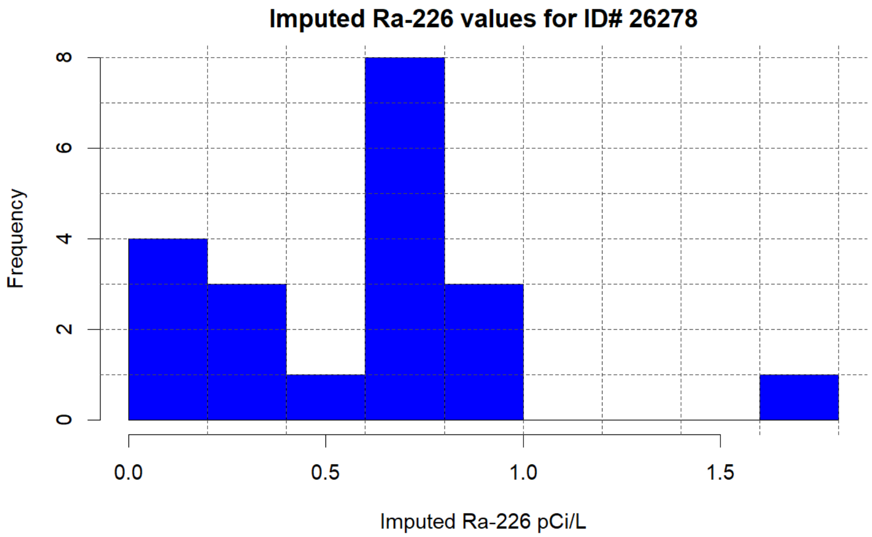


Figure 7. Histogram of Imputed Ra-226 values for MNAR observation ID 26278.

Dichotomizing the W.Type variable

There were an insufficient number of TG wells (6) for it to remain as a third category of the factor variable W.Type. The objective of dichotomizing was to eliminate the TG category. This involved using the *mice()* function to assign well types to the 6 TG wells by logistic regression imputation. The function *mice()* assigned all 6 of these as CHC and 0 as SG in every imputation set that was constructed directly from the original dataset. Consequently, a new category of W.Type, "CHC/TG," replaced the well types "CHC" and "TG", and W.Type became a dichotomous categorical variable with values "CHC/TG" and "SG."

Multiple Imputation

Of the candidate variables for multiple imputation (LRa, LBa, LFeT, W.Type, LNa, LCa, LCl, LTDS and pH), the function *redun()* identified the following as highly correlated ($R^2 > 0.9$): LNa, LCa, LCl and LTDS . The set of independent predictors recommended by *redun()* was LRa, pH, LBa, LFeT, and W.Type. I rejected pH as a variable because it had 69% missingness and retained LTDS instead (26 % missingness). The function *redun()* was executed again with redundant variables eliminated; Table 5 provides the missingness rates and adjusted coefficients of linear determination, R^2, for the chosen set of variables. Figure 8 illustrates the missingness pattern present in the dataset. In Figure 8, blue squares represent values that are not missing while red ones represent missing values. The numbers on the left margin represent the number of observations that fit each pattern while the numbers of missing values for each variable are given on the bottom margin. The numbers along the right margin are the number of variables that are missing values. For example there are: (1)

120 units with zero variables missing, (2) 42 units that have only LFeT missing, and (3) 91 units where LBa is missing. The scatterplot matrix of the variables that were still considered viable after MI is provided as Figure 9.

Table 5. Missingness and coefficients of linear determination.

Attribute	Number of NA out of 254 observations	Adjusted R^2 for prediction
LRa	1	0.802
LTDS	67	0.825
LBa	91	0.688
LFeT	120	0.576
W.Type	0	0.713

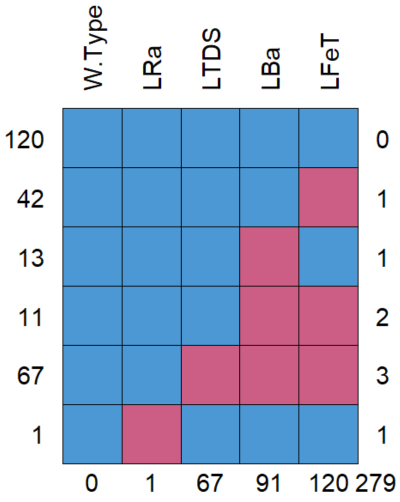


Figure 8. Missingness pattern in the dataset (red=missing, blue = not missing).

The imputation order should be unimportant provided that the MICE algorithm has converged. However, the imputation order can affect the speed of convergence (van Buuren, 2018). By default, mice() employs a “left to right” imputation order. The base case imputation order in this study was “reverse monotone,” which causes the variables with the most missingness to be imputed first. A sensitivity case used the opposite imputation order, “monotone,” to explore the effect of imputation order on the resulting model.

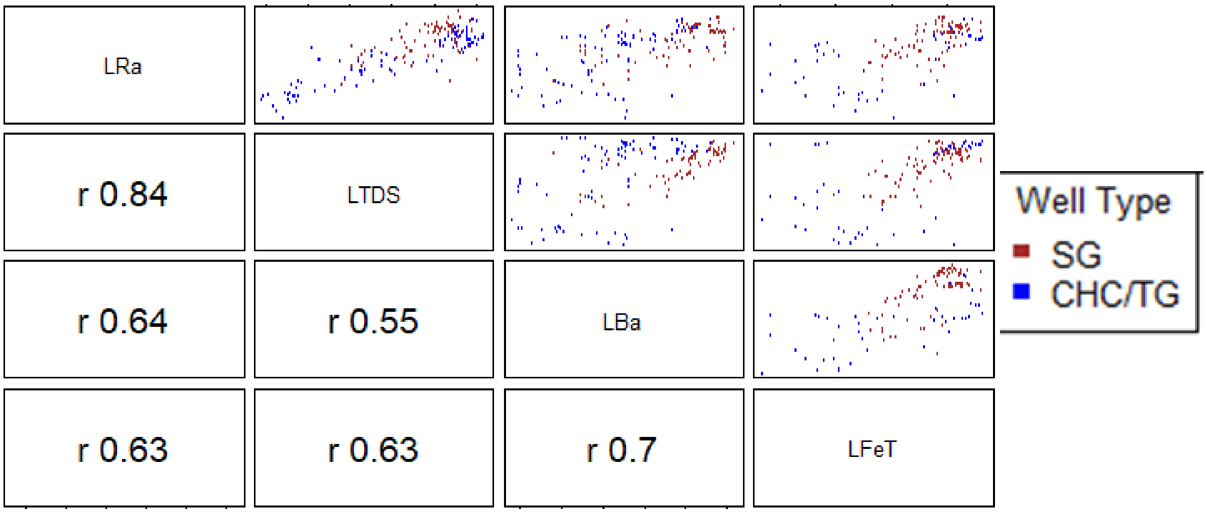


Figure 9. Scatterplot matrices for variables used in MI.

Multiple Linear and Logistic Regressions

Regression Modeling Using MI within BS Resampled Datasets

The models initially considered were:

(Eq 2.) $LRa \sim LTDS + LBa + LFeT + W.Type$,

(Eq 3.) $Log\big(Odds\ (GT60 = 1)\big) \sim LTDS + LBa + LFeT + W.Type$, and

(Eq 4.) $Log\big(Odds\ (GT600 = 1)\big) \sim LTDS + LBa + LFeT + W.Type$.

The regression models were fitted to each of the 200 bootstrap resampled datasets. For each bootstrap resample, *mice()* produced a set of 20 multiply imputed datasets. The *fit.mult.impute()*, and *ols()* and *lrm()* functions from the R packages **Hmisc** and **rms** were used to do the fits and apply Rubin's rules (Harrell Jr., 2021a, 2021b). The *fastbw()* function was used to identify the significant coefficients. The preliminary models and their performance measures are provided in Table 6. The first rows of each column of Table 6 provide the proportion of resamples that the various coefficients were judged significant when the *fastbw()* function in the RMS package was used. *Fastbw()* used a stopping rule based on minimization of the Akaike information criterion (AIC). Table 7 provides the resulting "final model" fits from the MI within BS regressions and their performance measures.

Based on the MI within BS regressions, the general forms of the final models are:

(Eq 5.) $LRa \sim LTDS + LBa,$

(Eq 6.) $Log\big(Odds\ (GT60 = 1)\big) \sim LTDS + LBa,$ and

(Eq 7.) $Log\big(Odds\ (GT600 = 1)\big) \sim LTDS + LBa + W.Type$

Complete Case Modeling Using BS Resampled Datasets

Based on the original dataset, bootstrap resampling was employed to create n=200 resampled datasets. Table 8 provides regression model results for bootstrapped datasets. All fits were performed by the *ols()* or *lrm()* functions from the **RMS** package. The values provided in Table 8 are averages of the sets of coefficients and performance measures obtained from fitting the resampled datasets. There were 120 complete cases in each resampled dataset for the GT600 logistic regression involving the quartet GT600 – LFeT –

Table 6. Preliminary models from MI within BS (n=200) and their performance measures.

```
MLR Model
Factor Significance Rate
 LBa  LTDS  LFeT  W.Type
   1    1     0       0
Coefficients and Measures
              Mean      SD
Intercept   -3.7111  0.3886
LTDS         1.1392  0.0864
LBa          0.1366  0.0650
LFeT         0.0864  0.0666
W.Type=SG    0.2962  0.0913
n          254.0000  0.0000
Model L.R. 359.5401 35.8100
d.f.         4.0000  0.0000
R2           0.7546  0.0339
g            1.0451  0.0625
Sigma        0.5636  0.0345
```

```
GT60 Logistic Model
Factor Significance Rate
 LBa   LFeT   LTDS  W.Type
 0.70   0.05   1.00   0.05
Coefficients and Measures
              Mean     SD
Intercept  -13.9933  2.8301
LTDS         3.0435  0.6169
LBa          0.6436  0.3036
LFeT         0.0089  0.3762
W.Type=SG    0.0576  0.5173
Obs        254.0000  0.0000
Max Deriv    0.0000  0.0000
Model L.R. 156.3816 19.8632
d.f.         4.0000  0.0000
P            0.0000  0.0000
C            0.9420  0.0207
Dxy          0.8839  0.0413
Gamma        0.8841  0.0413
Tau-a        0.3160  0.0245
R2           0.6931  0.0617
Brier        0.0695  0.0135
g            2.8778  0.4753
gr          20.9028 16.7177
gp           0.3159  0.0233
```

```
GT600 Logistic Model
Factor Significance Rate
 LBa   LFeT   LTDS  W.Type
 0.55  0.81   1.00    0.97
Coefficients and Measures
               Mean      SD
Intercept  -17.3152  2.7827
LTDS         2.7234  0.5508
LBa          0.3921  0.1930
LFeT         0.9531  0.3024
W.Type=SG    1.3670  0.3689
Obs        254.0000  0.0000
Max Deriv    0.0000  0.0000
Model L.R. 135.9382 14.7123
d.f.         4.0000  0.0000
P            0.0000  0.0000
C            0.8805  0.0190
Dxy          0.7610  0.0379
Gamma        0.7611  0.0379
Tau-a        0.3695  0.0203
R2           0.5570  0.0446
Brier        0.1343  0.0109
g            3.3338  0.4302
gr          32.4536 15.7145
gp           0.3730  0.0193
```

Table 7. Final Models from MI within BS (n=200) and Their Performance Measures.

```
MLR model
Factor significance proportion
     LBa          LTDS
      1             1
Coefficients and Measures
                Mean        SD
Intercept   -3.6696  0.3777
LTDS         1.1479  0.0830
LBa          0.2158  0.0436
n          254.0000  0.0000
Model L.R. 339.8978 34.6513
d.f.         2.0000  0.0000
R2           0.7350  0.0356
g            1.0312  0.0648
Sigma        0.5833  0.0335
```

```
GT60 Logistic Model
Factor Significance Proportions
     LBa          LTDS
    0.96            1
Coefficients and Measures
                Mean        SD
Intercept  -13.4774  2.5227
LTDS         2.9256  0.5427
LBa          0.6607  0.2255
Obs        254.0000  0.0000
Max Deriv    0.0000  0.0000
Model L.R. 153.7376 20.0332
d.f.         2.0000  0.0000
P            0.0000  0.0000
C            0.9397  0.0222
Dxy          0.8794  0.0443
Gamma        0.8795  0.0443
Tau-a        0.3144  0.0251
R2           0.6845  0.0629
Brier        0.0715  0.0134
g            2.7702  0.4502
gr          18.3016 13.7656
gp           0.3136  0.0237
```

```
GT600 Logistic Model
Factor significance proportion
     LBa        LTDS       W.Type
   0.970       1.000      0.985
Coefficients and Measures
                Mean       SD
Intercept  -17.9376  2.9747
LTDS         3.0818  0.5715
LBa          0.5969  0.1861
W.Type=SG    1.3491  0.3513
Obs        254.0000  0.0000
Max Deriv    0.0000  0.0000
Model L.R. 125.6480 18.3182
d.f.         3.0000  0.0000
P            0.0000  0.0000
C            0.8663  0.0244
Dxy          0.7327  0.0488
Gamma        0.7328  0.0488
Tau-a        0.3562  0.0266
R2           0.5234  0.0574
Brier        0.1427  0.0130
g            3.0960  0.4893
gr          26.3955 17.7877
gp           0.3599  0.0253
```

Table 8. Final Models from BS (n=200) of Complete Cases and Their Performance Measures.

```
MLR Model                                       GT60 Logistic Model                             GT600 Logistic Model
                    Mean        SD            SD                     Mean       SD                                   Mean        SD
Intercept         -4.046     0.375            Intercept         -14.218    2.642            Intercept    -18.745    4.738
LTDS                1.22     0.088            LTDS              3.05       0.58             LTDS         3.198      0.89
LBa                0.213     0.048            LBa               0.622      0.325            LBa          0.803      0.304
n                    162         0            Obs               162        0                W.Type=SG    0.944      0.678
Model L.R.       242.927    27.521            Max Deriv         0          0                Obs          162        0
d.f.                   2         0            Model L.R.        103.927    16.515           Max Deriv    0          0
R2                 0.774     0.038            d.f.              2          0                Model L.R.   87.929     14.554
g                  1.088     0.091            P                 0          0                d.f.         3          0
Sigma              0.556     0.038            C                 0.942      0.027            P            0          0
                                              Dxy               0.884      0.054            C            0.876      0.03
                                              Gamma             0.884      0.054            Dxy          0.752      0.06
                                              Tau-a             0.339      0.039            Gamma        0.752      0.06
                                              R2                0.694      0.07             Tau-a        0.371      0.029
                                              Brier             0.075      0.015            R2           0.559      0.069
                                              g                 2.86       0.548            Brier        0.134      0.017
                                              gr                20.79      15.525           g            3.416      0.781
                                              gp                  0.335      0.036          gr           45.026     60.897
                                                                                            gp           0.376      0.028
```

LTDS – W.Type, and 162 complete cases in each resampled dataset for the regressions involving the quartet LRa (or GT60 or GT600) – LTDS – LBa – W.Type. The result for the GT600 – LFeT – LTDS – W.Type logistic model are provided as Table B-6.

