Modelling Hydrocarbon Concentrations in Groundwater Monitoring Networks

Josh Cowley

Thesis submitted for the degree of

Doctor of Philosophy

School of Mathematics, Statistics & Physics

Newcastle University

Newcastle upon Tyne

United Kingdom

Under joint supervision of Daniel Henderson

and Colin Gillespie

June 2024

— The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.

John W. Tukey (1986)

Abstract

Groundwater networks provide a critical resource across the world by supplying fresh water for a wide array of scenarios including extraction for drinking water and irrigation. Protection of these naturally occurring geological features is an important component of the wider climate problem. Pollution to groundwater networks can occur in many forms, including nitrates, radioactive material and the focus of this thesis, hydrocarbons. Due to their carcinogenic nature, data is collected at groundwater monitoring sites for regulatory compliance and to ensure safe concentration levels are not exceeded. Data collection involves extraction of a water sample, *in situ*, to be later analysed in a laboratory capable of measuring hydrocarbon concentrations above a certain "non-detection" limit. This process is less than desirable as our data is left-censored at laboratory-dependent thresholds and it requires the construction of several groundwater monitoring wells. Furthermore, observations may be missed due to faulty wells, unsafe working conditions and other potential obstructions.

Hence, the aim of this thesis is to investigate whether statistical modelling of hydrocarbon concentrations based on measurements of predictors that are easier to obtain can provide more insight with less information. Models proposed in this thesis take the form of a regression where the dependent variable is a left-censored analyte of interest and the regressors are indicators of water quality such as temperature, pH and dissolved oxygen that could be more feasibly obtained using sensors and telemetry in the future.

An application with such complexity requires an inter-disciplinary approach and this thesis presents an exploratory data analysis, machine learning methods and mechanistic transport models based on physical laws. Following these results, we

propose models that avoid replacing censored data with half the detection limit; leverage the high correlation between analytes; apply mixture models to deal with non-linearity and a varying intercept model that makes use of the spatial aspect of the wells from which the data are sampled.

Acknowledgements

I would like to thank several people without whom this thesis would not have been possible. To my supervisors, Daniel Henderson and Colin Gillespie, thank you both for all your direction, key insights and sage advice. My personal growth and confidence in my ability would not have been the same without your mentorship over these past few years.

Thank you to all those working with industry partners such as Wayne Jones (Shell Research Ltd.) and Luc Rock (Shell Global Solutions International BV., now at Shell Global Solutions Canada Inc.) who supplied the motivating problem and data and made time to discuss the research contained within this thesis.

On a personal note, I appreciate the unwavering confidence in my abilities from my parents and sister Megan and the everlasting support from my partner Jonathan who has helped me through all the stress and difficulties endured and I only hope I can return the favour.

Declaration

All work within this thesis represents original research that is my own work under the supervision and guidance of Dr Daniel Henderson (Newcastle University) and Dr Colin Gillespie (Newcastle University).

The views and opinions expressed in this body of work are those of the authors and do not necessarily reflect the views, policy or positions of any associated companies including, but not limited to, Shell plc.

Table of contents

Chapter 1

Introduction

1.1 Motivation

Groundwater monitoring of hydrocarbons is a key element to risk assessment and remediation for companies who deal with transportation, storage and distribution of refined petroleum products and crude oil, making these analytes of particular interest (CL:AIRE, 2017).

Groundwater monitoring sites are concerned with a variety of possible pollutants that affect groundwater quality and must establish sustainable practises during operation with minimal social and environmental disruption. This could take the form of extensive data collection, monitoring and understanding to aid early detection and proactive action. Alternatively, after an incident has taken place and groundwater has been impacted, there is a need to delineate and evaluate the evolution of groundwater quality over time.

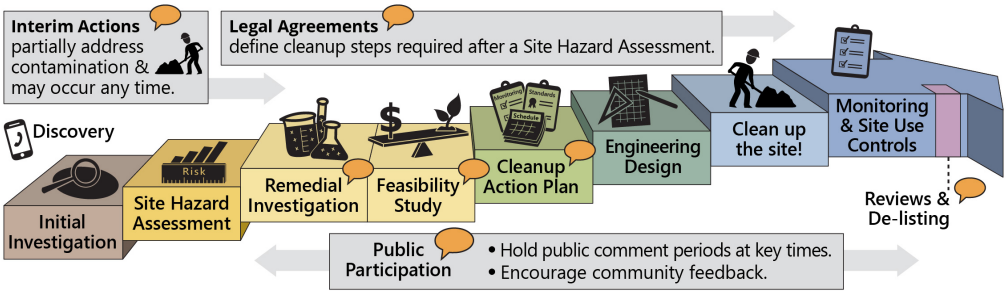

Figure 1.1: Model Toxics Control Act (Department of Ecology, State of Washington).

Once a location is retired and no longer used for processing raw materials containing hydrocarbons, it must be remediated to allow the land to be used for other purposes. Figure 1.1 shows a typical strategy for this process with emphasis on public participation. Due to the increase of extreme weather events (Pörtner *et al.*, 2022) it is increasingly important that any remediation measures are also resilient to any unexpected changes (Interstate Technology & Regulatory Council, 2021).

Most commonly, hydrocarbon groundwater monitoring methodology involves data collection of hydrocarbons, which we will refer to as 'analytes' as described in Appendix A, where samples are collected from multiple wells within a site that are to be analysed later. Photoionisation detectors are one such example in widespread use (Adamson *et al.*, 2012) that are capable of measuring the volatile organic compounds we are interested in. However, technologies such as field portable gas chromatographs or the UviLux fluorometer, by Chelsea technologies, allow for *in situ* measurements with the trade-off of a higher cost investment for the equipment and requisite training (Adamson *et al.*, 2012).

During this groundwater monitoring it is common to take measurements of water quality parameters such as, but not limited to, pH and temperature. Each of these

variables can be classified as physical, chemical or biological as described in Summers (2020) but we make no such distinction. While these data are collected from the same sample used for the analyte variables, or in the field, the measurements may at times occur several days apart leading to some inter-sample variation. The focus of this work will be on the water quality (WQ) parameters: electrical conductivity (EC), oxidation reduction potential (ORP), dissolved oxygen (DO), pH and temperature. These WQ parameters are routinely measured during groundwater sampling campaigns, and may allow for a more cost effective solution where analyte concentrations are inferred from remote measurements that are supplied on-line from a sensor or internet of things (IOT) infrastructure. The term 'predictors' will be used in this thesis to refer to the WQ parameters: conductivity (EC), oxidation reduction potential (ORP), dissolved oxygen (DO), pH and temperature. More details on each variable used can be found in Appendix A.

By more efficiently monitoring analytes, there is potential to develop an early detection system that would alert groundwater site employees to a potential event of interest allowing for more prompt and effective intervention.

Alternatively, groundwater sampling costs incurred can be reduced via optimising well sampling schedules, identifying optimal locations for additional wells or removing redundant wells that offer little extra information about a site (McLean *et al.*, 2019).

Hydrocarbon pollution of groundwater is one of many concerns within a greater context where other pollutants are also under study. Toxic elements, arsenic and barium, are considered in Sahoo & Hazra (2021) and radionuclides such as uranium are investigated in Schmidt *et al.* (2018). Most, if not all, models and methodology in this thesis can be applied to a general groundwater contamination context, when the distinct properties of each contaminant are taken into account.

1.2 Literature Review

Understanding the mechanisms and states of groundwater is a highly complex issue with many advocating for a multidisciplinary approach. Since two identical looking sites can have clearly heterogeneous geologies beneath the ground, such as different rock type, average temperature, aquifer shape and many more complexities, no single model will work in all applications. Any statistical model we propose is to fit within a larger toolbox of models that may arise from machine learning, chemistry, geology, hydrology, fluid dynamics or any discipline of interest. Therefore, it is useful to understand what modelling techniques are already available for this set of problems. Note that some of this literature review is deferred to Chapter 3.

1.2.1 GWSDAT

The groundwater spatial data analysis tool (GWSDAT), created during a collaboration between the University of Glasgow and Shell Global Solutions International BV, is a software package and accompanied Shiny application with Excel add-in (Jones *et al.*, 2014). It is able to analyse hydrocarbon groundwater monitoring data and present visualisations such as concentration maps and time series, enact trend analysis and determine contamination plume characteristics through the use of spatiotemporal smoothing (Jones *et al.*, 2015).

The scope of the software is to model analyte data independently of other data collected from that water sample and focus solely on the spatiotemporal nature of the data. Prior to GWSDAT, spatiotemporal modelling was often achieved by fitting a spatial model at successive time slices, however, the spatiotemporal model employed by the GWSDAT appears to "borrow strength" across time when compared to a simpler spatial model (McLean *et al.*, 2019).

The methodology of GWSDAT relies on penalised splines (P-splines), as proposed by Eilers & Marx (1996). P-splines are described as a flexible data smoother that act by minimising the penalised least squares, subject to some choice of penalty. For a more detailed overview see Mclean (2018) and Evers *et al.* (2015) and for discussion about splines in general see Eubank (1999).

An advantage to the spatiotemporal model is that less data is lost per well removed (McLean *et al.*, 2019), allowing reduced operational risks and costs for little to no information loss. GWSDAT has been implemented in multiple case studies including Malander (2016) where a consultant applied GWSDAT to groundwater monitoring data involving detections of light non-aqueous phase liquid (LNAPL), to be described in Section 1.5, groundwater elevation and hydrocarbon concentrations from analytes such as gasoline additive MTBE (methyl tert-butyl ether) and those denoted in BTEX (benzene, toluene, ethylbenzene, xylenes). From these analyses, recommendations were made on an individual well-by-well basis that contributed to a decrease in the number of wells to be sampled and sampling frequency for some wells.

1.2.2 Telemetry

The original scope of our research included the potential of real-time prediction of analyte concentrations using the aforementioned predictors. Trivially, such a procedure would require models that can be fit efficiently to ensure the predictions can be composed as quickly, if not quicker, than the rate of sampling of conductivity (EC), oxidation reduction potential (ORP), dissolved oxygen (DO), pH and temperature. Such an approach is defined as telemetry where data collection is automated by sensors.

The advantages of telemetry for groundwater monitoring applications are plentiful as data collection is at a high cost due to

- well construction and maintenance;
- labour costs associated with water sample extraction;
- analysis of the water sample to measure key water quality indicators.

Hence, it should not be surprising that a similar approach has been presented such as Schmidt *et al.* (2018) where *in-situ* "water quality data" is used to "estimate contaminant concentrations". The contaminant in Schmidt *et al.* (2018) refers to tritium and uranium concentrations and we shall investigate if such techniques can be reproduced with hydrocarbon concentrations.

Alternatively, instead of separating these analytes and predictors into distinct groups as we do in regression models described in Chapter 5, one could jointly model all types of water quality variables, including our analytes and predictors, as multivariate species (Gong *et al.*, 2021). Gong *et al.* (2021) applies this approach to air pollution over much larger distances than seen in data pertinent to our application but reveals how we can fit a spatially motivated multivariate regression to "borrow" information from variables of a similar nature.

The idea of leveraging sensors is not an original idea proposed in this research and has been considered for groundwater monitoring in detail for Morocco, over a decade prior (Taffahi *et al.*, 2013). Taffahi *et al.* (2013) highlights that this is not only a statistical task but also poses engineering problems with how the sensors are to be built to a reliable standard and a business problem to assess the business feasibility. Of great interest to us are the sensors for pH and electrical conductivity with more details given on extracting pressure measurements which could be a worthwhile predictor for our data to include in the future, if feasible.

1.3 Thesis Structure

Throughout this thesis we introduce our approach of modelling the log concentrations of analytes methyl tert-butyl ether (MTBE) and benzene using the predictors electrical conductivity (EC), oxidation reduction potential (ORP), dissolved oxygen (DO), pH and temperature for a specific site using a variety of tools and models.

Chapter 2 introduces our case study dataset that was shared with us under the terms of pseudonymisation and will be described throughout as case study A. Before any modelling is enacted we perform an exploratory data analysis to better understand patterns in these data, some of which are specific to this site and others can hopefully be generalised to other groundwater monitoring sites. A mechanistic dataset is presented in Chapter 3 as a representation of our expectations under a reactive transport model (RTM) that is based on well established physical laws, for example conservation of energy and conservation of solute mass. In conjunction with our exploratory analysis, we apply a black box machine learning method in Chapter 4 to potentially elicit any quantitative relationships that are easily missed in approaches used in the exploratory data analysis. We explain how these random forest models operate but more detailed explanations can be found elsewhere such as Breiman (2001).

The remaining chapters in this thesis begin with regression models where imputation methods and Tobit models (Tobin, 1958) are compared and then applied to our case study, as appropriate. Each subsequent chapter aims to build upon these baseline models and improve parameter estimation or predictive power as described in Section 1.6.

Chapter 6 aims to leverage high prior correlation between analytes with:

1. multivariate normally distributed analytes;
2. matrix-variate normally distributed analytes.

However, these generalisations are also presented as a trade-off since a multivariate censored regression (Tobit) model would require data augmentation to be described in Section 1.5.4 or a multivariate probability density function (PDF) that we have not considered here.

Chapter 7 considers the apparent non-linearity and "phase"-like behaviour of these data as suggested by RTM simulations. Leveraging mixture modelling, specifically mixture of regressions with concomitant variables, allows us to fit a clustering step and regression step into a single mixture of experts (MoE) model framework (Gormley & Frühwirth-Schnatter, 2019) and estimate piecewise linear relationships in a Bayesian context.

Chapter 8 applies a more pragmatic approach for our case study with a renewed focus on the variation within a single hydrocarbon groundwater monitoring site. This is done through a special case of a linear mixed effects model where the intercept term is defined to be different per well. With these varying intercept models, prediction for observed wells can be improved and inter-well variation is directly modelled leading to more accurate parameter estimates (Revie *et al.*, 2017).

By repeating the same prediction scenarios, to be defined in Section 1.6, for each of these comparable models we can employ model comparison techniques to find the most suitable model within each chapter. Moreover, we can compare the 'best' models from each chapter and quantify how much each generalisation improves over the baseline censored regression model introduced in Chapter 5. The results of which are included in the concluding chapter, Chapter 9.

1.4 Data Structure

Data arising from hydrocarbon groundwater monitoring wells is very typical of environmental statistics with key idiosyncrasies including the spatiotemporal component and left-censoring, to be defined in Section 1.5. Note that all data has associated metadata including the observation time and spatial location, the difference for these data is that the spatiotemporal data has a tangible impact on the observed values (Oliver *et al.*, 2015). To better visualise these data, consider Figure 1.2 where we plot left-censored analyte benzene from the dataset to be fully explored in Chapter 2. Immediately one can observe the temporal nature of these data, the more frequent data collection at "Well-01" and the high degrees of censoring at remote locations such as "Well-46". For groundwater data we wish to model, this is extended as there may be as many as 50 wells, each with associated spatial coordinates not shown. Similarly, several analytes such as MTBE, benzene, toluene and several predictors such as conductivity (EC) and pH are observed per water sample that is extracted from the monitoring site.

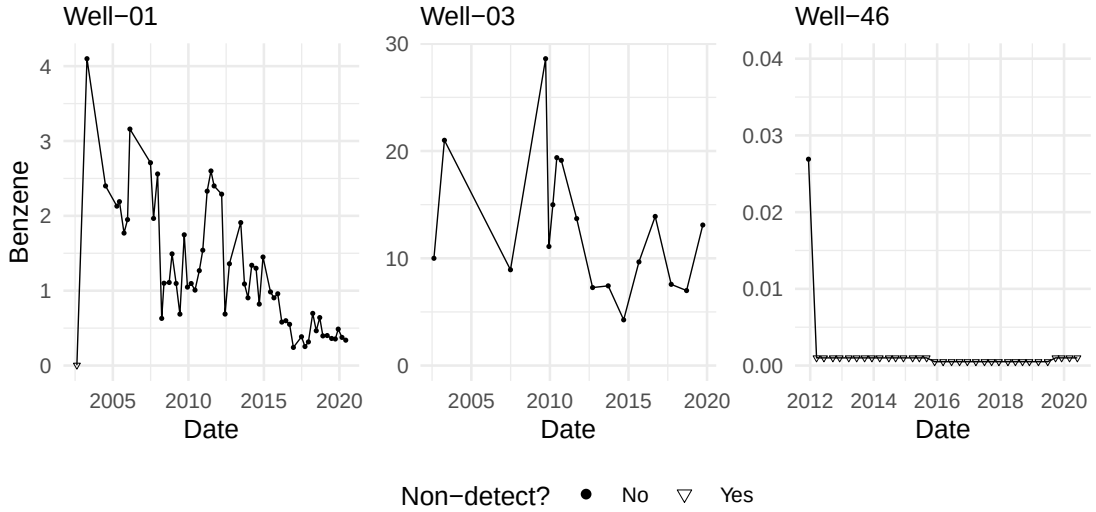

Figure 1.2: Time series of benzene concentrations for three specified wells.

Suppose we have n_s observations that are uniquely identified by the sample identifiers, a combination of the well name and sample date, where the frequency may vary by well from quarterly to yearly. Within each observation, we observe at most n_x predictors and n_y analytes. It is common for sample dates between these sets to differ by a few days. As a simplification, we combine these measurements into a single sample if they are from the same well and the difference in sample dates is less than 1 week for the datasets discussed in Chapter 2.

Our complete analyte data takes the form of a matrix

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1n_y} \\ \vdots & \ddots & \vdots \\ y_{n_s1} & \cdots & y_{n_sn_y} \end{pmatrix}, \tag{1.1}$$

where each element y_{ij} represents the log concentration for the i^{th} sample and j^{th} analyte variable that is typically left-censored. We apply a log transformation due to the high level of right skew in these data. Figure 1.3 shows evidence of this by visualising all analyte data from the site to be explored in Chapter 2. If one was to also include the censored observations, the data would only get more skewed but these data have been omitted as they typically occur as the same value many times within a site, by design. Further investigation into detection limits and censoring levels can be found in Section 2.3.

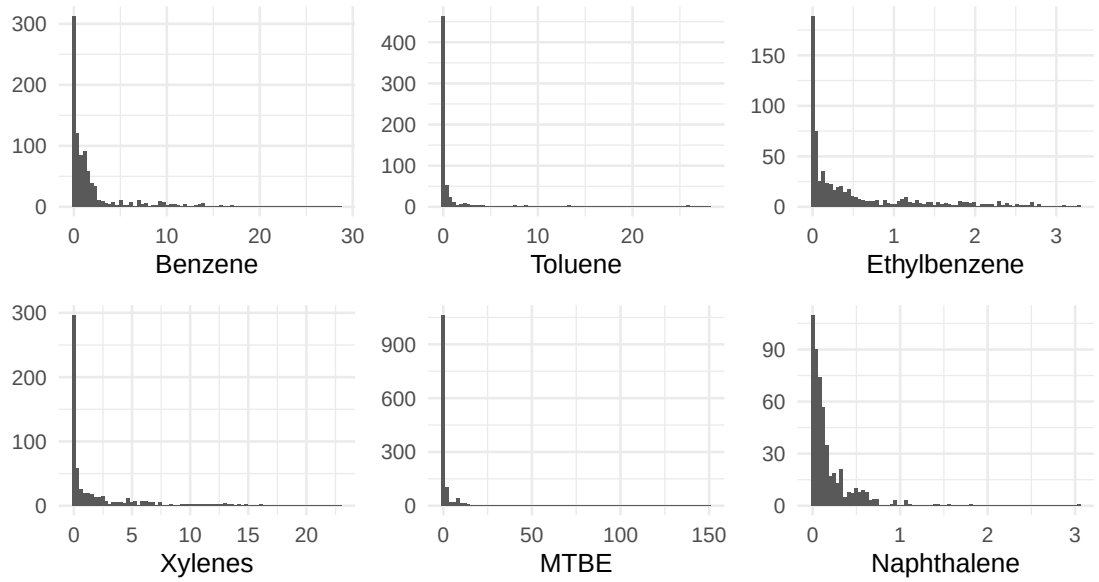

Figure 1.3: Histogram of all uncensored observations within hydrocarbon groundwater site, y-axis represents frequency.

As with the analyte data, the i^{th} sample measurement of the j^{th} predictor is denoted x_{ij} and we compactly collate these data in a single matrix

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1n_x} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n_s1} & \dots & x_{n_sn_x} \end{pmatrix},$$

with an "intercept" column of ones prepended, unless stated otherwise. It is often necessary to standardise predictors to assist with interpretation within a regression context and to ensure varying scales do not impact inferences (Gelman *et al.*, 1995).

We express metadata of each sample as a coordinate of length d, say $\mathbf{s}_i = (s_{i1}, \dots, s_{id})^T$. For groundwater monitoring applications considered in this thesis, well locations and sample dates yield spatiotemporal coordinates $(s_{ix}, s_{iy}, s_{it})^T$, however, this highlights a key data gap in these data where depth information is missing.

1.5 Censoring

Left-censoring is a key idiosyncrasy of environmental data that occurs when the concentration to be measured is less than some detection limit (DL), sometimes referred to as a reporting level (RL). This can be for many reasons, such as equipment tolerance or dilution of samples causing harder to detect concentrations. Since the observation is known to be at a safe level, substantially below any regulatory limit, no follow-up analysis is done and the information given is that the data point must lie in the interval $(0, \mathrm{DL}]$ for some detection limit DL.

Further complicating this issue is the fact that this limit of detection may change on a per-sample basis as samples may be sent to different facilities for analysis and facilities may replace their equipment over time with a higher or lower detection limit. Statistical methods have been developed for single DL datasets (Zoffoli *et al.*, 2013) and multiple DL datasets (Helsel, 2005). Multiple DL data can be recoded as single DL by combining all censored data to be censored at the maximum detection limit (MDL) but this will lead to a loss of information. Since data considered in Chapter 2 contains instances where DLs may be different even at the same well, we make no such assumption that all DLs are at the same level.

In a hydrocarbon context, our data should be thought of as interval-censored since a high concentration will be obscured by the presence of non-aqueous phase liquid (NAPL). For instance, the pure-phase solubility of benzene (100% single component) is 1790 mg/L, however a 'typical European fuel' containing other components will have an effective solubility of around 18 mg/L, 1% of the original solubility (Tomlinson *et al.*, 2014). By definition, these values represent the maximum groundwater concentration to be expected adjacent to a release of light non-aqueous phase liquid (LNAPL) so any concentrations exceeding their effective solubility are right-censored

observations. However, since effective solubility is situational and highly dependent on the composition of the contaminant it is difficult to model this in practice. Hydrocarbon groundwater monitoring investigations record any signs of LNAPL at the time of sample extraction. For our data, this is rarely a concern.

Suppose the log of the level of detection when measuring the j^{th} analyte at the i^{th} sample, determined by the measuring equipment, is denoted by $c_{ij} > 0$. There exists some theoretically uncensored (log) concentration, y^*_{ij}, that we are interested in, but we only observe the pairing (y_{ij}, δ_{ij}) for each sample $i = 1, \dots, n_s$ and analyte $j = 1, \dots, n_y$ where

$$y_{ij} = \begin{cases} c_{ij}, & \text{if } y^*_{ij} < c_{ij} \\ y^*_{ij}, & \text{otherwise} \end{cases},$$

and censoring indicator $\delta_{ij} = \mathbb{I}(y^*_{ij} < c_{ij})$ where $\mathbb{I}$ is the indicator function taking the value of one when the supplied argument is true and zero otherwise.

This convention of coding the indicator to one when the value is a "non-detect" is antonymic to the convention used in survival analysis where observed values are coded with a one and censored values with a zero (Kleinbaum *et al.*, 2012).

1.5.1 Substitution Methods

Left-censored data are often seen as sub-quality data but the statistical field of survival analysis demonstrates rigorous and principled ways of dealing with censored data (Kleinbaum *et al.*, 2012). One method that is often presented as a worst-case scenario (George *et al.*, 2021) is the naïve removal of the data, also known as complete cases only. A key flaw in removal is that the data that has been censored is, by definition, missing not at random (MNAR); the probability of an observation being censored is directly dependent on the underlying concentration. Expectedly,

removing censored data leads to highly biased estimates (George *et al.*, 2021; Zoffoli *et al.*, 2013) and should not be suggested in any scenario.

A more common approach is that of substitution; each censored value is replaced by a function of the detection limit, $y_{ij} = f(c_{ij})$. This idea has been used in the literature in various forms:

- replacement by half DL, $f(c_{ij}) = c_{ij}/2$, (Jones *et al.*, 2015);
- replacement by factor of DL, say $f(c_{ij}) = c_{ij}/\sqrt{2}$, (Hornung & Reed, 1990);
- replacement by the maximum value, $f(c_{ij}) = c_{ij}$;
- replacement by the minimum value, $f(c_{ij}) = \log(\epsilon)$, for some chosen minimum feasible value $\epsilon > 0$;
- stochastic replacement, $f(c_{ij}) = u$ where $u \sim U(0, c_{ij})$, (Baize *et al.*, 2009).

Replacement by half of the detection limit, which we refer to as the "DL/2" method, is one of the most frequently used ways of dealing with censoring (Helsel, 2005; Singh & Nocerino, 2002). It has been claimed that the DL/2 method may be accurate in some instances where the proportion of censored data does not exceed 10% (Helsel, 2005), however, the literature overwhelmingly concludes that substitution methods lead to unsatisfactory inferences and results in the majority of cases. Early research on the topic claimed a "significant loss of information" for DL/2, replacement by DL and replacement by 0 (Helsel & Cohn, 1988). Evidence that all considered substitution methods result in biased and inaccurate estimates to varying degrees, except in very specific cases, can be found in Helsel (2005) and Zoffoli *et al.* (2013). The prevailing theory is that the DL/2 method in widespread use is biased (Singh & Nocerino, 2002; Helsel, 2011; George *et al.*, 2021).

The main issue with DL/2 is that the fraction of the detection limit is arbitrary with no justification, as evidenced by the use of $1/\sqrt{2}$ in some instances (Hornung

& Reed, 1990). Helsel (2011) enacts a simulation study where the fraction of the DL used varies between 0 and 1; it is found that the estimates of the mean, standard deviation, correlation coefficient, regression slopes and t-tests are all highly sensitive to this choice.

Additionally, when considering the DL/2 method it may be, and usually is, the case that better methods exist. For example, when dealing with left-censored environmental data, rank-sum tests are recommended over t-tests by the United States Geological Survey (Helsel *et al.*, 2020). Usage of DL/2 can be justified as a trade-off where efficiency is gained at the cost of analysis-induced bias which may be appropriate for some applications as discussed in George *et al.* (2021). For example, GWSDAT provides a smoothed surface of an entire groundwater site in a reasonable time frame using DL/2 (Molinari, 2014; Mclean, 2018).

1.5.2 Regression on Order Statistics

Regression on order statistics, as proposed by Helsel & Cohn (1988), is a parametric method that combines an assumption about the distribution of censored data (or entire dataset) and the uncensored data to estimate likely values. Following these calculations, estimates of censored values can then be imputed in a similar way to any substitution method.

The intuition behind this method is to construct "plotting positions", yet to be defined, that represent theoretical quantiles commonly used in a normal Q-Q plot. A linear model is then fit to the (log) uncensored observations where the only explanatory variable is plotting positions. Imputation is enacted by using this linear model where censored data predictions, based on plotting positions, can be used as imputation values. We can ensure these imputed values are strictly positive by log transforming before fitting and exponentiating any predictions from the model.

To make this process clear, we apply regression on order statistics (ROS) to the same example given in Helsel & Cohn (1988), that is,

$$< 1, < 1, < 1, < 1, < 1, < 1, 3, 7, 9, < 10, < 10, < 10, 12, 15, 20, 27, 33, 50$$

Suppose we have m detection limits, $0 < d_1 < \cdots < d_m$, and then for each detection limit calculate

1. number of uncensored observations between d_j and d_{j+1}, denoted A_j;
2. number of all observations below d_j, denoted B_j;
3. number of censored observations only known to be below detection limit d_j, denoted C_j.

It follows that the conditional probability of exceeding some threshold, d_j is given iteratively,

$$p_j = p_{j+1} + \frac{A_j}{A_j + B_j}(1 - p_{j+1}),$$

for $j = m, m-1, \ldots, 1$ where we set $p_{m+1} = 0$. In the worked example where $d_2 = 10$ and $d_1 = 1$, we count observations such that $A_1 = 3, A_2 = 6, B_1 = 6, B_2 = 12, C_1 = 6, C_2 = 3$ and then derive

$$p_2 = p_3 + \frac{A_2}{A_2 + B_2}(1 - p_3) = \frac{1}{3} \approx 0.333$$
$$p_1 = p_2 + \frac{A_1}{A_1 + B_1}(1 - p_2) = \frac{15}{27} \approx 0.556.$$

Plotting positions use the complement of the probability of exceedance, say $q_j = 1 - p_j$, as we desire a quantile function. By using Weibull plotting positions (Helsel & Cohn, 1988) we assign plotting positions to be equidistant points within their respective interval.

For uncensored observations,

$$PP(i)^{(uncensored)} = q_j + (q_{j+1} - q_j)\left(\frac{r}{A_j + 1}\right),$$

where r is the rank of the i^{th} uncensored observation above the j^{th} detection limit, for which there are A_j total. Whereas, for censored observations,

$$PP(i)^{(censored)} = q_j\left(\frac{r}{C_j + 1}\right),$$

where r is the rank of the i^{th} censored observation below the j^{th} detection limit, for which there are C_j total.

We plot the results of this worked example in Figure 1.4 where the key plotting position boundaries, q_1 and q_2 are marked. Plotting positions for uncensored observations are equidistant between the respective marked boundaries, whereas plotting positions for censored observations are equidistant between the respective upper marked boundary and zero. This has the unintentional effect that some observations may share a plotting position at 0.5 as shown. Arrows denoting a change from detection limit to imputed value shows how each new value may move above the original detection limit. While undesirable, imputing by a value above the detection limit is intentional to extend the (log) linearity of the uncensored data to the censored data and demonstrates how an intentional inaccuracy for specific observations may be used for the benefit of the overall shape of these data. A further disadvantage of this approach is the assumption of normality which the method makes use of extensively.

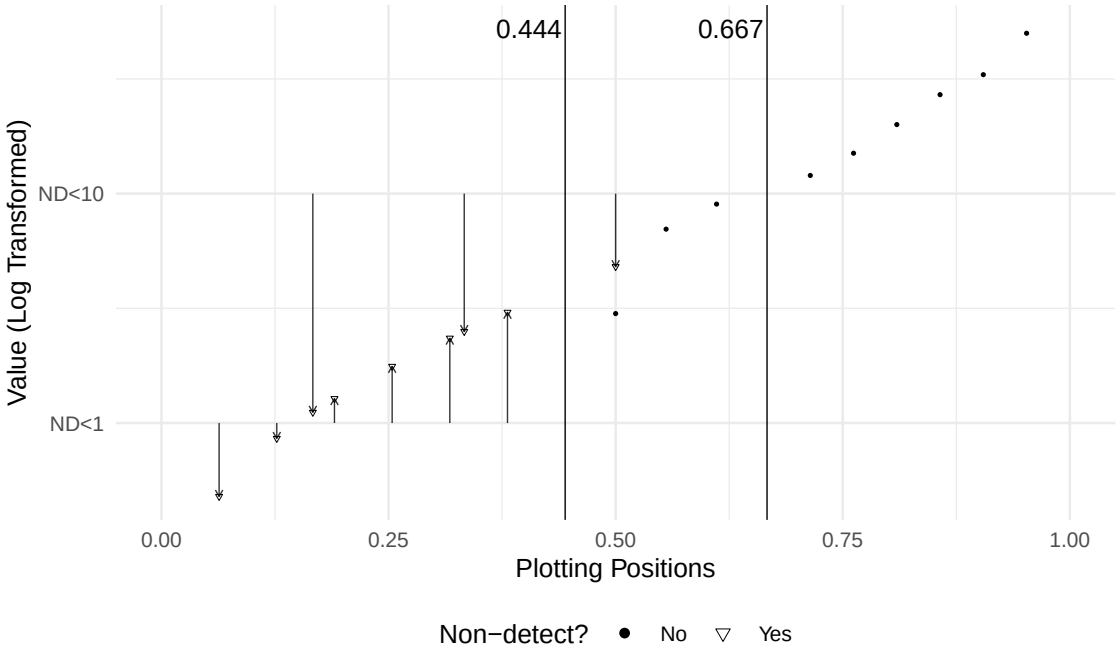

Figure 1.4: Quantile plot of worked example showing ROS imputation by arrow.

1.5.3 Maximum Likelihood Estimation

We can avoid substitution entirely if we follow a parametric approach where the distribution of the underlying data is assumed known. Constructing a likelihood based on censored data can be done using some cumulative distribution function (CDF), say $F(\cdot)$, for censored observations and using a probability density function (PDF), say $f(\cdot)$, for uncensored data. That is,

$$L(\Theta) = \prod_{i=1}^{n_s} f(y_i)^{1-\delta_i} F(y_i)^{\delta_i}, \tag{1.2}$$

where y_i are the recorded values whose nature depends on the censoring indicator δ_i and Θ are parameters of the distribution.

Rather than choosing $F(\cdot)$ and $f(\cdot)$ functions directly, they are derived from the assumed distribution. For our groundwater monitoring data to be introduced in Chapter 2, a good candidate distribution would be the normal distribution with

mean θ_μ and standard deviation θ_σ. By maximising (1.2) with respect to θ_μ and θ_σ we can

1. estimate the mean and standard deviation of these data using

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, L(\Theta);$$

2. fit a likelihood-based model to the data as in Section 5.3;

3. evaluate the log likelihood to perform likelihood-ratio tests (Helsel, 2011).

This approach is not without disadvantages; the MLE approach has been shown to perform poorly with small datasets of around 25-50 total observations (Helsel, 2011). Similarly, if there are too few uncensored observations, due to a high degree of censoring, the inferred parameters of the assumed distribution will be unreliable and this approach should not be used. We see this limitation explicitly in a similar approach known as the method of median semi-variance where parameters of a half-normal distribution are estimated using the upper half of the data (Zoffoli *et al.*, 2013). For trivially apparent reasons, this method is not to be used on data that exceeds 50% censoring.

1.5.4 Data Augmentation

In general, data augmentation (DA) is a set of techniques where artificial data is generated from existing data to improve data quality. For our purposes, DA is the imputation of incomplete data at each iteration of a Markov chain Monte Carlo (MCMC) algorithm through the use of a conditional distribution; see Appendix B for a primer on Markov chain Monte Carlo (MCMC) methods. Incomplete observations in our application are solely censored data but may also include missing observations (Lockwood *et al.*, 2004).

This approach has been leveraged in spatial statistics (Fridley & Dixon, 2007) and specifically groundwater contamination in Lockwood *et al.* (2004) where a truncated normal was used as the conditional distribution for each censored observation. One can consider this method as assigning each censored observation to be an unknown parameter with a restricted domain *a priori*, say $(0, \mathrm{DL})$ for some detection limit.

1.6 Model Prediction

There are numerous *scenarios* within a groundwater monitoring site in which we could apply a model to produce predictions to be communicated. In this thesis we consider the task of predicting several *holdout wells*, to be chosen for our case study in Section 2.2. Once these wells to be predicted are identified, we consider two hypothetical scenarios that differ by how much of the holdout well data is made available for model fitting and hyperparameter tuning.

1. **Leave-multiple-well-out (LMWO)**, a generalisation of leave-one-well-out (LOWO) (Evers *et al.*, 2015) where analyte concentrations from the holdout wells are to be predicted;
2. **holdout future**, analyte concentrations are observed for all wells up to a specified date, after this point only non-holdout wells observe concentrations.

In both cases, we assume our predictors are fully observed at all samples and are therefore available for use in prediction.

Motivating the LMWO scenario is the concept of telemetry where we could feasibly be asked to predict analyte concentrations for many candidate locations for a new well, with predictor data being supplied by *in situ* sensors. The holdout future scenario is pragmatic because there are many reasons why we may have historical data up to some terminating point for a well. Well closures can occur for many rea-

sons including deterioration beyond safe working conditions and to increase financial feasibility of the groundwater monitoring site.

In both scenarios, it is imperative that the described out-of-sample data is independent of any model building decisions to mitigate the risk of overfitting to these site-specific data. The consequence of this decision is that any hyperparameter to be determined, such as number of available predictors in Section 4.3.2 and number of components in Chapter 7, must further split the data with techniques including cross-validation. We adopt the nomenclature typical in machine learning to make the role of each observation clear:

1. **training data** describes observed data to be used in all model fitting steps, for example calculating maximum likelihood estimates (MLE) or executing an MCMC algorithm;
2. **validation data** denotes out-of-sample data used during hyperparameter tuning. Models are compared to decide on a "best" hyperparameter value;
3. **test data** forms the unseen data to be predicted that will provide an unprejudiced evaluation of a final model.