For comparison, the results of the complete case MLR (executed with *ols()* and *validate()*) are provided in Table 9 based on the final form model given in Eq 5. Figure 10 provides diagnostic plots and Figure 11 provides a histogram of standardized residuals for the fit. The Shapiro-Wilk test statistic for normality of the standardized residuals, W, was 0.988 (p-value = 0.1654). They exhibited a Kurtosis of 3.52.

Regression Modeling Using the MI – Validate() Procedure

The MLR models presented in Tables 10 and 11 were produced using the "reverse monotone" (base case) and "monotone" (sensitivity case) imputation orders respectively, to gauge its impact on model coefficients and measures of fit. The impact of "monotone" vs. "reverse monotone" imputation order for the MLR model is commented on in the conclusions section but was not explored further with additional modeling. Both of these MLR fits are based on the final form of the MLR model provided in Eq 5. The results were obtained by running *ols()* within *fit.mult.impute()*, and then executing the *validate()* function, which provides bootstrap estimates of model performance measures. Figure 12 provides diagnostic plots for the MLR model while Figure 13 provides confidence and prediction intervals for the MLR model.

Table 9. Results and Performance Measures Complete Case MLR Model.

```
MLR Model Complete Cases

Model:  LRa ~ LTDS + LBa

Frequencies of Missing Values Due to Each Variable
 LRa LTDS  LBa
   1   67   91

                 Model Likelihood    Discrimination
                       Ratio Test           Indexes
 Obs     162    LR chi2    240.30    R2       0.773
 sigma0.5634    d.f.            2    R2 adj   0.770
 d.f.    159    Pr(> chi2) 0.0000    g        1.095

           Coef    S.E.   t      Pr(>|t|)
 Intercept -4.0971 0.3398 -12.06 <0.0001
 LTDS       1.2316 0.0769  16.01 <0.0001
 LBa        0.2086 0.0371   5.62 <0.0001

                Analysis of Variance          Response: LRa

 Factor     d.f. Partial SS MS         F      P
 LTDS         1   81.31640  81.3164027 256.18 <.0001
 LBa          1   10.03498  10.0349770  31.61 <.0001
 REGRESSION   2  171.97352  85.9867613 270.90 <.0001
 ERROR      159   50.46902   0.3174152
             Effects              Response : NA

          index.orig training   test optimism index.corrected   n
R-square      0.7731   0.7735 0.7679   0.0056          0.7676 250
MSE           0.3115   0.3046 0.3187  -0.0141          0.3256 250
g             1.0950   1.0874 1.0951  -0.0077          1.1027 250
Intercept     0.0000   0.0000 0.0062  -0.0062          0.0062 250
Slope         1.0000   1.0000 0.9968   0.0032          0.9968 250

Variance Inflation Factors
      LTDS and LBa:  1.41

Kurtosis: 3.52

Shapiro-Wilk normality test: W = 0.98767, p-value = 0.1654
```

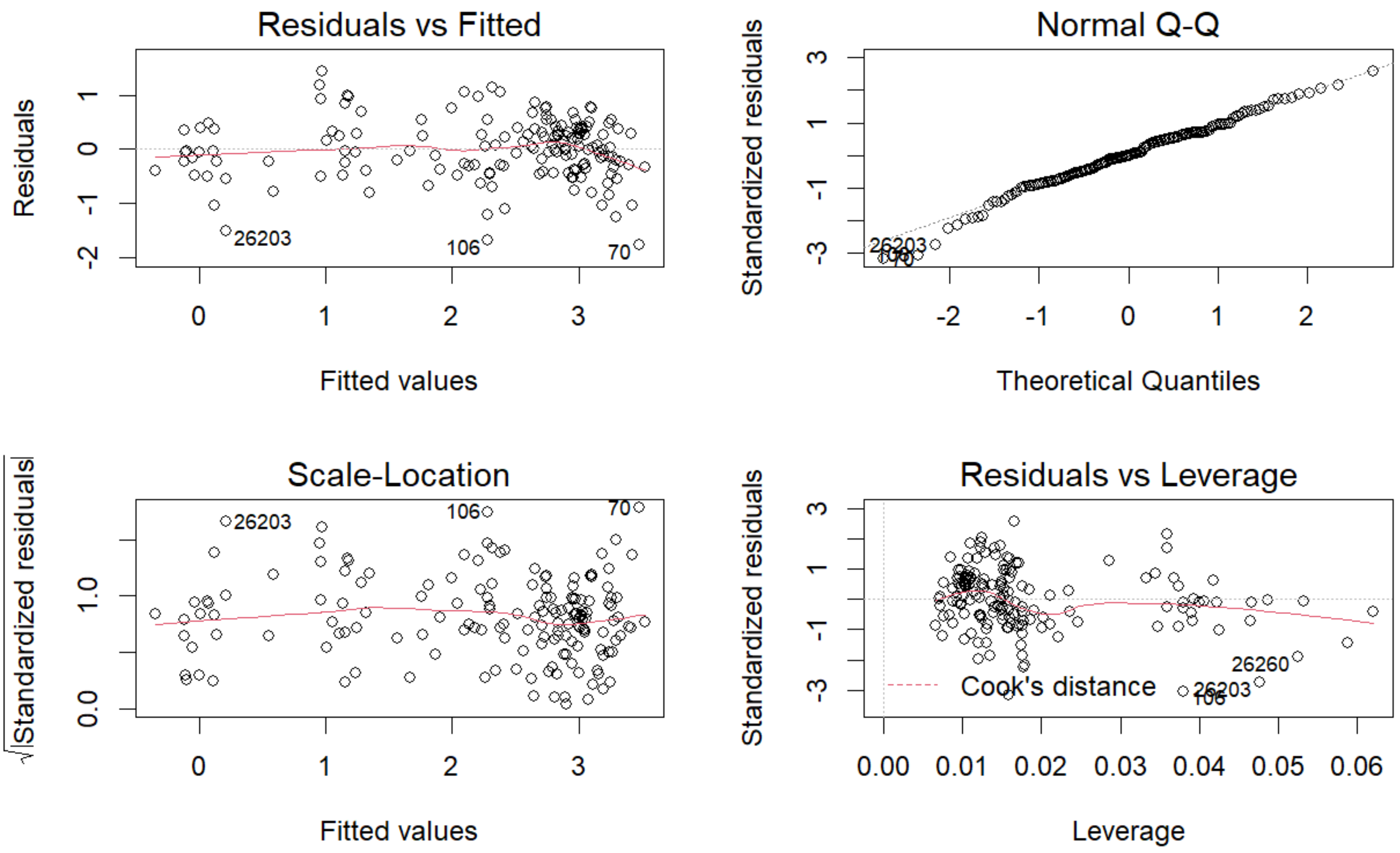


Figure 10. Diagnostic Plots for the Complete Case MLR Model (LRa ~ LTDS + LBa).

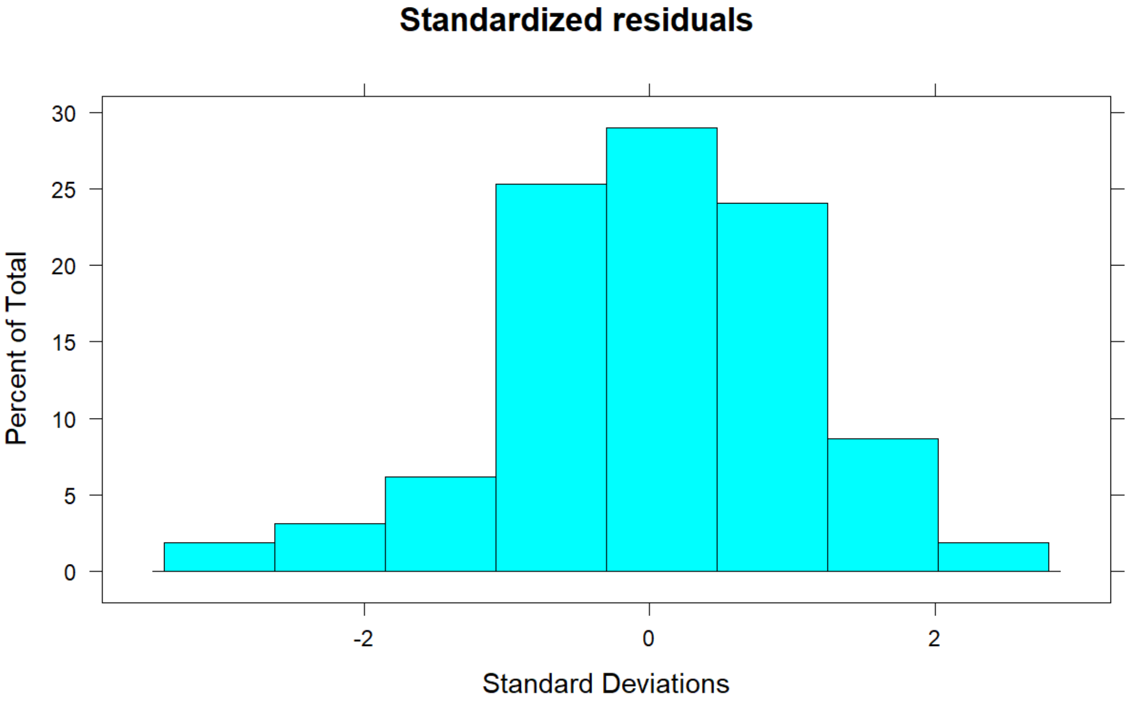


Figure 11. Standardized Residuals Histogram for the Complete Case MLR Model.

Table 10. Results and performance measures for the MLR regression (MI – *validate()*).

```
Final Model:  LRa ~ LTDS + LBa
Imputation Order: reverse monotone
                 Model Likelihood     Discrimination
                       Ratio Test     Indices
 Obs     254    LR chi2    333.92     R^2      0.731
 sigma0.5906    d.f.           2     R^2 adj  0.729
 d.f.    251    Pr(> chi2) 0.0000     g       1.029

           Coef    S.E.   t      Pr(>|t|)
 Intercept -3.6822 0.3086 -11.93 <0.0001
 LTDS       1.1522 0.0709  16.25 <0.0001
 LBa        0.2075 0.0389   5.33 <0.0001

                Analysis of Variance      Response: LRa

 Factor     d.f. Partial SS     MS       F        P
 LTDS         1   92.104915  92.1049145 264.09 <.0001
 LBa          1    9.913769   9.9137691  28.43 <.0001
 REGRESSION   2  200.834496 100.4172481 287.93 <.0001
 ERROR      251   87.538433   0.3487587

          index.orig training    test optimism index.corrected   n
R-square      0.7254   0.7241  0.7217   0.0024          0.7230 200
MSE           0.3551   0.3515  0.3600  -0.0085          0.3636 200
g             1.0357   1.0315  1.0357  -0.0042          1.0400 200
Intercept     0.0000   0.0000 -0.0210   0.0210         -0.0210 200
Slope         1.0000   1.0000  1.0087  -0.0087          1.0087 200

Variance Inflation Factors:

      LTDS   LBa
      1.47   1.47

Shapiro-Wilk normality test of regression residuals:

      W = 0.99346, p-value = 0.3336
```

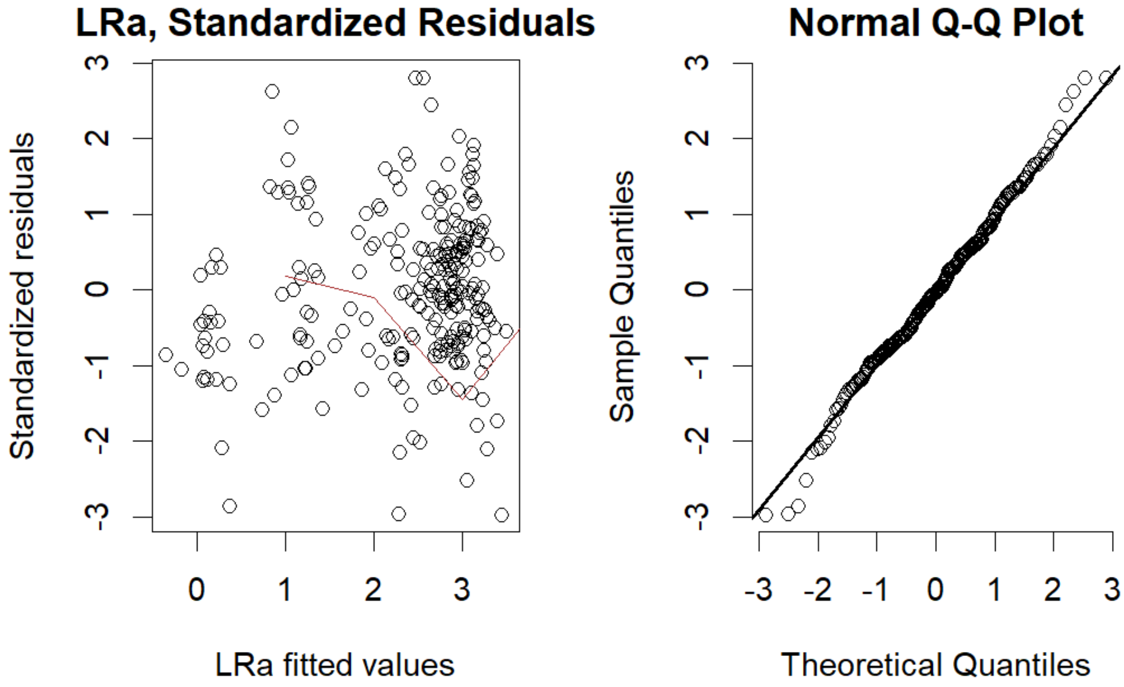


Figure 12. Diagnostic Plots for MLR regression (MI – *validate()*).

Table 11. Results and performance measures for the MLR regression (MI – *validate()*), Monotone Imputation Order.

```
Final Model: LRa ~ LTDS + LBa
Imputation Order: monotone

                 Model Likelihood    Discrimination
                       Ratio Test           Indexes
 Obs     254    LR chi2    342.11    R2       0.740
 sigma0.5817    d.f.            2    R2 adj   0.738
 d.f.    251    Pr(> chi2) 0.0000    g        1.036

           Coef    S.E.   t      Pr(>|t|)
 Intercept -3.7439 0.3022 -12.39 <0.0001
 LTDS       1.1630 0.0684  17.02 <0.0001
 LBa        0.2125 0.0334   6.36 <0.0001

                Analysis of Variance          Response: LRa

 Factor     d.f. Partial SS MS         F      P
 LTDS         1   97.95041   97.950408 289.52 <.0001
 LBa          1   13.67524   13.675240  40.42 <.0001
 REGRESSION   2  206.61383  103.306916 305.35 <.0001
 ERROR      251   84.91807    0.338319

          index.orig training    test optimism index.corrected   n
R-square      0.7334   0.7349  0.7299   0.0051          0.7283 200
MSE           0.3443   0.3373  0.3489  -0.0116          0.3559 200
g             1.0345   1.0327  1.0347  -0.0020          1.0365 200
Intercept     0.0000   0.0000 -0.0093   0.0093         -0.0093 200
Slope         1.0000   1.0000  1.0046  -0.0046          1.0046 200

Variance Inflation Factors
      LTDS   LBa
      1.38   1.38

Shapiro-Wilk normality test of regression residuals
      W = 0.99437, p-value = 0.4681
```

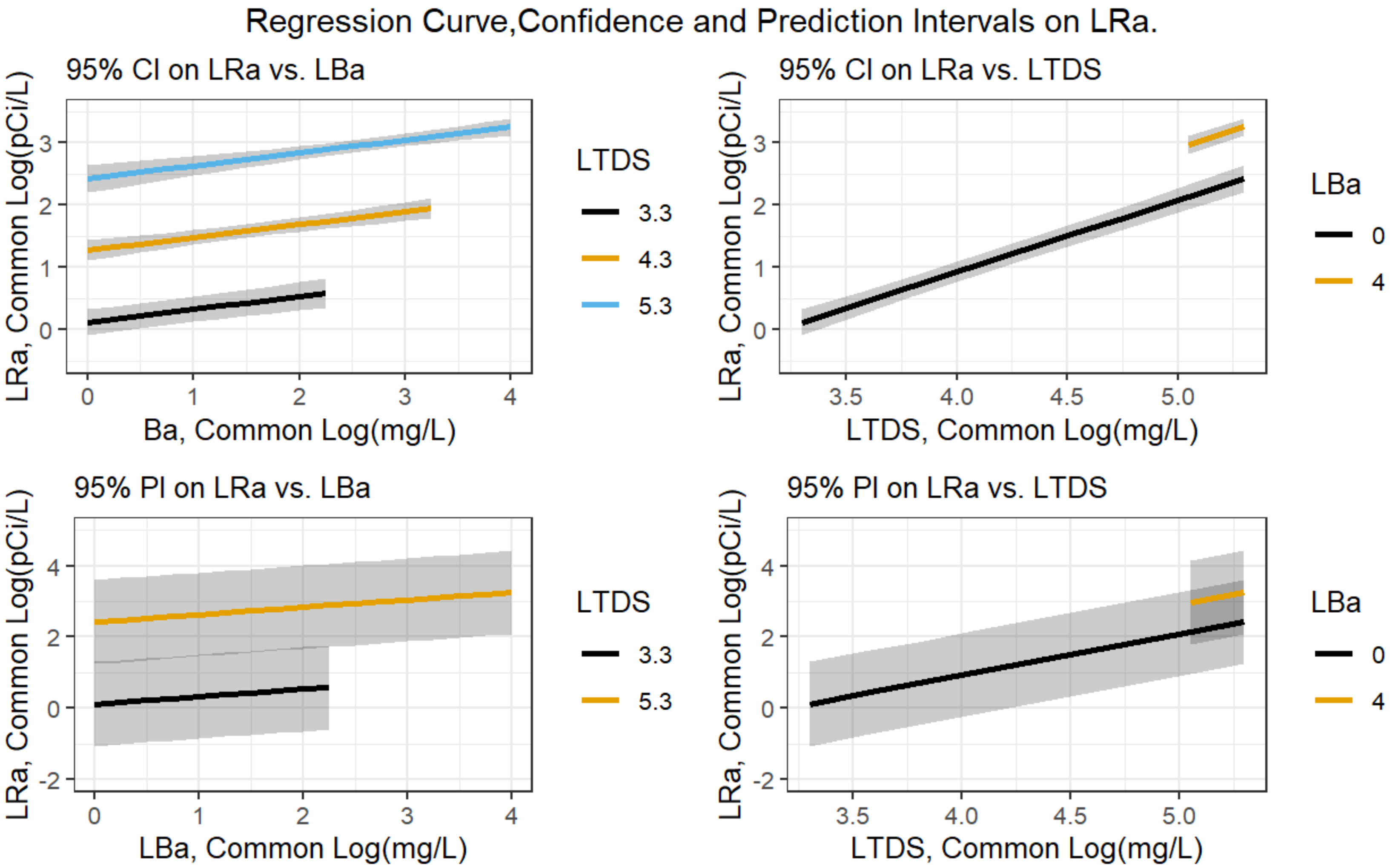


Figure 13. Regression Curves for MLR Model (MI – *validate()*).

Table 12 and Table 13 provide coefficients and measures of model fit for the GT60 and GT600 logistic models, respectively. The final model forms that these fits are based on are as given in Eq 6 and 7, respectively. An alternative GT600 model is provided in Table B-5. Figure 14 provides plots of log(Odds(Ra-226 > 60pCi/L)) vs LBa and probability(Ra-226 > 60pCi/L)) vs LBa for the GT60 logistic regression. The receiver operating characteristic curve (ROC) for the resulting GT60 model is provided in Figure 15. The sensitivity and specificity for the GT60 logistic regression was 0.73 and 0.92 respectively as are the positive and negative predictive values. Figure 16 provides plots of log odds versus the continuous predictor variables for the GT60 logistic model. Figure 17 provides plots of log(Odds(Ra-226 > 600pCi/L)) vs LBa and probability(Ra-226 > 600pCi/L)) vs LBa for the GT600 logistic model. Figure 18 is the ROC plot for the GT600 logistic model. A plot of log(odds(Ra-226 > 600)) vs. continuous predictors is not provided because it looks very similar to Figure 16.

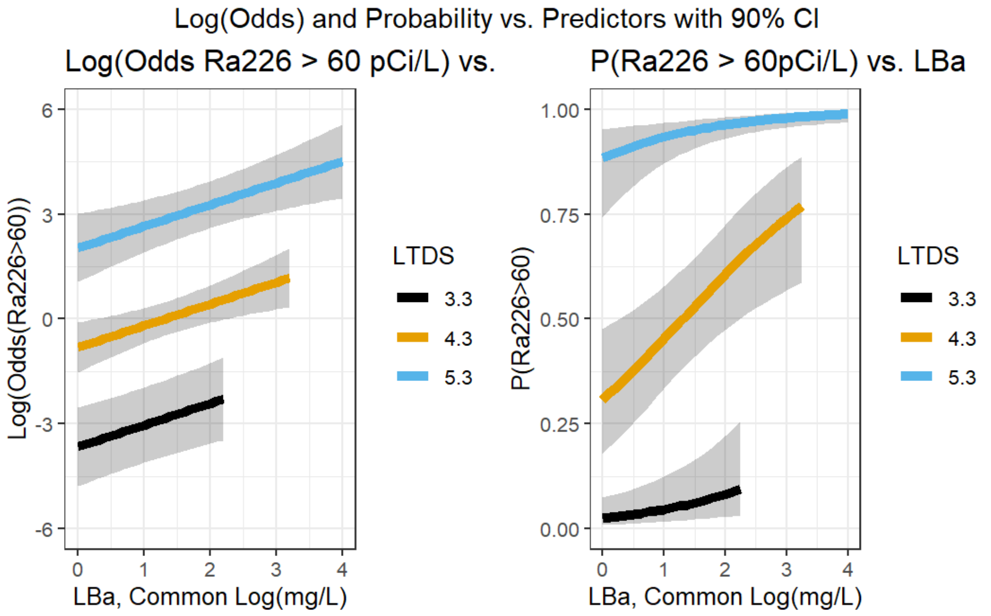


Figure 14. Log(Odds) and Probability Plots for GT60 Logistic Model (MI -*validate()*).

Table 12. Regression results for the GT60 Logistic Model (MI – *validate()*).

```
Final Model: GT60 ~ LTDS + LBa

                    Intercept    LTDS      LBa
VIF (Imputation)     1.18        1.16     1.45
VIF                              1.06    1.06
Missing Information   0.15       0.14     0.31
df Coefficients     821.77     957.32   196.58

                        Model Likelihood    Discrimination
                              Ratio Test           Indexes
 Obs           254    LR chi2      153.93    R2       0.687    C       0.941
  0             59    d.f.              2    g        2.676    Dxy     0.881
  1            195    Pr(> chi2) <0.0001    gr      14.675    gamma   0.881
 max |deriv| 5e-06                           gp       0.315    tau-a   0.316
                                            Brier    0.071

           Coef     S.E.   Wald Z Pr(>|Z|)
 Intercept -13.0484 2.1437 -6.09  <0.0001
 LTDS        2.8458 0.4733  6.01  <0.0001
 LBa         0.6215 0.2315  2.68  0.0073

                Wald Statistics          Response: GT60

 Factor     Chi-Square d.f. P
 LTDS       36.15      1    <.0001
 LBa         7.21      1    0.0073
 TOTAL      53.57      2    <.0001

          index.orig training   test optimism index.corrected   n
Dxy           0.8524   0.8538 0.8498   0.0040          0.8484 200
R2            0.6392   0.6432 0.6311   0.0121          0.6271 200
Intercept     0.0000   0.0000 0.0354  -0.0354          0.0354 200
Slope         1.0000   1.0000 0.9687   0.0313          0.9687 200
Emax          0.0000   0.0000 0.0133   0.0133          0.0133 200
D             0.5460   0.5526 0.5368   0.0158          0.5302 200
U            -0.0079  -0.0079 0.0002  -0.0080          0.0002 200
Q             0.5539   0.5604 0.5366   0.0238          0.5300 200
B             0.0809   0.0791 0.0827  -0.0035          0.0844 200
g             2.3778   2.4688 2.3633   0.1055          2.2723 200
gp            0.3047   0.3050 0.3028   0.0022          0.3025 200
```

Table 13. Regression results for the GT600 Logistic Model (MI - *validate()*).

```
Model: GT600 ~ LTDS + LBa + W.Type

                       Intercept  LTDS       LBa W.Type=SG
 VIF (Imputation)     1.35      1.36      1.35      1.09
 VIF                            1.29      1.12      1.42
 Missing Information 0.26      0.26      0.26      0.08
 d.f. Coefficients 285.68    274.71    288.07   2926.51

                           Model Likelihood        Discrimination
                                 Ratio Test               Indexes
 Obs            254    LR chi2      121.69    R2       0.512    C       0.862
  0             149    d.f.              3    g        3.013    Dxy     0.725
  1             105    Pr(> chi2) <0.0001    gr      21.008    gamma   0.725
 max |deriv| 1e-06                            gp       0.356    tau-a   0.353
                                              Brier    0.146

           Coef     S.E.   Wald Z Pr(>|Z|)
 Intercept -17.6586 3.7192 -4.75  <0.0001
 LTDS        3.0471 0.7064  4.31  <0.0001
 LBa         0.5335 0.1902  2.81  0.0050
 W.Type=SG   1.3775 0.4063  3.39  0.0007

           Wald Statistics     Response: GT600

 Factor     Chi-Square d.f. P
 LTDS       18.60      1    <.0001
 LBa         7.87      1    5e-03
 W.Type     11.49      1    7e-04
 TOTAL      36.52      3    <.0001

          index.orig training    test optimism index.corrected   n
Dxy           0.7455   0.7463  0.7346   0.0117          0.7338 250
R2            0.5251   0.5346  0.5169   0.0177          0.5075 250
Intercept     0.0000   0.0000 -0.0165   0.0165         -0.0165 250
Slope         1.0000   1.0000  0.9563   0.0437          0.9563 250
Emax          0.0000   0.0000  0.0126   0.0126          0.0126 250
D             0.4901   0.5039  0.4802   0.0236          0.4665 250
U            -0.0079  -0.0079  0.0033  -0.0112          0.0033 250
Q             0.4980   0.5118  0.4769   0.0348          0.4631 250
B             0.1417   0.1390  0.1451  -0.0061          0.1478 250
g             3.1216   3.2271  3.0240   0.2031          2.9185 250
gp            0.3612   0.3636  0.3578   0.0058          0.3554 250
```