The diagram in Figure 1.5 presents a simplified case of an approach to be applied to the random forest model (Section 4.5.1). An initial split is made based on wells and then any subsequent hyperparameter tuning can only act on the reduced data. Here, candidate models are fit and used to make predictions on the corresponding validation group for a total of 5 times; by evaluating a yet to be defined metric on these predictions, a single *validation score* is aggregated per candidate model. A final model using the "best" hyperparameter is fit to the initial training data and makes predictions on the test data resulting in a test score, an indicator of model predictive performance when facing previously unseen data.

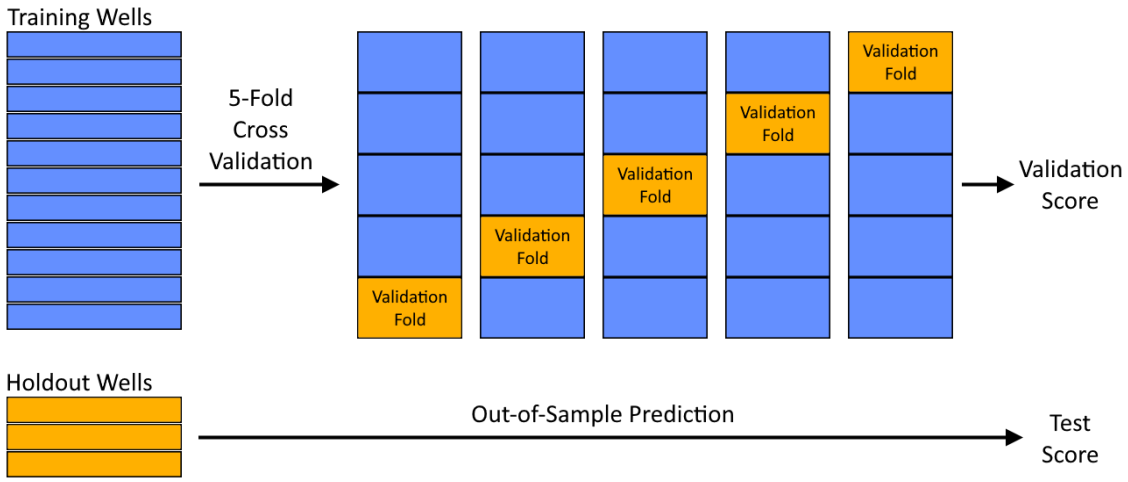

Figure 1.5: Diagram of training, validation and test data with 5-fold cross validation.

1.7 Model Comparison

Throughout this thesis, we assess predictions qualitatively using the holdout wells defined in Section 1.6. However, we also want methodology to quantify each model's predictive performance using a chosen metric relative to some other comparable model. Possible options include the Akaike information criterion (AIC) (Watanabe & Opper, 2010) and the related Bayesian information criterion (BIC) (Schwarz, 1978). In this thesis, we use the log pointwise predictive density (LPD) for goodness-of-fit tests and a combination of the widely applicable information criterion (WAIC) and the arguably more robust Pareto-smoothed importance sampling (PSIS) (Vehtari *et al.*, 2017).

For models described in this thesis, the primary goal is to accurately predict completely new (out-of-sample) data subject to constraints. Hence, WAIC is utilised since it is an estimate of pointwise out-of-sample predictive accuracy and will asymptotically converge to the leave-one-out (LOO) measure of predictive accuracy (Watanabe & Opper, 2010). Calculating LOO directly would require the

computationally infeasible task of fitting each model to be compared for as many observations as exist in the dataset while permuting the OOS data to be a single observation. Additionally, the pointwise nature of LPD, WAIC and PSIS allow for per-observation diagnostics that can identify influential observations and detect when other methods such as K-fold cross validation should be used. We opt to compare models using PSIS due to the robustness of these diagnostics whereas WAIC highlights issues in the predictive posterior using empirical values (Vehtari *et al.*, 2017).

Suppose we want a measure of model accuracy using n_s^* known analyte log concentrations, say $\mathbf{y}^* = (y_1^*, \dots, y_{n_s^*}^*)$. For our use-case, $\mathbf{y}^*$ either denotes our *in-sample* data, where the same data is used to fit the model and produce the accuracy metric (so $\mathbf{y}^* = \mathbf{y}$), or *out-of-sample* data where data used to asses the model's accuracy is different from data used to fit the model. For our cross-validation approach, we make use of the two scenarios, LMWO and hold-out future, as described in Section 1.6.

The aim of each metric is to estimate the expected log pointwise predictive density, a measure of predictive accuracy for n_s^* potentially unseen data points (Vehtari *et al.*, 2017). It is defined as

$$\text{ELPD} = \sum_{i=1}^{n_s^*} \int \log(\pi(y_i^*|\mathbf{y}))\pi_t(y_i^*)dy_i^*,$$

where $\pi(y_i^*|\mathbf{y})$ is the posterior predictive density but the true data generating process, $\pi_t(y_i^*)$, is the unknowable truth so this quantity cannot be calculated and will have to be estimated. We now consider several options.

1.7.1 LPD

One option is to consider the log pointwise predictive density (LPD),

$$\text{LPD} = \sum_{i=1}^{n_s^*} \log \pi(y_i^*|\mathbf{y}) = \sum_{i=1}^{n_s^*} \log \int \pi(y_i^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta},$$

under the assumption y_i^* independent to $\mathbf{y}|\theta$. LPD can be estimated using the Monte Carlo estimate for posterior simulations, $s = 1, \ldots, S$ of the model parameters, $\boldsymbol{\theta}$. That is,

$$\widehat{\text{ELPD}}_{\text{LPD}} = \sum_{i=1}^{n_s^*} \log \left\{ \frac{1}{S} \sum_{s=1}^{S} \pi(y_i^*|\boldsymbol{\theta}^{(s)}) \right\}. \tag{1.3}$$

When supplying the in-sample data to (1.3), we produce an overestimate of the ELPD for future data because it is evaluated on the data from which the model was fit (Vehtari *et al.*, 2017). That is, solely focusing on maximising this estimate will likely lead to an overfitted model that predicts observed data very well but may struggle with novel data.

1.7.2 WAIC

The widely applicable information criterion (WAIC) estimate subtracts a penalty term from the log pointwise predictive density estimate,

$$\widehat{\text{ELPD}}_{\text{WAIC}} = \widehat{\text{LPD}} - \hat{p}_{\text{WAIC}},$$

where the penalty is a Monte Carlo estimate of the effective number of parameters, given by the sum of the sample variances (over posterior simulations) of each data point,

$$\hat{p}_{\text{WAIC}} = \sum_{i=1}^{n_s^*} \text{Var}\left\{\log \pi(y_i^*|\boldsymbol{\theta}^{(s)})\right\}.$$

The WAIC estimate will asymptotically equal the leave-one-out (LOO) cross validation estimate of out-of-sample prediction (Watanabe & Opper, 2010), however for finite n_s^* the estimate may be unreliable and empirical evidence shows that the estimate is unreliable when any of the summands that make up $\widehat{p}_{\text{WAIC}}$ exceed 0.4 (Vehtari *et al.*, 2017).

1.7.3 PSIS

Pareto-smoothed importance sampling (PSIS) is more robust than WAIC as it provides an estimate of ELPD as well as formalised diagnostics to check the validity of the estimate on a per-observation basis.

Suppose the model is trained on all available data, say $\mathbf{y}$, except for the i^{th} observation, we denote this subset by $\mathbf{y}_{-i}$. Raw importance sampling would estimate the LOO predictive distribution by

$$\pi(y_i^*|y_{-i}) \approx \frac{\sum_{s=1}^{S} r_i^{(s)} \pi(y_i^*|\boldsymbol{\theta}^{(s)})}{\sum_{s=1}^{S} r_i^{(s)}},$$

with importance ratios

$$r_i^{(s)} \propto \frac{\pi(\boldsymbol{\theta}^{(s)}|\mathbf{y}_{-i})}{\pi(\boldsymbol{\theta}^{(s)}|\mathbf{y})}.$$

We avoid raw importance sampling as these ratios may have high or even infinite variance and so use PSIS where this estimate is altered by smoothing the top 20% of importance ratios with a generalised Pareto distribution as described in Vehtari *et al.* (2015). Given the importance ratios are smoothed into new importance ratios, say $w_i^{(s)}$, the PSIS estimate of the ELPD is

$$\widehat{\text{ELPD}}_{\text{PSIS}} = \sum_{i=1}^{n_s^*} \log \left\{ \frac{\sum_{s=1}^{S} w_i^{(s)} \pi(y_i^*|\boldsymbol{\theta}^{(s)})}{\sum_{s=1}^{S} w_i^{(s)}} \right\}.$$

The upshot of this procedure is that the estimated shape parameter of the generalised Pareto distribution, $\hat{k}$, is representative of the shape of the upper 20% of importance ratios where exceeding certain values of $\hat{k}$ imply the importance ratios have infinite mean or infinite variance due to the properties of the Pareto distribution. That is,

- $\hat{k} < 0.5$ implies finite variance of the importance ratios so a reliable estimate;
- $0.5 \leq \hat{k} \leq 1$ implies the ratios may have infinite variance but a mean exists;
- $\hat{k} > 1$, ratios have no mean and infinite variance (Vehtari *et al.*, 2017).

One can also use the per-observation shape parameter estimates as a measure of leverage and produce diagnostic plots such as Figure 5.9.

1.7.4 Application

Since each of the aforementioned metrics estimate a (pointwise) log density, it is expected that each pointwise contribution and the sum will be negative; in special cases where a pointwise density evaluates to a value higher than 1, the log density contribution will be positive meaning all ELPD estimates may take any real value. Alternatively, some may prefer to multiply each ELPD estimate by -2 to convert the metrics to a deviance scale where smaller values imply a better model (Vehtari *et al.*, 2017). As we present the metrics on the original scale, the model that produces the greatest value, typically the least negative value, is asserted to be the 'best' predictive model based on the model, training data and test data provided.

For comparable models, we show the LPD, WAIC and PSIS relative to the 'best' value for that metric since the absolute values have little interpretability and we only describe model performance as 'good' or 'bad' when compared to the performance of another model. In other cases where models are incomparable, such as an identical model applied to different analytes, each metric is given on the original scale.

All metrics are calculated with the `loo` R package (Vehtari *et al.*, 2023) and presented alongside a standard error (SE). There is nuance in the interpretation of these measures of variability. When the metric is shown on the original scale, the errors shown are given by

$$\mathrm{SE}\left(\widehat{\mathrm{ELPD}}\right) = \sqrt{n_s \ \mathrm{Var}\left(\widehat{\mathrm{ELPD}}_i\right)},$$

where $\widehat{\mathrm{ELPD}}$ can be any of our aforementioned metrics with associated pointwise elements $\widehat{\mathrm{ELPD}}_i$ for $i = 1, \ldots, n_s$. Intuitively, this quantity represents the standard deviation of n_s independent components. The shortcomings of this approach are two-fold (Vehtari *et al.*, 2017),

1. components are not strictly independent as they are all computed from the same posterior samples $\boldsymbol{\theta}^{(s)}$;
2. the terms $\widehat{\mathrm{ELPD}}_i$ may follow a highly-skewed distribution which implies using variance as a measure of uncertainty may not be advisable.

During model comparison, it is favourable to present the standard error of the difference in models, as opposed to the difference of the standard errors. Hence, when model comparisons are presented relative to the 'best' model, the standard error is calculated using a paired estimate,

$$\mathrm{SE}\left(\widehat{\mathrm{ELPD}} - \widehat{\mathrm{ELPD}}^{(best)}\right) = \sqrt{n_s \ \mathrm{Var}\left(\widehat{\mathrm{ELPD}}_i - \widehat{\mathrm{ELPD}}_i^{(best)}\right)}.$$

We make use of each of these metrics and standard errors to compare models within a chapter and then again to compare models from different chapters in Chapter 9.

Chapter 2

Case Study

2.1 Data Availability

For this project, data from two hydrocarbon groundwater monitoring sites were provided by Shell Global Solutions International BV. As these are commercially sensitive data, we anonymise both sources as Site A and Site B; all modelling and data is assumed to be from site A unless stated otherwise. Following data cleaning for each site, a common problem is not enough overlap between observations of analyte concentrations and predictors defined in Section 1.1. This leads to a small number of of observations where a relationship can be inferred. In the sites we have seen, analyte concentrations can often be found for the entire lifespan of the groundwater monitoring site but corresponding data for our predictors have not been as prevalent with fewer historical observations as shown in Figure 2.1, and in some cases the data was not digitised. As such, we focus our modelling efforts on a single case study referred to as site A that has been pseudonymised; these data serve as representations of a site dealing with constituents of particular interest where our predictors were also recorded for most samples.

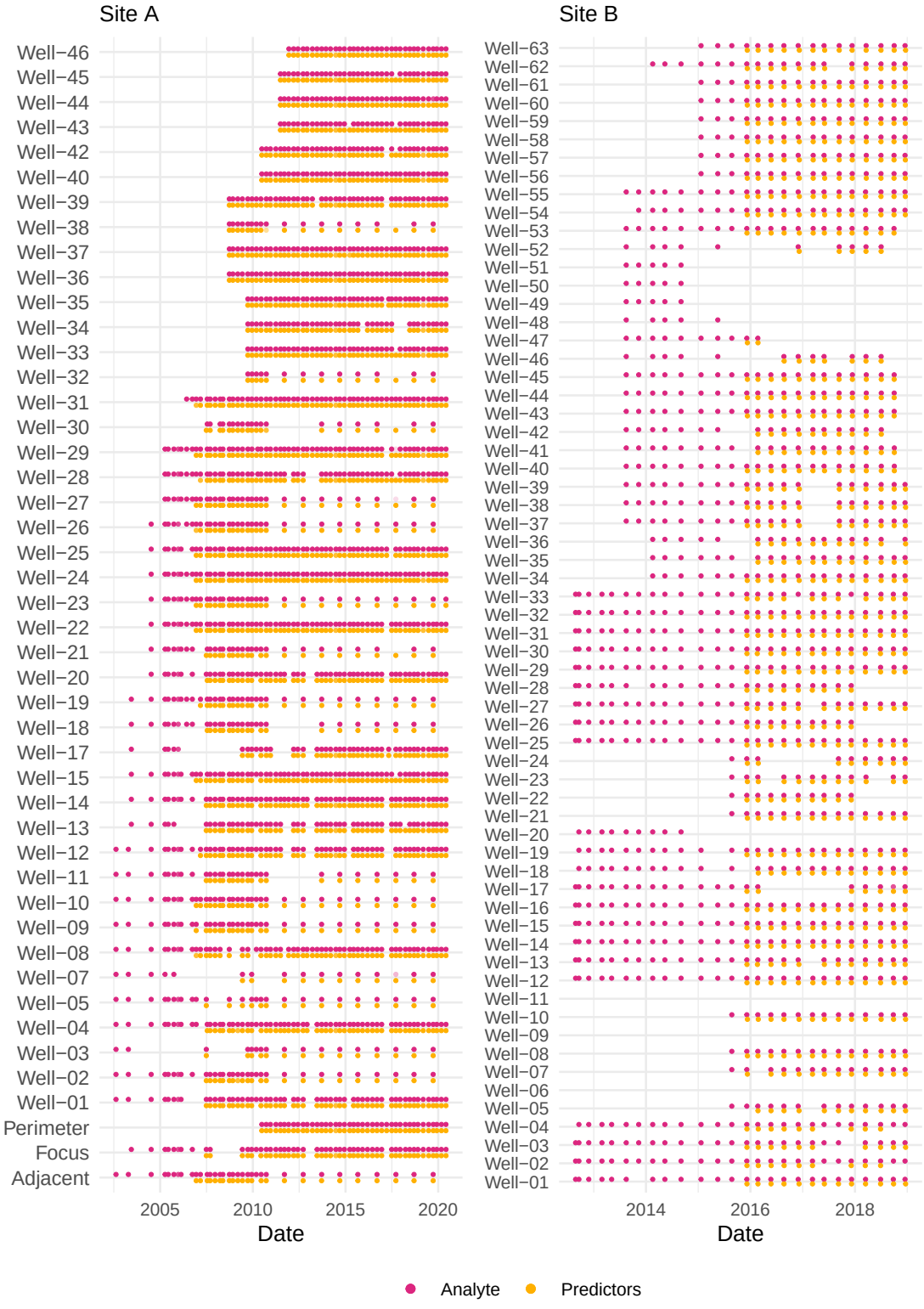

Figure 2.1: Sample alignment between analyte concentrations and predictors.

Figure 2.1 shows the variation in sample dates described in Section 1.4 where measurements of analyte and predictor data may be taken days apart. At site A, 1259 observations had both variable groups measured on the same day, 398 were 1 to 4 days apart and the remaining 252 observations found no match and are removed at the modelling stage. Site B is shown as a contrast where 43% of analyte measurements have no predictor measurement counterpart, there exists more re-alignment of dates and some wells such as "Well-06" have no data despite them being listed in the coordinates data. Leaving a sample, even for a few days, can drastically impact the composition of the water sample and introduces unnecessary variation; we ignore these issues by *fuzzy* joining of the data where analyte observations are matched to predictor observations when the sample dates are sufficiently 'close' and not necessarily equal. Note that directly modelling this added measurement error could be an avenue for future work.

Unique to site A, we were also supplied with multiple analyte measurements for the same sample, potentially arising from different laboratories, as part of a quality assurance and quality control procedure. We reduce these duplicates into a single observation and concede losing information about the measurement error, but this is not straightforward when dealing with censored data. As such, we propose the following pragmatic approach to be applied to each analyte variable independently in the data cleaning stage to ensure sample uniqueness. That is, for each set of observations coming from the same sample,

- replace groups of uncensored observations with the mean of the measurements as an uncensored observation;
- replace groups of censored observations with the lowest detection limit as a censored observation;

- discard any censored observations exceeding, or equal to, an uncensored observation, for example ND < 0.2 and 0.1 become 0.1;
- otherwise, the point is *conflicting* and we take the maximum uncensored measurement as a censored observation, for example ND < 0.3 and 0.4 become ND < 0.4.

While the latter option is less than ideal due to conflicting information, it is expected to happen when the true value is close to the detection limit relative to the measurement error. Thankfully, this happens infrequently in our data and of the $11,413$ analyte measurements made in Site A, $10,192$ (89.3%) were uniquely assigned to a single sample identifier, 671 (5.9%) were duplicated but uncensored, 513 (4.5%) were duplicated but wholly censored. Only 37 measurements were duplicated with different censoring indicators, 20 (0.18%) were in the congruent third scenario shown above and the remaining 17 (0.15%) were conflicting measurements.

2.2 Well Analysis

In some spatial and spatiotemporal studies we see well locations drawn uniformly at random on a unit square (Molinari, 2014) or by using a pattern such as a Latin square or regular grid (Fridley & Dixon, 2007; Sahoo & Hazra, 2021). Our data reveals a more pragmatic approach where well locations tend to cluster around high-risk locations such as facility buildings or storage facilities with fewer perimeter wells sparsely surrounding the general area as shown in Figure 2.2. Further investigation on the optimal well placement and by extension the sampling frequencies of these wells is beyond the scope of this project but is actively considered elsewhere (Sreekanth *et al.*, 2017; Mclean, 2018).

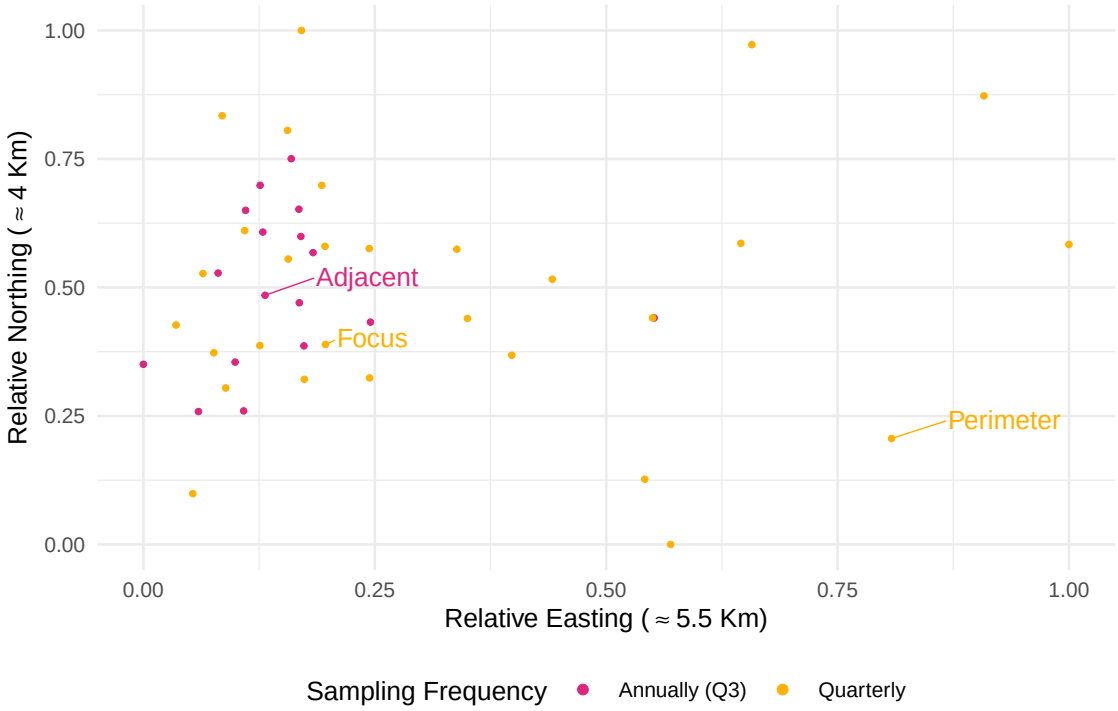

Figure 2.2: Well locations with site-relative coordinates.

As described in Section 1.6, we want to select several wells to be our *holdout wells* to form the basis of our prediction. Our choice of wells highlights key edge cases to reflect the main aims of this research; the best model is one that is able to predict extreme cases without sacrificing predictions elsewhere. The choice was also based on well locations and prior expectations communicated by the data providers, that is,

- **adjacent** is a well that is on-site and could be susceptible to spikes in hydrocarbon concentrations by proximity;
- **focus** is the nearest neighbour to the well with the most activity in the historical data;
- **perimeter** is a perimeter well that is expected to report low concentrations unless a contamination plume moves a 'large' distance.

2.2.1 Inter-Well Variation

Inter-well variation is a key component of the total variation of these data and describes the difference between two observations of the same quantity, taken at the same time from two separate wells that are some defined distance apart. To investigate this, we plot overlapping time-series for a specific quantity where each colour represents a well with distinct identifier and model inter-well variation explicitly in Chapter 8. If most of the series coalesce in these visualisations, as is the case for temperature (Figure 2.4), it suggests that there is minimal variation between wells for that variable. The converse also holds true, where seemingly independent time series would imply little to no spatial correlation. These deductions are consistent with the more principled approach, described in Section 2.5, involving empirical variograms.

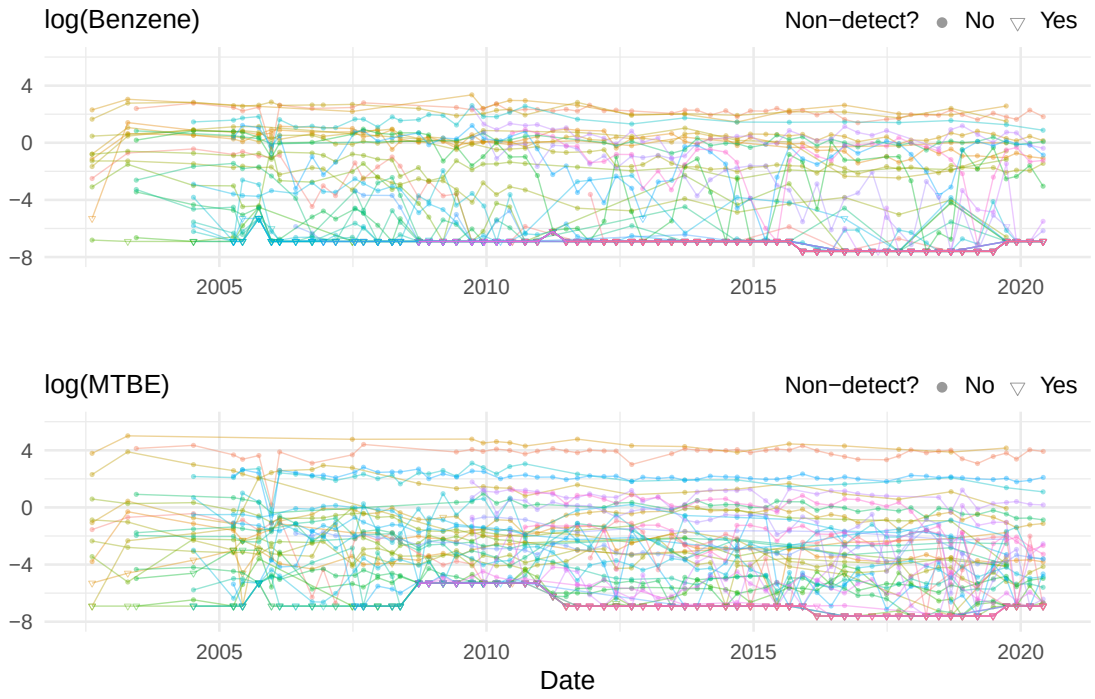

Figure 2.3: Multiple time series visualisation for analytes. Each colour represents a different well.

Figure 2.3 suggests that most wells with uncensored data could be spatially heterogeneous with no clear global trend or seasonality, separate time series plots (not shown) reveals that uncensored observations are heterogeneous in average concentration and in behaviour as some wells present a decreasing concentration over time while others appear to stabilise. On the other hand, censored data will often overlap perfectly due to the assumed site-wide detection limits. These visualisations also illuminate several key attributes of these data. Censoring levels differ over time more frequently than they differ by well and it is indeed the case that these detection limits can increase or decrease as time increases. High correlation in analytes is supported by the fact that the wells with the highest average log benzene concentrations are also shown to have the highest average log MTBE concentrations, although this is not always the case. The overlapping data that follows the detection limits implies that there are several wells (specifically, perimeter wells) that report only censored values or very few uncensored data.

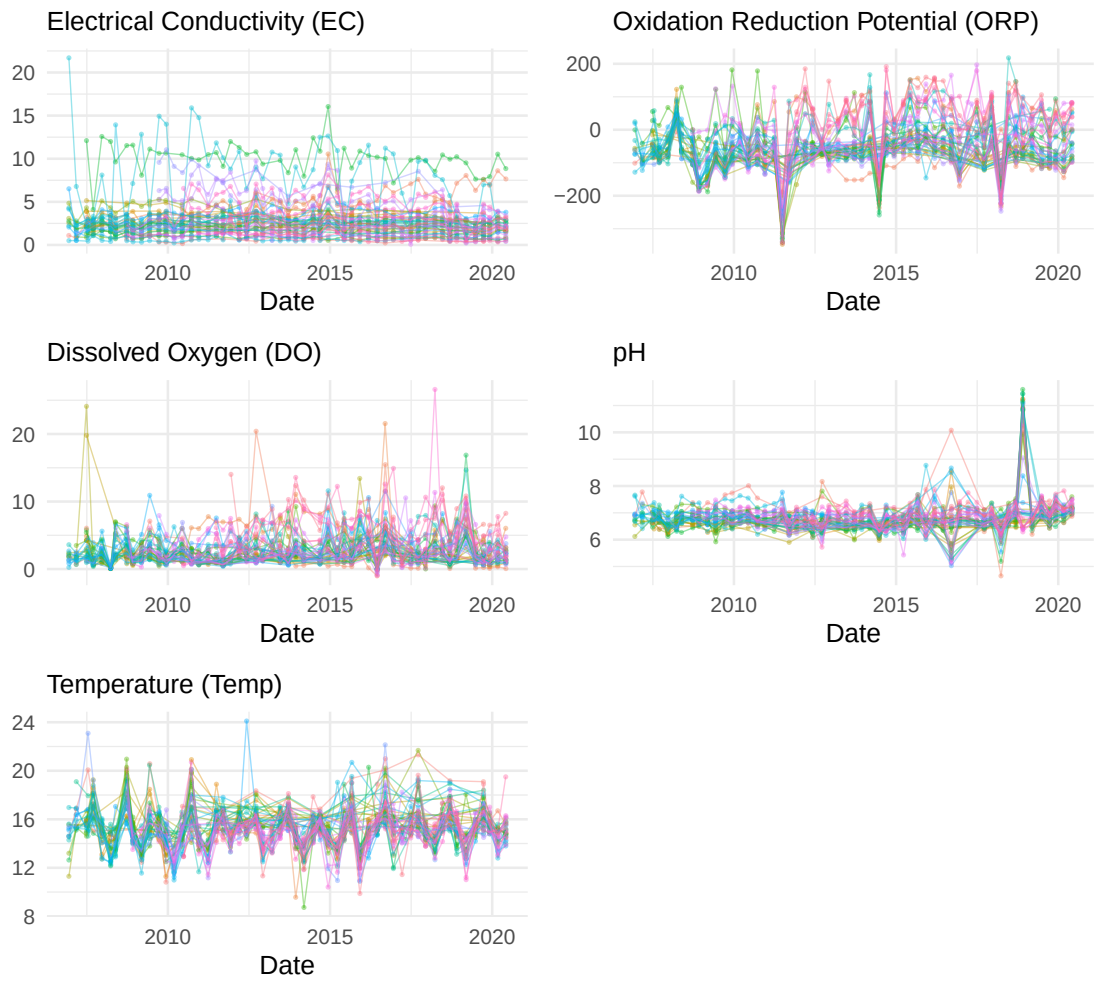

Figure 2.4: Multiple time series visualisation for predictors. Each colour represents a different well.

Figure 2.4 foreshadows which explanatory variables are most likely to be statistically significant in explaining analyte variation. There is a clear seasonal trend with temperature that is likely to overshadow any relationship our models would hope to infer. For conductivity, we see more inter-well variation with potentially multimodal data as most values exist in the $[0, 5]$ range but some wells contain data that is almost double these values, on average. Potential site-wide outliers exists for ORP, three clear troughs that occur for a large proportion of the wells, and pH, a singular spike in the final quarter of 2018. The data providers of site A investigated and

checked historical calibration reports and found no reason to believe these values are erroneous.

A key difference between Figure 2.3 and Figure 2.4 is the x-axis and range of sample dates, showing analyte observations several years before the first predictor observations. As mentioned in Section 2.1, a key challenge of telemetry within groundwater monitoring sites is ensuring sufficient data availability across both variable groups.

Figure 2.5: Multiple time series for analytes with holdout wells highlighted.

A weakness of such visualisations is the lack of well information shown beyond arbitrary colouring. Figure 2.5 highlights the holdout wells to be predicted and shows the heterogeneity of these wells. Furthermore, we observe that not all wells have existed since the inception of the groundwater monitoring site and it is not infeasible that extra monitoring wells may be added over time.

2.3 Censoring

The aim of any hydrocarbon groundwater monitoring, either during or after site operation, is to firstly understand the human impact on the groundwater and then to assist in the act of keeping this environment clean from pollution through preventive or remedial measures. It is then expected that the censoring level, defined as the percentage of observations that are left-censored, would increase if the site is being managed effectively. Combine this with the fact that new wells may be built over time and in site A, these wells are predominantly highly-censored perimeter wells. To validate these expectations, we can plot the censoring level, aggregated by the year recorded in the sample date, as shown in Figure 2.6.

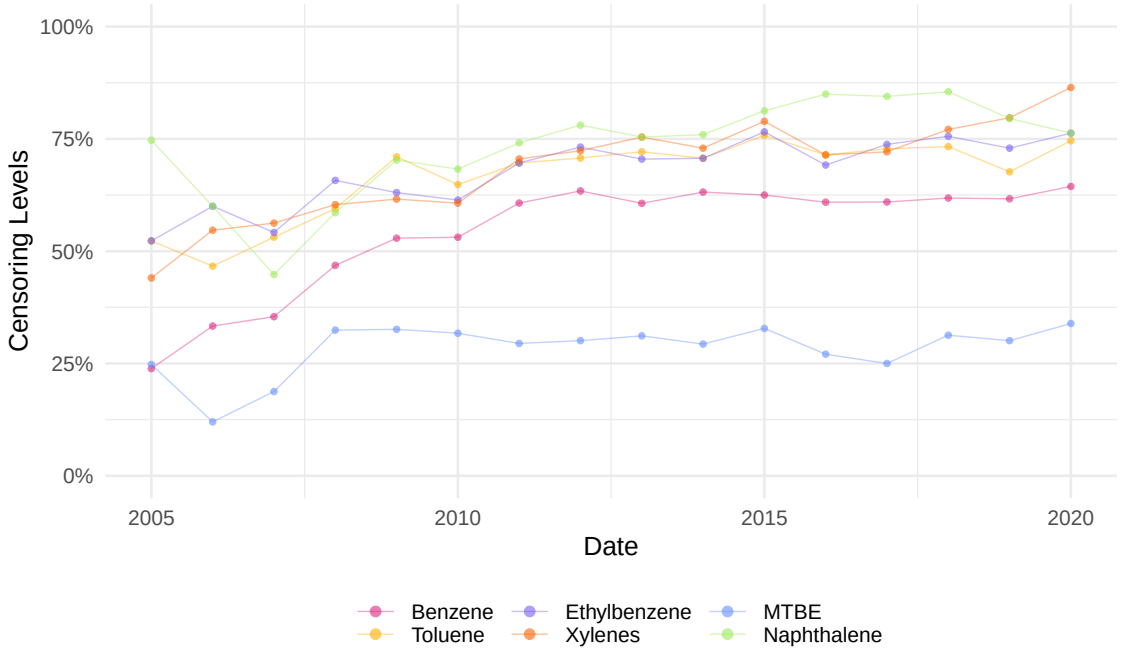

Figure 2.6: Percentage of censored observations per year, data before 2005 is omitted due to small sample size.

Since we have data for each analyte separately, variable groups common in these type of data such as BTEX (benzene, toluene, ethylbenzene and total xylenes) and TPH

(total petroleum hydrocarbons) have been removed in the data cleaning process. For Site A, the remaining analytes are the BTEX hydrocarbons, the polycyclic Naphthalene and methyl tert-butyl ether (MTBE) which showed the highest proportion of uncensored observations as shown in Figure 2.6.

Throughout this project we focus on the analyte with the lowest censoring level (MTBE) and the ubiquitous hydrocarbon, benzene, that has been formally classified as a carcinogen (World Health Organization, 2010). The argument behind this approach is that we are predominantly interested in modelling BTEX, however the average censoring level across all BTEX analytes is 63.4%, so we aim to leverage the correlated MTBE that exhibits less censoring overall. A simpler approach that would also be viable would be to model all analytes as dependent variables, this would similarly benefit from high correlation at the cost of more computationally intensive models. We explore this approach in Chapter 6.

2.4 Correlation

Computing a measure of correlation between variables where left-censoring may occur is less than straightforward. There exists three possible scenarios:

1. doubly-censored data (analyte, analyte);
2. singly-censored data (analyte, predictor);
3. fully observed data with no censoring (predictor, predictor).

The Pearson correlation coefficient is a function of the data and the means of both variables, this is only non-trivial in the latter case where no censoring occurs. A straightforward solution to this issue would be to impute the analyte data using DL/2, ROS or other substitution methods but this would be less than satisfactory due to poor estimation discussed in Section 1.5.1.

2.4.1 Doubly-Censored Data

In the case where we are interested in the correlation between two analytes such as benzene and toluene, we have doubly censored data. Helsel (2011) recommends using an adapted version of Kendall's Tau for these data as an alternate measure of correlation. Brown Jr *et al.* (1973) adapted the original rank-based method to work with censored data where one would assume overlapping intervals such as ND < 1 and 0.864 are a tie. This method has several advantages including the lack of an assumption about distribution and the metric is unaffected by monotonic transformations such as the log transformation we apply to the analytes.