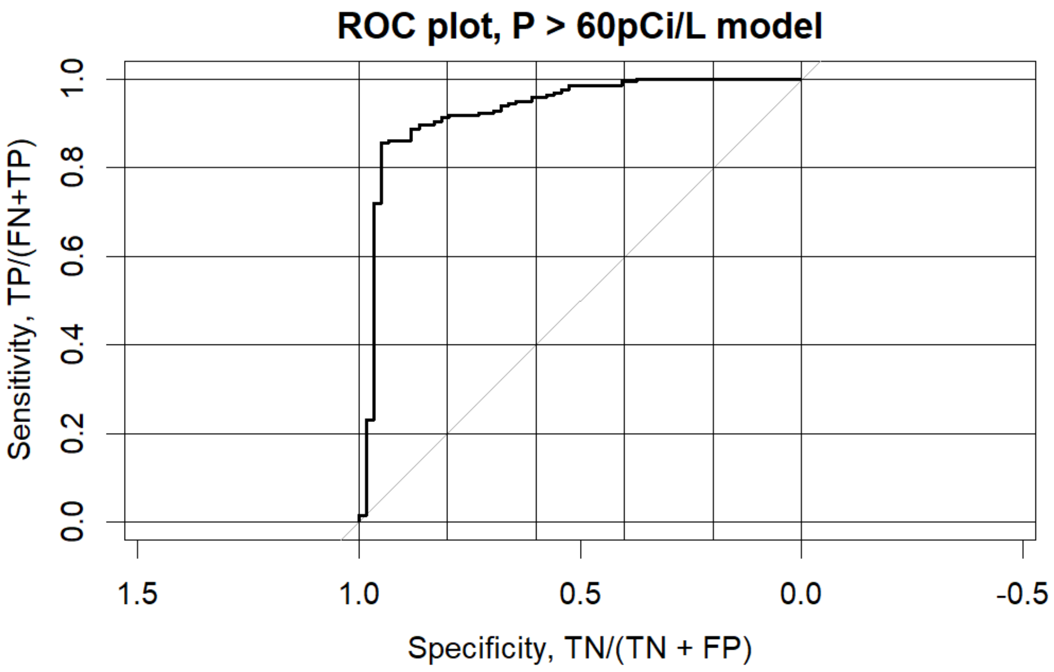


Figure 15. ROC Curve for the GT60 Logistic Model (MI -*validate()*).

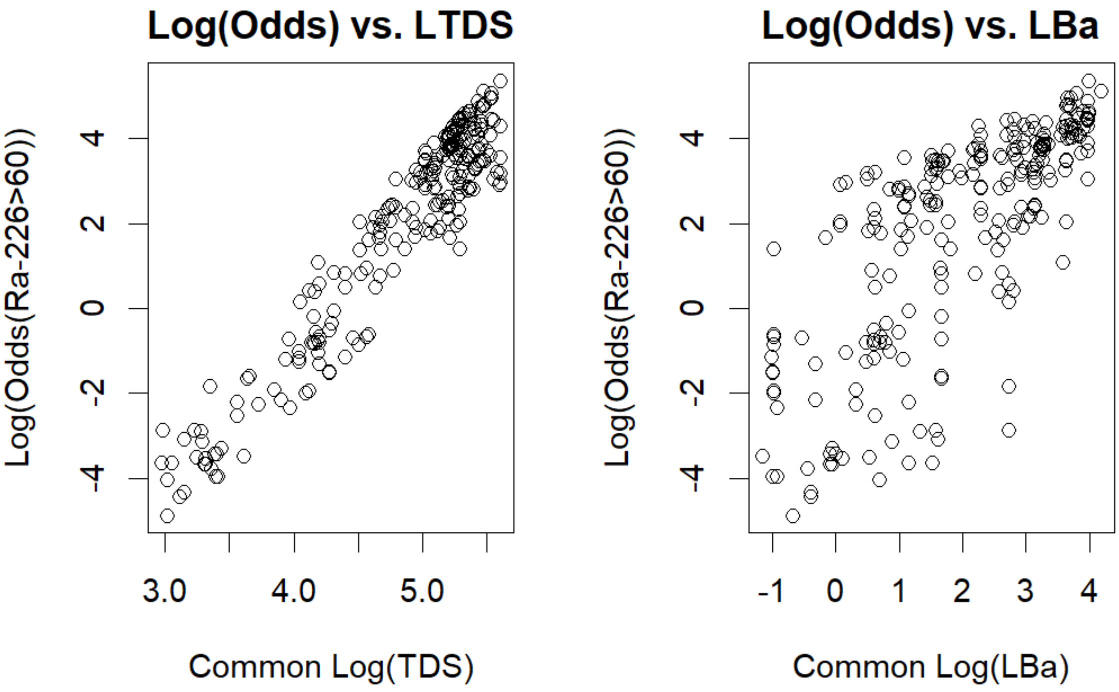


Figure 16. Log Odds Versus Predictors for GT60 Logistic Model (MI -*validate()*).

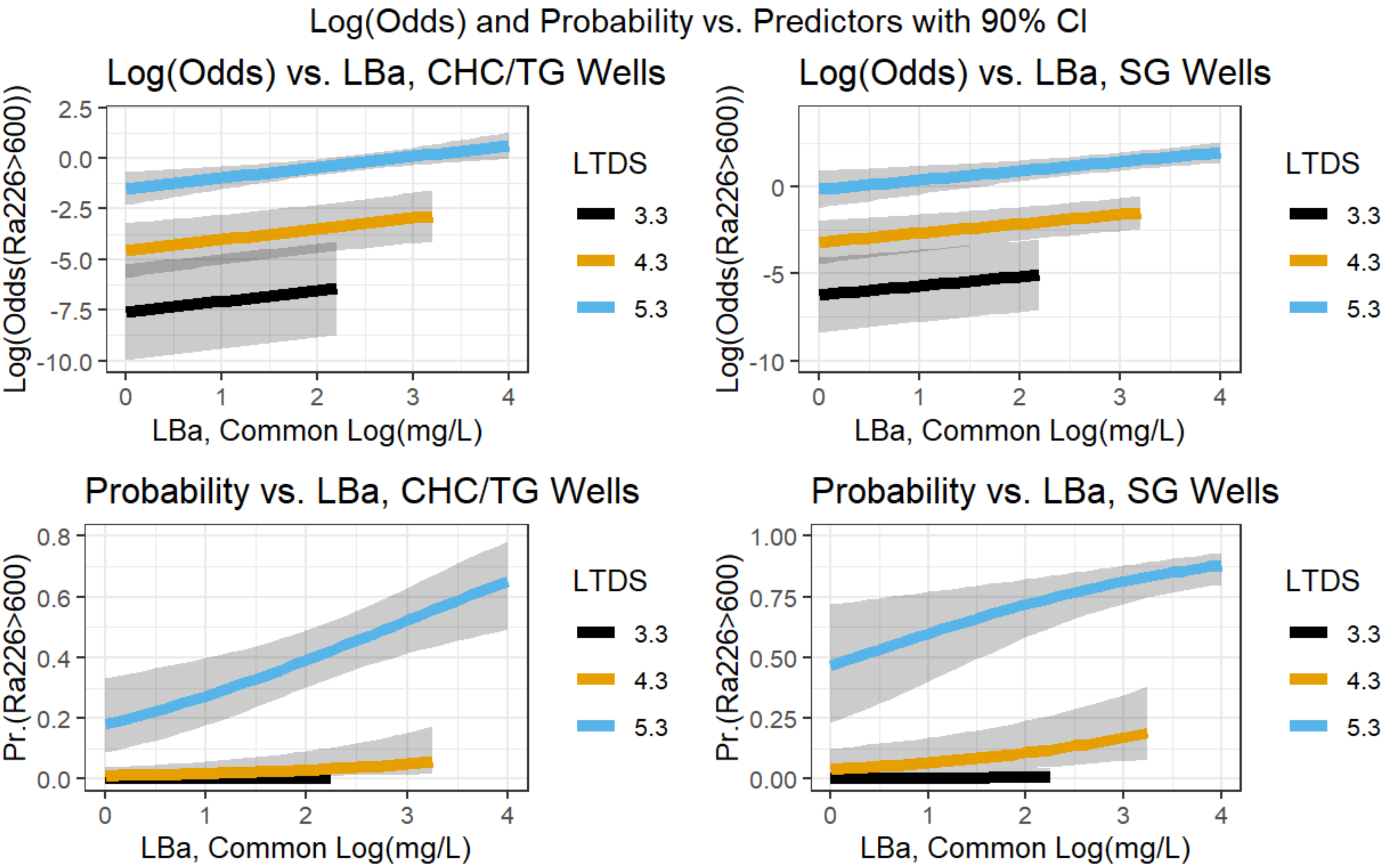


Figure 17. Log(Odds) and Probability Plots for GT600 Logistic Model (MI – *validate()*).

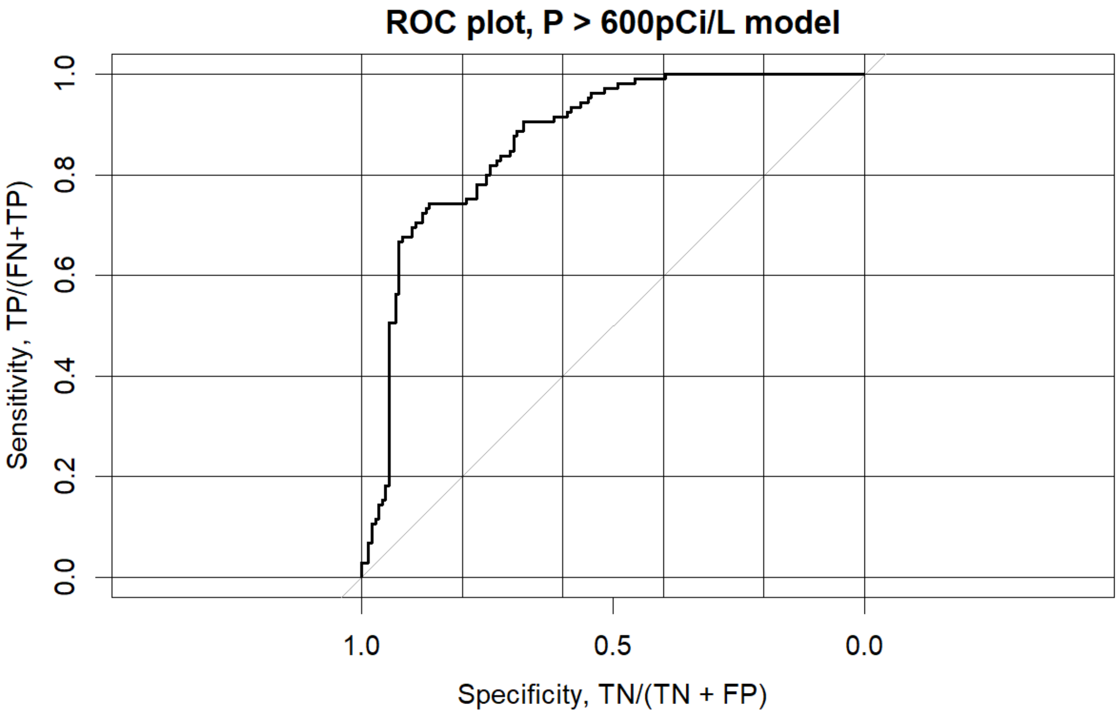


Figure 18. ROC plot for GT600 Logistic Model (*MI validate()*).

Chapter 6: Discussion and conclusions

This section provides a discussion of results and conclusions.

Discussion of Results

The datasets analyzed by MLR, and logistic regression(GT60 and GT600) differed only in that MLR required a continuous variable, LRa, while logistic regression required a dichotomous response variable (either GT60 or GT600). The dichotomous response variables were constructed by passive imputation from LRa using a deterministic formula, and these were not allowed to be used in any subsequent MI step.

General Model Form Results

The general form of the MLR, GT60 and GT600 models was determined by the MI within BS procedure, which involved using a loop that would:

1. do a bootstrap resample with replacement,
2. execute *mice()* to obtain a set of multiply imputed datasets,
3. use *fit.mult.impute()* to execute the standard linear model fitter *ols()* or logistic model fitter *lrm()* and apply Rubin's rules, based on the preliminary model forms given in Eq 2, 3, and 4,
4. Then fastbw() function from the **RMS** package was executed to identify significant coefficients for main effects models. All three models, MLR, GT60 and GT600, were fit in each resample and the vectors of their resulting model coefficients and

performance measures were appended as records in their respective tables for future reference,

5. The loop was repeated for each of 200 resamples.
6. After exiting the loop, the proportion of bootstrap resamples was obtained from the tables generated in step 4, in which each model parameter in the preliminary model was significant. This resulted in selection of the final MLR models form of the type given in Eq 5, the final form of the GT60 logistic model given in Eq 6. The GT600 logistic model suggested by this procedure would have had the predictors W.Type, LTDS and LFeT, and is provided in Table B-4. I rejected it in favor of a logistic model, Eq 7 with predictors W.Type, LTDS, and LBa that provided nearly the same performance measures and had significantly lower rates of missingness of predictor variables.

This entire process was repeated using the final model forms given in Eq 5, 6, and 7, to produce the final models provided in Table 7. The question of final model forms was not revisited in subsequent analyses.

MLR Result from MI – Validate() Procedure

The MLR model as presented in Table 10 is based on the general model form in Eq 5. The process involved MI, model fits, applying Rubin's rules, then bootstrap evaluation of model performance optimism using the **RMS** package function *validate().* Unfortunately, this procedure does not yield a table of bootstrap coefficients, but it produces a data object that facilitates producing graphics.

The QQ-plot for this model is slightly skewed (Figure 12, and the standardized residuals exhibit a kurtosis of 3.37, which is slightly broader than is expected if they were normally distributed. A Shapiro-Wilk test of normality of the standardized residuals gave a test statistic of W=0.993 and p-value of 0.334; thus, there was strong evidence that they are approximately normally distributed. For the Table 10 model, the variance inflation factors for LTDS and LBa were 1.47, which suggests that there is not strong collinearity between the predictors.

The optimism in the R^2 value for the MLR fit was 0.0024 as estimated by the *validate()* function, which is quite small compared to R^2value for the model fit, 0.731

Figure 13 provides the MI-*validate()* MLR regression lines, confidence intervals and prediction intervals for selected LTDS and LBa values. The figures take into account that the barium never exceeds approximately 5% of the TDS concentration. LTDS is a much stronger predictor of LRa than LBa. Figure 13 suggests that a new observation from the data distribution with an LTDS of:

- 3.3 (approximately 7% the TDS of seawater) is unlikely to have an LRa value greater than about 1.8 (63pCi/L), and
- 5.3 (approximately 6 times the TDS of seawater) is unlikely to have an LRa value of less than 1.3 at low barium concentrations and 2 at high barium concentrations.

Tables 10 and 11 are MLR models produced by the MI – *validate()* procedure, and were run under identical conditions, but with different imputation order. The differences in the performance measures and coefficients for the two models were slight. The coefficients of linear determination, after adjustment for optimism, were 0.7283 and 0.7230 for the monotone and reverse monotone cases respectively. The monotone imputation order yielded a model with slightly smaller standard errors of coefficients but the regression coefficients for each model agreed to within a fraction of their standard errors. Overall the imputation order had a small effect on model effect. The effect of imputation order on the rate of convergence of the predictive mean matching algorithm was not explored further.

Comparison of MI within BS and MI-validate() MLR models

The MLR models in Table 7, column 1 and Table 10 may be compared. The regression coefficients are very similar. The standard errors of coefficients obtained with the MI-*validate()* model are approximately 10 to 20% smaller than those for the model produced by MI within BS technique and provided in Table 7 column one. The R^2 values are nearly identical, 0.731 vs. 0.735, with the MI within BS procedure having the higher score.

Plots, like those provided in Figure 13, but based on the MI within BS modeling are expected to have slightly broader prediction intervals. For making predictions, Figure 13 is useful for crude estimates, but it would be preferable to produce estimates of $E(LRa_{new\ obs}|LTDS, LBa)$ and its and confidence and prediction intervals based on the table of bootstrap coefficients that was produced during step 4 of the section entitled “General Form Results.”

Complete Case MLR Model

The complete case dataset (n=162) is addressed by Table 8, column 1, (as a result of modeling with a bootstrap loop) and in Table 9, based on running the model fitter *ols()* and *validate()* functions. In both cases, the model fitted was based on the final model form given by Eq 5. The average of the standard deviations of regression coefficients obtained from the explicit bootstrap calculation were 10 to 30% larger than those obtained from running *ols()* and *validate().* The reported R^2 values were virtually identical, 0.773 vs. 0.774. The diagnostic plots for the *ols() – validate()* model appear to indicate an approximately normal distribution of standardized residuals. The Shapiro-Wilk statistic for the *ols() - validate()* model was W=0.987 (p-value = 0.165), which indicates that there was strong evidence that the distribution of standardized residuals was approximately normal. Generally, the intercept and all of the coefficients of the ols() – validate() MLR model in Table 9 were further from zero that were those from the MI – *validate()* MLR model in Table 10.

The *ols() – validate()* MLR model considers only complete cases, and the missing data pattern is clearly not MCAR. Consequently, it may be more biased than the models that addressed missing data and employed MI.

GT60 and GT 600 Logistic Model Results

The procedure for developing the GT60 and GT600 models is as described in the section entitled "General Form Results." Generally, it involved performing multiple imputation, model fitting, applying Rubin's rules and then accumulating coefficients and model fit measures on a series of bootstrap resamples within a loop, with final model results provided in Table 7. This process yields a table of coefficients that can be used to provide bootstrap estimated confidence intervals for new observations.

The GT60 and GT600 final model forms, Eq 6 and 7, were also fitted using the procedure described in the section entitled "*MLR Result from MI – Validate() Model*" with two exceptions. They were fitted as generalized linear models using the **RMS** function, *lrm()* and the response variable LRa was replaced by corresponding dichotomous response variables, GT60 or GT600. The **RMS** package function *validate()* was then used to bootstrap estimates of model fit and optimism, but not regression coefficients. This procedure also facilitated the preparation of graphics, which would be more difficult to produce using data from the MI within BS procedure.

The variance inflation factors for the GT60 model in Table 12 (LTDS, LBa =1.06) suggests that collinearity of predictor variables is minimal. The area under the receiver operating characteristic curve, C = 0.941 is rather close to the ideal limiting value of 1, Figure 15. The estimated optimism of 0.002 in this statistic is slight, suggesting that overfitting is not a serious problem. It also is nearly identical to the average C value reported in Table 7, column 2, 0.940.

The GT60 logistic model produced by the MI – *validate()* procedure (Table 12) had coefficients that were a bit closer to zero than the model fitted by the MI within BS procedure (Table 7 column 2).

The Shapiro-Wilk test statistic for normality of standardized residuals for the model was not determined because residuals from logistic regression are not required to be normally distributed or to exhibit homoskedasticity. There is an approximately linear relationship between log(odds(GT60)) and the individual predictor variables as Figure 16 demonstrates.

Overall, the GT60 logistic model should have good power at discriminating between PW with concentrations above 60pCi/L and below, provided that new observations are drawn from a distribution that is identically distributed to the data that was used to fit the model.

The practical interpretation of the results for the GT60 logistic model is that it has a sensitivity of 0.729 and a specificity of 0.918. That is to say:

- of those units that truly have a Ra-226 concentration greater than 60pCi/L, about 73% should have a positive test; and
- of those units that truly have a Ra-226 concentration less than 60pCi/L, about 92% should have a negative test.

The practical interpretation of Figure 14 (right panel) is:

- At low concentrations of TDS in PW, the concentration of radium-226 is unlikely to exceed 60pCi/L,
- At concentrations of TDS in PW that are about 60% of that in seawater (*i.e.,* LTDS ~ 4.3), the probability of radium-226 exceeding 60pCi/L strongly depends on the barium concentration, and

- At concentrations of TDS in PW that are about 6 times that found in seawater (*i.e.,* LTDS = 5.3) , the probability of radium-226 exceeding 60pCi/L is high, regardless of the barium concentration.

The variance inflation factors for the GT600 model in Table 13 (LTDS = 1.29, LBa =1.12, W.Type=1.42) suggest that collinearity of predictor variables is minimal. The area under the receiver operating characteristic curve, C = 0.862 is reasonably close to the ideal limiting value of 1, Figure 18. The estimated optimism of 0.006 in this statistic is slight, suggesting that overfitting is not a serious problem. It also is nearly identical to the average C value reported in Table 7, column 3, 0.866.

The GT600 logistic model produced by the MI – *validate()* procedure (Table 12) had coefficients for the intercept and continuous variables that were a bit closer to zero than the model fitted by the MI within BS procedure (Table 7 column 2). There is an approximately linear relationship between log(odds(GT600)) and the individual predictor variables, but the graphic was not provided because it looks very similar to Figure 16. Overall, the GT600 logistic model has a reasonable amount of power at discriminating between PW with concentrations above 600pCi/L and below, provided that new observations are drawn from a distribution that is identically distributed to the data that was used to fit the model.

The practical interpretation of the results for the GT600 logistic model is that it has a sensitivity of 0.879 and a specificity of 0.714. That is to say:

- of those units that truly have a Ra-226 concentration greater than 600pCi/L, about 88% should have a positive test; and
- of those units that truly have a Ra-226 concentration less than 60pCi/L, about 71% should have a negative test.

The practical interpretation of Figure 17 is dependent on well type. For SG wells:

- At low concentrations of TDS (~2000mg/L), the concentration of radium-226 is unlikely to exceed 600pCi/L,
- At concentrations of TDS in PW that are about 60% of that in seawater (~20,000ppm), the probability of radium-226 exceeding 600pCi/L are well below 50%, and
- At concentrations of TDS in PW that are about ~200,000mg/L or 6 times that found in seawater, the probability of radium-226 exceeding 600pCi/L is significantly dependent on the barium concentration.

For CHC/TG wells:

- At low or intermediate concentrations of TDS (~2000mg/L and ~20,000mg/L respectively), the concentration of radium-226 is unlikely to exceed 600pCi/L,
- At concentrations of TDS in PW that are about ~200,000mg/L the probability of radium-226 exceeding 600pCi/L is markedly dependent on the barium concentration.

Practical Utility of Results: MLR Examples

The example well chemistry report in Appendix A identifies the well as #7 on the Adelaide lease in the Erath Field, operated by Phillips Oil Company, and the sampling date is given as September 25, 1986. No radium-226 concentration is provided, but the TDS and barium were reported as 132,173mg/L and 235mg/L respectively. No well serial number or API number is provided for the well; either of these would have been powerful identifiers.

The Louisiana Department of Natural Resources' SONRIS web database Operator History by Well page (LA DNR, 2022) has an entry for a well 7 in the Erath Field on the Adelaide Lease, operated by Phillips Petroleum Company. Drilling began on December 26, 1958, and the well was plugged and abandoned on November 19, 1987. This is almost certainly the well in question. No information is provided in SONRIS about the well type possibly because it was drilled 66 years ago and nearly all hydrocarbon wells being drilled at that time were conventional hydrocarbon.

The coefficient table from the MLR modeling using the MI within BS procedure has 200 sets of coefficients for intercept, LTDS, and LBa, these coefficients were used to calculate 200 estimates of LRa. The resulting vector of 200 LRa realizations is expected to be asymptotically normally distributed. A Shapiro-Wilk normality test of the vector yields a test statistic, W, of 0.9931 (p-value =0.5019), which is strong evidence that the calculated LRa values are indeed normally distributed. The mean value of LRa is 2.7218 with variance 0.04088, which corresponds to 527pCi/L. The standard deviation for prediction, *SP*, is calculated per Eq 8.

(Eq 8.) $SP = ((Variance) + 199 * Variance)^{1/2}$

The 90% prediction bounds are calculated per Eq 9.

(Eq 9.) $10^{mean\ LRa\ \pm 1.645 * SP}$,

which corresponds to the interval (59 to 4,709pCi/L).

The 90% and 80% lower prediction limits are calculated per Eq 10.

(Eq 10.) $10^{mean\ LRa - Z * SP}$,

Where Z is 1.282 and 0.8422 respectively. These correspond to 96 and 172pCi/L radium-226 respectively. Consequently, one can conclude that to a reasonable degree of scientific certainty, more probably than not, the concentration of radium-226 in the sample is well above the drinking water standard of 5pCi/L total radium. This conclusion is subject to the caveat that the water sample is from a population that is identically distributed as the PW data that was used to develop the MLR model.

Similar confidence interval and confidence bound calculations can be performed using the GT60 and GT600 models to calculate the point estimate and confidence intervals on the probability of an observation from an identically distributed populating exceeding 60 or 600pCi/L

Conclusions

The following concerns and conclusions from this study are offered below:

1. Most of the data (76%) used in the analysis comes from the Appalachian Basin, Table 2. Table B-3 provides a further breakdown of the Appalachian Basin data by geologic unit, and it indicates that 98 of those Appalachian Basin observations (51%) come from a single formation, the famous Marcellus Shale. Overall, 39% of the 254 observations in the study come from the Marcellus Shale. Consequently, the Appalachian Basin in general and the Marcellus Shale in particular may be overrepresented in the data, and the dataset might not be a simple random sample of all hydrocarbon wells in the United States.
2. There were only 6 TG wells in the dataset of 254 observations but this type of well has become increasingly important in the last 40 years, so this type of well is almost certainly underrepresented.
3. The dataset suffered from extreme missingness rates for variables that could be important, Table 4. These include attributes such as temperature, pressure, hydrogen sulfide, bisulfide, bicarbonate and carbonate. There were also high missingness rates for species such as pH, chloride, calcium, magnesium, strontium, sodium and potassium. The high missingness rates prevented calculation of ionic strengths, activity coefficients and generally a more elegant analysis.
4. Many of the continuous variables such as TDS, pH, sodium, chloride, and calcium proved to be highly correlated, as Figure 2 shows. Most of these variables also had unacceptable rates of missingness. Using principal components analysis to reduce the dimensionality of these variables was considered and rejected since the intended audience, attorneys and other non-statisticians, would find it too

confusing. It also would have required preparation of a much more complicated prediction matrix.