A parametric approach to doubly-censored data can be found in Newton & Rudel (2007) where maximum likelihood estimates (MLE) are used. Recall from Section 1.5.3 that for a single censored variable with recorded values y_i and censoring indicator δ_i, for $i = 1, \ldots, n_s$, we can express the likelihood as

$$L(\Theta) = \prod_{i=1}^{n_s} f(y_i)^{1-\delta_i} F(y_i)^{\delta_i}. \tag{2.1}$$

Extending (2.1) to a bivariate case can be done using the properties of the bivariate normal distribution where the parameters, Θ, are the mean and standard deviation of both variables and a correlation parameter. That is,

$$L(\Theta) = \prod_{i=1}^{n_s} G_i \tag{2.2}$$

where

$$G_i = \begin{cases} f(x_i, y_i) & \text{if neither censored,} \\ f(x_i)F(x_i|y_i) & \text{if only } x_i \text{ is censored,} \\ F(x_i|y_i)f(y_i) & \text{if only } y_i \text{ is censored,} \\ F(x_i, y_i) & \text{otherwise,} \end{cases}$$

where the parameters of each PDF and CDF are omitted for the sake of brevity and $f(\cdot,\cdot)$, $F(\cdot,\cdot)$ are the bivariate generalisations of the aforementioned PDF and CDF.

Maximising (2.2) subject to the 5 parameters can lead to issues such as convergence on local, rather than global, maxima or non-convergence in a feasible time frame (Newton & Rudel, 2007). A further disadvantage to this approach is the bias that occurs when these data are multiply censored, as is the case in a typical groundwater monitoring site. A possible solution to convergence includes using (2.1) to estimate both variables to give variables' mean and standard deviation separately and then consider the quantities fixed in maximising (2.2). When the physical size of the water sample is known, bias can be mitigated using a technique described in Newton & Rudel (2007).

Our approach initially involved using data augmentation in a Bayesian linear model with a potentially censored response and explanatory variable. At each iteration of the MCMC-based Gibbs sampling algorithm, we impute censored data in the response variable and single explanatory variable iteratively using the derived full conditional distributions (FCD); a conjugate prior was assumed to produce analytical FCDs. By fitting this model we obtain a distribution representing correlation by calculating the Pearson's correlation coefficient between each imputed data, where uncensored data are unchanged throughout the imputation step. We present no results from this method in this thesis as preliminary models showed a bias when the data has multiple detection limits.

2.4.2 Singly-Censored Data

With the specification of the likelihood in (2.1), one can fit a censored regression model between a censored dependent variable and an arbitrary number of real-valued explanatory variables; more detail is given in Chapter 5. Using this linear model, we can estimate the correlation between censored and uncensored data using the coefficient of determination as a measure of how well these data fit to a straight line (Helsel, 2011). Alternatively, in a Bayesian paradigm, one can use data augmentation and calculate the correlation between imputed data as described in Section 2.4.1.

2.4.3 Uncensored Data

The most straightforward case is computing the correlation between two predictors since we can calculate Pearson's sample correlation coefficient directly due to lack of censoring. Relationships between these water quality predictors are studied in various water systems including groundwater networks, but also rivers and larger bodies of water (Summers, 2020). One must be cautious of generalising these relationships to all water systems.

Further complicating these results is the variety of environments in which groundwater systems inhabit. Silva *et al.* (2017) found highly significant ($r = 0.70$, $p < 0.001$) correlation between oxidation reduction potential (ORP) and dissolved oxygen (DO) during the warm rainy season whereas the other colder periods and the data taken as a whole showed no significant correlation. Figure 2.7 shows the same positive correlation between ORP and DO but is not limited in describing the complex relationships. We explore a simulation-based model in Chapter 3 to better describe our expectations of the relationships that occur in a groundwater network.

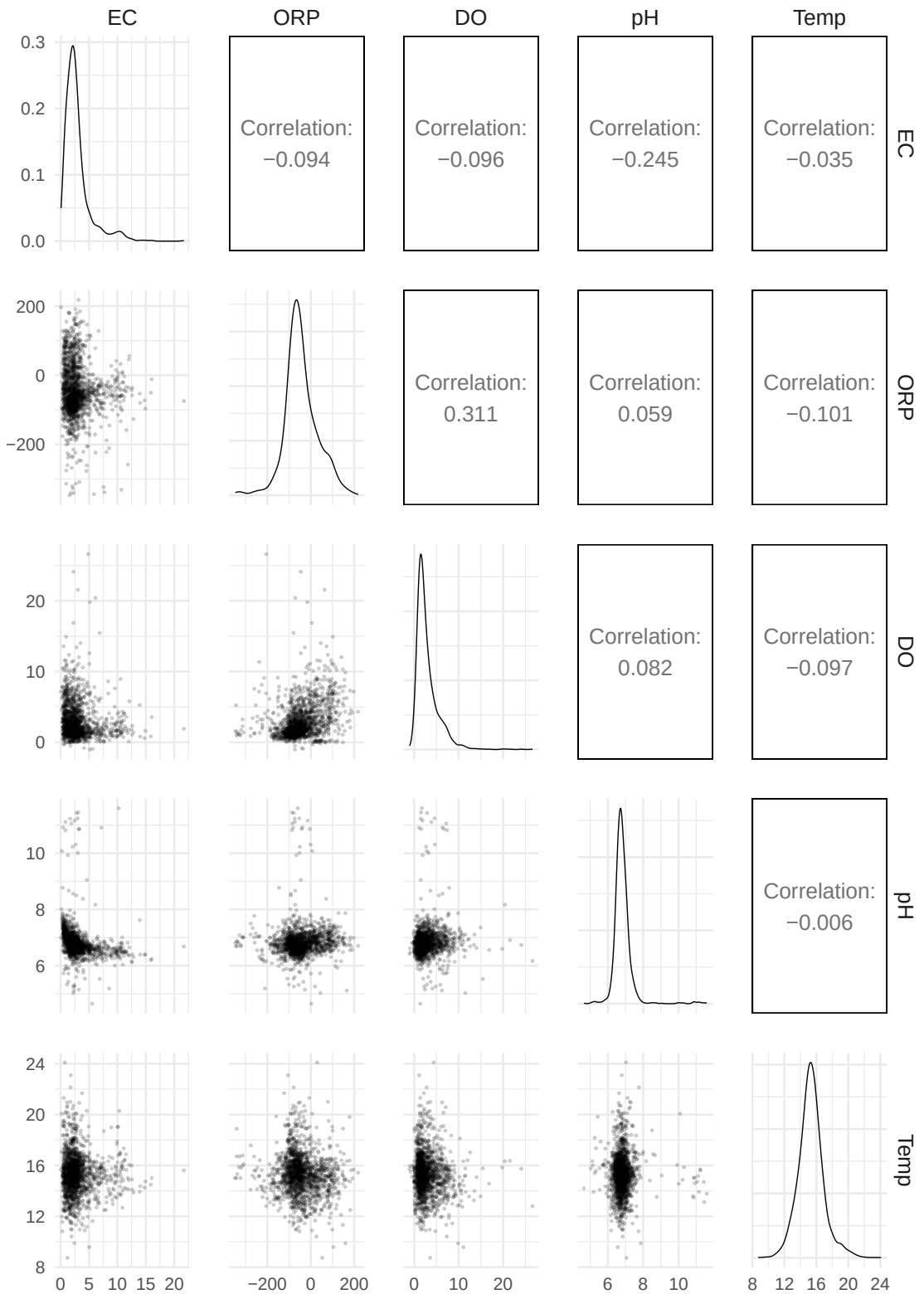

Figure 2.7: Pairs visualisation between unnormalised predictors showing correlation values, scatter and density plots.

2.5 Variograms

To understand the spatial correlation of these data it is important to recall the first law of geography,

> "… everything is related to everything else, but near things are more related than distant things," (Tobler, 1970).

And so our aim is to quantify when the spatial correlation between two arbitrary points is not non-existent, but negligible. To achieve this, we make use of a statistical visualisation known as the empirical variogram and also fit a variogram model for demonstrative purposes. Suppose our input is a spatial sample such as $\{y_{11}(\mathbf{s}_1), \dots, y_{n_s 1}(\mathbf{s}_{n_s})\}$ denoted to be a sample of our first analyte as a function of the corresponding spatial information, $\mathbf{s}_i = (s_{ix}, s_{iy})$. Note that we have dropped the temporal information, s_t, from $\mathbf{s}$ for this section only as our focus here is purely spatial, not spatiotemporal.

Our task is then to estimate the semi-variance between our random variable observed at arbitrary spatial coordinates $\mathbf{s}$, and another location $\mathbf{s} + \mathbf{h}$ where we define $\mathbf{h}$ to be the distance between the two locations. That is,

$$\gamma(\mathbf{h}) = \frac{1}{2}\operatorname{Var}(Z(\mathbf{s}) - Z(\mathbf{s} + \mathbf{h})), \tag{2.3}$$

where Z is a random variable (Oliver *et al.*, 2015). For our groundwater monitoring data, these vectors are 2-dimensional with x and y coordinates but (2.3) applies to higher dimensional $\mathbf{s}$ and $\mathbf{h}$. An assumption of second order stationarity, where the random variable has a constant mean at $\mathbf{s}$ and $\mathbf{s} + \mathbf{h}$, would typically be required for (2.3) and such an assumption may not be reasonable in certain applications.

A weaker assumption of *intrinsic stationarity* (Matheron, 1963),

$$E(Z(\mathbf{s}) - Z(\mathbf{s} + \mathbf{h})) = 0,$$

where the expected differences are 0 can also be used to ensure the validity of (2.3). Thus, the variogram is a plot of the semi-variance, $\gamma(\mathbf{h})$, against specified lag distances $\mathbf{h}$.

The structure of a variogram can be described in three main quantities,

1. the **nugget** denotes the estimated semi-variance at lag $\mathbf{h} = \mathbf{0}$;
2. the **range** is the estimated lag distance at which semi-variance "levels off";
3. the **sill** denotes the estimated semi-variance for all lags greater than or equal to the range.

In this thesis, we present results from the `gstat` R package (Pebesma, 2004) that estimates empirical variograms based on log imputed analyte data and normalised predictors. A spherical variogram is also shown as a solid line in Figure 2.8 to make the nugget, sill and range more clear; better choices may exist since this model assumes our data is isotropic, which may not be true for our groundwater application. No variogram model is shown for Figure 2.9. The power of these models are made clear when combined with Gaussian processes or Kriging, where the estimated nugget, sill and range can be used to specify kernel function parameters. In this thesis, variogram results are not directly used in any later analyses and are presented for demonstration only.

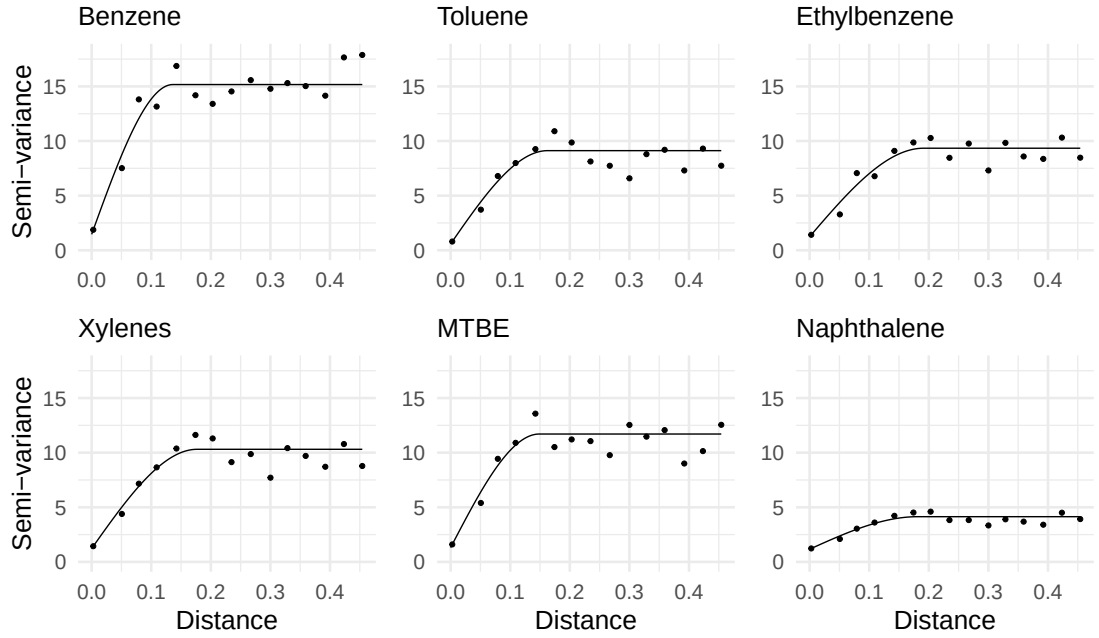

Figure 2.8: Empirical variogram of imputed (log) analyte data with spherical variogram model line.

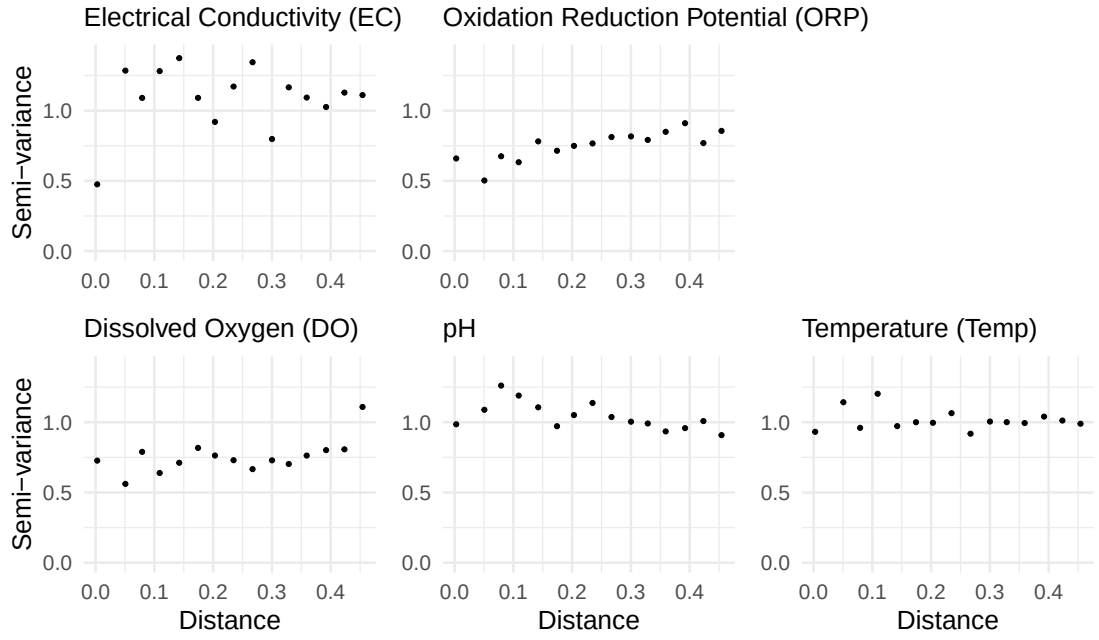

Figure 2.9: Empirical variogram of normalised water quality predictors.

As the spatial coordinates are scaled to $[0, 1]$ in both directions, any distances are relative to the overall length and width of the site given in Figure 2.2. We see that the variograms for the analyte data roughly follow a similar pattern, expected due to the high correlation of these data. Based on Figure 2.8, we expect the variance between two points to "level off" after a *range* of approximately 0.1 implying that our data only exhibits non-negligible spatial correlation for distances up to 10% of the groundwater monitoring site.

The empirical variograms in Figure 2.9 suggest a nugget-only model where the fitted line would be horizontal, the range would be effectively 0 and the sill would be equal to the nugget. Conductivity may have a similar correlation structure to the analytes over short distances. DO, ORP both appear to have increasing spatial variance over increased lag distances but we mostly observe little spatial correlation for these data.

For groundwater monitoring data, we have no reason to expect that the spatial variance will be identical in both directions due to the complex geological landscape of the underground aquifers. Further work in this area could involve different types of variograms, altering specified lag distances and investigating potentially anisotropic data; this can be achieved with a bi-directional variogram as demonstrated in Oliver *et al.* (2015). Alternatively, in models yet to be described in Section 6.2 and Chapter 8 we define length-scales for both s_x and s_y and would expect similar posterior distributions for each if these data are isotropic.

Variogram results presented here suggest a spatial correlation in the analytes, whereas the predictors show no evidence of substantial spatial correlation. This could be due to several reasons including the measurement error overpowering any spatial effect or the range at which spatial correlation "levels off" is much smaller than we expected.

2.6 Conclusion

Throughout this chapter we have considered one of the supplied datasets from a groundwater monitoring site as a case study, presenting our insights alongside the exploratory analysis and visualisations. Subsequent chapters will use these data as an example for how one could apply the respective methodology to any groundwater monitoring site. While there is a clear heterogeneity between sites that should be accommodated and taken into account, there are also key features that we would expect to see at similar sites.

As with many other industries, logistical challenges arise and produce irregular sampling strategies. For groundwater monitoring sites, this takes the form of extracting water samples from the environment up to several times per year, but tends to result in irregular time intervals between observations from the same well. Sample collection can also be sporadic as wells may be decommissioned due to asset damage or changing business requirements. Further complications include inter-sample variation, which may be considerable due to different data collection strategies where measurements of analytes and water quality predictors may occur several days apart.

Censoring is a major component of these data as intermediate steps are required for application of models including linear regression and random forests. Multiple detection limits are present and must be accounted for. While censoring levels are high and occur at various detection limits, further investigation has shown this to be skewed by perimeter wells that are necessary for assurances that the site is not contaminating areas beyond the site's operational area. That is, increases in censoring shown in Figure 2.6 can be partly explained by the construction of new perimeter wells. Moreover, excessive censoring in analytes may be indicative of smaller concentrations which makes those analytes less of a concern.

Correlation between analytes is known *a priori* and this is reflected in these data. However, quantifying the correlation between two random variables with censoring at multiple detection limits is a non-trivial task and has not been fully explored in this thesis. Another difficult task that is more pertinent to our aims is understanding correlation between each analyte and each predictor. Such a relationship does not appear to be straightforward and is thought to be highly sensitive to external factors such as geological composition or rainfall (Newton & Rudel, 2007).

Groundwater monitoring sites follow a similar well construction scheme with most wells concentrated around an area of interest with sparse wells surrounding the perimeter. This can make for interesting challenges in estimating spatial correlation as the inter-well correlation may not be the same over larger distances. Analyte variograms show spatial autocorrelation which can be explained by groundwater flow. Unfortunately, we have not been able to leverage groundwater flow data and similar inferences have not been reproduced for the water quality predictors.

Chapter 3

Reactive Transport Model

3.1 Introduction

Within such a complex environment such as groundwater, there is a clear need for interdisciplinary analysis and collaboration. This is resonated in water resources research where groundwater flow and transport model simulations have already been combined with statistical spatial models to assist in the problem of optimal placements of groundwater wells (Sreekanth *et al.*, 2017). Therefore, we consider a mechanistic model beyond the typical scope of solely statistical research, the reactive transport model (RTM). Built by combining several key scientific laws, for example conservation of momentum, RTM models are able to describe coupled physical, chemical and biological processes in Earth systems at a range of spatial and time scales (Steefel *et al.*, 2005).

RTM models have been applied to areas of research very closely aligned with our hydrocarbon monitoring application (Ng *et al.*, 2015) in the critical zone, typically characterised as where "rock meets life". The critical zone is the veneer between the planet surface and the bottom of any groundwater networks (Li *et al.*, 2017)

where many random geophysical properties such as groundwater, atmosphere and rock interact creating the potential for complex chemical, physical, and biological interactions. Therefore, a key advantage to the RTM is that we not only utilise equations representative of reality from other disciplines, such as those in the following list (Steefel *et al.*, 2005), but we also describe the coupling between these processes:

- conservation of energy;
- conservation of momentum (Navier-Stokes, Darcy's Law, Cauchy Equation);
- conservation of mass;
- conservation of solute mass.

An appropriately designed RTM model, with appropriate initial values can simulate a wide range of processes within a hypothetical aquifer (Li, 2023) including:

- fluid flow (single or multiphase);
- solute transport (advective, dispersive, and diffusive transport);
- geochemical reactions (e.g. precipitation);
- biogeochemical processes (e.g. oxidation–reduction reactions).

The ability to simulate these heterogeneous environments means we can analyse a groundwater monitoring facility with any desired sampling scheme that would be infeasible practically, and then analyse these simulated data. However, the statistical analysis of any simulated data will be limited by the closeness of the simulations to reality and, given the complexity and numerous possible chemical reactions, this is no trivial task. One method to improve the parity between simulation and reality would be to extend the processes and laws known to apply, *a priori*, to consider application-specific processes, for example, effect of partial oil saturation through the

application of relative water permeability (Ng *et al.*, 2015). Alternatively, extracting more information about the groundwater monitoring site such as aquifer composition or location may also improve simulations, but further work is required to confirm the efficacy of this approach.

A key application of interest to us is the ability to model the transport of hydrocarbon chemical species through porous media; Ng *et al.* (2015) applies an RTM to the infamous Bemidji site in an effort to understand secondary water quality impacts. While similar to our aims, the paper focuses on the impact on groundwater quality that arise from remediation techniques that introduce organic matter to the system that will enhance biodegradation but could do so to the detriment of the overall health of the environment. We present simulations from Delft University of Technology (TU Delft) in Section 3.2 based on the Bemidji, Minnesota case study.

3.2 Simulation

To produce a RTM-based dataset we collaborated with Wetsus, European centre of excellence for sustainable water technology, and TU Delft. The dataset provided was created to represent the Bemidji case study (Essaid *et al.*, 2011) and developed iteratively where each revision to the model should increase the accuracy of the simulated data. In these simulated mechanistic data, the sampling scheme can be arbitrarily changed and all spatial or temporal units are only to help with comparison to the real world application. As such, we observe data from 5 wells, each with 3 different screening depths, over a period of 42 years with a monthly sampling frequency. We label these wells numerically as shown in Figure 3.1 and use the three suffixes S (shallow), M (medium) and D (deep) to denote screening depths of 4, 8 and 12 metres. Multiple screening depths arise from sites where there is a benefit

in collecting samples from the same well at different depths due to the geological landscape causing different concentrations and behaviour at different depths.

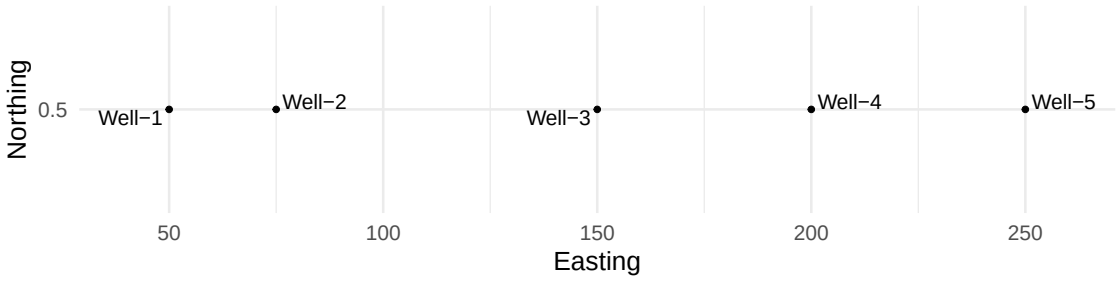

Figure 3.1: Well locations of simulated RTM dataset.

Simulated variables differ from what we observe in groundwater monitoring sites. Volatile organic compounds such as the hydrocarbons in BTEX are simulated alongside non-volatile organic compounds, short chain alkanes and long chain alkanes. Predictors arising from these simulations include the aforementioned pH, electrical conductivity (EC), dissolved oxygen (DO) and a new predictor, *pe*, that represents the tendency of a compound to either gain or lose electrons and is therefore expected to be correlated with oxidation reduction potential (ORP). For our purposes, we consider inorganic compounds to be potential predictors of hydrocarbon concentrations that would require further data collection. Hence, we extend our set of predictors to include iron, manganese, sulfur and calcium. The RTM models compounds such as iron in various oxidation states, for example Fe^{2+} and Fe^{3+}, but we combine these measurements.

Performing an exploratory data analysis on the RTM dataset reveals a clear pattern to the data that we will describe as "phases". In these data, wells either have two phases, such as Figure 3.2, or three phases as seen in Figure 3.3. Recall that simulated reactions inform the system through differential equations based on physical laws described in Section 3.1. Of all feasible reactions, one is expected to dominate

the system and become the main influence on hydrocarbon concentrations and predictor observations until some critical point. It is this dominating behaviour that informs the aforementioned phases.

Built into these data generating processes is the introduction of hydrocarbon species, as may occur at groundwater monitoring sites, therefore all hydrocarbon concentrations report 0 before this inflection point. A consequence of synthetically introducing the hydrocarbons at a single point is that we observe different analyte concentrations and behaviour due to a different well location or different depth. Hydrocarbons are introduced at the surface so deeper wells are further away and must wait the longest time to detect any non-zero hydrocarbon concentrations.

All wells share a starting state where there is some level of dissolved oxygen that depletes over time from some initial value. The physical interpretation of this is that when there is sufficient dissolved oxygen in the groundwater system, an aerobic degradation is able to occur up to some point in time when there is insufficient dissolved oxygen. Following this point in time, we then observe an anaerobic reaction that continues to decrease the hydrocarbon concentration in a reaction pathway involving some number of inorganic compounds including iron and manganese.

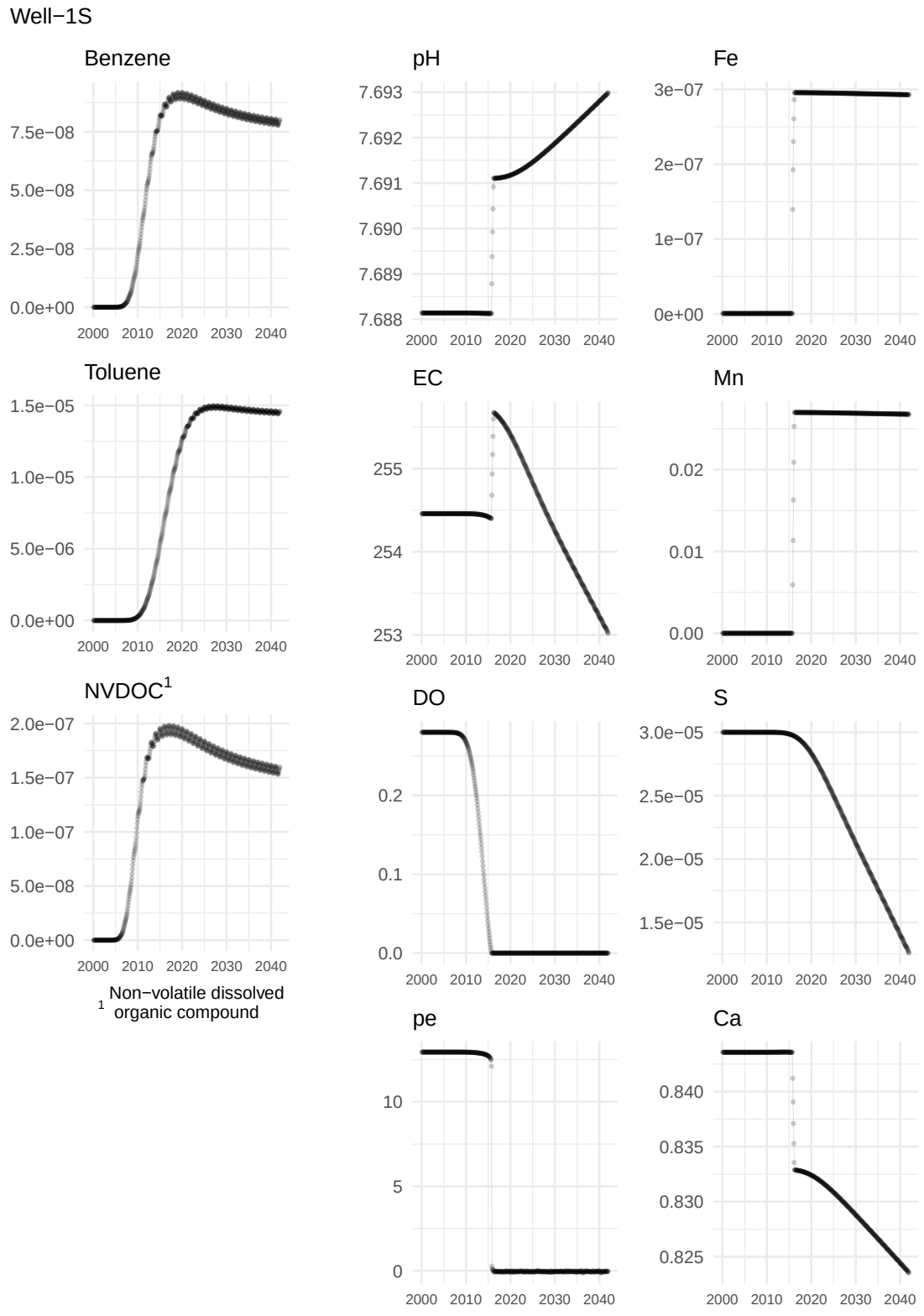

Figure 3.2: Time series visualisations of key variables. Well-1S, RTM.

Care must be taken with these visualisations due to the differences in the y-axis limits. In some cases, a relationship may be present but due to the minimal change in a predictor, say an increase in pH of 0.004, it is unlikely that this relationship will be observed due to the measurement error and other stochastic processes occurring within the groundwater system.

Due to the omission of measurement error, the variable *pe* appears very similar to a discontinuous step function as shown in Figure 3.2 and more clearly in Figure 3.3. The nature of this quantity would make a poor predictor in the typical regression sense, to be described in Chapter 5, as there are many values of benzene per one 'true' value of *pe*. Instead, we may be able to utilise some of these predictors to cluster our observations according to several data generating processes to help deal with the non-linearity suggested by the RTM model. A statistical model motivated by this concept is defined and considered in Chapter 7.

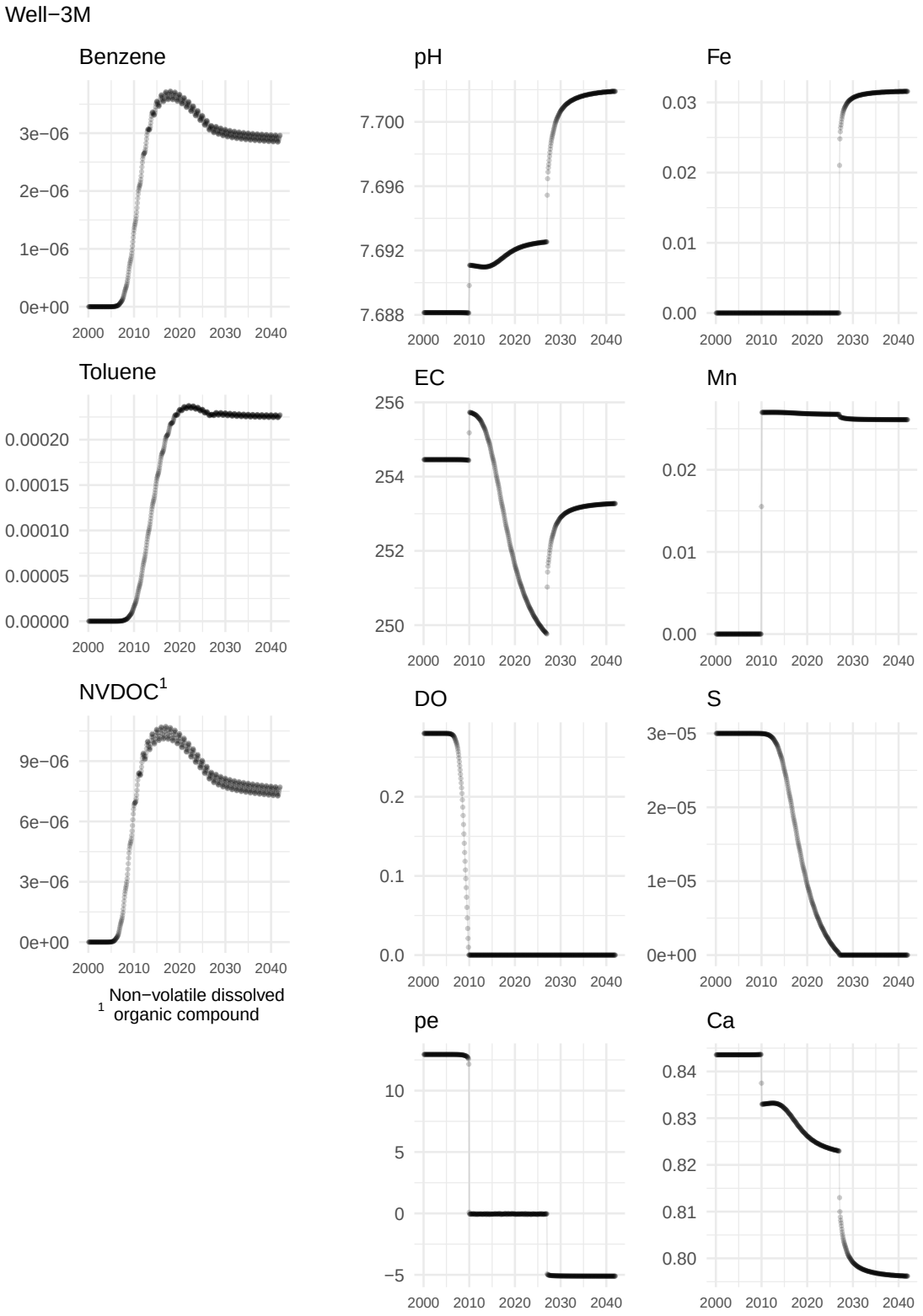

Figure 3.3: Time series visualisations of key variables. Well-3M, RTM.

3.3 Conclusion

Geological modelling aims to calculate the composition of a system at equilibrium. This is achieved by aiming to find the fewest independent variables that are capable of fully describing the state of equilibrium (Appelo & Postma, 2005). This is similar to our statistical approach where we want to find the fewest independent variables to describe a state of equilibrium, subject to some random, but quantifiable, variation. Consider the task of clustering hydrocarbon concentrations within a groundwater monitoring network: some geological models may require in excess of 20 complex governing equations required to describe the system, clustering may reduce this number by combining some clusters and ignoring others. That is, many of our goals can be achieved by leveraging a sub-model and it may not be worth adding complexity if no benefit is added.

The mechanistic model presented and analysed in this chapter is invaluable and can be used in several ways.

1. Description of elemental and nutrient fluxes between major Earth reservoirs (Steefel *et al.*, 2005);
2. better understanding of a groundwater site by simulating a *synthetic* site designed to mirror reality;
3. simulate forward in time based on some initial conditions;
4. elicit expert opinion on reactions that are key to the groundwater monitoring network.

Furthermore, this approach has motivated our use of the Mixture of Experts (MoE) model, to be introduced in Chapter 7, where data is clustered and then fit to separate regressions. Key modelling decisions like choosing the number of components is based on lessons learned from these mechanistic models like the RTM.

Chapter 4

Random Forests

4.1 Introduction

Motivated by the desire to understand the limits of how well we can explain our analyte concentrations using our chosen predictors, we turn to machine learning methods for comparison and to act as a baseline. One such method, 'random forests' (RF), are ensemble learning methods often used for classification problems but can also be applied to a regression context (Breiman, 2001). The motivating idea is to use tree models, that often result in low bias but high variance, then reduce the variance by averaging over multiple trees. We model each dependent variable separately using the same predictors.

A further complication is the left-censored nature of our data. For this chapter we impute our data using the DL/2 imputation method described in Section 1.5.1 and accept the imperfections of this method in the name of pragmatism. Further work on random forests applied to groundwater monitoring data could include an extension to the case where the dependent variable is left-censored. Oblique random survival forests may be used for situations where the dependent variable is right-censored,

however, the feasibility of such methods is not completely understood (Jaeger *et al.*, 2023).

The output of each RF model is multi-faceted and will be considered once we have completed all necessary tuning of the model. Model artefacts include

- feature sampling that yields **importance measures** offering insight;
- **partial dependence plots (PDP)** which visualise marginal effects between a single analyte and single predictor;
- **predictions** to be generated given some new explanatory data.

The main disadvantage is the opaqueness of such a black-box method where a misunderstanding of the methodology can lead to applying these models in an unsatisfactory way. As such, we cover the basic manner in which these models are constructed in Section 4.2.

4.2 Methodology

Even though we leverage the random forest as a black-box tool to understand complex relationships at a preliminary stage, there is some benefit in understanding some simple concepts related to this method. Moreover, even partially understanding machine learning techniques can help when applying these concepts. We build up the concept of a random forest starting with a single tree.

4.2.1 CART

Classification and regression trees (CART) were originally proposed by the same author of the random forests approach (Breiman *et al.*, 1984). These tree-based models operate by assigning all observations to a single root node and then deciding on some

'best' split where the data is partitioned into a left and right node. Such a process is repeated recursively until some stopping criteria is met such as all terminating nodes have fewer than m observations or fewer than k classes in the classification context. When a node is terminating, it can contain one or several observations with known dependent variable values. Prediction can then be executed by finding a terminating node using the new explanatory data, then taking a summary of the dependent data contained within that node; in regression we typically use the mean and classification would use the most popular vote.