5. There were only four predictor variables in the prediction matrix: LBa, LTDS, LFeT and W.Type plus the continuous response variable, LRa. Out of 254 observations, 67 (26%) had no information for LBa, LTDS, and LFeT. This led to a dilemma: to discard or to keep. The 67 observations were ultimately retained.
6. A considerable amount of data is missing and the missingness clearly is not MCAR. I do not know for certain why values are missing for key analytes such as FeTot (total iron), barium, and total dissolved solids (TDS). Some operators may have had reasons for collecting or failing to collect certain types of data. For example, operators may have known that barite or celestine scaling was not a practical concern for PW from some formations; that knowledge might have caused some barium and strontium data to be MNAR. PW is regulated state by state, and in some instances, PW could have been analyzed in a certain way to demonstrate compliance with state regulations or to satisfy facility waste acceptance criteria; this could lead to a MAR pattern. Finally older data was generated when PW was less regulated and older data would also be more likely to be from CHC wells, as is evident from Tables 3 and 4.
7. The method of dichotomizing the W.Type variable, logistic regression, worked in a satisfactory manner, and it consistently categorized TG wells as CHC wells. This is consistent with how one would have re-categorized TG wells given that the

category had to be eliminated and they had only looked at the group average concentrations given in Table 3.

8. The left-censored radium-226 value that was MNAR was treated as if it were MAR, and MI yielded satisfactory estimates that were in line with detection limits that have been observed for a high TDS dataset.
9. Reasonably feasible internal validation measures have been performed by standard methods that include bootstrap resampling of observations and the use of the **RMS** function *validate()*. *Validate()* appears to resample residuals instead of observations. Unfortunately, the USGS DB already includes all or almost all of the data that is reasonably available. External validation would be beneficial, but it does not appear feasible unless the European Union or some other entity publishes a similar database.
10. Log transformation of the original data resulted in MLR models where the standardized residuals were approximately normally distributed.
11. The prediction intervals for the MLR models are very broad once the $\log_{10}$(Ra) values are transformed to conventional units, but still appear to be potentially useful.
12. The logistic models appear to satisfy the assumption that there be an approximately linear relationship between log odds of the response variables and individual predictors.
13. The conclusion that there is an approximately linear relationship between log(Ra-226) and log(TDS) has been reported previously by Kraemer (Kraemer & Reid, 1984).

14. The conclusion that there log(barium) is a predictor of log(radium-226) partly makes sense and partly is baffling. Using it as a predictor is an acknowledgement that available radium and available barium in a formation will behave very much like one another. However, the amount of radium-226 in a formation is usually a function of its uranium-238 concentration, not its barium concentration due to the serial radioactive decay of uranium-238 to radium-226 through a series of intermediates. This may partly explain the large amount of noise in the scatterplot of LRa vs LBa in Figure 9.

References

Andridge, R. R., & Little, R. J. A. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review = Revue Internationale De Statistique*, *78*(1), 40–64. https://doi.org/10.1111/j.1751-5823.2010.00103.x

BC. (undated). *Conventional versus Unconventional Oil and Gas*. British Columbia Ministry of Natural Gas Development and Minister Responsible for Housing. https://www2.gov.bc.ca/assets/gov/farming-natural-resources-and-industry/natural-gas-oil/petroleum-geoscience/conventional_versus_unconventional_oil_and_gas.pdf

Bell, K. G., Goodman, C., & Whitehead, W. L. (1940). Radioactivity of Sedimentary Rocks and Associated Petroleum. *Bull. Am. Assoc. of Petroleum Geologists*, *24*(9), 1529–1547.

Blackwell, R. S., Burgard, K., Clark, R., & Cuadra, J. D. (2021). *Implementation Manual for Management of Naturally Occurring Radioactive Material (NORM) , Final Draft*. Louisiana Department of Environmental Quality, Emergency and Radiation Services Division.

Breger, I. A., & Whitehead, W. L. (1951). *Radioactivity and the Origin of Petroleum in Proceedings Third World Petroleum Congress-Section I*. 421–427.

Curie, P., Curie, M., & Gustave, B. (1898). Sur une nouvelle substance fortement radio-active, contenue dans la pechblende (On a new, strongly radioactive substance contained in pitchblende). *Comptes Rendus*, *127*, 1215–1217.

Drever, J. L. (1997). *The Geochemistry of Natural Waters: Surface and Groundwater Environments* (3rd ed.). Prentice Hall.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Engle, M. A., Saraswathula, V., Thordsen, J. J., Morrissey, E. A., Gans, K. D., Blondes, M. S., Kharaka, Y. K., Rowan, E. L., & Reidy, M. E. (2019). *U.S. Geological Survey National Produced Waters Geochemical Database v2.3* [Data set]. U.S. Geological Survey. https://doi.org/10.5066/F7J964W8

Eriksson, G., & O’Hagan, L. A. (2021). Selling “Healthy” Radium Products With Science: A Multimodal Analysis of Marketing in Sweden, 1910–1940. *Science Communication*, *43*(6), 740–767. https://doi.org/10.1177/10755470211044111

Escott, P. (1984). *NRPB Report for the Department of Energy London: The Occurrence of Radioactive Contamination on Offshore Installations.*

Gott, G. B., & Hill, J. W. (1951). *Radioactivity of Some Oil Fields of Southeastern Kansas. Trace Elements Investigations Report 121*. United States Geological Survey.

GWB. (2021). *The Geochemist's Workbench®*. Aqueous Solutions LLC. https://www.gwb.com/

Haaker, R. F. (2021). *Personal Communication concerning data from Kern Broussard et al. V HilCorp Energy Company, et al.*

Harrell Jr., F. E. (2021a). *Package Hmisc: Harrell Miscellaneous*. https://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf

Harrell Jr., F. E. (2021b). *Package rms: Regression Modeling Strategies, Version 6.2.0*. https://hbiostat.org/R/rms/, https://github.com/harrelfe/rms

IAEA. (1990a). *The Environmental Behavior of Radium, Vol. 1. Technical Reports Series No. 310*. International Atomic Energy Agency, Vienna.

IAEA. (1990b). *The Environmental Behavior of Radium, Vol. 2, Technical Reports Series No. 310*. https://inis.iaea.org/collection/NCLCollectionStore/_Public/21/052/21052628.pdf

IAEA. (2016, September 5). *The Environmental Behaviour of Radium: Revised Edition, Technical Report Series 476* [Text]. IAEA.

https://www.iaea.org/publications/10478/the-environmental-behaviour-of-radium-revised-edition

Keith Schneider. (1990a, December 3). *Radiation Danger Found in Oilfields Across the Nation*. https://go-gale-com.libproxy.unm.edu/ps/retrieve.do?tabID=News&resultListType=RESULT_LIST&searchResultsType=MultiTab&hitCount=1&searchType=AdvancedSearchForm&currentPosition=1&docId=GALE%7CA175610611&docType=Article&sort=Relevance&contentSegment=ZXAY-MOD1&prodId=OVIC&pageNum=1&contentSet=GALE%7CA175610611&searchId=R1&userGroupName=albu78484&inPS=true

Keith Schneider. (1990b, December 24). *2 Suits on Radium Cleanup Test Oil Industry's Liability—Document—Gale In Context: Opposing Viewpoints*. https://go-gale-com.libproxy.unm.edu/ps/retrieve.do?tabID=News&resultListType=RESULT_LIST&searchResultsType=MultiTab&hitCount=1&searchType=AdvancedSearchForm&currentPosition=1&docId=GALE%7CA175605356&docType=Article&sort=Relevance&contentSegment=ZXAY-MOD1&prodId=OVIC&pageNum=1&contentSet=GALE%7CA175605356&searchId=R1&userGroupName=albu78484&inPS=true

Kraemer, T. F., & Reid, D. F. (1984). The occurrence and behavior of radium in saline formation water of the U.S. Gulf Coast region. *Chemical Geology*, *46*(2), 153–174. https://doi.org/10.1016/0009-2541(84)90186-4

LA DNR. (2022, January 21). *Louisiana Department of Natural Resources SONRIS Data Portal: Operator History by Well*. https://sonlite.dnr.state.la.us/pls/apex/f?p=108:9035:8487722954200:::::

Landis, J. D., Sharma, M., & Renock, D. (2018). Rapid desorption of radium isotopes from black shale during hydraulic fracturing. 2. A model reconciling radium extraction with Marcellus wastewater production. *Chemical Geology*, *500*, 194–206. https://doi.org/10.1016/j.chemgeo.2018.08.001

Landis, J. D., Sharma, M., Renock, D., & Niu, D. (2018). Rapid desorption of radium isotopes from black shale during hydraulic fracturing. 1. Source phases that control the release of Ra from Marcellus Shale. *Chemical Geology*, *496*, 1–13. https://doi.org/10.1016/j.chemgeo.2018.06.013

Langmuir, D., & Riese, A. C. (1985). The thermodynamic properties of radium. *Geochrmlca Y Cosmorhrmrca Acfa*, *49*, 1593–1601.

Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, *110*, 63–73. https://doi.org/10.1016/j.jclinepi.2019.02.016

O’Brien, R. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, *41*, 673–690. https://doi.org/10.1007/s11135-006-9018-6

Radionuclides Rule: A Quick Reference Guide, EPA 816-F-01-003. (2001). USEPA, Office of Water. https://nepis.epa.gov/Exe/ZyPDF.cgi?Dockey=30006644.txt

Rubin, D. B. (1986). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Smith, S. H. (2015). *Crude Justice: How I Fought Big Oil and Won, and What You Should Know About the New Environmental Attack on America*. BenBella Books.

T. F. Kraemer. (1987). *Radium Content of Central Mississippi Salt Basin Brines. Open File Report 87-694*.

Taylor, W. (1993). NORM In Produced Water Discharges in the Coastal Waters of Texas. *SPE\EPA Exploration &. Production Environmentlll Conference in San Antonio. Teoc_. March 7-10. 1993.*, 9.

Thoya, D., Waititu, A., Magheto, T., & Ngunyi, A. (2018). Evaluating Methods of Assessing “Optimism” in Regression Models. *American Journal of Applied Mathematics and Statistics*, *6*(4), 126–134. https://doi.org/10.12691/ajams-6-4-2

Unknown. (2021). *PHREEQC Version 3: PHREEQC--A Computer Program for Speciation, Reaction-Path, Advective Transport, and Inverse Geochemical Calculations*. United States Geological Survey. https://wwwbrr.cr.usgs.gov/projects/GWC_coupled/phreeqc.v1/index.html

US EPA, O. (2014). *Health Risk of Radon* [Overviews and Factsheets]. https://www.epa.gov/radon/health-risk-radon

USEIA. (2020). *Natural gas explained—U.S. Energy Information Administration (EIA)*. https://www.eia.gov/energyexplained/natural-gas/

USEIA. (2021, October 8). *Where our natural gas comes from—U.S. Energy Information Administration (EIA)*. U. S. Energy Information Administration Independent Statistics & Analysis. https://www.eia.gov/energyexplained/natural-gas/where-our-natural-gas-comes-from.php

USEPA, EMSL. (1980). *“Method 903.1: Radium-226 in Drinking Water Radon Emanation Technique.” Prescribed Procedures for Measurement of Radioactivity in Drinking Water, EPA/600/4/80/032.* USEPA.

van Buuren, S. (2018). *Flexible Imputation of Missing Data, 2nd Edition*. CRC Press Taylor & Francis Group.

van Buuren, S. & et al. (2022). *mice source: R/mice.R*. https://rdrr.io/cran/mice/src/R/mice.R

van Buuren, S., & Groothuis-Oudshoorn. (2021). *Package “mice” Multivariate Imputation by Chained Equations, version 3.14.0*. https://github.com/amices/mice, https://amices.org/mice/

WOGCC. (2021). *WOGCC Data Explorer concerning API # 49-007-23320*. https://dataexplorer.wogcc.wyo.gov/

List of Appendices

Appendix A: Example Water Chemistry Report

PETROLITE

Petrolite Corporation

369 Marshall Avenue • St. Louis, Missouri 63119
314 961-3500 • Telex: 44-2417

WATER ANALYSIS REPORT

Company: PHILLIPS OIL CO SOUTH EARTH FIELD	Sampling Date: 09/25/86 Analysis Date: 10/07/86 Sample ID: F22601

RESERVOIR
REXBY #3

Sample Source	Submitted by: CLIFTON, W.J.
➤ Lease: ADELAIDE	Sampled by:
Well: #7	Chem. Treatment:
Sample Pt: WELL HEAD	Sample Condition: OG/TRBD

ANALYTICAL RESULTS

pH at the time of sampling:	6.50	
pH at the time of analysis:	7.60	
Density:	1.085	
Hydrogen Sulfide (H2S):		
TDS: Calculated	132173.3	mg/L

CONSTITUENT		mg/L	meq/L	method	comments
ANIONS					
*Bicarbonate	HCO3-	240.0	3.93	FLD	
Boron	B(OH)4-	300.3	3.81	ICP	
*Carbonate	CO3--	.0	.00	N.A.	
*Chloride	Cl-	78200.0	2205.74	FIA	
Phosphate	PO4---	0.0	0.00	ICP	DL= 2.020
*Sulfate	SO4--	26.4	.55	FIA	
	SUM OF ANIONS=		2214.03		
CATIONS					
Aluminum	Al+++	0.0	0.00	ICP	DL=10.100
*Barium	Ba++	235.0	3.42	ICP	
*Calcium	Ca++	2653.0	132.39	ICP	
Chromium	Cr+++	0.0	0.00	ICP	DL=10.100
Copper	Cu++	0.0	0.00	ICP	DL= 2.020
*Iron	Fe++	0.0	0.00	ICP	DL= 2.020
Lead	Pb++	0.0	0.00	ICP	DL=10.100
Lithium	Li+	0.0	0.00	N.A.	
*Magnesium	Mg++	362.0	29.79	ICP	
Manganese	Mn++	5.6	.20	ICP	
Nickel	Ni++	0.0	0.00	ICP	DL= 2.020
Potassium	K+	469.0	12.00	ICP	
Silica	SiO2	7.9	0.00	ICP	
*Sodium	Na+	49450.0	2150.94	ICP	
*Strontium	Sr++	224.0	5.11	ICP	
Vanadium	V++	0.0	0.00	N.A.	
	SUM OF CATIONS=		2333.84		

Ratio of ANIONS:CATIONS .95

PETROLITE

Petrolite Corporation

369 Marshall Avenue • St. Louis, Missouri 63119
314 961-3500 • Telex: 44-2417

SATURATION INDEX TABLE

Sample ID: F22601
pH (at 25.0 deg C): 6.50

Temperature		Scale Component				
		CaCO3	CaSO4	CaSO4*2H2O	SrSO4	BaSO4
deg F	deg C	(Calcite)	(Anhydrite)	(Gypsum)	(Celestite)	(Barite)
32.00	.00	-.116	-2.828	-2.074	-1.737	1.914
68.00	20.00	.020	-2.617	-2.190	-1.800	1.534
77.00	25.00	.059	-2.568	-2.208	-1.807	1.444
104.00	40.00	.187	-2.426	-2.247	-1.811	1.184
140.00	60.00	.374	-2.240	-2.264	-1.786	.858
176.00	80.00	.573	-2.051	-2.255	-1.739	.553
212.00	100.00	.781	-1.852	-2.230	-1.678	.266

S.I.=SATURATION INDEX
S.I.=log(Product of activities of component ions/Ksp)

S.I. less than 0 — The water is undersaturated and indicates a non-scaling situation.
S.I. near or equal to 0. — The water is saturated and scale formation is likely.
S.I. greater than 0 — The water is supersaturated and favors scale formation.

POSSIBLE SCALE FORMATION

Temperature		Scale Component (mg/1000 g H2O)				
		CaCO3	CaSO4	CaSO4*2H2O	SrSO4	BaSO4
deg F	deg C	(Calcite)	(Anhydrite)	(Gypsum)	(Celestite)	(Barite)
32.00	.00	0.	0.	0.	0.	67.
68.00	20.00	2.	0.	0.	0.	66.
77.00	25.00	6.	0.	0.	0.	65.
104.00	40.00	21.	0.	0.	0.	63.
140.00	60.00	44.	0.	0.	0.	57.
176.00	80.00	70.	0.	0.	0.	47.
212.00	100.00	97.	0.	0.	0.	29.

The POSSIBLE SCALE FORMATION predicts the maximum amount of any one scale component that could precipitate from the water as analyzed. As precipitation progresses, these predictions become less accurate.

To estimate the POSSIBLE SCALE FORMATION in lbs/1000 barrels (US 42 gal) use the following:
APPROXIMATE lbs/1000 barrels = (mg/1000g H2O) x 0.35

Appendix B: Supplementary Tables and Figures

The USGS data dictionary is provided as Table B.1. Table B.2 provides missing records information for the 254 observations used in this analysis.

Table B.1. Data Dictionary and Percent of n=114,943 Records Missing (Engle *et al.*, 2019).

Variable Name	Description	Percent Missing
IDUSGS	Unique ID in this database	0%
IDORIG	ID in original database or publication	0%
IDDB	ID (name) of input database	0%
SOURCE	Source of data	34%
REFERENCE	Publication	94%
LATITUDE	Latitude	10%
LONGITUDE	Longitude	9%
LATLONGAPX	Description if LATITUDE or LONGITUDE are approximate	80%
API	API well number, 14 digits	36%
USGSREGION	USGS Region	0%
BASIN	Basin	0%
BASINCODE	Basin Code	39%
STATE	State	0%
STATECODE	State Code	0%
COUNTY	County	27%
COUNTYCODE	County Code	29%
FIELD	Field	16%
FIELDCODE	Field Code	51%
WELLNAME	Well name	13%
WELLCODE	Well Code	86%
WELLTYPE	Well type	0%
TOWNRANGE	Township, Range, Section, Quarter	80%
REGDIST	Regional District	83%
LOC	Location	96%
QUAD	Quad	100%
TIMESERIES	Order of time-series data	100%
DAY	Sample day of time-series data	98%
DATECOMP	Date of well completion	94%
DATESAMPLE	Date of sample collection	26%
DATEANALYS	Date of analysis	91%
METHOD	Sample Method	41%

Variable Name	Description	Percent Missing
OPERATOR	Well operator	73%
PERMIT	Well permit holder	93%
DFORM	Geologic formation name of greatest depth	79%
GROUP	Geologic group name	100%
FORMATION	Geologic formation name	0%
MEMBER	Geologic member name	98%
AGECODE	Geologic Age code	52%
ERA	Geologic Era name	0%
PERIOD	Geologic Period name	0%
EPOCH	Geologic Epoch name	80%
DEPTHUPPER	Upper perforation depth, ft. Depth added here if non-specific.	29%
DEPTHLOWER	Lower perforation depth, ft	41%
DEPTHWELL	Reported Total depth of well, ft	63%
ELEVATION	Elevation of well, ft	82%
LAB	Laboratory that analyzed the results	89%
REMARKS	Remarks or comments	93%
LITHOLOGY	Lithology	75%
POROSITY	Porosity, % reported	100%
TEMP	Temperature, deg F reported	97%
PRESSURE	Pressure, psi reported	99%
SG	Specific Gravity, reported or calculated (see text)	31%
SPGRAV	Specific Gravity, reported	46%
SPGRAVT	Temperature of Specific Gravity measurement, deg F	73%
RESIS	Resistivity, Ohm m	43%
RESIST	Temperature of Resistivity measurement, deg F	50%
PH	pH	25%
PHT	Temperature of pH measurement, deg F	99%
EHORP	Eh / Oxidation Reduction Potential, mV	100%
COND	Conductivity, µS/cm	99%
CONDT	Temperature of Conductivity measurement, deg F	100%
TURBIDITY	Turbidity	100%
HEM	Oil and Grease	100%
MBAS	Surfactants and Detergents	100%
UNITS	mg/L or ppm, applies to all chemistry unless specified	0%
TDSUSGS	Total Dissolved Solids, calculated (see text)	4%
TDS	Total Dissolved Solids, measured	15%
TDSCALC	Total Dissolved Solids, calculated, as reported in reference	98%
TSS	Total Suspended Solids	99%
CHARGEBAL	Charge balance of major ions, %, reported	97%
chargebalance	Charge balance of major ions, %, calculated	5%
Ag	Silver	100%
Al	Aluminum	99%

Variable Name	Description	Percent Missing
As	Arsenic	100%
Au	Gold	100%
B	Boron	96%
BO3	Borate	100%
Ba	Barium	89%
Be	Beryllium	100%
Bi	Bismuth	100%
Br	Bromide	94%
CO3	Carbonate	91%
HCO3	Bicarbonate	14%
Ca	Calcium	6%
Cd	Cadmium	100%
Cl	Chloride	5%
Co	Cobalt	100%
Cr	Chromium	98%
Cs	Cesium	100%
Cu	Copper	99%
F	Fluoride	99%
FeTot	Iron, total	76%
FeIII	Iron, 3+	100%
FeII	Iron, 2+	99%
FeS	Iron sulfide	100%
FeAl	Iron plus Aluminum, reported as elements	100%
FeAl2O3	Iron plus Aluminum, reported as oxides	100%
Hg	Mercury	100%
I	Iodine	97%
K	Potassium	73%
KNa	Potassium plus Sodium	93%
Li	Lithium	95%
Mg	Magnesium	10%
Mn	Mangansese	97%
Mo	Molybdenum	100%
N	Nitrogen, total	100%
NO2	Nitrite	100%
NO3	Nitrate	97%
NO3NO2	Nitrate plus Nitrite	100%
NH4	Ammonium	99%
TKN	Kjeldahl Nitrogen	100%
Na	Sodium	16%
Ni	Nickel	100%
OH	Hydroxide	100%
P	Phosphorus	100%

Variable Name	Description	Percent Missing
PO4	Phosphate	100%
Pb	Lead	100%
Rh	Rhodium	100%
Rb	Rubidium	99%
S	Sulfide	100%
SO3	Sulfite	100%
SO4	Sulfate	19%
HS	Bisulfide	100%
Sb	Antimony	100%
Sc	Scandium	100%
Se	Selenium	100%
Si	Silica	97%
Sn	Tin	100%
Sr	Strontium	93%
Ti	Titanium	100%
Tl	Thallium	100%
U	Uranium	100%
V	Vanadium	100%
W	Tungsten	100%
Zn	Zinc	99%
ALKHCO3	Alkalinity as HCO3	99%
ACIDITY	Acidity as CaCO3	100%
DIC	Dissolved Inorganic Carbon	100%
DOC	Dissolved Organic Carbon	100%
TOC	Total Organic Carbon	100%
CN	Cyanide	100%
BOD	Biochemical Oxygen Demand	100%
COD	Chemical Oxygen Demand	100%
BENZENE	Benzene	99%
TOLUENE	Toluene	99%
ETHYLBENZ	Ethybenzene	100%
XYLENE	Xylene	100%
ACETATE	Acetate	99%
BUTYRATE	Butyrate	100%
FORMATE	Formate	100%
LACTATE	Lactate	100%
PHENOLS	Phenols	100%
PERC	Tetrachloroethylene	100%
PROPIONATE	Propionate	100%
PYRUVATE	Pyruvate	100%
VALERATE	Valerate	100%
ORGACIDS	Total Organic Acids	100%

Variable Name	Description	Percent Missing
Ar	Argon gas	100%
CH4	Methane gas	100%
C2H6	Ethane gas	100%
CO2	Carbon Dioxide gas	99%
H2	Hydrogen gas	100%
H2S	Hydrogen Sulfide gas	97%
He	Helium gas	100%
N2	Nitrogen gas	100%
NH3	Ammonia gas	100%
O2	Oxygen gas	100%
ALPHA	Alpha particles, pCi/L	100%
BETA	Beta particles, pCi/L	100%
dD	δH, per mil	99%
H3	Tritium, 3H, tritium units	100%
d7Li	δ7Li, per mil	100%
d11B	δ11B, per mil	100%
d13C	δ13C, per mil	100%
C14	14C, pCi/L	100%
d18O	δ18O, per mil	99%
d34S	δ34S, per mil	100%
d37Cl	δ37Cl, per mil	100%
K40	40K, pCi/L	100%
d81Br	δ81Br	100%
Sr87Sr86	87Sr/86Sr	99%
I129	129I/I, parts per quadrillion	100%
Rn222	222Rn, pCi/L	100%
Ra226	226Ra, pCi/L	99%
Ra228	228Ra, pCi/L	100%
cull_PH	“X” if pH < 4.5 or pH > 10.5	98%
cull_MgCa	“X” if Mg > Ca	96%
cull_KCl	“X” if K > Cl	100%
cull_K5Na	“X” if K > 5xNa	100%
cull_chargeb	“X” if charge balance > 15%	79%

Table B.2 Data Dictionary and Percent Missingness, n=254 (after Engle *et al*, 2019).

Variable Name	Description	% Missing
IDUSGS	Unique ID in this database	0
IDORIG	ID in original database or publication	0
IDDB	ID (name) of input database	0
SOURCE	Source of data	1.2
REFERENCE	Publication	15.7
LATITUDE	Latitude	3.1
LONGITUDE	Longitude	3.1
LATLONGAPX	Description if LATITUDE or LONGITUD	98
API	API well number, 14 digits	53.1
USGSREGION	USGS Region	0
BASIN	Basin	0
BASINCODE	Basin Code	100
STATE	State	0
STATECODE	State Code	0
COUNTY	County	44.9
COUNTYCODE	County Code	100
FIELD	Field	77.6
FIELDCODE	Field Code	100
WELLNAME	Well name	30.3
WELLCODE	Well Code	100
WELLTYPE	Well type	0
TOWNRANGE	Township, Range, Section, Quarter	77.6
REGDIST	Regional District	44.5
LOC	Location	100
QUAD	Quad	98
TIMESERIES	Order of time-series data	100
DAY	Sample day of time-series data	100
DATECOMP	Date of well completion	100
DATESAMPLE	Date of sample collection	77.6
DATEANALYS	Date of analysis	84.3
METHOD	Sample Method	34.6
OPERATOR	Well operator	77.6
PERMIT	Well permit holder	99.6
DFORM	Geologic formation name of greatest	100
GROUP	Geologic group name	100
FORMATION	Geologic formation name	0
MEMBER	Geologic member name	100
AGECODE	Geologic Age code	100
ERA	Geologic Era name	0
PERIOD	Geologic Period name	0
EPOCH	Geologic Epoch name	100

Variable Name	Description	% Missing
DEPTHUPPER	Upper perforation depth, ft. Depth	79.1
DEPTHLOWER	Lower perforation depth, ft	79.1
DEPTHWELL	Reported Total depth of well, ft	100
ELEVATION	Elevation of well, ft	100
LAB	Laboratory that analyzed the result	84.3
REMARKS	Remarks or comments	69.3
LITHOLOGY	Lithology	71.7
POROSITY	Porosity, % reported	100
TEMP	Temperature, deg F reported	90.6
PRESSURE	Pressure, psi reported	100
SG	Specific Gravity, reported or calculated	91.3
SPGRAV	Specific Gravity, reported	91.3
SPGRAVT	Temperature of Specific Gravity meas.	100
RESIS	Resistivity, Ohm m	97.6
RESIST	Temperature of Resistivity measurement	100
PH	pH	69.7
PHT	Temperature of pH measurement, deg	98
EHORP	Eh / Oxidation Reduction Potential,	98
COND	Conductivity, Î¼S/cm	67.3
CONDT	Temperature of Conductivity measure	73.6
TURBIDITY	Turbidity	100
HEM	Oil and Grease	85.4
MBAS	Surfactants and Detergents	85
UNITS	mg/L or ppm, applies to all chemist	0
TDSUSGS	Total Dissolved Solids, calculated	26.4
TDS	Total Dissolved Solids, measured	30.3
TDSCALC	Total Dissolved Solids, calculated,	95.7
TSS	Total Suspended Solids	94.1
CHARGEBAL	Charge balance of major ions, %, re	100
chargebalance	Charge balance of major ions, %, ca	31.5
Ag	Silver	100
Al	Aluminum	81.5
As	Arsenic	93.7
Au	Gold	100
B	Boron	93.7
BO3	Borate	100
Ba	Barium	35.8
Be	Beryllium	100
Bi	Bismuth	100
Br	Bromide	57.5
CO3	Carbonate	96.9
HCO3	Bicarbonate	89.4
Ca	Calcium	31.5