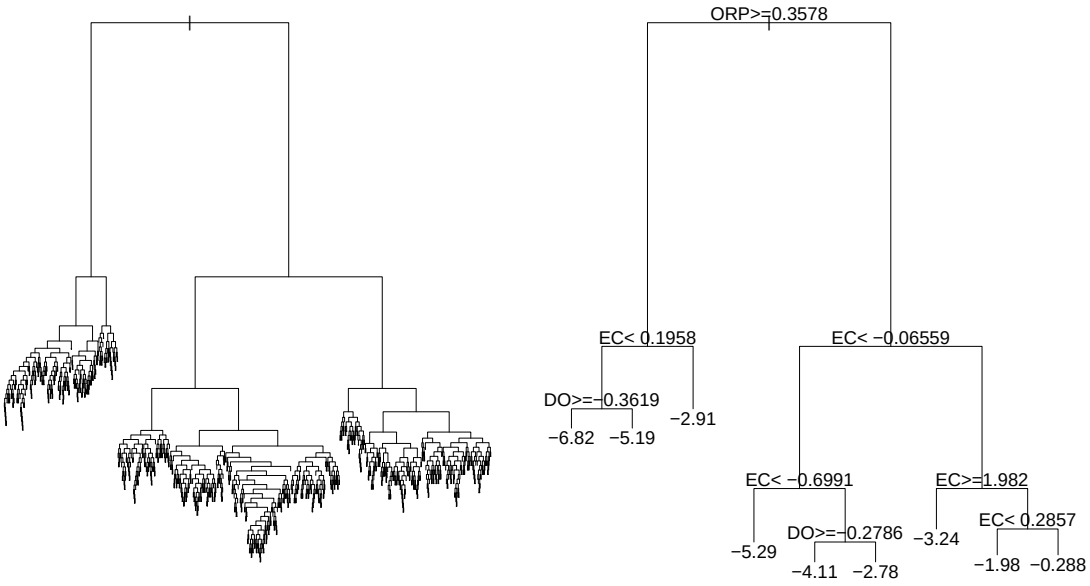

Figure 4.1: CART output showing maximal (left) and pruned with labels (right) trees. MTBE imputed by DL/2.

Figure 4.1 shows an example of a *maximal tree* when applied to our Site A data without the three wells to be predicted in Section 4.5. A pruned tree is a simplified version of the original tree with minimal loss in predictive performance. While we do not consider pruning here since it is not used in random forest models, we can apply pruning using the `rpart` R package (Therneau & Atkinson, 2022) to show a much simpler tree, with labels.

Choosing the best split is done by minimising a cost function, $C(t)$, that is representative of the *impurity* of node t. Suppose we have n observations at an arbitrary node, t, to be split into a left node, t_L, and right node, t_R. Let n_L and n_R represent the sample sizes of the left and right node respectively. Then, we define the within-node impurity to be minimised as

$$\frac{n_L}{n} C(t_L) + \frac{n_R}{n} C(t_R).$$

Equivalently, we can think of this problem as the maximisation of increase in purity caused by the split, where that quantity is

$$C(t) - \left(\frac{n_L}{n} C(t_L) + \frac{n_R}{n} C(t_R) \right).$$

For regression, we simply consider the cost function to be the variance defined as

$$V(t) = \frac{1}{n} \sum_{i: y_i \in t} (y_i - \bar{y}_t)^2,$$

where $y_i \in t$ if and only if observation i is allocated to node t and $\bar{y}_t$ denotes the mean of the n observations at node t (Genuer *et al.*, 2020). While the misclassification rate would be a viable cost function for classification trees, uniqueness of the 'best' split would not be guaranteed (Genuer *et al.*, 2020) and so we advise minimising the Gini impurity function,

$$\phi(t) = \sum_{k=1}^{K} \hat{p}_t^k (1 - \hat{p}_t^k),$$

where $\hat{p}_t^k$ is the probability of picking an observation with label k at node t, which is calculated for all K possible labels.

One of the main issues of CART models suffer is a high sensitivity to the training data. That is, a slight change in the training data may result in a different root node and the recursive nature of tree-based models would develop a vastly different

maximal tree further causing the optimal tree after pruning to be different. The random forest model originally proposed in Breiman (2001) mitigates this via a combination of bagging, explained in Section 4.2.3, and using random input trees in place of CART trees.

4.2.2 Random Input Trees

A random input (RI) tree is constructed in a very similar way to a CART tree with two distinct differences.

1. No pruning is performed on a RI tree;
2. each potential node split can only use a randomly sampled subset of the available predictors.

With the introduction of the latter point, we have a new tuning parameter for these trees, denoted p_f, that represents the number of randomly selected predictors; we discuss how to choose a value of p_f in Section 4.3. In a CART model, we produce several *competing splits* based on each predictor and pick the best split to improve node purity. Whereas, for RI, the exact predictors to be made available will be a subset of all predictors, randomly sampled without replacement. It is important to note that while each node of the RI tree will have the same number of predictors available, the predictors that can be used for splitting will be different per node.

4.2.3 Random Forest Random Input

Random forest models are ensemble learning methods where the predictions of many models are collated to form a final prediction, ideally with less variance than any single model. This approach is appealing when dealing with high-variance models such as CART or RI. One way to allow our model to benefit from ensemble learning

is the application of bootstrap aggregation, otherwise known as *bagging*. Bootstrap sampling is a method where all observations are randomly sampled, with equal probability and with replacement, to produce a new dataset of the same size. For observations indexed by $\{1, 2, 3, 4, 5\}$, examples of bootstrap samples include $\{1, 1, 2, 5, 5\}$, $\{5, 4, 3, 3, 2\}$, and $\{4, 2, 4, 4, 3\}$. The motivation for applying this method to each tree-based model within a RF is to encourage heterogeneity between the models and to improve the accuracy of the final aggregated estimate.

Randomness is further perpetuated by the use of RI trees instead of CART trees. We use the `randomForest` R package (Liaw & Wiener, 2002) to implement this particular version of the random forest model based on the original Fortran code developed by Leo Breiman.

A preliminary model attempted to incorporate spatiotemporal information by defining a new predictor to be some function of the sample's spatiotemporal data, say $f(x, y, t)$ in our case. In hindsight, this approach applied to a tree-based model would likely lead to clustering of the data based on the associated well or sample date. One can see how this would lead to overfitting and since our aims are to predict previously unseen information, this model was swiftly deprecated.

4.3 Parameter Tuning

4.3.1 Number of Trees

There is a trade-off when choosing the number of trees. Too many trees will increase the duration of each step such as model fitting and calculating partial dependence in Section 4.4.3, but too few trees lead to poor estimation. In the extreme case when only one tree is used, problems present in CART models may also present themselves.

We take the number of trees to be 500 in all random forest models with the justification that the results do not change substantially as more trees are added. Conversely, the model only becomes sensitive to the choice of the number of trees when considerably fewer trees are used.

4.3.2 Number of Predictors

When fitting a random forest with random input tress we have to set tuning parameter, p_f, that controls the number of predictors that are randomly sampled at each split. A reasonable rule of thumb, for regression, is to let $p_f = n_x/3$, however, this may not always be the most optimal value (Genuer *et al.*, 2020) and so we opt to use grid search of all possible values, $p_f = 1, \ldots, n_x$ and K-fold cross validation to tune this quantity. As described in Section 2.2, we do not include our hold-out wells at the model fitting stage or at any parameter tuning stage.

We split the data to be used for parameter tuning into K groups, called 'folds', with the intention of fitting several candidate random forest models to be defined. For each model fit, a single fold is allocated to the *validation* set and all other folds are allocated to the *training* set. Each model is then trained on its respective training data and constructs predictions for the validation set that can be assessed qualitatively or quantitatively using model comparison metrics.

Since the random forest model does not define a likelihood, we can not use LPD, WAIC or PSIS and instead we obtain a model performance metric in the form of the root mean square error,

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n_s}(y_i - \hat{y}_i)^2}{n_s}}, \tag{4.1}$$

where $\hat{y}_i$ is the predicted concentration.

One concern of this cross-validation methodology is the sensitivity of the results due to its inherent randomness. To alleviate this we can repeat the process M times. Hence, each candidate model has to be fit and produce predictions for $K \times M$ distinct partitions, consequently the impact of any especially favourable splits are reduced.

A random forest model is valid if the number of permutable predictors is $p_f \in \{1, \ldots, n_x\}$. Since there are relatively few predictors for our dataset, we execute an exhaustive grid search over the whole parameter space and fit a total of $K \times M \times n_x$ random forest models for each analyte to be modelled.

4.4 Leave-One-Well-Out (LOWO)

To demonstrate the importance and partial dependence output of the random forest models, we fit several random forest models in a leave-one-well-out (LOWO) scenario (Evers *et al.*, 2015). Splitting the data into partitions based on the corresponding well identifier allows us to fit models to hypothetically predict a single 'left out' well using the data from all other wells for training the model. This approach should highlight any high-influence wells while serving as an illustrative example of importance and partial dependence. Additionally, the impact on our models of adding or removing a well is highlighted which could be useful information when optimising well placement and sampling frequency of groundwater monitoring networks (Mclean, 2018).

4.4.1 Pseudo R-Squared

For the regression version of the random forest model, we obtain a goodness-of-fit metric in the pseudo R-squared statistic, defined as

$$R^2 = 1 - \frac{(\text{RMSE})^2}{\text{Var}(y)}, \tag{4.2}$$

where RMSE denotes the root mean square error as defined in (4.1) and $\text{Var}(y)$ is the sample variance of all log concentrations $y_1, \ldots, y_{n_s}$.

Unlike the coefficient of determination, this quantity can be outside the conventional range of $[0, 1]$ where negative values imply that the mean of the data provide a better fit to the data than the model. Hence, the interpretation that (4.2) is the proportion of explained variation is spurious. Using too few trees, say fewer than 10, will lead to poor R^2 values and therefore a poor model fit. On the other hand, adding more trees to an existing forest will lead to a fit that is typically no worse than a fit with fewer trees, according to the aforementioned pseudo R^2.

4.4.2 Importance

Variable importance ranks all predictors, ordered by the impact they have on the dependent variable. We consider two importance metrics as calculated in Liaw & Wiener (2002).

To construct the first importance score, out-of-bag (OOB) evaluation is used. Since bagging is used, there often exists a subset of the data that was not used to create the tree and we define this to be the out-of-sample subset for this tree. We also formulate an alternate version of these data, where the values of the j^{th} predictor are randomly permuted. Then the importance contribution of a single tree for the j^{th} predictor is the difference in prediction error when the original out-of-sample data is used and when the perturbed out-of-sample data is used (Liaw & Wiener, 2002). For this process, we require a function to represent the prediction error, for example misclassification rate for classification. We use mean square error in our regression context.

An alternative importance metric can also be constructed by considering the effect an arbitrary predictor has on the change in node purity. At each node where the predictor in question is used to make a split, we observe an increase in node purity that is solely due to that predictor that can be recorded during creation of each tree. By calculating the weighted mean of this change in node purity, weighted by the proportion of observations in the node, we can generate an importance contribution from each tree (Breiman, 2001).

4.4.3 Partial Dependence Plots (PDP)

The motivation of partial dependence plots (PDP) is to provide an insightful visualisation of the relationship between the predicted concentrations and the predictors. When the predictors of interest, say $\mathbf{x}_l$, are single dimensional this is straightforward using a scatter plot, and a contour plot can be used for two dimensional inputs; higher dimensions of $\mathbf{x}_l$ become very problematic to visualise. The proposed solution is to estimate the marginal effect that some subset of our explanatory variables has on our predictions (Friedman, 2001). A PDP can be made for any "black box" method including, but not limited to, neural networks, support vector machines and random forests.

Suppose we are able to make predictions using some arbitrary model and a vector of predictors, say $\mathbf{x}$; we denote these predictions as $\hat{f}(\mathbf{x})$. By partitioning our predictors into a chosen subset $\mathbf{x}_l$ of size l and the complementary $\mathbf{x}_{\_l}$, these functions can be expressed as

$$\hat{f}(\mathbf{x}) = \hat{f}(\mathbf{x}_l, \mathbf{x}_{\_l}),$$

where we intend to marginalise over the complement predictors to obtain a function of chosen predictors only.

In fact, Friedman (2001) posits that if the dependence on the complement set is not too strong, then a useful measure of the partial dependence is defined as the expectation

$$\bar{f}_l(\mathbf{x}_l) = E_{\mathbf{x}_{-l}}[\hat{f}(\mathbf{x})] = \int \hat{f}(\mathbf{x}_l, \mathbf{x}_{-l}) p_{-l}(\mathbf{x}_{-l}) d\mathbf{x}_{-l},$$

where $p_{-l}(\mathbf{x}_{-l})$ is the marginal probability density of the complement set.

We estimate this quantity by averaging over the same data that trained the model, that is,

$$\bar{f}_l(\mathbf{x}_l) \approx \frac{1}{n_s} \sum_{i=1}^{n_s} \hat{f}(\mathbf{x}_l, \mathbf{x}_{-l}^{(i)}),$$

where $\mathbf{x}_{-l}^{(i)}$ denotes the complement predictors for observation $i \in 1, \ldots, n_s$. Intuitively, we can create a partial dependence plot by supplying values for the chosen subset and then calculating the average prediction where the values for the complement set of predictors are based on the training data.

In practice, we leverage the `partial` R package (Greenwell, 2017) and loop through each predictor so the chosen subset is always of size 1. As a consequence, all plots produced are only two-dimensional scatter plots showing the marginal effect of varying a single predictor by plotting these estimated concentrations against the input values. We also include a rug plot below the plotting area to highlight values observed in the training data to avoid unnecessary extrapolation.

4.5 Groundwater Application

We apply the models described in this chapter in three scenarios, each with different motivations.

1. Cross validation, with $K = 5$ folds and $M = 10$ repetitions, is applied to all non-holdout wells in Section 4.5.1 to discern the number of predictors hyperparameter for all other models;
2. LOWO models are applied to datasets that are constructed by specifying a single well to be predicted and all non-holdout wells are available for model fitting. Section 4.5.2 highlights the influence of each well by visualising key model artefacts including importance and partial dependence plots (PDP);
3. final models, shown in Section 4.5.3, imagine a scenario where data to be predicted is truly unseen as prediction is enacted on the holdout wells that were not used in choosing model hyperparameters.

Decisions common to all scenarios include the choice of the number of trees and available predictors. All random forests consist of 500 trees, as discussed in Section 4.4.1, and assume that the variables to be randomly sampled from at each node split are the entire collection of available predictors: pH, conductivity (EC), temperature, dissolved oxygen (DO) and oxidation reduction potential (ORP).

4.5.1 Number of Predictors

To determine the number of predictors to be sampled at each node split we apply the approach described in Section 4.3.2 with $K = 5$ folds and $M = 10$ repetitions. Each model provides a value denoting our model comparison metric, say RMSE, that is then plotted.

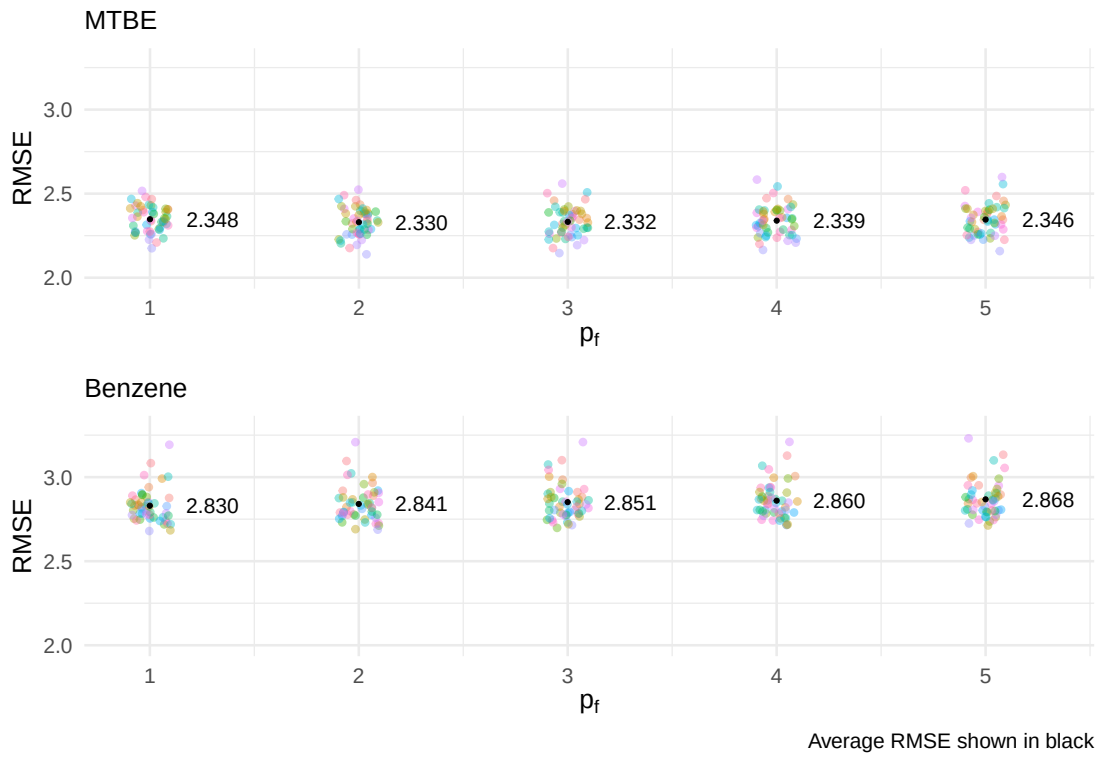

Figure 4.2: RMSE for all tuning random forest fits. Both analytes imputed by DL/2.

By clustering points according to the choice of number of predictors and colouring points to denote the specific data partition we can see if any choice of this hyperparameter produces a consistently better performing model. We see from Figure 4.2 that the difference in RMSE is relatively negligible, for both analytes, in almost all cases suggesting that these models are highly insensitive to the choice of p_f. Moreover, since we have used colour to distinguish different splits in the data, we observe that the model with the lowest RMSE for each candidate model tends to come from the same data partition. We opt to take the tuning parameter that minimises average RMSE, per analyte, and therefore assume $p_f = 2$ for MTBE and $p_f = 1$ for benzene are the best choices for random forest models applied to site A. A further simplification we have not explored would be to choose a common p_f value for all analytes.

4.5.2 LOWO Models

Many visualisations designed for a single random forest can be applied to our collection of random forest models by allowing each LOWO data partition to be denoted with a different colour. As such, we present model outputs and diagnostics for multiple partitions formed by changing the well missing from the training data but present in the validation data.

4.5.2.1 Pseudo R-Squared

By plotting this metric against the number of trees used one can see that increasing the number of trees leads to diminishing returns. In the LOWO models when there are very few trees used, almost all models have a R^2 value less than zero as shown in Figure 4.3. Even as the number of trees used increases, we see that each model struggles to explain more than 50% of the variation in analyte concentrations.

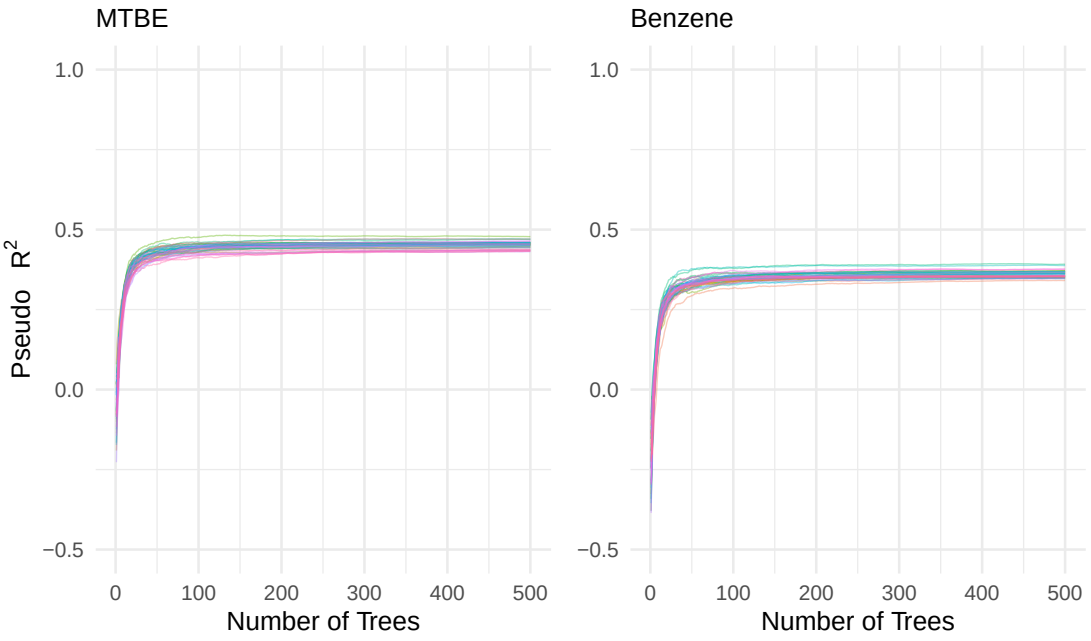

Figure 4.3: Pseudo R^2 against the number of trees for each LOWO model. Each colour represents a different LOWO model. Both imputed by DL/2.

4.5.2.2 Importance

Recall from Section 4.4.2 that we have two metrics to convey predictor importance. To read the importance plots, the y-axis represents the change in mean square error when the corresponding predictor is perturbed, whereas the x-axis denotes the average change in node purity across all splits that use the corresponding predictor.

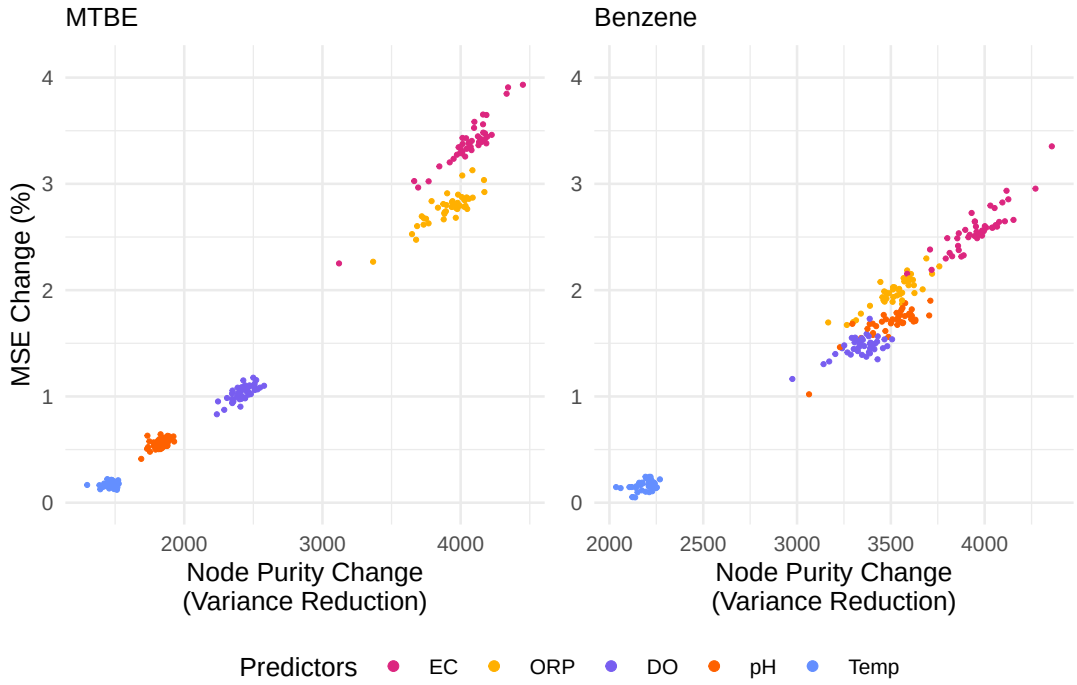

Figure 4.4: Importance metrics for each LOWO model. MTBE and benzene imputed by DL/2.

We see many similarities between the importance plots for both analytes. The impact of changing the well not used in model training is minimal in all but a few outlier cases. For the MTBE plot in Figure 4.4, two atypical points with node purity change between $3,000$ and $3,500$ correspond to wells "Well-22" and "Well-29" which are located to the east-side of the "Focus" holdout well with respective relative easting and northing of $(0.24, 0.32)$ and $(0.35, 0.44)$, as shown in Figure 2.2; a priori, we know the contamination event occurred close to the "Focus" holdout

well and MTBE had a tendency to be transported eastwards. Similarly, for benzene, two atypical points close to 1% MSE change both correspond to the LOWO model where the left-out well is "Well-04" which is on the west side of the operating area with a relative easting and northing of $(0.13, 0.39)$, as shown in Figure 2.2. These "middling" wells showing a decreased importance for some predictors behave most like the "Adjacent" holdout well with concentrations much higher than perimeter wells but an order of magnitude less than wells near the contamination event like "Focus".

Conductivity (EC) appears to have the greatest average importance to both analyte concentrations. On the other hand, temperature is seemingly unimportant in all models for both analytes which foreshadows the regression coefficients discussed in Chapter 5. One reason for this could be the noise and seasonality around the temperature data where minor fluctuations are overshadowed by bigger changes that are irrelevant to the hydrocarbon concentrations. While the ordering of importance appears to be fairly similar, DO and pH are more alike in terms of importance for benzene than MTBE.

Further work on importance metrics could be useful, for example suppose our interests were fixed on interaction effects, effects from two or more predictors that may be more or less than the sum of the effects separately. Then, we could look into extending the permutations of a single predictor to multiple (Gregorutti *et al.*, 2015).

4.5.2.3 Partial Dependence Plots (PDP)

Figure 4.5 and Figure 4.6 show partial dependence plots for MTBE and benzene, respectively. As before, we combine plots using colour to discern LOWO models.

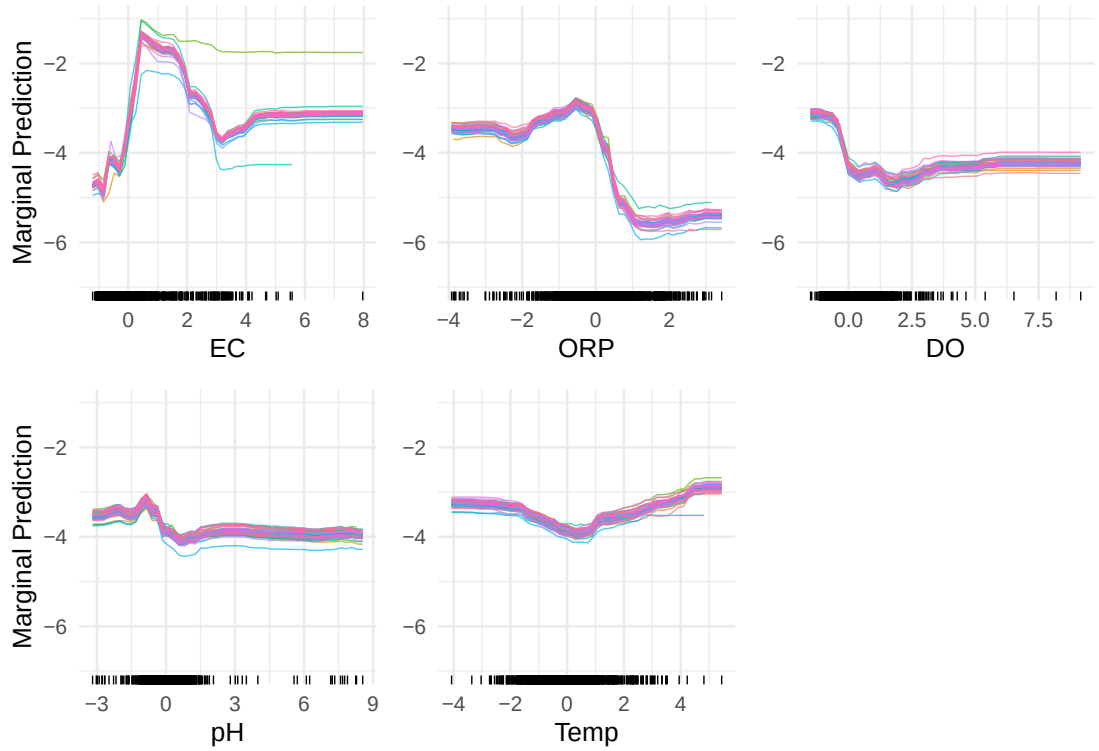

Figure 4.5: Partial dependence plots with observed data marked below. Each colour represents a different LOWO model. MTBE, imputed by DL/2.

Other than temperature, most partial dependence plots are showing non-monotonically increasing functions implying that any relationships in the data may be non-linear. For example, pH appears to increase slowly then drop harshly around the mean value of zero and then increases afterwards. Since the predictors are normalised, the change-point value of zero corresponds to the average pH observation at this site: 6.82. Considering the physical meaning of this predictor, predictions where pH is negative are more likely to be basic and positive pH implies a higher chance that the water sample was acidic.

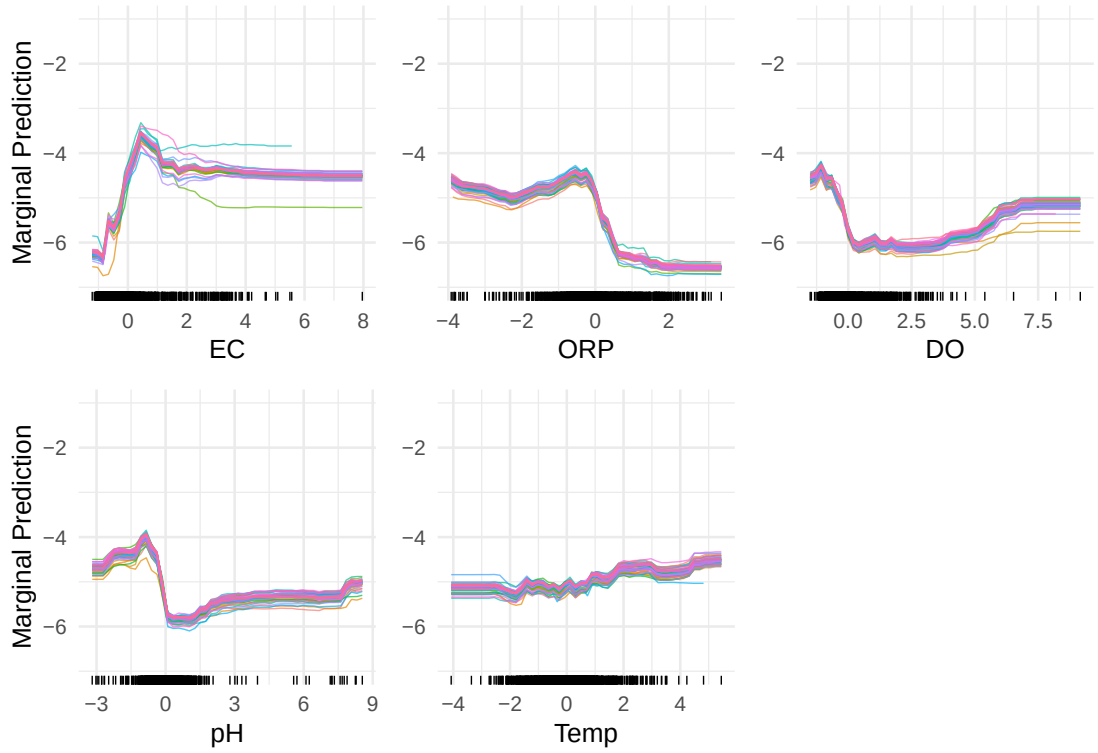

Figure 4.6: Partial dependence plots with observed data marked below. Each colour represents a different LOWO model. Benzene, imputed by DL/2.

The predictor DO also mirrors the non-linear trend but one must exercise caution with the higher values as extrapolation into outlier values are not as reliable; the rug plot below each PDP showing a line for each observed value should assist with this.

4.5.3 Final Models

In all random forest models fitted so far, we have not used any data from the three holdout wells as described in Section 2.2. This is intentional as we want to get out-of-sample predictions based on completely new data, which we can only obtain if the hyperparameter tuning was completed without those out-of-sample data as we have done in Section 4.3. Recall that we sample two predictors at each split in

each tree when modelling MTBE, so $p_f = 2$ and only a single predictor is sampled at each split for benzene, so $p_f = 1$.

For each of the predictions, calculating uncertainty is not trivial but can be done using quantile regression forests (Meinshausen & Ridgeway, 2006) or the empirical distribution of out-of-bag errors (Zhang *et al.*, 2019). We have not explored such methods with a clear opportunity for further work. A naïve approach would involve recognising that RFs are ensemble learning algorithms and so each prediction shown is the average of several predictions, one from each tree. Using all trees to elicit some predictive uncertainty may produce sound results in some cases but, by construction, these trees were formed with heterogeneity in mind and are unlikely to yield rigorous prediction intervals.

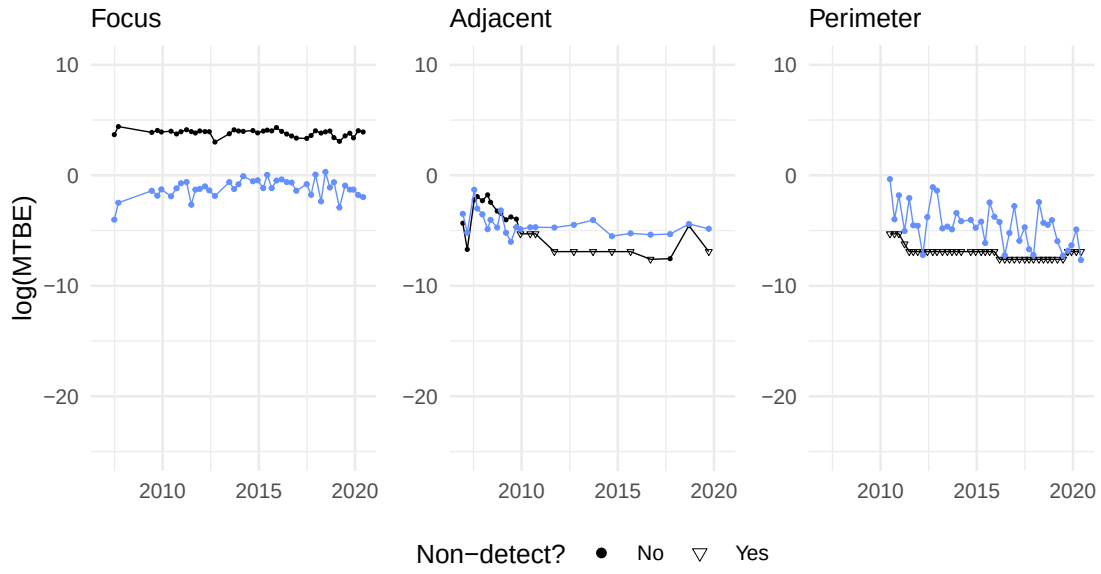

Figure 4.7: Random forest predictions with comparison to truth in black. MTBE imputed by DL/2, LMWO.

Figure 4.7 shows how the heterogeneity of groundwater monitoring wells lead to unsatisfactory predictions for all three holdout wells. Recall, the 'Focus' well is an edge case where all analyte concentrations are the second highest on average, and

so most models including RFs will underpredict the truth for this well. However, all 'Focus' well predictions exceed almost all predictions and true values from the other two wells implying this model may be able to detect a problem without giving a sense of severity; further investigation is required.

We observe that predictions for all three holdout wells of similar value and that value appears to be close to the average MTBE concentration (after imputation). Such predictions are expected to be similar to the null model where no predictors are supplied and the linear predictor is replaced by a common mean parameter. By comparing to prediction of the censored regression model to be introduced in Section 5.3, we see very similar predictions with a less-opaque model. All points that are penitent to MTBE also apply to the results when fitting a RF to benzene; Figure 4.8 is included for completeness.

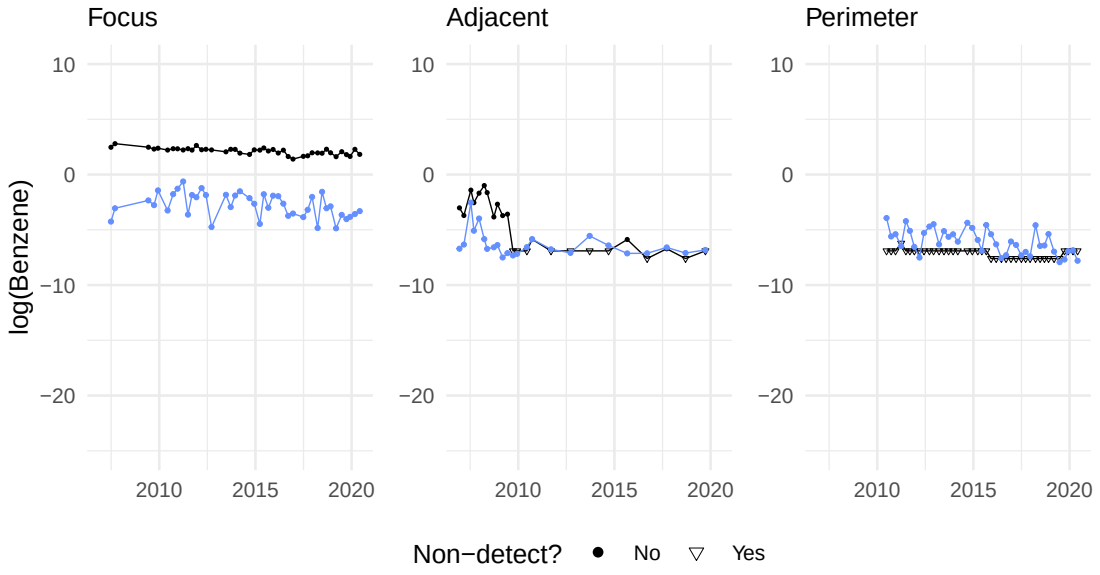

Figure 4.8: Random forest predictions with comparison to truth in black. Benzene imputed by DL/2, LMWO.

4.6 Conclusion

The attraction of a random forest model is to find underlying patterns that may be undetectable to human perception or classical models and to explain most of the variation where possible. A model suspected of overfitting would still afford us insight into how each analyte at this specific groundwater site, site A, are correlated with our chosen predictors. Instead, we have not been able to explain over half of the variation as shown in Figure 4.3 where R^2 values never exceed 0.5 for either analyte. We argue this is could be due to poor signal-to-noise ratio between the analyte concentrations and predictors or a limitation of random forest models not allowing left-censored data and therefore requiring biased imputation methods such as DL/2.

As an illustrative example, consider how temperature is correlated with MTBE concentrations. While it is expected that any chemical reaction increasing or decreasing the analyte concentration will affect the temperature of the groundwater network, it still remains unclear if this would be detectable over the natural seasonality of the temperature reading caused by regional changes in weather and temperature. Similar concerns around other predictors persist but these models have suggested that we would see the most information from conductivity (EC) and ORP through the importance measures produced.

4.6.1 Further Work

By understanding these models, we could hypothesise a RF model where left-censored data is used. This would require some measure of location to aggregate observations on the terminating nodes and a measure of spread to be assigned to node impurity and the cost function to be minimised by each split. Potentially we

could adopt ROS as described in Section 1.5.2 and Helsel & Cohn (1988), however, preliminary models showed little benefit over DL/2 for random forest models. A more rigorous investigation with a variety of datasets would be enlightening and could allow the RF model to be applied to a broader range of applications.

Random forests are still an instrumental tool for any analysis and the PDP output have called into question our assumption of linearity as we continue with regression models in Chapter 5 and Chapter 6.

Chapter 5

Regression Models

5.1 Introduction

The approach to modelling throughout this project is to start with a linear regression model within a Bayesian context and address any incorrect assumptions or shortcomings of this model with various extensions to be introduced in subsequent chapters. Advantages of these regression models are numerous:

- relative computational simplicity allows us to fit these models to our site A data in minutes and easily scale to bigger sites with more data;
- interpretability of regression coefficients empower us to quantify change in analyte log concentration per unit change in standardised predictors;
- given only predictor values, we can predict analyte concentrations.

These benefits are aligned with our motivations of better understanding the general relationship between analyte concentrations and predictors. This is in contrast to time series models which may produce more accurate localised predictions, but may not generalise to other groundwater monitoring sites.

As discussed in Section 1.5, the approaches to left-censored data are plentiful and each model discussed makes it clear which approach is being used in their respective chapters.

5.2 Univariate Linear Regression

Our initial approach is a multiple linear regression,

$$y_{ij} = \mathbf{x}_i^T \boldsymbol{\beta}_j + \epsilon_{ij}, \tag{5.1}$$

for each analyte, indexed by $j = 1, \dots, n_y$, where $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \tau_j^{-1})$ for $i = 1, \dots, n_s$. Hence, the parameters of this model are the regression coefficients $\boldsymbol{\beta}_j = (\beta_{0j}, \dots, \beta_{n_x j})^T$ and precision parameters τ_j.

Advantageous to these models is the interpretability of the parameters, where β_{kj} is the impact of the k^{th} predictor on the concentration of the j^{th} analyte. Similarly, we have a precision parameter, τ_j, per analyte where higher values represent a smaller degree of measurement error in the recording of the concentrations of the j^{th} analyte. To further understand the role of τ_j, suppose there exists a true log concentration y_{ij}^* and observed log concentration $y_{ij} = y_{ij}^* + \epsilon_{ij}$, not subject to censoring, where $\epsilon_{ij} \in \mathbb{R}$ denotes the additive measurement error on the log scale. On the original scale,

$$\exp(y_{ij}) = \tilde{\epsilon}_{ij} \cdot \exp(y_{ij}^*),$$

where we define $\tilde{\epsilon}_{ij} \coloneqq \exp(\epsilon_{ij}) > 0$ to be the multiplicative measurement error. In reality, measurement error is indeterminable and we have used a random variable, $\epsilon_{ij} \sim N(0, \tau_j^{-1})$, to describe this uncertainty. It follows that $\tilde{\epsilon}_{ij} \sim LN(0, \tau_j^{-1})$ and we know from the properties of this log normal distribution that $(1-\alpha)\%$ of plausible

values lie in the, not necessarily symmetric, range

$$0 \pm \exp\left(\tau_j^{-\frac{1}{2}}\, \Phi^{-1}\left(\frac{\alpha}{2}\right)\right),$$

where Φ^{-1} is the inverse cumulative distribution of the standard normal distribution $N(0,1)$. For small values of τ_j, say 0.2, a 95% highest density interval (HDI) would be $(0.01, 80.05)$ implying observations may be wrong by several orders of magnitude. For larger values of τ_j, say 2, a 95% HDI would be $(0.25, 4.00)$ implying an observed concentration of 1 mg/L could be up to 4 times as large in truth.

To model each analyte independently, we express model (5.1) as a collection of multivariate normals,

$$\mathbf{y}_j | X\boldsymbol{\beta}_j, \tau_j \sim N_{n_s}(X\boldsymbol{\beta}_j, \tau_j^{-1} I_{n_s}), \tag{5.2}$$

for all $j = 1, \ldots, n_y$ where X is the n_s by $n_x + 1$ design matrix, intercept column included, introduced in Section 1.4.

Using the compact (5.2), we can express the likelihood of the model parameters, $\Theta = \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_j, \tau_1, \ldots, \tau_j$, as

$$L(\Theta) = \prod_{j=1}^{n_y} \pi(\mathbf{y}_j | X, \boldsymbol{\beta}_j, \tau_j),$$

where the likelihood contribution of the j^{th} analyte is, up to proportionality,

$$\pi(\mathbf{y}_j | X, \boldsymbol{\beta}_j, \tau_j) \propto \tau^{n_s/2} \exp\left\{-\frac{\tau_j}{2}(\mathbf{y}_j - X\boldsymbol{\beta}_j)^T(\mathbf{y}_j - X\boldsymbol{\beta}_j)\right\}.$$

5.2.1 Prior

Before defining our prior choice, note that a conjugate prior exists for this model. By this we mean for any choice of prior $\pi(\boldsymbol{\theta})$ in some family of distributions $\mathscr{F}$, the posterior must also be inside this family, $\pi(\boldsymbol{\theta}|\mathbf{y}_j, X) \in \mathscr{F}$. For (5.2), the conjugate prior takes the form of a multivariate normal gamma distribution where $\boldsymbol{\beta}_j|\tau_j$ is conditionally multivariate normally distributed and the precision τ_j follows a gamma distribution or alternatively a Chi-squared distribution (Gelman *et al.*, 1995).

While this conjugacy is an appealing property, we have no reason to believe the impact of each predictor is conditional on the variation of the analyte and our prior assumptions should reflect this. As such, we opt for a conditionally conjugate (or semi-conjugate) prior distribution where the prior and posterior for each parameter arise from the same family, conditional on a further random variable. While conditionally conjugacy is a weaker assumption than full conjugacy, an advantage of this property is that we are guaranteed standard full conditional distributions (FCD) for use in a Gibbs sampling algorithm; see Appendix B for a primer on Markov chain Monte Carlo (MCMC) methods. Hence, we assume *a priori*,

$$\begin{aligned} \boldsymbol{\beta}_j &\sim N_{n_x+1}(\mathbf{m}_\beta, V_\beta), \\ \tau_j &\sim Ga(a_\tau, b_\tau), \end{aligned} \tag{5.3}$$

where $\mathbf{m}_\beta$, V_β, a_τ and b_τ are hyperparameters to be chosen and we have assume independence between these distributions and hence $\pi(\boldsymbol{\beta}_j, \tau_j) = \pi(\boldsymbol{\beta}_j)\pi(\tau_j)$.

5.2.2 Bayesian Inference

Since we described a semi-conjugate model and do not have full conjugacy, we implement a specific type of Markov chain Monte Carlo (MCMC) algorithm known as the Gibbs sampling algorithm. For conditional conjugacy it must be the case

that the conditional posterior for $\boldsymbol{\beta}_j$ is multivariate normal, moreover

$$\begin{aligned} \boldsymbol{\beta}_j|\mathbf{y}_j, X, \tau_j &\sim N_{n_x+1}(\mathbf{m}^*_\beta, V^*_\beta), \\ \mathbf{m}^*_\beta &= \left(V^{-1}_\beta + \tau_j X^T X\right)^{-1} (V^{-1}_\beta \mathbf{m}_\beta + \tau_j X^T \mathbf{y}_j), \\ V^*_\beta &= \left(V^{-1}_\beta + \tau_j X^T X\right)^{-1}. \end{aligned} \tag{5.4}$$

To see why this is the case, we require an alternate form of a multivariate normal probability distribution function.

Proposition 5.1. *If $\boldsymbol{\theta} \in \mathbb{R}^p$ has a density, given up to proportionality,*

$$\pi(\boldsymbol{\theta}) \propto \exp\left\{\boldsymbol{\theta}^T\mathbf{b} - \frac{1}{2}\boldsymbol{\theta}^T A\boldsymbol{\theta}\right\},$$

then $\boldsymbol{\theta} \sim N_p(A^{-1}\mathbf{b}, A^{-1})$ given some invertible matrix A and column vector b.

Proof. Assume $\boldsymbol{\theta} \sim N_p(\boldsymbol{\mu}, \Sigma)$ for some mean matrix $\boldsymbol{\mu}$ and covariance matrix Σ to be determined. Then,

$$\begin{aligned} \pi(\boldsymbol{\theta}) &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right\}, \\ &\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right)\right\}, \\ &\propto \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^T(\Sigma^{-1}\boldsymbol{\mu})\right\}, \\ &\propto \exp\left\{\boldsymbol{\theta}^T\mathbf{b} - \frac{1}{2}\boldsymbol{\theta}^T A\boldsymbol{\theta}\right\}, \end{aligned}$$

where $A = \Sigma^{-1}$ and $\mathbf{b} = \Sigma^{-1}\boldsymbol{\mu}$. It is then clear that $\boldsymbol{\mu} = A^{-1}\mathbf{b}$ and $\Sigma = A^{-1}$ and the distribution must be multivariate normal.

□

Rather than prove (5.4) directly, we can prove the more general case in Proposition 5.2 and then set $\Sigma = \tau_j^{-1} I_{n_s}$.

Proposition 5.2. *Suppose* $\mathbf{y}|X, \boldsymbol{\beta}, \Sigma \sim N_n(X\boldsymbol{\beta}, \Sigma)$ *takes the form of a univariate linear regression likelihood and* $\boldsymbol{\beta} \sim N_p(\mathbf{m}, V)$ *takes the form of a general normal prior parametrised by mean vector* $\mathbf{m}$ *and covariance matrix* V*. It then follows that*

$$\begin{aligned}\boldsymbol{\beta}|\mathbf{y}, X, \Sigma &\sim N(\mathbf{m}^*, V^*)\\ m^* &= (V^{-1} + X^T\Sigma^{-1}X)^{-1}(V^{-1}\mathbf{m} + X^T\Sigma^{-1}\mathbf{y})\\ V^* &= (V^{-1} + X^T\Sigma^{-1}X)^{-1}.\end{aligned}$$

Proof. By Bayes' theorem,

$$\begin{aligned}\pi(\boldsymbol{\beta}|\mathbf{y}, X, \Sigma) &= \frac{\pi(\mathbf{y}|X, \boldsymbol{\beta}, \Sigma)\pi(\boldsymbol{\beta})}{\pi(\mathbf{y}|X, \Sigma)}\\ &\propto \pi(\mathbf{y}|X, \boldsymbol{\beta}, \Sigma)\pi(\boldsymbol{\beta})\\ &\propto \exp\left\{-\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta})^T\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta})\right\}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{m})^T V_\beta^{-1}(\boldsymbol{\beta} - \mathbf{m})\right\}\\ &\propto \exp\left\{\boldsymbol{\beta}^T X^T\Sigma^{-1}\mathbf{y} - \frac{1}{2}\boldsymbol{\beta}^T X^T\Sigma^{-1}X\boldsymbol{\beta} - \frac{1}{2}\boldsymbol{\beta}^T V^{-1}\boldsymbol{\beta} + \boldsymbol{\beta}^T V^{-1}\mathbf{m}\right\}\\ &\propto \exp\left\{\boldsymbol{\beta}^T\left(V^{-1}\mathbf{m} + X^T\Sigma^{-1}\mathbf{y}\right) - \frac{1}{2}\boldsymbol{\beta}^T\left(V^{-1} + X^T\Sigma^{-1}X\right)\boldsymbol{\beta}\right\}.\end{aligned}$$

and then using Proposition 5.1 we recognise the normal density with the mean and covariance parameters stated above.

□

Similarly, we can derive the full conditional distribution for the measurement precision parameter,

$$\begin{aligned}\tau_j|\mathbf{y}_j, X, \boldsymbol{\beta}_j &\sim Ga(a_\tau^*, b_\tau^*),\\ a_\tau^* &= a_\tau + \frac{n_s}{2},\\ b_\tau^* &= b_\tau + \frac{1}{2}(\mathbf{y}_j - X\boldsymbol{\beta}_j)^T(\mathbf{y}_j - X\boldsymbol{\beta}_j),\end{aligned} \tag{5.5}$$

as proven in Proposition 5.3.

Proposition 5.3. *If*

$$\pi(\tau|\mathbf{x}, \boldsymbol{\mu}) \propto \pi(\mathbf{x}|\boldsymbol{\mu}, \tau)\, \pi(\tau),$$

where $\mathbf{x}|\boldsymbol{\mu}, \tau \sim N_n(\boldsymbol{\mu}, \tau^{-1} I_n)$ *is a normally distributed n-length vector and* $\tau \sim Ga(a, b)$*, then it must be that*

$$\tau|\mathbf{x}, \boldsymbol{\mu} \sim Ga\left(a + \frac{n}{2}, b + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T(\mathbf{x} - \boldsymbol{\mu})\right).$$

Proof. By direct calculation, we observe

$$\begin{aligned}
\pi(\tau|\mathbf{x}, \boldsymbol{\mu}) &\propto \pi(\mathbf{x}|\boldsymbol{\mu}, \tau)\, \pi(\tau) \\
&\propto \tau^{\frac{n}{2}} \exp\left\{-\frac{\tau}{2}(\mathbf{x} - \boldsymbol{\mu})^T(\mathbf{x} - \boldsymbol{\mu})\right\} \tau^{a-1} \exp(-\tau b) \\
&\propto \tau^{a+\frac{n}{2}-1} \exp\left\{-\tau\left[b + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T(\mathbf{x} - \boldsymbol{\mu})\right]\right\},
\end{aligned}$$

is a Gamma density and hence $\tau|\mathbf{x}, \boldsymbol{\mu}$ follows the Gamma distribution given.

□

Once these distributions have been derived we can implement the Gibbs sampling algorithm in Algorithm 1; executing the algorithm several times with different initial values for $\tau_j^{(0)}$ produces several "chains". As described in Appendix B, M draws are generated but we must discard some amount of draws as "burn-in" to yield only $M^* < M$ posterior samples. Details of each MCMC algorithm including burn-in and thinning are presented alongside results.

Algorithm 1 Univariate linear regression (conditionally conjugate)

1: Initialise $\tau_j^{(0)}$
2: **for** $s = 1, \ldots, M$ **do**
3: compute $\mathbf{m}_\beta^*$ and V_β^* using (5.4) where $\tau_j = \tau_j^{(s-1)}$
4: sample $\boldsymbol{\beta}_j^{(s)} \sim N_{n_x}(\mathbf{m}_\beta^*, V_\beta^*)$
5: compute a_τ^* and b_τ^* using (5.5) where $\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^{(s)}$
6: sample $\tau_j^{(s)} | \mathbf{y}_j, X, \boldsymbol{\beta}_j \sim Ga(a_\tau^*, b_\tau^*)$
7: **end for**

5.3 Censored Linear Regression

All regression techniques covered so far assume uncensored data and, as such, we have to choose some imputation method as described in Section 1.5. None of these methods adequately express the data accurately since each imputation replaces observations that are known only up to some range with a set value such as DL/2.

Throughout this project our aims involve modelling the censored data, with their intrinsic uncertainty, within a Bayesian paradigm where censored observations are often treated as unknown quantities. We will achieve this either by specifying the likelihood directly or using data augmentation as introduced in Section 1.5.4.

Consider a censored regression model as a generalisation of the Tobit model (Tobin, 1958). In the original Tobit model, it is assumed that all censored observations are censored at a common detection limit and further that limit must be zero. While this was sufficient for the application of economic surveys of households where some expenditures were left-censored at zero, hydrocarbon groundwater monitoring data contains non-zero detection limits that vary by observation or analyte and are determined by availability of facilities and equipment.

Instead, we define the likelihood as in (1.2) such that each contribution from an uncensored observation uses the probability density function (PDF) and each contribution from a censored observation uses the cumulative distribution function (CDF). In this case, the likelihood contribution for each modelled response variable is

$$\pi(\mathbf{y}_j|\boldsymbol{\delta}_j, X, \boldsymbol{\beta}_j, \tau_j) = \prod_{i=1}^{n_s} f(y_{ij}|\mathbf{x}_i^T, \boldsymbol{\beta}_j, \tau_j)^{1-\delta_{ij}} F(y_{ij}|\mathbf{x}_i^T, \boldsymbol{\beta}_j, \tau_j)^{\delta_{ij}}, \tag{5.6}$$

where $\mathbf{y}_j = (y_{1j}, \ldots, y_{1n_s})^T$ represents the observed analyte concentration and $\boldsymbol{\delta}_j = (\delta_{1j}, \ldots, \delta_{1n_s})^T$ is a vector of censoring indicators introduced in Section 1.5 where δ_{ij} equals 1 when y_{ij} is censored and 0 when y_{ij} is uncensored. As we assume normality in these data, we use the normal PDF and CDF for $f(\cdot)$ and $F(\cdot)$ respectively, both parameterised by mean $\mathbf{x}_i^T\boldsymbol{\beta}_j$ and precision τ_j.

5.3.1 Bayesian Inference

Our prior information for this model is identical to the priors stated in Section 5.2

$$\boldsymbol{\beta}_j \sim N_{n_x+1}(\mathbf{m}_\beta, V_\beta),$$
$$\tau_j \sim Ga(a_\tau, b_\tau),$$

for $j = 1, \ldots, n_y$. Due to the change to the likelihood to explicitly include censoring, we can no longer assume conjugacy or even conditional conjugacy. As such, we resort to more general Markov chain Monte Carlo (MCMC) algorithms to sample from the posterior distribution for this model.

We use JAGS (Plummer *et al.*, 2003), a program for Bayesian Graphical modelling that allows the user to fit advanced models in a declarative programming language similar in syntax to R. One of the main advantages of JAGS is the ability to declare a model, data and prior information and leave the specifics of the algorithm to the software. For example, JAGS will recognise the conditional conjugacy of the models

described in this chapter and use a Gibbs sampler as we have explained. For the censored regression model, where conjugacy is not present, JAGS will use a slice sampling algorithm (Neal, 2003).

When the data is both censored and multivariate, as is the case in the hydrocarbon groundwater monitoring application we are interested in, expressing the likelihood in a closed form is non-trivial. For example, benzene and MTBE have different censoring indicator values for approximately 30% of the observations within Case Study A. Modelling multiple correlated dependent variables that are potentially censored is not considered in this thesis in great detail but has been applied to traffic accident data by Anastasopoulos *et al.* (2012) and can also be improved through the use of varying effects within that particular application (Zeng *et al.*, 2017). One would expect parameter estimates of the multivariate model to be similar to the respective parameters of each univariate model but jointly modelling analytes could improve overall prediction due to a high correlation.

5.4 Simulation Study

Of the two models proposed in this chapter, the substitution of censored values by half their detection limit takes substantially fewer computational resources but may introduce bias (Helsel & Cohn, 1988). Therefore, we will enact a simulation study to compare the relative strengths and weaknesses of the models,

1. Bayesian univariate multiple linear regression with imputed response data, as described in Section 5.2;
2. Bayesian univariate multiple censored regression, as described in Section 5.3.

Simulated data is used throughout this thesis to verify the validity of models, mitigate any errors during the model development stage and improve overall confidence

in each respective model. For the sake of brevity, only a single demonstrative simulation study is shown.

We simulate $n_s = 100$ data points where $x_{ik} \overset{\text{iid}}{\sim} N(0,1)$, for $i = 1, \ldots, 100$ and $k = 1, 2, 3$ denote our explanatory data; normalisation is applied such that each explanatory variable has zero mean and unit variance. A single response variable is simulated according to (5.2) where the regression coefficients, $\boldsymbol{\beta}_1$, are arbitrarily set. In this case,

$$\boldsymbol{\beta}_1 = (3, 4, 1, 2)^T.$$

The precision parameter, τ_1, is chosen to be 0.1 based on preliminary investigations into our groundwater application data, shown in Chapter 2. Once the response variable is simulated, we post-process the output by artificially censoring; all values lower than the 60^{th} percentile of the data are replaced with non-detect observations with a detection limit equal to this calculated percentile. Note that this mechanism yields singly censored data unlike the multiply censored data that arises from groundwater monitoring networks. The univariate multiple linear regression with imputed response data does not allow for these censored data and replaces all censored observations with half the detection limit.

For both models, assume vague prior information using (5.3) where

$$\begin{aligned} \boldsymbol{\beta}_1 &\sim N_{n_x+1}(\mathbf{0}_{n_x+1}, 0.01^{-1} I_{n_x+1}), \\ \tau_1 &\sim Ga(1, 1). \end{aligned}$$

We obtain $10,000$ posterior samples for each model by running the respective MCMC algorithm for $70,000$ samples then discarding the first $20,000$ as burn-in and then "thinning" by 5 (discarding all but every 5^{th} sample).

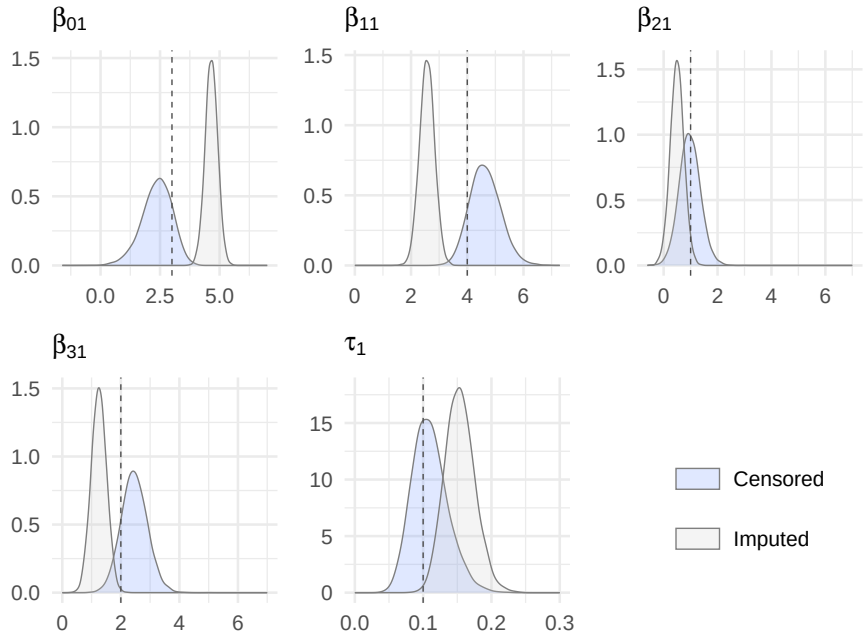

Figure 5.1: Marginal posterior densities of $\boldsymbol{\beta}_1 = (\beta_{01}, \ldots, \beta_{31})$ and τ_1, true values marked by dashed lines.

Figure 5.1 shows overlapping posterior distributions of the precision parameter for the models but the location of the true value, shown as a dashed line, is within the posterior distribution produced from the censored regression and even outside the 95% interval of the imputed model's posterior samples for the quantity. The models differ in regression parameters, shown in Figure 5.1, where the posterior uncertainty for the censored regression appears larger while also showing a larger density around the 'true' value of each parameter. It is worrying that the imputation methods are not only incorrect in the inference of the regression parameters but also confident enough to show a smaller credible interval around said incorrect value.

5.5 Groundwater Application

Here, we only apply the censored regression model (5.6) and not the imputed linear model, (5.2), because of the simulation study results, which show imputed models struggle even in more ideal circumstances. The complexity of groundwater systems with varying geological landscapes between sites and a multitude of possible reactions to be modelled means we must have more confidence in our models and their uncertainty.

For Site A, introduced in Chapter 2, we focus on two main analytes:

- benzene, the quintessential aromatic hydrocarbon that is a confirmed carcinogenic and of most concern to regulators (World Health Organization, 2010);
- methyl tert-Butyl Ether (MTBE), the least censored (28.5%) analyte.

These data must be pre-processed to be used in a regression model including steps to match analyte and predictors by closest date and summarising repeated analyte measurements with a single observation, as discussed in Chapter 2.

Applications such as hydrocarbon groundwater monitoring have many personnel that are responsible for data at a site or sometimes a group of sites. As such, these experts can offer valuable insight within a Bayesian paradigm through the use of prior elicitation such as the web-based tool MATCH (Morris *et al.*, 2014) or an expert aggregation tool such as SHELF (Williams *et al.*, 2021). However, a drawback of these methods is the time requirement from each expert which was unfortunately not secured for this project. Therefore, we implement a vague prior approach, for all analytes $j = 1, \dots, n_s$,

$$\begin{aligned} \boldsymbol{\beta}_j &\sim N_{n_x+1}(\mathbf{0}_{n_x+1}, 0.1^{-1} I_{n_x+1}), \\ \tau_j &\sim Ga(2, 0.1). \end{aligned} \tag{5.7}$$

A symmetrical distribution around 0 is used for our regression coefficients where the diagonal of the covariance matrix quantifies our prior uncertainty. We have found that our posterior draws are fairly insensitive to the choice of prior precision with vague and precise priors leading to very similar results.

Since the gamma distribution has a strictly positive support and is conjugate in the non-censored case, it is a satisfactory distribution to describe our prior beliefs for τ_j. Rather than choosing shape and rate parameters, it is more intuitive to describe the distribution in terms of mean and variance. Recall from Section 5.2 that each τ_j describes the feasible amount of multiplicative measurement error on the original scale; setting $\tau_j = 20$ corresponds to 95% HDI, (0.65, 1.55), of viable measurement errors. This interval seems reasonable given our knowledge of measurement uncertainty and the data collection process in groundwater monitoring and so we take 20 to be the prior mean and choose the final free parameter, variance, as a measure of our prior uncertainty. As with regression coefficients, changing the prior to say $\tau_j \sim Ga(20, 1)$ has very little impact on the results to be shown and highlights a certain degree of prior insensitivity.

We fit the censored regression model to our analytes of interest within site A, benzene and MTBE, in two different scenarios as discussed in Section 1.7. These splits of the data differ by the training data used:

- **LMWO** uses all data except the holdout wells;
- **hold-out future** uses all data except samples taken after January 1^{st}, 2015 from the holdout wells.

5.5.1 Regression Parameters

A key advantage of this model is the clear interpretability of linear regression coefficients. We consider changes in our predictors to be based on the sample standard deviation change and not the unit change as we have normalised the predictors; our inferences are independent of the choice of units for each predictor, decided in the data collection stage. For example, $\beta_{\mathrm{DO\,MTBE}}$, represents the expected change in log concentration of MTBE per increase in DO by one sample standard deviation, and similarly for other analyte and predictor combinations. The intercept terms are sensitive to the units used as they represent the expected concentration when all predictors take their mean value, which is 0 after normalisation. All analytes are reported in the same units within the data.

Figure 5.2 shows a substantial prior to posterior update and regression coefficients for each analyte tend to have identical sign, as expected given the high correlation. As expected, the intercept for MTBE is greater than the benzene intercept as the MTBE concentrations are typically higher within this specific groundwater monitoring site. Discussions with those involved in generating the RTM data revealed a preliminary expectation that conductivity (EC) would have a positive correlation with any analyte which is supported by this model that estimates a significant positive coefficient for the predictor.

Interestingly, the impact of temperature on MTBE as shown in Figure 5.2 has a posterior density with a mode at 0, implying no significant effect, but with a much higher peak than the prior. This is not quite the case with benzene, but any effect from temperature on benzene is quite small relative to the intercept and other coefficients. From this we can infer that temperature has very little utility in describing the variation of MTBE or benzene as suggested from the random forest results in Section 4.4.2.

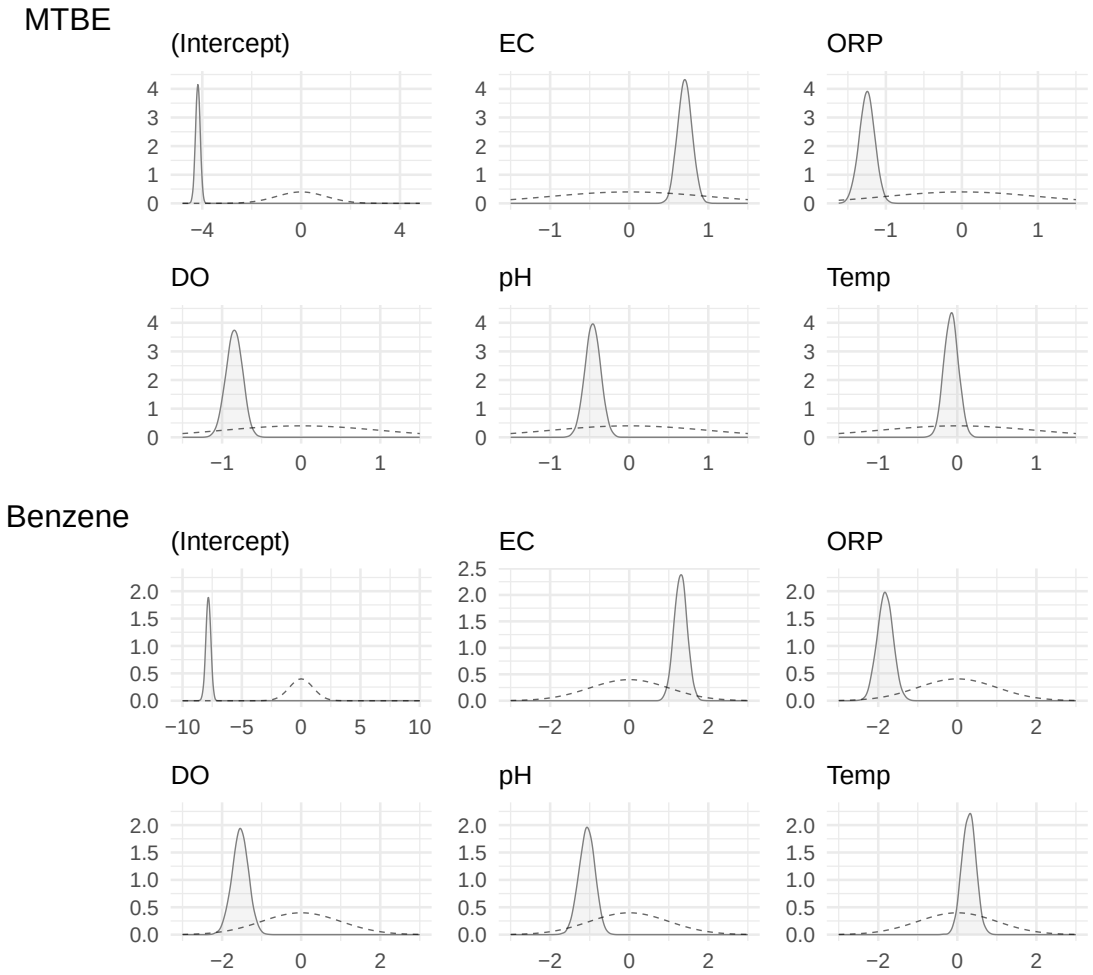

Figure 5.2: Marginal posterior density of regression coefficients, prior distribution shown by dashed line. LMWO, censored regression.

We also consider the joint posterior distributions of the regression coefficients by presenting a series of scatter plots. Individual plots below the main diagonal of Figure 5.3 visualise the bivariate posteriors of $\boldsymbol{\beta}_1$ (MTBE) and the remaining plots visualise bivariate posteriors of $\boldsymbol{\beta}_2$ (benzene). Many of the plots show a reasonably circular pattern which implies little to no correlation between these posterior distributions. Some relationships may be weakly correlated such as $\beta_{\text{ORP MTBE}}$ and $\beta_{\text{DO MTBE}}$ which displays a very slight downward trend.

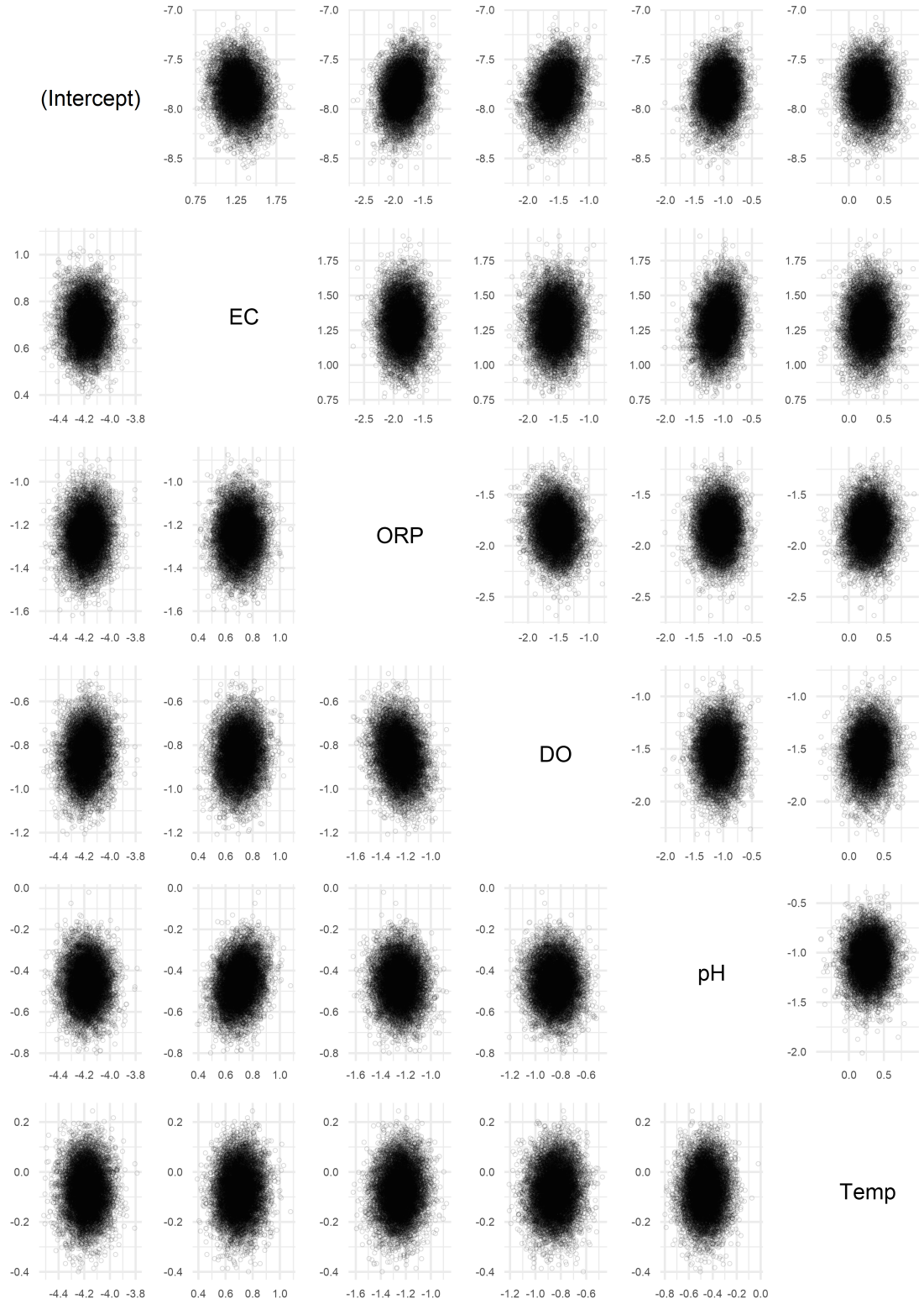

Figure 5.3: Joint posterior plots of regression coefficients; MTBE (benzene) shown in lower (upper) triangular region. LMWO, censored regression.