Variable Name	Description	% Missing
Cd	Cadmium	92.9
Cl	Chloride	29.5
Co	Cobalt	98.8
Cr	Chromium	89
Cs	Cesium	93.3
Cu	Copper	92.1
F	Fluoride	92.9
FeTot	Iron, total	47.2
FeIII	Iron, 3+	100
FeII	Iron, 2+	100
FeS	Iron sulfide	100
FeAl	Iron plus Aluminum, reported as elemental	100
FeAl2O3	Iron plus Aluminum, reported as oxides	100
Hg	Mercury	97.2
I	Iodine	91.3
K	Potassium	74.4
KNa	Potassium plus Sodium	100
Li	Lithium	60.2
Mg	Magnesium	46.1
Mn	Manganese	61
Mo	Molybdenum	96.1
N	Nitrogen, total	100
NO2	Nitrite	100
NO3	Nitrate	99.6
NO3NO2	Nitrate plus Nitrite	97.2
NH4	Ammonium	93.7
TKN	Kjeldahl Nitrogen	94.1
Na	Sodium	44.5
Ni	Nickel	92.1
OH	Hydroxide	100
P	Phosphorus	100
PO4	Phosphate	99.6
Pb	Lead	87
Rh	Rhodium	100
Rb	Rubidium	93.3
S	Sulfide	98
SO3	Sulfite	99.6
SO4	Sulfate	66.9
HS	Bisulfide	100
Sb	Antimony	100
Sc	Scandium	100
Se	Selenium	98
Si	Silica	98.4

Variable Name	Description	% Missing
Sn	Tin	99.6
Sr	Strontium	53.5
Ti	Titanium	100
Tl	Thallium	99.6
U	Uranium	100
V	Vanadium	100
W	Tungsten	100
Zn	Zinc	74.4
ALKHCO3	Alkalinity as HCO3	86.2
ACIDITY	Acidity as CaCO3	96.9
DIC	Dissolved Inorganic Carbon	98
DOC	Dissolved Organic Carbon	93.3
TOC	Total Organic Carbon	99.6
CN	Cyanide	100
BOD	Biochemical Oxygen Demand	94.5
COD	Chemical Oxygen Demand	94.1
BENZENE	Benzene	100
TOLUENE	Toluene	100
ETHYLBENZ	Ethybenzene	100
XYLENE	Xylene	99.6
ACETATE	Acetate	93.7
BUTYRATE	Butyrate	100
FORMATE	Formate	100
LACTATE	Lactate	100
PHENOLS	Phenols	98.8
PERC	Tetrachloroethylene	100
PROPIONATE	Propionate	100
PYRUVATE	Pyruvate	100
VALERATE	Valerate	100
ORGACIDS	Total Organic Acids	100
Ar	Argon gas	100
CH4	Methane gas	100
C2H6	Ethane gas	100
CO2	Carbon Dioxide gas	100
H2	Hydrogen gas	100
H2S	Hydrogen Sulfide gas	98
He	Helium gas	100
N2	Nitrogen gas	100
NH3	Ammonia gas	93.3
O2	Oxygen gas	98.4
ALPHA	Alpha particles, pCi/L	63
BETA	Beta particles, pCi/L	64.2
dD	H, per mil	90.9

Variable Name	Description	% Missing
H3	Tritium, 3H, tritium units	100
d7Li	7Li, per mil	100
d11B	11B, per mil	98
d13C	13C, per mil	98
C14	14C, pCi/L	100
d18O	18O, per mil	90.9
d34S	34S, per mil	99.6
d37Cl	37Cl, per mil	100
K40	40K, pCi/L	81.1
d81Br	81Br	100
Sr87Sr86	87Sr/86Sr	98.4
I129	129I/I, parts per quadrillion	100
Rn222	222Rn, pCi/L	100
Ra226	226Ra, pCi/L	0
Ra228	228Ra, pCi/L	28.3
cull_PH	X if pH < 4.5 or pH > 10.5	100
cull_MgCa	X if Mg > Ca	100
cull_KCl	X if K > Cl	100
cull_K5Na	X if K > 5xNa	100
cull_chargeb	X if charge balance > 15%	50

Table B-3. Frequency of observations by formation for the Appalachian Basin.

Formation	Appalachian (N=192)
Bass Islands Dolomite	5 (2.6%)
Bradford Gp	1 (0.5%)
Catskill & Lock Haven Groups	4 (2.1%)
Fifty Foot Sand	1 (0.5%)
Helderberg Ls	2 (1.0%)
Huntersville Chert	3 (1.6%)
Kane Sand	1 (0.5%)
Lock Haven Fm	3 (1.6%)
Marcellus Shale	98 (51.0%)
Medina Gp	36 (18.8%)
Onondaga Ls	1 (0.5%)
Oriskany Ss	7 (3.6%)
Queenston Shale	5 (2.6%)
Red Valley Sand	1 (0.5%)
Theresa Fm	3 (1.6%)
Tuscarora Fm	1 (0.5%)
Unknown	10 (5.2%)
Upper Devonian	5 (2.6%)

Formation	Appalachian (N=192)
Venango Gp	4 (2.1%)
Warren Sand	1 (0.5%)

Figure B-1. Variable Behavior during Multiple Imputation.

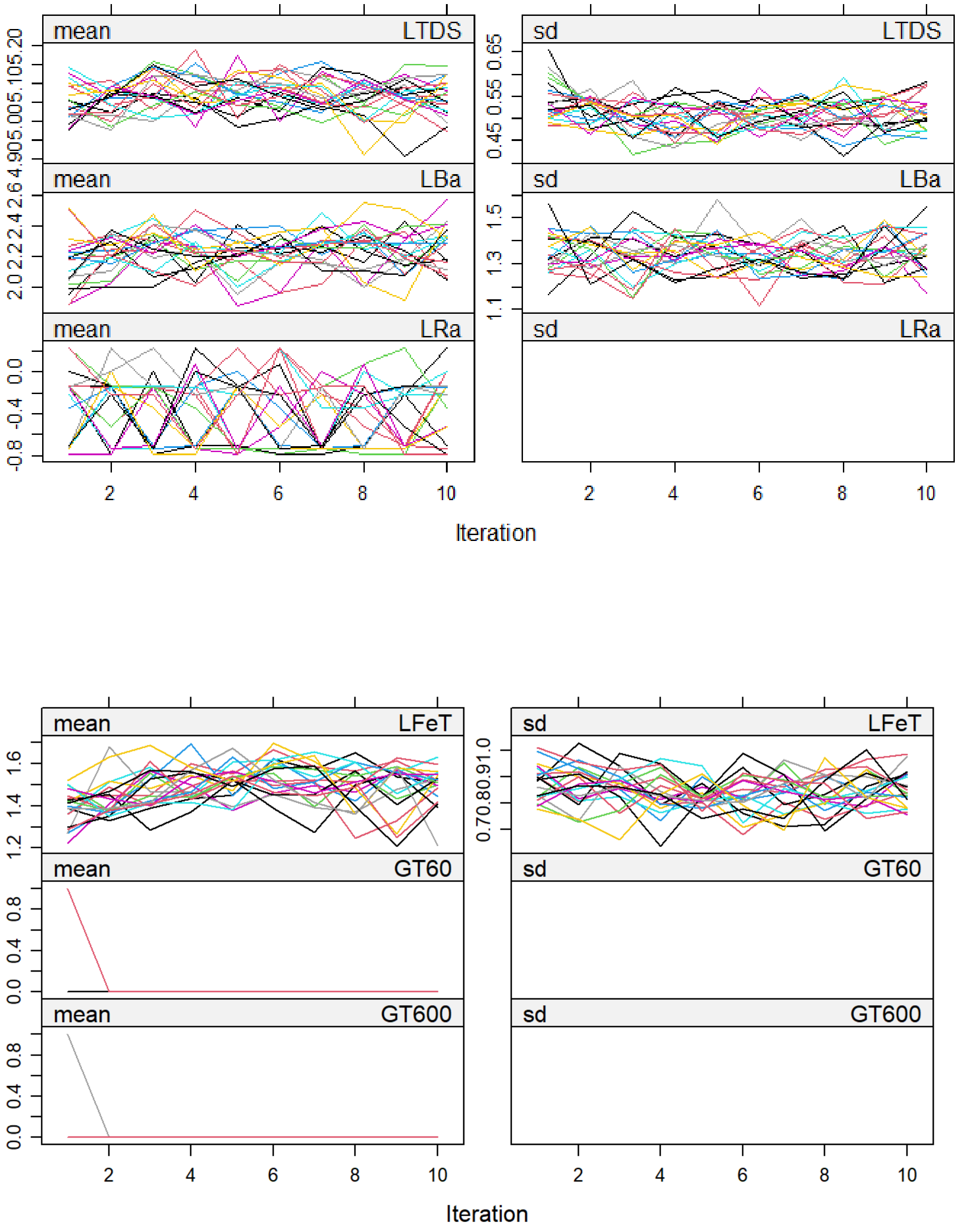


Table B-4. Alternative GT600 Logistic Model, BS(n=200)/MI.

```
GT600 Logistic Model
Factor significance proportion
LFeT        LTDS       W.Type
0.99        1          1
```

```
Coefficients and Measures
               Mean      SD
Intercept  -17.0559  2.5636
LTDS         2.7367  0.5281
LFeT         1.2384  0.3014
W.Type=SG    1.6536  0.3777
Obs        254.0000  0.0000
Max Deriv    0.0000  0.0000
Model L.R. 129.7270 13.5978
d.f.         3.0000  0.0000
P            0.0000  0.0000
C            0.8723  0.0185
Dxy          0.7447  0.0369
Gamma        0.7448  0.0369
Tau-a        0.3616  0.0198
R2           0.5376  0.0422
Brier        0.1395  0.0100
g            3.1851  0.3797
gr          27.1594 11.3453
gp           0.3652  0.0186
```

Table B-5. Alternative GT600 Logistic Model, MI without Prior Bootstrap.

```
Logistic Regression Model, MI without bootstrap.

 fit.mult.impute(formula = GT600 ~ LTDS + LFeT + W.Type.

                         Model Likelihood        Discrimination
                            Ratio Test                  Indexes
 Obs           254     LR chi2     127.54    R2       0.531    C       0.870
  0            149     d.f.             3    g        3.129    Dxy     0.740
  1            105     Pr(> chi2) <0.0001    gr      23.810    gamma   0.740
 max |deriv| 1e-06                           gp       0.364    tau-a   0.360
                                             Brier    0.142

           Coef     S.E.   Wald Z Pr(>|Z|)
 Intercept -16.9564 3.6125 -4.69  <0.0001
 LTDS        2.7442 0.7104  3.86  0.0001
 LFeT        1.1682 0.4918  2.38  0.0175
 W.Type=SG   1.6382 0.3976  4.12  <0.0001

Variance inflation factors:  1.38, 1.10, 1.27
```

Table B-6. Alternative GT600 Logistic Model, Bootstrap (n=200) without MI.

```
 GT600 Logistic Model
                  Mean       SD
 Intercept     -15.036    4.506
 LTDS             1.61    0.886
 LFeT            2.951    0.795
 W.Type=SG       2.217    0.967
 Obs               120        0
 Max Deriv           0        0
 Model L.R.     77.899   12.407
 d.f.                3        0
 P                   0        0
 C               0.895    0.032
 Dxy              0.79    0.064
 Gamma           0.791    0.064
 Tau-a           0.394    0.033
 R2              0.635    0.071
 Brier           0.117     0.02
 g               4.254    0.922
 gr             129.11  344.401
 gp              0.405    0.027
```

Table B-7. Results and Performance Measures for the MLR Regression (MI – *validate()*); Observations with LTDS= NA excluded.

```
Model: LRa ~ LTDS + LBa
Imputation Order: reverse monotone
                Model Likelihood    Discrimination
                      Ratio Test           Indexes
 Obs     187    LR chi2    270.63    R2        0.765
 sigma0.5810    d.f.            2    R2 adj    0.762
 d.f.    184    Pr(> chi2) 0.0000    g         1.131

          Coef    S.E.   t      Pr(>|t|)
 Intercept -3.8072 0.3129 -12.17 <0.0001
 LTDS       1.1718 0.0720  16.27 <0.0001
 LBa        0.2153 0.0370   5.82 <0.0001

                Analysis of Variance          Response: LRa

 Factor     d.f. Partial SS MS         F      P
 LTDS         1   89.30030  89.3003007 264.58 <.0001
 LBa          1   11.43527  11.4352703  33.88 <.0001
 REGRESSION   2  194.93946  97.4697298 288.79 <.0001
 ERROR      184   62.10241   0.3375131

           index.orig training    test optimism index.corrected   n
 R-square      0.7674   0.7698  0.7631   0.0067          0.7607 250
 MSE           0.3285   0.3217  0.3346  -0.0129          0.3414 250
 g             1.1327   1.1329  1.1320   0.0009          1.1317 250
 Intercept     0.0000   0.0000 -0.0037   0.0037         -0.0037 250
 Slope         1.0000   1.0000  1.0007  -0.0007          1.0007 250


Variance Inflation Factors:
       LTDS   LBa
       1.47   1.47

Shapiro-Wilk normality test of regression residuals:
       W = 0.98845, p-value = 0.1326
```