5.5.2 Precision Parameters

Figure 5.4 shows the trace-plot for the precision parameter, after thinning and removing burn-in, with a similar appearance to a "hairy caterpillar". Trace-plots for several chains and other parameters such as the regression coefficients and observed data log likelihood (based on in-sample data) have a similar appearance but are not shown. This visual quality gives us no evidence that the MCMC algorithm has not converged.

By inspection, the measurement precision parameter for the MTBE model fit is larger than the benzene counter-part. In both cases, the posterior mean for the precision is very low. For instance if we set $\tau_j = 0.1$ then the additive measurement error on the log scale, ϵ_{ij}, follows a normal distribution where approximately 95% of the draws are expected to lie between 2 standard deviations, $2\sigma_j = 6.32$ which is more than half the observed range of MTBE. The posterior distributions for τ_1 and τ_2 show the main issue with these data: the signal-to-noise ratio may be too low for us to meaningfully infer any association.

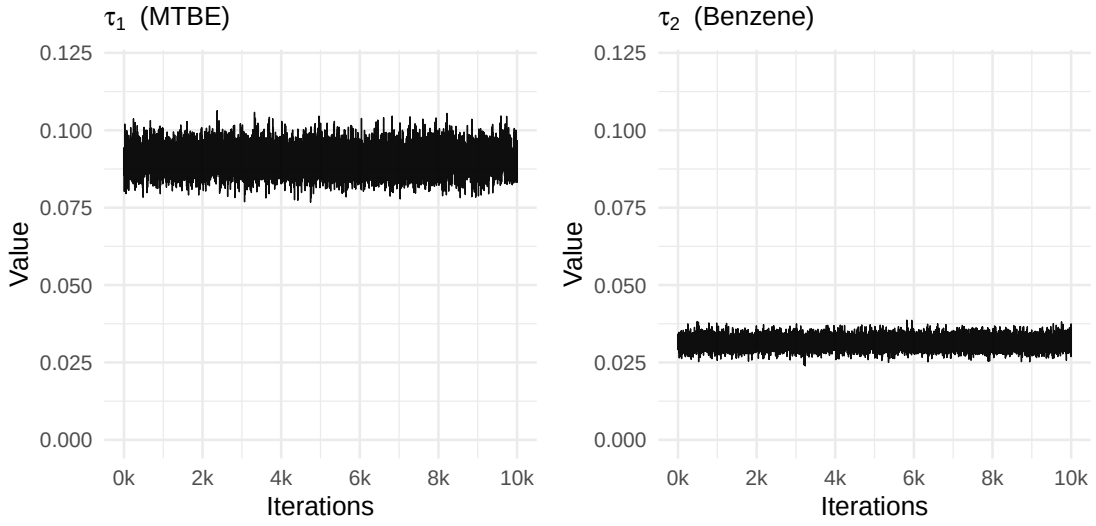

Figure 5.4: Marginal posterior trace-plots of $\boldsymbol{\tau}$. LMWO, censored regression.

5.5.3 R Squared

The coefficient of determination (R squared) of a model is an effective measure of how much variation in the analytes can be explained by our predictors, as utilised in Chapter 4. Within a Bayesian context, with a posterior sample of the unknown quantities $\boldsymbol{\theta}^{(s)}$, $\sigma^{2,(s)}$ for $s = 1, \ldots, S$, we can obtain a posterior sample of the coefficient of determination by applying the conventional definition,

$$R^{2\,(s)}_{\text{classic}} = \frac{\text{Var}(E[\mathbf{y}|X, \theta^{(s)}])}{\text{Var}(\mathbf{y})}, \tag{5.8}$$

for each posterior sample $s = 1, \ldots, S$. However, this definition is not ideal since the value can exceed 1 as noted in Gelman *et al.* (2019). Instead we opt to use the definition suggested by Gelman *et al.* (2019),

$$R^2_s = \frac{\text{Var}(E[\mathbf{y}|X, \theta^{(s)}])}{\text{Var}(E[\mathbf{y}|X, \theta^{(s)}]) + \text{Var}(\sigma^{2\,(s)})}, \tag{5.9}$$

where $\text{Var}(E[\mathbf{y}|X, \theta^{(s)}])$ and $\text{Var}(\sigma^{2\,(s)})$ represent the variance of the fitted values and variance of the residual variance respectively for each posterior sample $s = 1, \ldots, S$.

Figure 5.5 contains the densities of each R^2 distribution for both analytes and details the posterior mean. Both models observe a proportion of variance explained less than 35% which highlights the inability of this model to explain the data and suggests that there may not be enough of a signal in the chosen predictors.

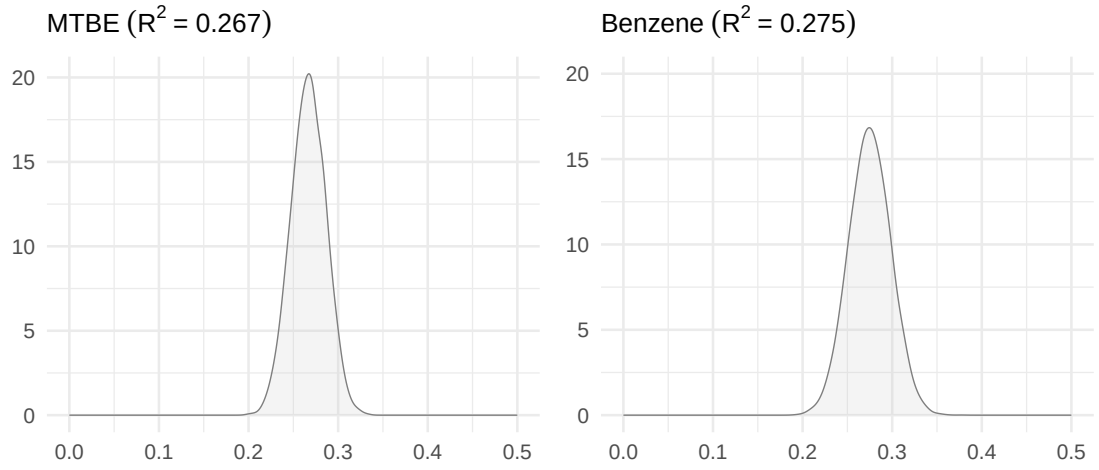

Figure 5.5: Bayesian R^2 posterior densities, posterior mean shown in title. LMWO, censored regression.

While Bayesian R^2 is a useful measure in a linear regression, it is not without limitations. Both R^2 score higher with overfitted models making them unsuitable for model comparison; model comparison in this thesis is based on metrics described in Section 1.7. Equation (5.9) assumes that the residual variance is consistent across observations, but this may not be the case such as in the mixture of experts model, to be introduced in Chapter 7. Moreover, the denominator in (5.9) is a function of the fitted values and residual variance and, in a Bayesian context, this is a function of both data and model; the denominator of (5.8) is purely data-based. Thus, the original interpretation as proportion of explained variation is not applicable here and it is better presented as a data-based estimate of this proportion (Gelman *et al.*, 2019) as is done here.

5.5.4 Prediction

Figures 5.6 and 5.7 show posterior predictive summaries alongside true analyte log concentrations in black for MTBE and benzene respectively. The posterior predictive mean is shown as a blue point, whereas pointwise 95% prediction intervals are denoted by the light blue shading. For the censored regression model, our point predictions are similar to the predictions made by the random forest model, but now we can quantify our uncertainty in these predictions. For the holdout well "Focus", where analyte measurements are higher than average, we see predictions are poor for both analytes where most 'true' observations are close to the upper boundary of the 95% prediction interval. At this well, both analyte predictions around 2012 are substantially lower than the other predictions of the same analyte; this reveals an issue in this model where any outlier or valid extreme value in a predictor can vastly affect predictions.

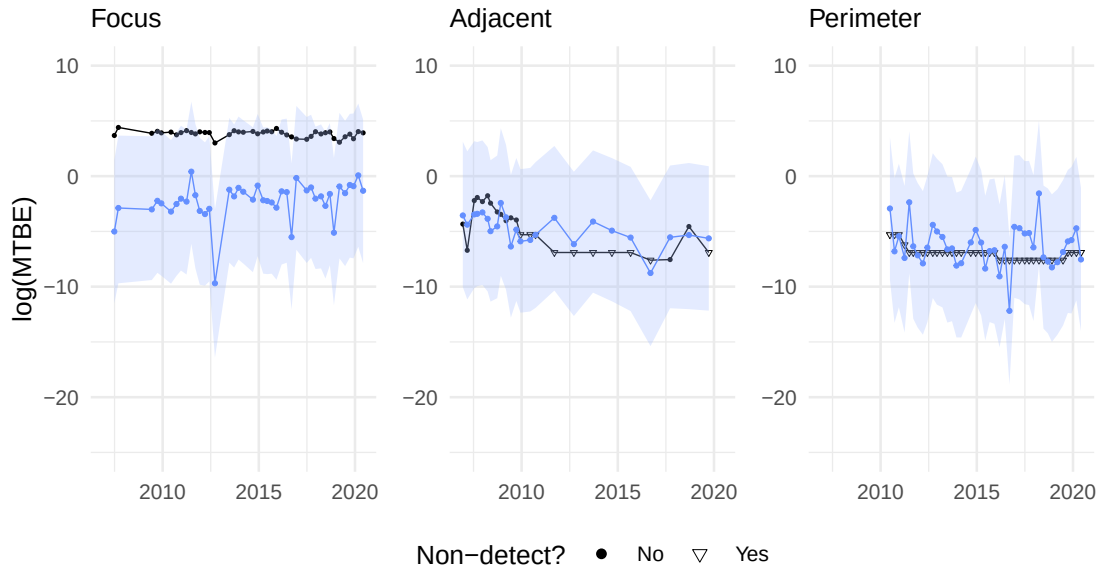

Figure 5.6: Predictions with comparison to truth in black. MTBE, LMWO, censored regression.

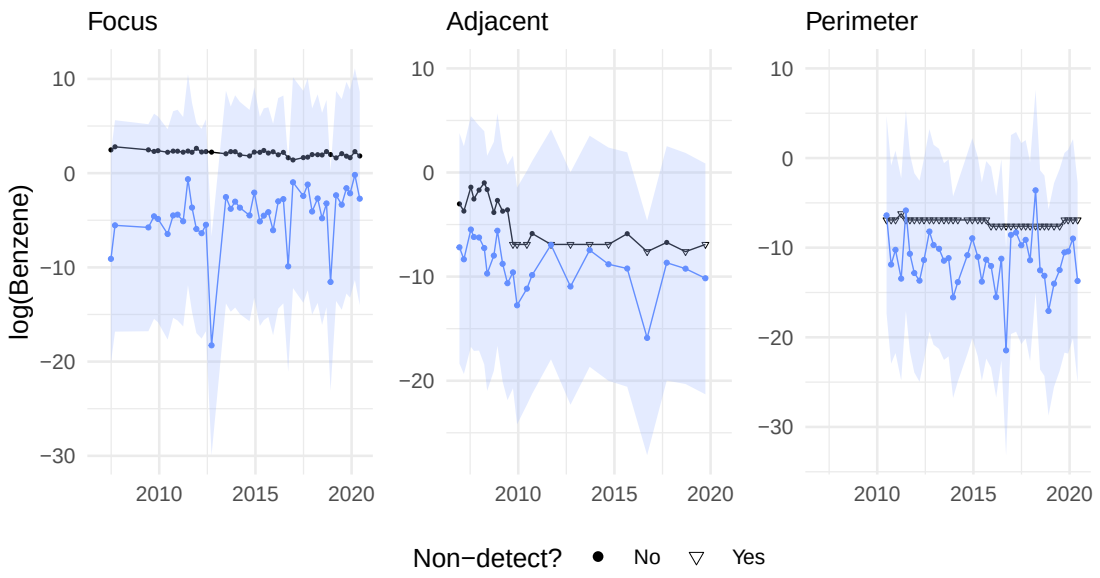

Figure 5.7: Predictions with comparison to truth in black. Benzene, LMWO, censored regression.

Figure 5.7 shows consistent underprediction for the more censored analyte benzene; this highlights one key shortcoming of this model. All data is assumed to come from the same process, however due to the nature of data collection in groundwater monitoring there may be oversampling of several low-activity wells to validate expectations of low concentrations. This leads to a skew in the data and a lower sample average which would substantially affect inference on the intercept term. The current model does not mitigate this unequal data collection through a more careful prior or by incorporating the spatial aspect of these data. As such, this model tends to underpredict concentrations and in the application of hydrocarbon monitoring, this would be dangerous and detrimental to any early detection system. We propose an extension to this model, to be introduced in Chapter 8, where the globally estimated mean, $\beta_{01}, \ldots, \beta_{0j}$, is replaced by a localised well-specific mean.

Another issue with the efficacy of these predictions, over both analytes, is the wide uncertainty shown by the 95% prediction intervals. All analyte data are on the log

scale so a range of $(-20, 5)$, as is common in Figure 5.7, corresponds to measurements of $(2 \times 10^{-9}, 148)$ mg/L where the highest measurement in the entire Case Study A is 150 mg/L from 2003.

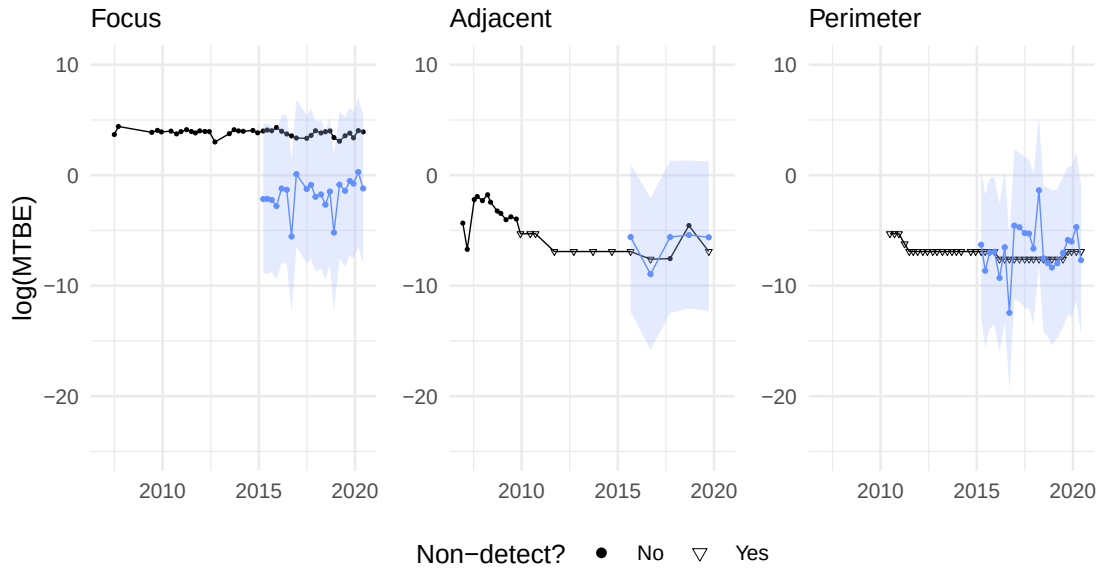

Figure 5.8: Predictions with comparison to truth in black. MTBE, holdout future, censored regression.

Furthermore, this model is unable to leverage data arising from the well to be predicted, as in Chapter 8, meaning the prediction for both LMWO and hold-out future are extremely similar since the model training data has only changed by a handful of observations. To see this clearly one can compare the latter predictions in Figure 5.6 with all predictions in Figure 5.8. The hold-out future predictions for benzene have been omitted for the sake of brevity, but follow a similar pattern.

5.5.5 Model Metrics

To quantify the quality of predictions we can calculate the metrics introduced in Section 1.7 where more negative values suggests a worse fit to that specific data. Further recall that a comparison between MTBE and benzene would be invalid and reflect very little about relative model quality. Each quantity is calculated for both the data the model is trained on (in-sample) and the data arising from the wells to be predicted (out-of-sample) and then collated in Table 5.1 and Table 5.2 respectively.

	LPD (SE)	WAIC (SE)	PSIS (SE)
MTBE, in-sample	-3159.2 (34.9)	-3168.5 (35.2)	-3168.5 (35.2)
Benzene, in-sample	-2434.0 (53.6)	-2445.2 (54.3)	-2445.3 (54.3)

Table 5.1: Model fit metrics based on in-sample data. LMWO, censored regression.

	LPD (SE)	WAIC (SE)	PSIS (SE)
MTBE, out-of-sample	-244.1 (16.1)	-246.4 (16.7)	-246.4 (16.7)
Benzene, out-of-sample	-205.0 (17.1)	-207.0 (17.7)	-207.0 (17.7)

Table 5.2: Model fit metrics based on out-of-sample data. LMWO, censored regression.

Metrics of incomparable model fits as shown in this chapter, are difficult to interpret as a measure of predictive quality and these metrics are better used for comparison. As each metric is based on the log score, less negative values indicate a better fit. Once all models are fully described and fit to the data we can form a complete comparison between all models in this thesis; such a comparison will take place in Chapter 9.

Note that the values for WAIC and PSIS are similar and tend to be so for most scenarios. As such, we shift our focus to PSIS estimates as a more rigorous version of the WAIC estimate where we can also leverage the fitted k parameter of the Pareto distribution to find high-leverage observations and report the reliability of the estimate. For example, we can plot these k values in Figure 5.9. We observe many values below the 0.7 and even 0.5 values that are highlighted in Vehtari *et al.* (2017) implying the estimates for this relatively straightforward model are reliable.

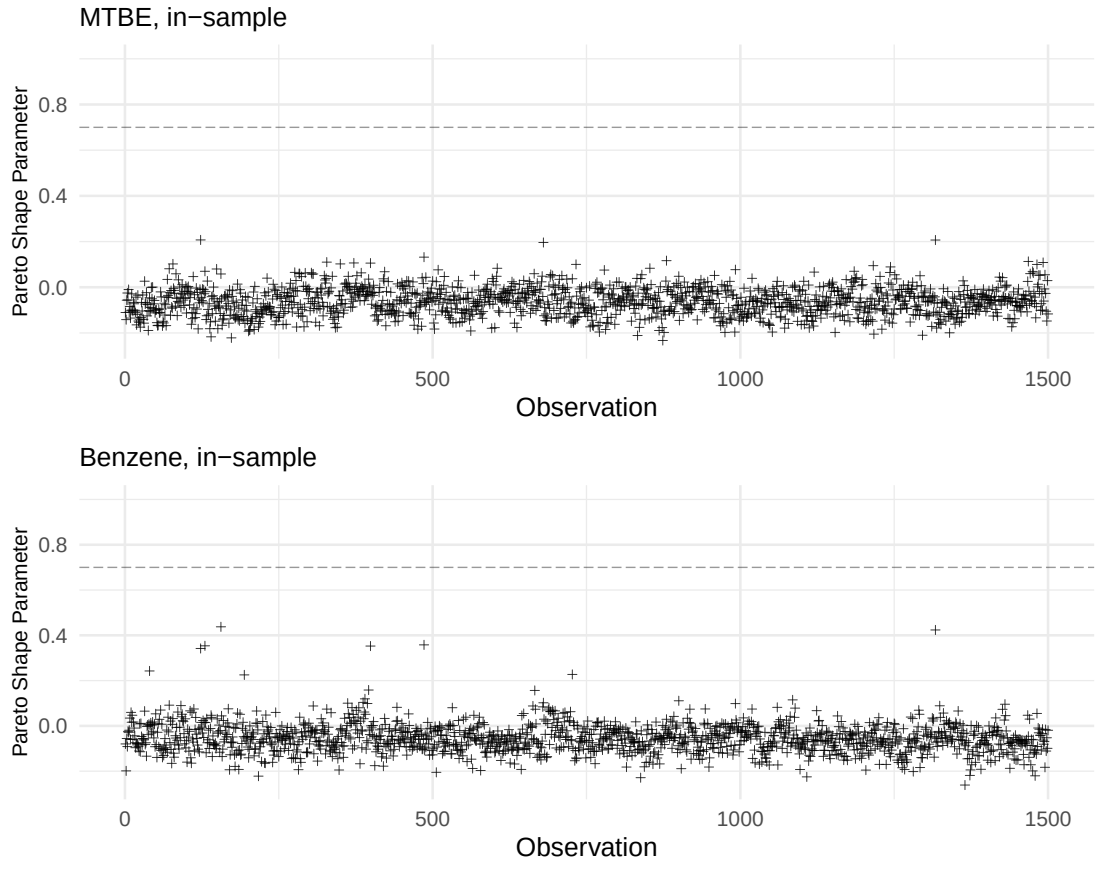

Figure 5.9: PSIS diagnostics based on observed data log likelihood. LMWO, censored regression.

5.6 Conclusion

In this chapter we have considered variations on a univariate multiple linear regression including

- imputing censored observations with half of the associated detection limit;
- censored regression (Tobit) model that modifies the likelihood for censored data.

Simulation results presented are consistent with the literature that claim the DL/2 method leads to inconsistent estimators. As such, we only present results for our case study from the censored regression and will use these results in the remainder of the thesis as a baseline to be compared to.

Regression coefficients provide interpretable parameters for each combination of our analytes, MTBE and benzene, and the chosen water quality predictors conductivity (EC), oxidation reduction potential (ORP), dissolved oxygen (DO), pH and temperature. The sign of each coefficient is consistent with our expectations but the magnitude of the true effects are not well established due to the heterogeneity of groundwater sites.

Similar to the random forest results in Chapter 4, we observe further evidence that the noise of the hydrocarbon groundwater monitoring system may be excessive. For these models, this appears in the form of a lower than expected precision parameter, τ_j, for both analytes considered.

Chapter 6

Multivariate Models

6.1 Multivariate Linear Regression

A clear criticism of (5.2) is that each analyte is modelled independently, whereas hydrocarbon concentrations are often highly correlated. To model this correlation structure directly, suppose we let $B = \{\beta_{ij}\}$ contain all previously defined regression coefficients for $i = 1, \ldots, n_x + 1$ and $j = 1, \ldots, n_y$ with the same interpretation described in Section 5.2. The model is then compactly denoted by

$$Y = XB + E, \tag{6.1}$$

where Y is the analyte data in matrix form as defined in (1.1) and E is a matrix formed by joining row vectors $\boldsymbol{\epsilon}_i^T \sim N_{n_y}(\mathbf{0}, \Sigma_y)$ as n_s independently and identically distributed rows parameterised by the same positive definite matrix Σ_y that we will refer to as the among-columns covariance matrix.

The diagonal of the covariance matrix is a measure of spread for each analyte, whereas the off-diagonal terms are to be interpreted as the covariance between each analyte and are assumed to be high, *a priori*, in hydrocarbon groundwater mon-

itoring. Analyte concentrations, specifically those with similar number of carbon atoms, are components of a greater mixture, say petroleum, and any introduction of an analyte into the closed groundwater system is expected to correspond with the introduction of several analytes at the same time. That is, a high benzene concentration implies that we are more likely to observe high concentrations of other hydrocarbons and vice versa. On the other hand, differences in the analytes' intrinsic properties that impact the flow of these concentrations, such as molecular mass, may lead to lower correlation over time as the physical distance between concentration 'hotspots' will increase.

The multivariate normal regression model described here generalises (5.2) to account for among-column covariance in Y by relaxing the assumption of independent errors among the columns of E. A further generalisation that relaxes the assumption of independent errors among the rows of E is described in Section 6.2 where a matrix-variate normal distribution is introduced.

Since a multivariate normal density is assumed, the likelihood is therefore

$$\pi(Y|X,B,\Sigma_y) = \prod_{i=1}^{n_s}(2\pi)^{-n_y/2}|\Sigma_y|^{-1/2}\exp\left(-\frac{1}{2}(Y-XB)^T\Sigma_y^{-1}(Y-XB)\right).$$

We can express this likelihood in an alternative form based on a special case of the matrix normal density, to be discussed in Section 6.2, with independent among-row covariance, $\Sigma_s = I_{n_s}$. Hence, using the properties of the trace operator and ignoring any constants of proportionality (Rossi *et al.*, 2012, p.32),

$$\pi(Y|X,B,\Sigma_y) \propto |\Sigma_y|^{-n_s/2}\,\mathrm{etr}\left(-\frac{1}{2}(Y-XB)^T(Y-XB)\Sigma_y^{-1}\right),$$

where $\mathrm{etr}(x) \equiv \exp(\mathrm{tr}(x))$.

6.1.1 Prior

Recall the univariate model introduced in Section 5.2 had a conjugate prior where the regression parameters were normal conditional on some measure of spread. Similarly, a normal-inverse-Wishart prior distribution is conjugate for the multivariate case (Gelman *et al.*, 1995) where we can define the prior distributions marginally, for $j = 1, \dots, n_y$, as

$$\boldsymbol{\beta}_j | \Sigma_y \sim N\left(\mathbf{m}_0, \Sigma_y\right),$$
$$\Sigma_y \sim IW\left(d_0, S_0\right).$$

The inverse Wishart (IW) distribution can be intuited as a generalisation of the inverse gamma distribution with positive-definite matrix support (Iranmanesh *et al.*, 2010). We say that Σ follows an inverse Wishart distribution, $\Sigma \sim W_q^{-1}(Q, a)$, with real positive definite scale matrix $Q \in \mathbb{R}^{q \times q}$ and degrees of freedom $a > 0$ if its probability density function is

$$f(\Sigma) = \frac{2^{-q(a+q-1)/2}}{\Gamma\{(a+q-1)/2\}} |Q|^{(a+q-1)/2} |\Sigma|^{-(a+2q)/2} \operatorname{etr}\left(-\frac{1}{2}\Sigma^{-1}Q\right).$$

where $\operatorname{etr}(A)$ is the exponential of the trace of square matrix A. For intuition, the expectation of Σ is $(a-2)^{-1}Q$ for $a > 2$. See Iranmanesh *et al.* (2010) for a more general inverse matrix gamma distribution for which the inverse Wishart distribution is a special case.

As with the univariate model, we assume the regression parameters have no dependence on the covariance matrix, *a priori*, and therefore use a conditionally conjugate prior instead as detailed in (6.2). For these distributions a prior mean and variance, $\mathbf{m}_\beta$ and V_β respectively, are to be chosen for the regression parameters, whereas a prior scale matrix and degrees of freedom, $S_{\Sigma_y}^{-1}$ and d_{Σ_y} respectively, are to be

chosen for the covariance parameter such that

$$\begin{aligned} \boldsymbol{\beta}_j &\sim N\left(\mathbf{m}_\beta, V_\beta\right), \\ \Sigma &\sim IW\left(S_{\Sigma_y}, d_{\Sigma_y}\right). \end{aligned} \tag{6.2}$$

6.1.2 Bayesian Inference

By repeating the methodology of Section 5.2.2, we are able to derive the full conditional distributions (FCDs) of all model parameters then construct a Gibbs sampling algorithm to generate our posterior samples. The semi-conjugate nature of our prior guarantees that the FCDs can be tractably sampled from and omits the need for a different Markov chain Monte Carlo (MCMC) algorithm. Iranmanesh *et al.* (2010) describes a conjugate multivariate normal regression using the inverse matrix gamma distribution as a prior distribution in place of the inverse Wishart. Hoff (2009, pp. 108) provides a derivation of FCDs for the non-regression case where a constant mean is used in place of the linear predictor.

Since we consider the multivariate normal regression model to be a special case of the matrix normal regression model, to be described in Section 6.2, we use the same software to fit both models. Stan is chosen over the aforementioned JAGS because of wider base functionality including evaluation of covariance functions used in Gaussian processes and potential, but not guaranteed, efficiency gains since it generates and compiles the necessary algorithm in C++ (Stan Development Team, 2023). Pertinent to this semi-conjugate model, both JAGS and Stan have great efficiency gains when a conjugate distribution is identified.

6.2 Matrix Normal Regression

In the previous section we generalised our model to account for covariance among our dependent variables. By leveraging the matrix-variate normal distribution, we can model our response variable as a compact matrix with two covariance matrices, among-rows and among-columns. A real random p by q matrix, X, follows a matrix normal distribution, $X \sim MN_{p,q}(M, P, Q)$, if its probability density function is

$$f(X) = (2\pi)^{-pq/2}|Q|^{-p/2}|P|^{-q/2}\exp\left[-\frac{1}{2}\operatorname{tr}\{Q^{-1}(X-M)^T P^{-1}(X-M)\}\right],$$

where M is a p by q real *mean* matrix, P and Q are positive definite *scale* matrices of dimensions p by p and q by q respectively. It is worth noting than an equivalent representation exists. That is,

$$X \sim \mathrm{MN}_{p,q}(M, P, Q),$$

if and only if,

$$\operatorname{vec}(X) \sim N_{pq}(\operatorname{vec}(M), Q \otimes P),$$

where $\otimes$ is the Kronecker product and $\operatorname{vec}(\cdot)$ is the vectorisation of a matrix by stacking each column of the matrix into a single column vector sequentially. Hence, suppose we have our data in compact form,

$$Y = XB + E,$$

and we make the assumption that

$$Y \sim \mathrm{MN}(XB, \Sigma_s(S,S), \Sigma_y), \tag{6.3}$$

where $\Sigma_s(S,S)$ is a among-row covariance matrix to be defined in Section 6.2.2 and Σ_y is the same among-column covariance matrix introduced in Section 6.1.

If one assumes model (6.1), then $\Sigma_s(S, S)$ is set to be the identity matrix of appropriate size and observations are assumed independent given the model parameters. The spatial matrix-variate normal regression model (6.3) relaxes this assumption and assumes our observations are correlated in space and time.

A naive approach would assume Σ_s is a stochastic parameter to be estimated, potentially with a prior from the same parametric family as the prior used for Σ_y. However, a considerable difference between these covariance matrices are the sizes of each. We observe 1500 observations in the leave-multiple-well-out (LMWO) prediction scenario and could fit up to 6 analytes for groundwater site A specifically. The number of parameters within a N by N covariance matrix is equal to $N + \frac{N\,(N-1)}{2}$ due to the symmetry; Σ_s would be a symmetric matrix of $1,125,750$ unknown parameters as opposed to the, at most, 21 unknown parameters required to construct Σ_y. To deal with this intractability we leverage Gaussian processes (GP) and the corresponding spatiotemporal data associated with each row in Y and X.

6.2.1 Gaussian Processes

Consider a Gaussian process (GP) as a generalization of the normal distribution. The support of the normal distribution can be scalar, a vector of scalar values or a matrix when considering the matrix normal distribution, whereas a GP is a distribution of continuous functions (Rasmussen *et al.*, 2006). In practical terms, we only ever evaluate these functions at a finite number of vector *inputs* denoted $\mathbf{s}$ which allows us the desirable property that the joint distribution of these finite elements is Gaussian. Therefore, we define the Gaussian process,

$$f(\mathbf{s}) \sim GP(m(\mathbf{s}), k(\mathbf{s}, \mathbf{s}')),$$

to be a collection of random variables following a multivariate normal distribution, that is completely defined by some mean function $m(\mathbf{s})$ and covariance function, $k(\mathbf{s}, \mathbf{s}')$, which we refer to as a *kernel* function.

We can make use of these distributions as a method for defining covariance matrices over a field, say $\mathbb{R}^d$, for the spatiotemporal data S described in Section 1.4. To this end, we consider two kernel functions that are both radial basis functions, a function that depends only on distance not location. We incorporate characteristic length-scales, that is a different length-scale parameter per input dimension, by defining each kernel function in terms of the weighted distance,

$$d = \left\| \frac{\tilde{\mathbf{s}} - \tilde{\mathbf{s}}'}{\boldsymbol{\gamma}} \right\| = \sqrt{\sum_{d=1}^{D} \frac{(\tilde{s}_d - \tilde{s}'_d)^2}{\gamma_d^2}},$$

for each input dimension $d = 1, \ldots, D$. The two kernels we have considered can then be expressed as

1. squared exponential kernel, with amplitude γ_α,

$$K_{SE}(d) = \gamma_\alpha^2 \exp\left(-\frac{d^2}{2}\right);$$

2. Matérn kernel, with $\nu = 3/2$ and amplitude γ_α,

$$K_{\text{Matérn}}(d) = \gamma_\alpha^2 \left(1 + \sqrt{3}d\right) \exp\left(-\sqrt{3}d\right).$$

These kernels are related since the Matérn kernel is $\nu - \frac{1}{2}$ times differentiable and the squared exponential is the limiting case where $\nu \to \infty$ (Beckers, 2021). Since we have no explicit prior knowledge about the existence of higher order derivatives, and we have found the Matérn to be more computationally stable, we opt to use the Matérn kernel function in our analyses.

The use of characteristic length-scales in the squared exponential kernel, where γ_d is a different parameter per dimension, is also referred to as the automatic relevance determination (ARD) extension (Rasmussen *et al.*, 2006). We describe in Section 8.2.2 how we can assume several, to be defined, well effects follow a GP with characteristic length-scales *a priori*.

6.2.2 Among-Row Covariance

For the matrix-variate regression model considered in this thesis, we assume there is a spatiotemporal effect that we can model directly over the n_s observations corresponding to the rows of Y, X and S. To this end, we require a function that can convert spatiotemporal data to a covariance matrix reflecting the spatial structure of the inputs. Thus, suppose we have two spatiotemporal matrices with each column corresponding to a different spatiotemporal dimension, say

$$S = \begin{pmatrix} \mathbf{s}_1 \\ \vdots \\ \mathbf{s}_{n_s} \end{pmatrix} = \begin{pmatrix} s_{1x} & s_{1y} & s_{1t} \\ \vdots & \vdots & \vdots \\ s_{n_s x} & s_{n_s y} & s_{n_s t} \end{pmatrix};$$

$$S' = \begin{pmatrix} \mathbf{s}'_1 \\ \vdots \\ \mathbf{s}'_{n_s} \end{pmatrix} = \begin{pmatrix} s'_{1x} & s'_{1y} & s'_{1t} \\ \vdots & \vdots & \vdots \\ s'_{n'_s x} & s'_{n'_s y} & s'_{n'_s t} \end{pmatrix}.$$

We then define $\Sigma_S(S, S')$ to be a function that maps S and S' to a covariance matrix such that the k, l^{th} element is given by

$$\{\Sigma_S(S, S^*)\}_{kl} = K(d_{kl}|\gamma_\alpha, \gamma_x, \gamma_y, \gamma_t),$$

where the kernel function K and GP hyperparameters $\Gamma = (\gamma_\alpha, \gamma_x, \gamma_y, \gamma_t)^T$ including the amplitude γ_α and characteristic length-scales per dimension are to be specified

For our data, the scaled distance is given by

$$d_{kl} = \sqrt{\frac{(s_{kx} - s'_{lx})^2}{\gamma_x^2} + \frac{(s_{ky} - s'_{ly})^2}{\gamma_y^2} + \frac{(s_{kt} - s'_{lt})^2}{\gamma_t^2}}.$$

As shown in (6.3), these matrices need not be distinct and assuming $S = S'$ results in a symmetric covariance matrix.

For the model described in (6.3) we avoid an identifiability issue by setting the amplitude parameter γ_α equal to 1. An alternate solution to be investigated further would fix Σ_y to be a correlation matrix with an appropriate prior specification such as using the LKJ distribution (Lewandowski *et al.*, 2009). To see why this is required, consider the covariance,

$$\text{Cov}(y_{ij}, y_{i'j'}) = \{\Sigma_y\}_{jj'} K(d_{ii'}|\Gamma), \tag{6.4}$$

where $\{\Sigma_y\}_{jj'}$, the j^{th} row, j'^{th} column of Σ_y, denotes the covariance between analytes j and j'. Then, we could produce the same covariance by multiplying this term by some constant as long as we divide the amplitude parameter by the same value. We present a model where the amplitude parameter is determined to be stochastic and estimated in Chapter 8.

The parameters of model (6.3) are B, Σ_y and the GP hyperparameters $\Gamma = (\gamma_\alpha, \gamma_x, \gamma_y, \gamma_t)^T$ since the among-row covariance is deterministic given the choice of kernel function and associated hyperparameters and observed spatiotemporal data S. Hence, the likelihood is given by

$$L(B, \Sigma_y, \Sigma_s) = (2\pi)^{-n_s n_y/2} |\Sigma_y|^{-n_s/2} |\Sigma_s|^{-n_y/2} \,\text{etr}\left(-\frac{1}{2}\Sigma_y^{-1}(Y - XB)^T \Sigma_s^{-1}(Y - XB)\right),$$

For the spatial matrix-variate normal regression model, we have assumed $\Sigma_s = \Sigma_s(S, S)$ and for the multivariate normal regression model, let $\Sigma_s = I_{n_s}$.

6.2.3 Prior

The priors for the parameters B and Σ_y remain unchanged from Section 6.1.1 due to the generalising nature of the matrix normal model. For a fully specified prior, we must also define our prior beliefs for each GP length-scale. Hence, we assume an identical gamma prior for each length-scale as there is little information *a priori* to discern each dimension and the strictly positive support reflects the strictly positive nature of length-scale parameters

$$\gamma_x, \gamma_y, \gamma_t \sim Ga(a_\gamma, b_\gamma) \text{ independently.}$$

We determine a reasonable prior mean by investigating the resultant correlation matrix for each choice of a_γ and b_γ until the correlation matrix seems appropriate. Prior variance is then chosen to reflect our confidence in the prior and can be changed to investigate the sensitivity of this choice. For example, preliminary analyses found $Ga(0.2, 1)$ gave too much prior weight to extreme values in the tail and we opted to refit the model with $Ga(2, 10)$.

6.2.4 Bayesian Inference

As described in Section 6.1, we use Stan (Stan Development Team, 2023) to fit these models due to the efficiency gains that make these models tractable. Another justification is that Stan already defines kernel functions as proprietary methods and allows user-supplied functions to extend these kernels as modular components. This is very desirable when investigating the impact of different kernels used.

6.3 Prediction

In this section we detail how to obtain the posterior predictive of new data for the matrix normal regression model. All steps also apply to the special case multivariate regression model where we can set $\Sigma_s = I_{n_s}$. Once the posterior predictive distribution is defined we can follow the steps described in Appendix B.1 to obtain posterior predictive samples to be presented in Section 6.4.

Suppose that in addition to our observed data, Y, X, S, which is modelled according to (6.3), we want to predict new data Y^* conditional on our posterior beliefs, new predictors X^*, and new spatiotemporal metadata S^*. We assume there are n_s^* new observations, but the number of predictors, n_x, and assumed spatiotemporal dimensions d are the same for both observed and new data. Similarly, we aim to compose predictions for all n_y analytes that appear in the observed data.

By using the equivalent multivariate normal parameterisation of the matrix-variate distribution, as described in Section 6.2, we can express (6.3) as

$$\text{vec}(Y) \sim N_{n_s n_y}(\text{vec}(XB), \Sigma_y \otimes \Sigma_s(S, S)),$$

where $\otimes$ is the Kronecker product and $\text{vec}(\cdot)$ is the vectorisation of a matrix by stacking each column of the matrix into a single column vector sequentially. We also assume this model for the data to be predicted, that is,

$$\text{vec}(Y^*) \sim N_{n_s^* n_y}(\text{vec}(X^*B), \Sigma_y \otimes \Sigma_s(S^*, S^*)).$$

Under this assumption, multivariate normal distributions of observed and new data can be combined into a joint multivariate normal distribution, we can then condition on the joint density in the usual way. Therefore, suppose

$$\begin{pmatrix} \text{vec}(Y) \\ \text{vec}(Y^*) \end{pmatrix} \sim N_{(n_s+n_s^*)n_y} \left(\begin{pmatrix} \text{vec}(XB) \\ \text{vec}(X^*B) \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right),$$

where

$$\begin{aligned} \Sigma_{11} &= \Sigma_y \otimes K(S, S), \\ \Sigma_{12} &= \Sigma_y \otimes K(S, S^*), \\ \Sigma_{21} &= \Sigma_y \otimes K(S^*, S), \\ \Sigma_{22} &= \Sigma_y \otimes K(S^*, S^*). \end{aligned}$$

Note that this constructed covariance matrix assigns the expected covariance we have defined in (6.4) but it is not identical to

$$\Sigma_y \otimes K \left(\begin{pmatrix} S \\ S^* \end{pmatrix}, \begin{pmatrix} S \\ S^* \end{pmatrix} \right).$$

Therefore, the posterior predictive distribution for the spatial matrix-variate regression model is

$$\begin{aligned} \text{vec}(Y^*)|Y = y, X, S, S^*, \Theta &\sim N_{n_s^* n_y} \left(\boldsymbol{\mu}_{Y^*|Y}, \Sigma_{Y^*|Y} \right), \\ \boldsymbol{\mu}_{Y^*|Y} &= \text{vec}(X^*B) + \Sigma_{12}\Sigma_{22}^{-1} \text{vec}(y - XB), \\ \Sigma_{Y^*|Y} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

6.4 Groundwater Application

We specify both models in Stan, to be fit with the Hamiltonian Monte Carlo (HMC) algorithm (Stan Development Team, 2023). When the spatial matrix-variate normal regression model is fit to our groundwater monitoring site, 1500 observations for the LMWO prediction scenario described in Section 1.6, we find evaluation of the log density takes between 1 and 3 seconds. The ramifications of this is that the model will not fit within a couple of days but will require weeks or even months because the HMC algorithm evaluates this gradient several times per MCMC iteration. Moreover, this algorithm would have to be run several times during the process of model checking, for example, fitting the model with various starting values to obtain different chains and improve confidence of convergence. Therefore, we present no results for matrix-variate models and instead advise the reader to use the varying intercept model as described in Chapter 8 for a spatially-based model. We concede further work on matrix-variate distributions such as efficiency analysis, approximations and emulation would lead to interesting results we did not investigate.

Accordingly, all results presented in this section are therefore produced from the Bayesian multivariate multiple linear regression as described in Section 6.1 where observations (water samples) are modelled independently but analyte covariance is modelled directly. We again assume independent and identically distributed standard normals with unit variance and zero-valued mean for each regression coefficient. For the inverse Wishart prior distribution, we fix the scale matrix

$$S_{\Sigma_y} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix},$$

to reflect our prior beliefs of positive correlation between the analytes and then allow the degrees of freedom, $d_{\Sigma_y} \geq n_y$, to be reflective our prior certainty.

Using the prior draws as a check, we have chosen $d_{\Sigma_y} = 20$. We also considered letting $d_{\Sigma_y} = 2$ and $d_{\Sigma_y} = 200$ *a priori* and also fit models (not shown) with these choices to investigate the sensitivity of this choice. We found no noticeable difference when the prior was made more vague, that is $d_{\Sigma_y} = 2$, but were able to very slightly affect the posterior distributions with a more precise prior, that is $d_{\Sigma_y} = 200$. However, these changes only decreased the log likelihood estimates by less than 0.5% on average.

As described in Section 5.5, the LMWO and holdout future prediction scenarios produce overlapping observed data since holdout future scenario is identical to the LMWO scenario with pre-2015 observations from holdout wells added. As expected, we do not observe any substantial impact on the parameter estimates between these inferences. Hence, output relating to parameter estimates for the model trained in the holdout future case are omitted.

6.4.1 Regression Parameters

When analyte covariance is directly modelled as is the case in the imputed multivariate multiple regression model, we notice that the regression coefficients associated with benzene are all less extreme than the univariate censored case shown in Figure 5.2. That is, our posterior beliefs increase in their certainty and hold more density closer to 0 as shown in Figure 6.1. We previously justified modelling MTBE as a proxy to better understand, typically left-censored ($> 50\%$), benzene concentrations and so we expected more informative regression coefficients here.

On the other hand, this bivariate relationship can not be the only reason that we see posterior densities "moving" closer to 0 as both intercepts and oxidation reduction potential (ORP) effects on both analyte concentrations are estimated to have a smaller impact in the multivariate case.

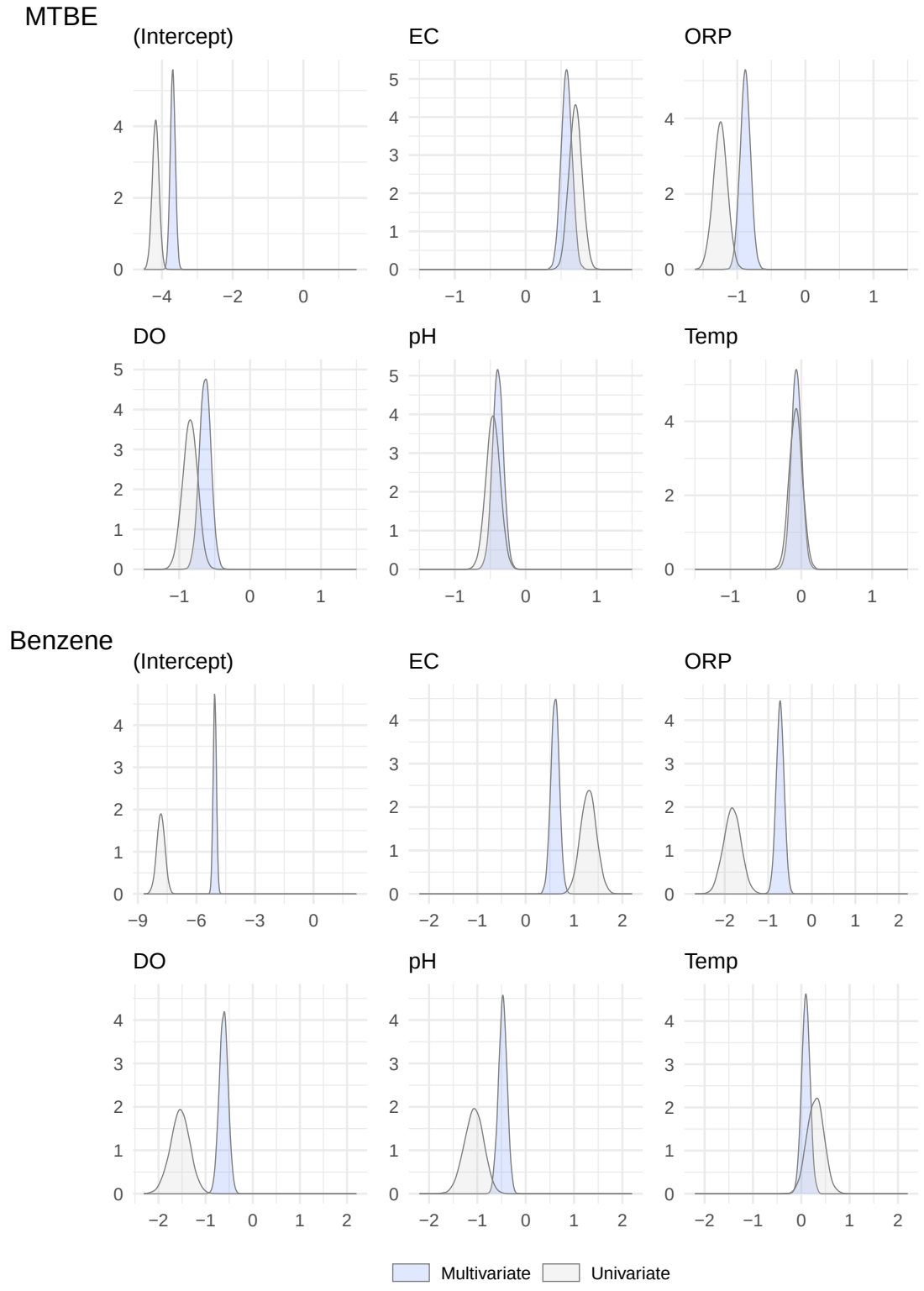

Figure 6.1: Marginal posterior density of regression coefficients, corresponding parameter from univariate regression shown for comparison. LMWO, imputed multivariate multiple linear regression.

6.4.2 Among-Column Covariance

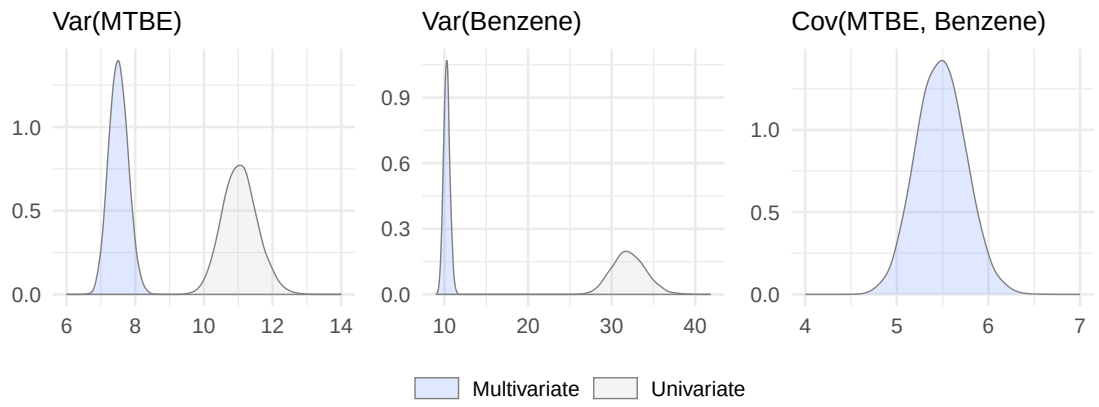

Figure 6.2: Posterior density of marginal variance and covariance terms in Σ_y LMWO, imputed multvariate multiple linear regression.

Since we have only modelled two analytes, MTBE and benzene, the model parameter Σ_y is a 2 by 2 matrix comprising three parameters. The main diagonal of Σ_y corresponds to the marginal variances and the non-diagonal entries are equal by symmetry and represent the marginal posterior distribution of the covariance between MTBE and benzene. Thus, all three marginal posterior distributions are shown as densities in Figure 6.2.

Comparing with univariate model parameters for measurement error shown as marginal precisions (inverse variance) in Figure 5.4, we notice that the posterior precisions for both analytes have increased. In particular, the posterior mean of the marginal precision associated with MTBE increased from 0.09 to 0.13, for benzene the posterior mean changed from 0.03 to 0.10. Residual variance not explained by the model is decreased for both analytes when jointly modelled, more so for the more frequently censored benzene concentrations.

We parameterise our model with a covariance parameter as opposed to a precision matrix to help with interpretability of the non-diagonal terms. Positive covariance

between analytes, as shown, reflect that our posterior beliefs are aligned with our prior beliefs of high correlation between these analytes. Furthermore, we could decompose the covariance parameter Σ_y into a correlation parameter $\tilde{\Sigma}_y$ for all MCMC iterations $m = 1, \ldots, M$ such that

$$\tilde{\Sigma}_y^{(m)} = D^{(m)} \Sigma_y^{(m)} D^{(m)},$$

where $D^{(m)}$ is a diagonal matrix where the diagonal elements are equal to the inverse square root of the diagonal elements in $\Sigma_y^{(m)}$. Using the draws from the off-diagonal of this posterior correlation matrix yields our estimated posterior correlation between MTBE and benzene; the multivariate multiple regression model estimates the correlation between our two analytes to be between $(0.592, 0.654)$ based on the 95% symmetric credible interval.

We have described in Section 2.4.1 how estimating analyte correlation is difficult when both variables are doubly censored. Future work could investigate the feasibility of using the Bayesian multivariate multiple regression model to estimate the correlation between two censored analytes; censoring would have to be dealt with using imputation or data augmentation. The approach could be compared to existing methods such as Kendall's Tau (Helsel, 2011) and bivariate maximum likelihood estimation (Newton & Rudel, 2007) as described in Section 2.4.1.

6.4.3 Prediction

In the univariate model predictions shown in Section 5.5.4, we observed a much lower prediction for benzene log concentrations with some prediction intervals lower than -20 on the log scale. By jointly modelling both analytes we would expect lower concentrations of benzene to be estimated to be closer to the MTBE concentration, given the moderate inter-analyte correlation in the posterior. Figure 6.3 shows that

this is the case for the LMWO prediction scenario where benzene is now predicted to be much higher when the model jointly predicts MTBE and benzene in one step using the bivariate normal posterior predictive distribution.

As with Chapter 5, we omit prediction results for holdout future scenarios as predictions after 2015 look identical to Figure 6.3.

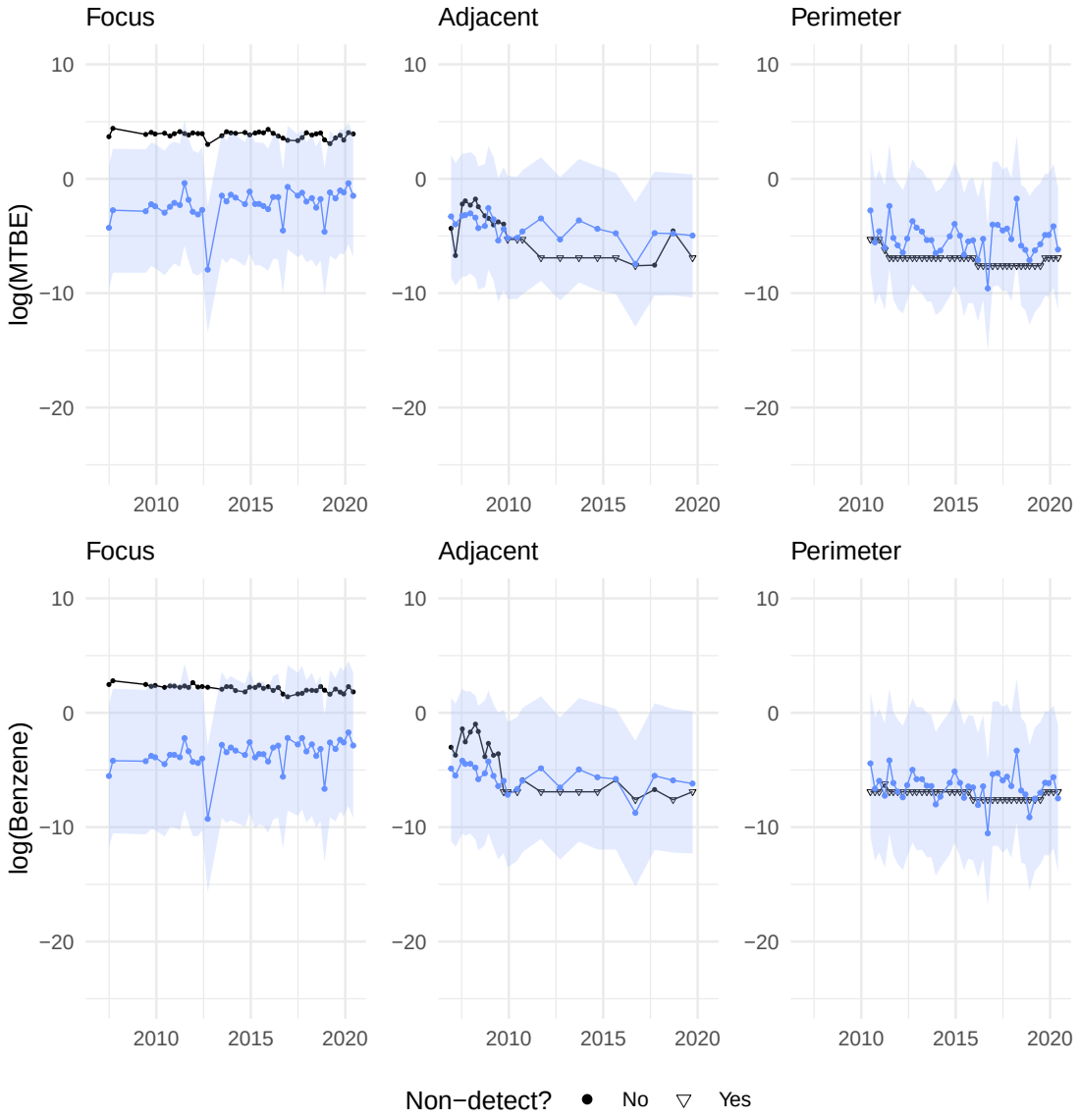

Figure 6.3: Predictions with comparison to truth in black. LMWO, imputed multivariate multiple linear regression.

6.4.4 Model Comparison

For our model comparison metrics described in Section 1.7 we must evaluate the log likelihood for each MCMC iteration. Previously, we have extracted the log pointwise predictive density (LPD) for each analyte per MCMC iteration. With the multivariate multiple regression model, the bivariate density in the log likelihood means we can only obtain a single contribution per observation and per MCMC iteration. If a matrix-variate regression was to be compared, we would not be able to evaluate the log pointwise predictive density (LPD) as the density is evaluated with the entire dataset Y, X and S. The upshot of this is that when comparing multivariate models, we are comparing the predictive performance of the model based on all analytes, as shown in Table 6.1 and Table 6.2. As described in Section 1.7.4, each metric value shown is relative to the 'best' model as we present the average standard error of the difference in models, not the difference of the standard errors; consequently, the best model is easily identified by a zero value.

	Δ LPD (SE)	Δ WAIC (SE)	Δ PSIS (SE)
Univariate	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Multivariate	-1562.5 (64.2)	-1560.4 (64.6)	-1560.4 (64.6)

Table 6.1: Model comparison metrics based on in-sample data. Both analytes, LMWO, imputed multvariate multiple linear regression.

	Δ LPD (SE)
Univariate	0.0 (0.0)
Multivariate	-98.7 (20.9)

Table 6.2: Model comparison metrics based on in-sample data. Both analytes, LMWO, imputed multvariate multiple linear regression.

We speculate that while benzene prediction may have improved due to the presence of MTBE, it could also be the case that prediction of MTBE has suffered due to the overly-censored benzene. Another difference between these two models is that censoring was not dealt with in a principled way and the imputation technique, DL/2, as defined in Section 1.5.1 was used for the multivariate model. Table 6.1 shows that the multivariate model is considerably worse on the in-sample data. Similarly, Table 6.2 highlights that the univariate model is likely better at predicting new data since the log pointwise predictive density (LPD) is approximately 5 standard errors greater.

On the other hand, one should be careful to draw conclusions from these results as the log density of the multivariate model must be evaluated on the imputed data. That means the univariate model has an unfair advantage since it is easier to predict an observation that "agrees" with ND < 1 yielding a higher CDF contribution than to predict a value close to half the detection limit, 0.5. To mitigate this issue, further work could consider assessing model performance based on an 'important' subset of the uncensored data and potentially using simpler metrics such as RMSE.

6.5 Conclusion

We have shown how jointly modelling analyte concentrations within hydrocarbon groundwater monitoring could improve upon the predictive power of our models. Qualitative assessments of prediction show more appropriate prediction intervals for benzene. On the other hand, model comparison metrics such as widely applicable information criterion (WAIC) and Pareto-smoothed importance sampling (PSIS) refute that prediction is improved, but this could be due to the differing nature of the log likelihoods for each model.

No regression coefficients were shown for the univariate imputed case due to the apparent bias demonstrated in the simulation study from Section 5.4. For the same reason, one should be sceptical of the regression coefficients presented in this chapter but we have highlighted how the true effect of each predictor may be smaller in magnitude than the univariate model suggests. Since these parameters represent our estimated effects of each of our predictors, electrical conductivity (EC), oxidation reduction potential (ORP), dissolved oxygen (DO), pH and temperature on the MTBE and benzene concentrations, we further explore the potential inaccuracy in Chapter 8.

To keep the number of modelled analytes general, such that n_y can be any positive integer, we have presented the multivariate multiple linear regression as a trade-off where directly modelling correlation is beneficial, but is done at the cost of needing to impute the censored data, as no appropriate general censored multivariate density can be used. Further work could investigate if a compromise where the number of analytes is restricted to at most 2 and a new likelihood is composed using the likelihood defined in Newton & Rudel (2007).

Alternatively, data augmentation as described in Section 1.5.4 could be employed to deal with the censoring where each censored observation is regarded as a random variable. This approach is more computationally expensive and may not be worth the gains of modelling the analytes jointly as evidenced by similar predictions in Figure 6.3 and Figures 5.6 and 5.7.

Chapter 7

Mixture of Experts

7.1 Motivation

In Chapter 3, we presented the reactive transport model (RTM) which highlighted potential non-linearity between analytes and predictors since each relationship was determined by a "phase", or current state, of the groundwater network that is being monitored. By generalising to a mixture modelling framework, we can improve model performance when applied to the idealised mechanistic data, simulated from the RTM. We then investigate if this generalisation improves our fit to case study A, introduced in Chapter 2.

To relax our assumption of a linear relationship between analytes and predictors, we assume a mixture model where observations are split into components to be determined, and within each component there is a local linear relationship. One could find component allocations for each observation using K-means clustering (MacQueen *et al.*, 1967) or a similar algorithm, but here we opt to integrate the partitioning of the observations within the model description. The Mixture of Experts (MoE) model was originally described in the neural network literature (Jacobs *et al.*,

1991) but is also referred to as the *concomitant variable mixture regression model* (Wedel, 2002) in the statistical literature. A more explanatory description found in Frühwirth-Schnatter (2006) refers to the model as

> "a finite mixture of regressions models with observation-dependent weight distribution."

In our model, the weight distribution represents the different phases of the groundwater system and we are making the assumption that these phases can be inferred using observations of our water quality predictors. Hence, one can think of these models as a two-stage approach,

- split the data into K discrete components;
- perform regressions within each of these components.

The power of this model comes from the ability to describe complex and potentially non-linear relationships using multiple linear regression coefficients. These models are very suitable for our aims of predicting analyte concentrations using the previously defined water quality predictors such as conductivity (EC) and oxidation reduction potential (ORP) because our predictors that will inform the regressions and component membership are intrinsic to the geological system. That is, we reasonably expect that the EC, ORP or even pH of a water sample is indicative of the fluctuations in the groundwater system affecting components, whilst also presenting a quantifiable impact on specific analyte concentrations.

7.2 Model Specification

Consider a finite mixture model (McLachlan & Peel, 2000), which can be expressed as a convex combination of K finite parametric distributions, $f_1, \dots, f_K$,

$$f(y_i|\boldsymbol{\eta}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \eta_k f_k(\cdot|\boldsymbol{\theta}_k), \tag{7.1}$$

for each observation $i = 1, \dots, n_s$, where $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)^T$ and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\}$. The set of parameters for the k^{th} distribution are denoted by $\boldsymbol{\theta}_k$, for $k = 1, \dots, K$, and $\eta_1, \dots, \eta_K$ are non-negative weights that sum to one. We refer to each distribution as a "component", where the total number of components, K, is assumed to be unknown *a priori* and discussed further in Section 7.4.

If f_k also depends on covariates, $\mathbf{x}_i$, Equation (7.1) describes a mixture of regressions model. Allowing a mixture of regressions model to possess weights that also depend on these covariates yields a MoE model as described in (7.2) (Frühwirth-Schnatter *et al.*, 2019)

$$f(y_i|\mathbf{x}_i, \boldsymbol{\omega}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \eta_k(\mathbf{x}_i|\boldsymbol{\omega}_k) f_k(\cdot|\mathbf{x}_i, \boldsymbol{\theta}_k), \tag{7.2}$$

where a *gating function*, η_k, performs clustering based on weighting parameters $\boldsymbol{\omega}_k$ and covariates $\mathbf{x}_i$. Each parametric distribution, otherwise known as an *expert* (Jacobs *et al.*, 1991), depends on component-specific distribution parameters $\boldsymbol{\theta}_k$ and covariates.

7.2.1 Likelihood

Using the general model (7.2), we can extend the original censored regression model (5.2) to allow for clustering. In a groundwater application, we restrict component membership to be shared across analytes, $j = 1, \dots, n_y$, meaning y_{ij} corresponds to component k if and only if $y_{ij'}$ corresponds to component k for all $j' = 1, \dots, n_y$.

Given the physical interpretation of the components as phases of groundwater network systems, this restriction is appropriate.

The censored mixture of experts model for multiple analytes, is given by the likelihood

$$f(\mathbf{y}_i|\mathbf{x}_i, \Omega, \tilde{B}, T) = \sum_{k=1}^{K} \eta_k(\mathbf{x}_i|\boldsymbol{\omega}_k) f_k(\mathbf{y}_i|\mathbf{x}_i, B_k, \boldsymbol{\tau}_k), \tag{7.3}$$

for $i = 1, \ldots, n_s$, where $\eta_k(\mathbf{x}_i|\boldsymbol{\omega}_k)$ represents the probability that observation i is associated with component k based on choice of weighting function η, covariates $\mathbf{x}_i$ and weighting parameters $\boldsymbol{\omega}_k$. For our left-censored dependent variables, f_k denotes the normal probability density function (PDF) when y_{ij} is uncensored and the normal cumulative distribution function (CDF) otherwise, an approach introduced in the censored regression model (5.6). The parameters of this multivariate model are

- $\Omega = (\boldsymbol{\omega}_1 \ \ldots \ \boldsymbol{\omega}_K)$, a $n_x + 1$ by K matrix of weighting parameters. A greater ω_{jk} represents a greater probability than an observation is associated with component k, assuming all other weighting parameters are fixed and the j^{th} predictor is positive, the converse holds true for negative predictor values;
- $\tilde{B}$, an array of dimension $n_x + 1 \times n_y \times K$ that contains K regression coefficient matrices, B_k for $k = 1, \ldots, K$,
  - for convenience, we define $\boldsymbol{\beta}_{jk}$ to be the j^{th} column of B_k representing the impact of each intercept and predictor on the j^{th} analyte under component k;
- $T = (\boldsymbol{\tau}_1 \ \ldots \ \boldsymbol{\tau}_K)$, a $n_y \times K$ matrix of component-specific and analyte-specific precision parameters.

By design, the regression coefficients vary among components, but whether the precision parameters should be considered distinct across all components is more ambiguous as measurement error could be expected to be independent to the "phase"

of the groundwater network. We allow the measurement precision to vary by analyte and component with the remark that if these parameters are identical, the more general model will allow for this case. If all precision estimates appear to be identical, we can refine the model by imposing the restriction of a single precision parameter to improve estimation quality.

Note that it may be the case that the predictors used in the regression context could differ from the predictors used in the clustering context. For instance, consider the role of *pe*, a measure of electron activity, in the RTM data. This variable almost exactly informs the phase of the mechanistic system but seemingly has less utility in predicting the simulated concentration of any analyte. More nuanced variable selection is beyond the scope of this work and we incorporate all predictors for both weighting and regression where it is implied that $\mathbf{x}_i^T$ denotes the same predictors in both cases. Furthermore, our estimates of these coefficients provide a crude assessment of variable importance as we observe with temperature in Section 7.6.

7.2.2 Latent Variables

Suppose there exists latent allocation variables z_i for each observation $i = 1, \dots, n_s$ that determine which component each extracted water sample i belongs to, as shown in (7.4). Using these allocation variables, we can extend the univariate (5.2) and censored (5.6) linear regression models to have distinct parameters for each component and enact clustering with a multinomial logistic model. This alternative parameterisation can be a useful form as the generated quantity $\mathbf{z} = (z_1, \dots, z_{n_s})$ can assist in post-processing algorithms like label switching to be discussed in Section 7.3.

$$\begin{aligned} y_{ij}|\mathbf{x}_i, z_i = k &\sim N(\mathbf{x}_i^T \boldsymbol{\beta}_{jk}, \tau_{jk}^{-1}), \\ \Pr(z_i = k|\mathbf{x}_i) &= \eta_k(\mathbf{x}_i|\boldsymbol{\omega}_k). \end{aligned} \tag{7.4}$$

While a meaningful interpretation of $\mathbf{z} = (z_1, \dots, z_{n_s})^T$ is not required to be used effectively (Frühwirth-Schnatter *et al.*, 2019), we can perceive each z_i as nominal data that labels the phase of a complex geophysical system as shown in the RTM model, Chapter 3.

Obtaining maximum likelihood estimates (MLEs) from likelihood (7.3) is not straightforward but can be determined using an expectation maximisation (EM) algorithm (Dempster *et al.*, 1977). During the EM algorithm, the likelihood (7.3) is computed and then maximised with respect to the model parameters. This procedure is then enacted iteratively until the model parameters converge on a local maxima (Gormley & Frühwirth-Schnatter, 2019, pp.277). A possible alternative to the likelihood used in the "E" step is the conditional expectation of the complete data likelihood. The complete data likelihood is the likelihood we would construct if we assume the latent variable $\mathbf{z}$ was known. For a MoE model described in (7.4), the complete data likelihood is

$$L_C(\boldsymbol{\theta}, \boldsymbol{\omega}) = \prod_{i=1}^{n_s} \prod_{j=1}^{n_y} \prod_{k=1}^{K} \left\{ \eta_k(\mathbf{x}_i|\boldsymbol{\omega}_k) f_k(y_{ij}|\mathbf{x}_i, \beta_{jk}, \tau_{jk}) \right\}^{\mathbb{I}(z_i=k)},$$

where $\mathbb{I}(z_i = k)$ equals one if $z_i = k$ and zero otherwise.

7.2.3 Weighting Function

For this model to be fully defined, it remains to choose weighting function, η_k, that is capable of mapping our covariates to the K-dimensional simplex denoted $\mathscr{S}^K$. This restriction is required so the output can be used as probabilities. We generate a weighting linear predictor, $\mathbf{x}_i^T \Omega$, to be passed to the softmax function, otherwise known as the inverse multinomial logit function. Explicitly, the allocation

probabilities for components $k = 1, \dots, K$ are given by the column vector

$$\begin{aligned}\boldsymbol{\eta}_i &= (\eta_1(\mathbf{x}_i|\boldsymbol{\omega}_1), \dots, \eta_K(\mathbf{x}_i|\boldsymbol{\omega}_K))^T \\ &= \text{softmax}(\mathbf{x}_i\boldsymbol{\omega}_1, \dots, \mathbf{x}_i\boldsymbol{\omega}_K) \\ &= \left(\frac{\exp(\mathbf{x}_i\boldsymbol{\omega}_1)}{\sum_{k'=1}^{K} \exp(\mathbf{x}_i\boldsymbol{\omega}_{k'})}, \dots, \frac{\exp(\mathbf{x}_i\boldsymbol{\omega}_K)}{\sum_{k'=1}^{K} \exp(\mathbf{x}_i\boldsymbol{\omega}_{k'})} \right)^T.\end{aligned}$$

Further work could investigate better choices of η_k and take inspiration from neural networks where the rectified linear unit (RelU) function can be chosen over other "activation" functions due to more satisfactory properties (Goodfellow *et al.*, 2016).

Hence, the censored MoE model with latent variables (7.4) can be more compactly expressed as

$$\begin{aligned}\pi(y_{ij}|\delta_{ij}, \mathbf{x}_i, z_i = k) = f(y_{ij}|\mathbf{x}_i^T, \boldsymbol{\beta}_{jk}, \tau_{jk})^{1-\delta_{ij}} F(y_{ij}|\mathbf{x}_i^T, \boldsymbol{\beta}_{jk}, \tau_{jk})^{\delta_{ij}}, \qquad (7.5)\\ z_i \sim \text{Multinomial}_K(1, \boldsymbol{\eta}_i),\\ \boldsymbol{\eta}_i = \text{softmax}(\mathbf{x}_i^T\boldsymbol{\omega}_1, \dots, \mathbf{x}_i^T\boldsymbol{\omega}_K),\end{aligned}$$

where f and F are the PDF and CDF of the normal distribution, δ_{ij} is as defined in Section 1.5 where $\delta_{ij} = 1$ when y_{ij} is censored and zero otherwise.

7.2.4 Identifiability

Since $\boldsymbol{\eta}_i$ lies on the simplex and must sum to one, we are trying to estimate K parameters with only $K-1$ degrees of freedom leading to an issue with identifiability, a necessary condition for the existence of consistent estimators (Hennig, 2000). The MoE model is not identifiable but it is generically identifiable, up to label switching, as the "set of non-identifiable parameters has zero measure" (Allman *et al.*, 2009). To make this issue explicit, consider Proposition 7.1 where we infinitely many weighting parameters yield the same allocation probabilities, that is $\boldsymbol{\eta}_i^* = \boldsymbol{\eta}_i$.

Proposition 7.1. *Suppose that for any arbitrary constant, $C \in \mathbb{R}$, we let*

$$\boldsymbol{\omega}_k^* = \boldsymbol{\omega}_k + C\mathbf{e}_j,$$

for all $k = 1, \dots, K$ where $\mathbf{e}_j$ is a unit vector where the j^{th} element is 1 and all other elements are zero. Intuitively, we are adding C to all elements in the j^{th} row of the weighting matrix $\Omega = (\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_K)$. It follows that the allocation probabilities produced in (7.5) are identical.

Proof. To show $\boldsymbol{\eta}_i = \boldsymbol{\eta}_i^*$ it suffices to show that any k^{th} element is equal, that is,

$$\begin{aligned}
\eta_{ik}^* &= \frac{\exp(\mathbf{x}_i^T \boldsymbol{\omega}_k^*)}{\sum_{k'=1}^{k} \exp(\mathbf{x}_i^T \boldsymbol{\omega}_{k'}^*)} \\
&= \frac{\exp(\mathbf{x}_i^T \boldsymbol{\omega}_k + Cx_{ij})}{\sum_{k'=1}^{k} \exp(\mathbf{x}_i^T \boldsymbol{\omega}_{k'} + Cx_{ij})} \\
&= \frac{\exp(\mathbf{x}_i^T \boldsymbol{\omega}_k)}{\sum_{k'=1}^{k} \exp(\mathbf{x}_i^T \boldsymbol{\omega}_{k'})} \cdot \frac{\exp(Cx_{ij})}{\exp(Cx_{ij})} \\
&= \eta_{ik}.
\end{aligned}$$

□

We consider two approaches to solve the identifiability problem with the weighting coefficients which will help improve estimation. The first is a corner constraint as employed by Kruschke (2014) where for some choice of reference *pivot*, $r \in \{1, \dots, K\}$, we reparameterise $\boldsymbol{\omega}_k^* = \boldsymbol{\omega}_k - \boldsymbol{\omega}_r$. This has the effect that $\boldsymbol{\omega}_r$ is a vector of zeros and all interpretations of the weighting coefficients are relative to that reference component. Alternatively, one can impose a sum to zero constraint on each row of the matrix Ω. One way to achieve this is by subtracting the means of the corresponding rows leading to a different interpretation where each weighting coefficient is relative to a mean value.

A distinction between these two constraints that will become relevant in Section 7.4 is the property of *independence from irrelevant attributes* (Kruschke, 2014). In this context, independence from irrelevant attributes asserts the ratio of probabilities between two components is unchanged in the presence of more components. This property is desirable when our approach to choosing the number of components involves fitting the model at various candidate values of K. When using a corner constraint, the weighting variable is transformed into $\omega_{jk} - \omega_{jr}$ for some components k and reference component r and this quantity remains independent of ω_{jk^*} for $k^* \notin \{k, r\}$. Using the sum to zero constraint violates this property as the weighting variable is now transformed into $\omega_{jk} - \bar{\omega}_j^T$, where $\bar{\omega}_j^T$ would change in the presence of more components.

In a groundwater monitoring application, our interpretation of each component is a different possible phase of the system with very little meaning given to the average phase. For this reason and the desirable independence from irrelevant attributes property, we adopt the former approach and set $\boldsymbol{\omega}_1 = \mathbf{0}$ as the reference pivot. One could utilise a baseline phase representing an uncontaminated system at equilibrium as the pivot component, but this is a non-trivial task specific to each groundwater monitoring site.

7.2.5 Prior

For all regression coefficients in $\tilde{B}$, we assume independent normal distributions with mean 0 and specified precision *a priori*. This is consistent with previous models but we could manufacture a more precise prior if there was more information on the underlying assumed components that represent "phases" of the groundwater network. A similar reasoning produced independent and identical gamma prior distributions for the analyte-specific and component-specific precision parameters T.

Weighting coefficients, Ω, take on a similar role to $\tilde{B}$ as the stochastic component of a linear predictor and so we assume, *a priori*, independent and identical normals. However, the prior certainty of these distributions will commonly exceed the prior certainty of the regression coefficients since prior predictive checks (Stan Development Team, 2023) highlighted weighting coefficients greater in magnitude than 1 tends to dominate component probabilities. Our decision of reducing the prior variance for the weighting coefficients is reflective of our prior beliefs that the system is controlled by multiple phases. This is seen in Section 7.5 where Ω is an order of magnitude less than $\tilde{B}$ to ensure a comprehensive simulation study that contained equal data in each component.

7.2.6 Bayesian Inference

Once again, we adopt a Bayesian approach to inference and sample from the posterior distribution via an MCMC algorithm, as described in Appendix B. The particular MCMC algorithm is slice sampling (Neal, 2003) that is enacted through the use of the JAGS software (Plummer *et al.*, 2003) where we specify all distributions in (7.5) and all requisite prior distributions.

7.3 Label Switching

When dealing with nominal data such as $\mathbf{z} = (z_1, \dots, z_n)$ we assign numerical labels arbitrarily, say $1, 2, \dots, K$; trivially, a perfectly valid alternative configuration exists if we were to choose different labels. There are $K!$ permutations of this labelling configuration that would yield the same results and the corollary is that the likelihood has $K!$ modes, invariant to any relabelling. Within a Bayesian context, using a symmetric prior extends this issue to the posterior distribution. This leads to the *label*

switching problem within our MCMC algorithm where the algorithm may converge to one likelihood and **switch** to another of the equally valid $K!-1$ likelihoods within that chain or even between chains due to different initial values (Redner & Walker, 1984). If our interest was purely in prediction, using posterior predictive densities, there would be no problem due to invariance to these label changes. However, our interests in estimating the component-specific parameters, $\tilde{B}, \Omega, T$, require a solution since posterior draws of these parameters are vastly affected by label changes (Redner & Walker, 1984).

7.3.1 Existing Solutions

One such solution to the label switching problem is to impose an ordering constraint, with the reasoning being if the prior is asymmetric, the posterior should be also. However this method can lead to unsatisfactory results, as shown by Stephens (2000), where multimodality persists even after imposing an ordering constraint. Further consider that in the multivariate MoE case, ordering multidimensional parameters is not trivial.

Stephens (2000) follows up criticism of ordering constraints with an alternative decision theoretic approach. Defining a loss function that takes some relabelling action and true parameters and outputs a value that is representative of the loss incurred allows Stephens to rephrase the problem as a minimisation problem. Advantageous to this approach is the generality and wide range of problems this can be applied to including latent variables (Boys & Henderson, 2002). In particular, this motivated further methods including Boys & Henderson (2002) where a loss function represents dissimilarity to a rolling marginal posterior mode that is updated each MCMC iteration. For comparison of similar methods, the R software package `label.switching` (Papastamoulis, 2016) allows users to apply a range of solutions including Stephens

(2000); Papastamoulis (2016) also applies each solution to examples of label switching highlighting strengths and weaknesses of each.

A more automatic approach (Frühwirth-Schnatter, 2006) applies some form of K-means clustering to the MCMC output. This may be done sequentially allowing for on-line implementation (Celeux, 1998) or enacted once all raw MCMC iterations are found (Zens, 2019). One benefit of such a method is to assist in determining if the model is overfitting due to a poor choice of K. Section 3.7.7 of Frühwirth-Schnatter (2006) asserts that if the relabelling configurations produced from such an algorithm, say $\nu_m(k)$ for MCMC iterations $m = 1, \dots, M$, are not a permutation of $\{1, \dots, K\}$ this may be indicative of overfitting the number of components. That is,

$$\sum_{k=1}^{K} \nu_m(k) = \frac{K(K+1)}{2},$$

does not hold for a substantial fraction of iterations.

Rephrasing label switching as a by-product of these MCMC algorithms instead of a problem leads to an interesting solution where instead of enforcing convergence of a single posterior mode, one aims to explore the whole posterior; this is difficult for random walk algorithms (Jasra *et al.*, 2005). Neal (1996) proposes *tempered transitions* to jump between these distant modes of the multi-modal posterior distribution. Tempering is possible with JAGS (Plummer *et al.*, 2003) although we do not leverage this functionality in this thesis. Further discussion of tempering can be found in Neal (1996), Jasra *et al.* (2005).

Care must be taken when applying these methods in a MoE setting due to the corner constraint for reasons to be explained in Section 7.3.2. For implementation of a relabelling technique applied to a MoE model, see Zens (2019) where a random permutation is introduced to force balanced label switching and draws that do not create a unique permutation of $\{1, \dots, K\}$ are discarded. Both Zens (2019) and

Neal (1996) are interested in fully exploring the posterior sample space whereas our method, defined in Section 7.3.4, is mainly interested in estimation and convergence around a single posterior mode.

7.3.2 Further Consideration

In addition to dealing with the label switching problem that is present in a finite mixture model or even a hidden Markov model (Boys & Henderson, 2002), there is the further issue that our weighting parameters are invariant to translation as shown in Section 7.2.4. To see why this complicates the issue, suppose we are estimating

$$\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3) = (0, 2, -2),$$

where the pivot is the first label so $\omega_1 = 0$. If the labels 1 and 3 were to switch we would be estimating $\boldsymbol{\omega} = (-2, 2, 0)$ but due to the presence of the pivot, an estimate, say from some MCMC algorithm, would appear to be $\hat{\boldsymbol{\omega}} = (0, 4, 2)$.

Even in this simple example, a hypothetical scenario could be that the model fitting algorithm estimates 0, 2 and -2 for the first $M/2$ posterior samples and then estimates 0, 4 and 2 for the latter $M/2$ posterior samples. This would cause severe issues for the K-means post-processing as suggested by Frühwirth-Schnatter (2006) and used in Zens (2019).

7.3.3 Example

As an illustrative example of label switching, we simulate data from (7.5) with $K = 3$ assumed and do not apply any censoring. In the MCMC analysis, we notice some odd behaviour in the trace-plots that indicate the chains may not have converged. This is most clearly seen in Figure 7.1 where each column and colour represents a

different component and each row is a different simulated explanatory variable or the intercept term. A change-point, specifically a change in series variance, can be observed at around the same iteration index for the first two components.

Figure 7.1: Marginal posterior trace-plots of regression coefficients. Label switching example, MoE, $K = 3$.

Further supporting the hypothesis that label switching has occurred is the diagnostic provided by the allocation raster visualisation, Figure 7.2. Since the allocation variable data is discrete, a trace-plot would not be very useful and so we instead create a raster image where the rows are the unique sample identifiers, columns are MCMC iterations and the colour of each equally sized rectangle encodes the discrete allocation data. Figure 7.2 indicated label switching but also conveys the uncertainty around each allocation since any observation that is more "uncertain" will exhibit a more varied row of colours representing the components.

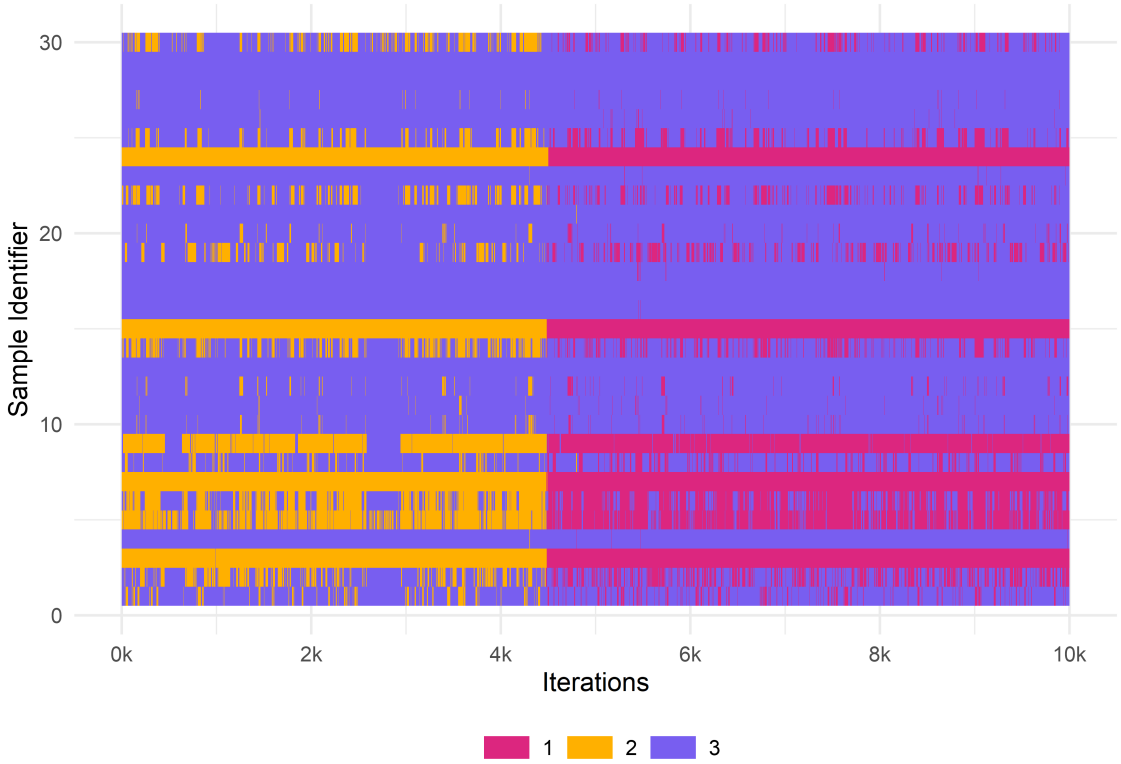

Figure 7.2: Component allocation plot. Label switching example, MoE, $K = 3$.

7.3.4 Proposed Solution

We propose a modified version of the relabelling method described in Boys & Henderson (2002) that allows for the corner constraint producing a label switching solution that accommodates the MoE model.

Suppose we fit our model to the data and produce raw MCMC output denoted as $\mathbf{z}^{(m)}, B^{(m)}, \Omega^{(m)}$ for iterations $m = 1, \dots, M$. For some chosen amount of burn-in iterations, m_0, where ideally no label switching occurs, we apply Algorithm 2 to iteratively reverse any label switching by comparing with the marginal posterior mode up to that point.

Algorithm 2 Relabelling algorithm for MoE models

1: **for** $m = m_0 + 1, \dots, M$ **do**

2: set $\hat{\mathbf{z}}^{(m-1)}$ equal to the marginal posterior mode, that is, for $i = 1, \dots, n$,

$$\hat{z}_i^{(m-1)} = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} \sum_{m^*=1}^{m-1} \mathbb{I}\left(z_i^{(m^*)} = k\right);$$

3: choose ν_m from the set of all $K!$ permutations to minimise 'disagreement',

$$D = -\sum_{i=1}^{n} \mathbb{I}\left(\nu_m(z_i^{(m)}) = \hat{z}_i^{(m-1)}\right);$$

4: apply permutation ν_m to output $\mathbf{z}^{(m)}, \Omega^{(m)}, B^{(m)}$;

5: translate $\Omega^{(m)}$ by subtracting modified reference column vector $\boldsymbol{\omega}_r^{(m)}$.

6: **end for**

By applying this algorithm to the example introduced in Section 7.3.3, we can revisit the trace-plot and diagnostic behaviour to verify the algorithm has had the intended effect. As such, we recreate Figures 7.1 and 7.2 in Figures 7.3 and 7.4 respectively.

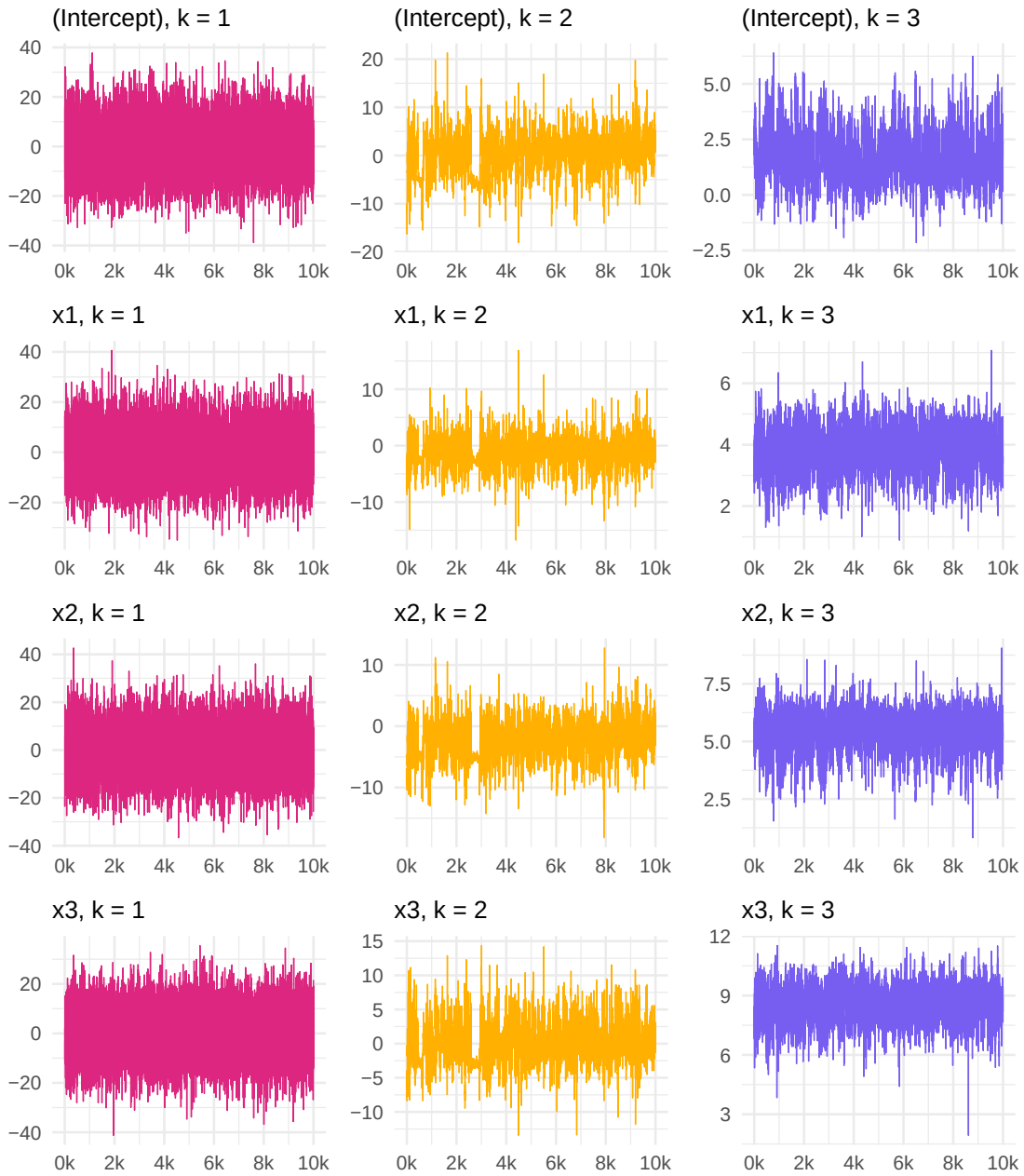

Figure 7.3: Marginal posterior trace-plots of regression coefficients, after relabelling. Label switching example, MoE, $K = 3$.

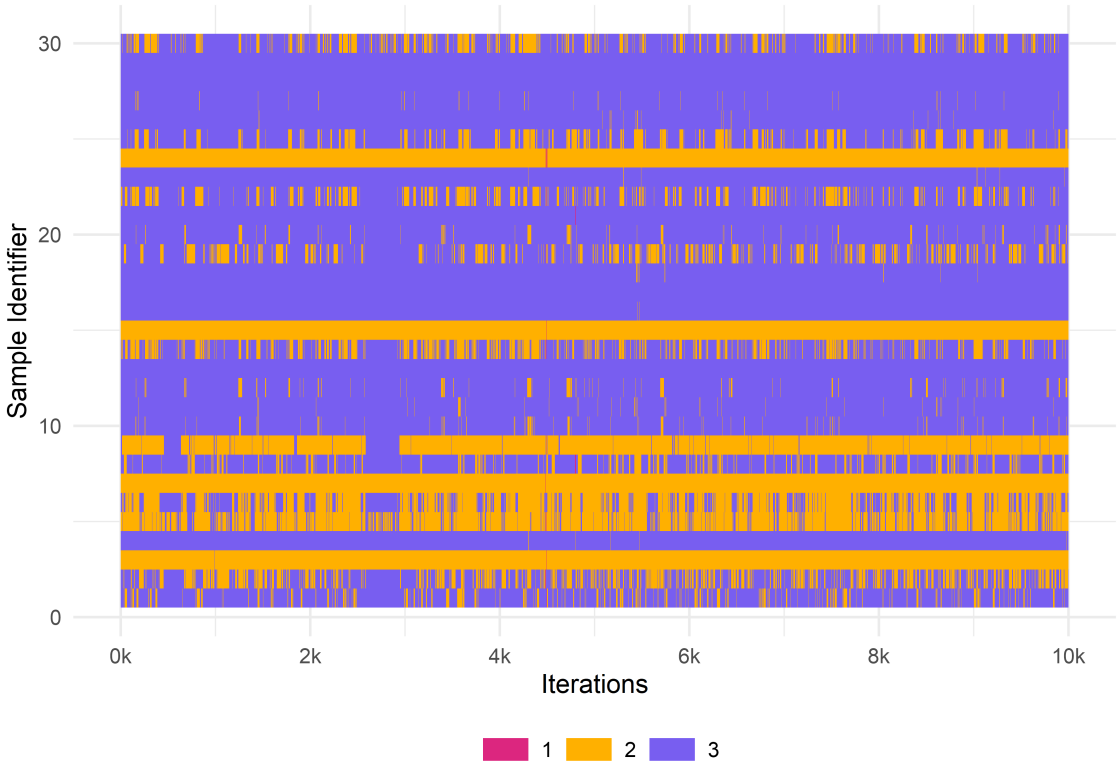

Figure 7.4: Component allocation plot, after relabelling. Label switching example, MoE, $K = 3$.

While Figure 7.4 highlights that the first component is underutilised and this model may be overfitted when $K = 3$ is assumed, we see that the algorithm has corrected any label switching issues and we are now able to compute posterior means of the regression coefficients. Not all issues are fixed by this procedure and one can see an inconsistency in the output for the second component near the 500^{th} and $2,500^{th}$ MCMC samples. These are indicative of potential bimodality, or even multimodality, within our simulated example which highlights a key complication where the MCMC algorithm may converge to none of the $K!$ posterior global modes and instead converge to a local maxima.

7.4 Choice of K

So far we have assumed that the number of components, K, is fixed and known. In any groundwater monitoring application there exist numerous possible reactions dominating the system and affecting the phase, meaning it is not a straightforward task to let K represent the total number of substantial system phases.

7.4.1 Existing Methods

Ideally, one would jointly infer the number of components and the model parameters in a single model, but this is difficult to achieve. A rigorous version of the expectation-maximisation (EM) algorithm has been developed that allows penalised maximum likelihood estimation in an unsupervised context without the requirement of knowing the number of components beforehand (Chamroukhi, 2016). The upshot of this methodology is the decreased sensitivity to initialisation when compared to a more standard EM algorithm. Within a Bayesian approach, one can leverage a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm that allows chains to traverse over parameter spaces and thus relax the requirement that K is known *a priori* (Richardson & Green, 1997). Practically, the RJMCMC algorithm allows two key steps:

1. splitting one component into two, or combining two in one;
2. the birth and death of an empty component,

where a superfluous component is declared empty when associated with too few observations or it is too similar to another component. RJMCMC acts upon the fact that when the number of fitted components exceeds the true number of components, as in Section 7.5, either parameter estimates will converge to similar values or the allocation probability will decrease towards zero (Frühwirth-Schnatter *et al.*, 2019).

In a Bayesian framework, there is no need to assume that K is known *a priori* and we can relax this assumption by making use of a Dirichlet process prior (Ferguson, 1973). A Dirichlet process, parameterised by a base distribution and concentration parameter, produces realisations that are also probability distributions. The advantage of this non-parametric approach is that one can apply a stick-breaking approach where each successive component takes some proportion, less than 1, from a unit stick for countably infinitely many components (Gelman *et al.*, 1995). Since each realisation is a random distribution with associated probability density function, we have generated a generalisation of (7.2) for infinitely many components. There are other approaches that we do not cover here for the sake of brevity.

Our approach is to treat the total number of components, K, as a hyperparameter to be "tuned". That is, fit the model to several candidate values of K and use model comparison metrics, such as those introduced in Section 1.7, to decide on the 'best' number of components (Huynh, 2019; Gormley & Frühwirth-Schnatter, 2019). Frühwirth-Schnatter *et al.* (2019) provides a thorough discussion of various metrics such as AIC, BIC and the deviance information criterion (DIC) used for this purpose. Candidate values can be obtained by non-parametric methods such as K-means clustering, preliminary analysis or fitting incrementally more components until signs of overfitting are found.

7.4.2 Signs of Overfitting

When choosing a value for K it is important to understand the interpretation of our weighting parameters and how these parameters will behave when there are too many components. Suppose, without loss of generality, there is only one explanatory variable, and no intercept is used. We choose our reference cluster to be $r = 1$ and, without knowing the true value of K, we wrongly assume $M > K$ components.

The true value of our weighting matrix is therefore

$$\Omega = \begin{pmatrix} 0 & \omega_2 - \omega_1 & \dots & \omega_K - \omega_1 \end{pmatrix}.$$

For $m = K+1, \dots, M$, it must be the case that $\eta_{im} = \Pr(z_i = m) = 0$ since these components are *empty*, but our model will not allow zero probabilities since the exponential of any finite value can never equal zero. Instead, the allocation probability of some arbitrary empty component, $\eta_{im} \to 0$, tends to zero as the respective linear predictor, $x_{i1}\omega_m^* \to -\infty$, becomes more negative, where $\omega_m^* = \omega_m - \omega_1$. Hence, an *empty* component can be inferred from extremely negative weighting coefficients. The argument here requires the property of independence from irrelevant attributes as described in Section 7.2.4. A negative weighting coefficient implies a unit increase in x_{i1} makes the event $z_i = m$ less likely, relative to the event $z_i = 1$; trivially, the converse holds true.

Further methods of empty component identification such as prior-to-posterior updates (Section 7.5.1) or convergence to similar values (Frühwirth-Schnatter *et al.*, 2019) should also be considered when assessing the output of these models.

7.4.3 Proposed Solution

We implement a quantitative approach that uses the model comparison metrics introduced in Section 1.7. If we were to assume our data to be predicted is a true *test* dataset, then our out-of-sample data would be completely hidden from the model decision making process and we would base our metric on the in-sample data only. The LPD based on in-sample data is the observed data log likelihood and would choose the most complex model meaning it is unreliable and will likely lead to overfitting; we expect the LPD score for each model to increase at a declining rate as more components are added. To account for increasing model complexity, we

make use of the PSIS measure as an approximation to leave-one-out cross validation.

Alternatively, one could view the data to be predicted as a *validation* dataset which can be used to choose an optimal value of some tuning parameter, in this case K. In this context we can use the LPD as a log-scoring rule (Gneiting & Raftery, 2007) to asses each model's predictive power and choose the optimal value of K accordingly. Further work could explore the impact of choosing a training-validation-test data split on reliably estimating the correct number of components.

These two approaches may not always agree, as shown in Section 7.5.1. Hence, we opt for a pragmatic approach to this decision problem where the implicit multi-attribute utility function depends not only on a model's predictive performance, but also computational time requirements and model interpretability.

7.5 Simulation Study

To demonstrate how the mixture of experts (MoE) model responds to various circumstances we investigate with multiple simulation studies where we simulate data from a univariate MoE model and fit using the same model. To simulate the data, we must choose a total number of components, K; we assume $K = 3$ with the caveat that we cannot use this fact in the model fitting stage and must instead use methods described in Section 7.4 to choose a suitable total number of components. We present two datasets of size $n_s = 1000$, where parameters B, Ω and allocation variables $\mathbf{z}$ are the same each time but we produce a different signal-to-noise ratio by setting a common precision parameter, $\tau_{1k} = 4$, or $\tau_{1k} = 36$,for all k. In this instance, we arbitrarily set

$$\boldsymbol{\beta}_{11} = \begin{pmatrix} 9 & 2 & 4 \end{pmatrix}^T, \qquad \boldsymbol{\beta}_{12} = \begin{pmatrix} 6 & 3 & 7 \end{pmatrix}^T, \qquad \boldsymbol{\beta}_{13} = \begin{pmatrix} 5 & 1 & 8 \end{pmatrix}^T,$$

where $\boldsymbol{\beta}_{1k}$ represents the regression coefficients of component k. We arbitrarily set Ω on a different scale to ensure each component has a reasonable sample size as shown in Table 7.1, that is,

$$\Omega = \begin{pmatrix} 0.0 & 0.6 & 0.0 \\ 0.0 & -0.2 & -0.4 \\ 0.0 & 1.2 & 0.5 \end{pmatrix}.$$

Data is then artificially censored at the 15% level, that is, all values less than the 15^{th} percentile are taken to be left-censored with a common detection limit set to the 15^{th} percentile. Note that $\boldsymbol{\tau}$ is kept constant across components in the simulation of these data, whereas the model to be fit allows for component-specific precisions.

7.5.1 Identifying K

As K is unknown at the model fitting stage, we fit candidate models with $K = 1, \ldots, 5$ to both datasets. Consider the fit where we erroneously assumed $K = 5$; two components are superfluous and our posterior beliefs should reflect this. Firstly, Figure 7.5 shows that the regression coefficients for all predictors in the final two components have a larger posterior variance relative to the other three components and, moreover, they tend to more closely resemble draws from the prior. One possible explanation is that there is a smaller volume of data in these superfluous components affecting the posterior estimation, although, a small prior to posterior change could be due to other issues with estimation.

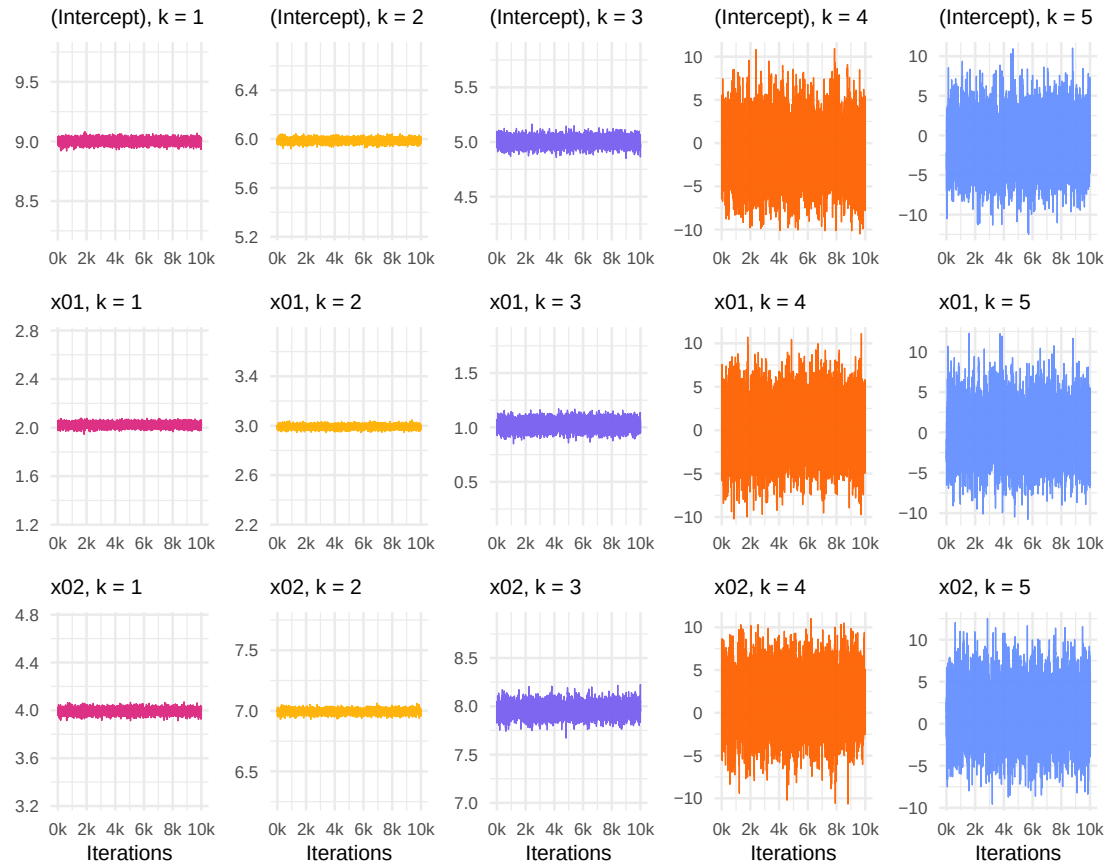

Figure 7.5: Marginal posterior trace-plots of regression coefficients. Simulated example, high precision, MoE, $K = 5$.

Further supporting the claim that this data is overfitted and the extra components may not be needed is the posterior allocation probabilities. Suppose we define our estimate of the allocation probability for observation i based on M posterior samples as

$$\hat{\eta}_{ik} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}(z_i^{(m)} = k),$$

for $k = 1, \dots, K$ and then let

$$\begin{aligned} \hat{\boldsymbol{\eta}}_k &= \frac{1}{n_s} \sum_{i=1}^{n_s} \hat{\eta}_{ik} \\ &= \frac{1}{n_s} \sum_{i=1}^{n_s} \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}(z_i^{(m)} = k), \end{aligned}$$

denote the estimated posterior total allocation probabilities for all observations $i = 1, \ldots, n_s$. Table 7.1 reflects this quantity when estimated from our posterior draws with the true latent variables that we store during the simulation stage. We see that the extra components, $K > 3$ have small probabilities with very few data associating with these components which explains the poor inference of each superfluous component; the remaining probabilities are close to the truth in this particular simulation study.

Component Index	Estimated Probability	True Probability
$\boldsymbol{\eta}_1$	28.36%	28.60%
$\boldsymbol{\eta}_2$	51.49%	49.40%
$\boldsymbol{\eta}_3$	19.07%	22.00%
$\boldsymbol{\eta}_4$	0.52%	0.00%
$\boldsymbol{\eta}_5$	0.56%	0.00%

Table 7.1: Estimated posterior total allocation probabilities. Simulated example, high precision, MoE, $K = 5$.

When assessing model comparison metrics to choose an appropriate value of K, we notice distinct patterns depending on the precision, relative to the regression coefficients. Models with a higher precision, $\tau_{jk} = 36$, are compared in Table 7.2 and we see that the best model has 3 components. While using more components than necessary results in a 'worse' model, it is preferred to fitting too few components as evidenced by the poor score of $K = 2$ and $K = 1$. We see similar scores between models for the low signal-to-noise model shown in Table 7.3 implying that adding more components while increasing model complexity, is not improving model performance.

	Δ LPD (SE)	Δ WAIC (SE)	Δ PSIS (SE)
K = 1	-716.1 (21.0)	-709.0 (21.4)	-708.9 (21.4)
K = 2	-377.9 (23.9)	-377.8 (24.4)	-377.8 (24.4)
K = 3	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
K = 4	-3.2 (0.4)	-3.3 (0.4)	-3.3 (0.4)
K = 5	-8.2 (0.6)	-8.2 (0.6)	-8.2 (0.6)

Table 7.2: Model comparison metrics based on in-sample data. Simulated example, high precision, MoE, $K = 5$.

	Δ LPD (SE)	Δ WAIC (SE)	Δ PSIS (SE)
K = 1	-357.3 (18.0)	-346.5 (18.6)	-346.4 (18.6)
K = 2	-98.4 (17.5)	-97.2 (19.0)	-97.1 (18.9)
K = 3	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
K = 4	-2.8 (0.4)	-3.6 (0.5)	-3.6 (0.5)
K = 5	-5.7 (0.8)	-6.5 (0.8)	-6.5 (0.8)

Table 7.3: Model comparison metrics based on in-sample data. Simulated example, low precision, MoE, $K = 5$.

Our simulations suggest that even in an idealised scenario with all model assumptions met, noise dominating the regression coefficients may lead to poor mixing and could lead to difficulty recovering the true value of K. Overfitting may cause poor mixing and unreliable posterior samples, but these symptoms can also be caused by other factors such as multimodality. It is therefore important to leverage further diagnostics to ensure this is not due to other issues with the model assumptions or data.

7.6 Groundwater Application

One advantage of modelling each analyte as conditionally independent, where independence is conditional on the latent variable z_i for all observations, is that we need only choose the number of components once per prediction scenario introduced in Section 1.6. Here, we only present results from the LMWO prediction scenario since the analyses appear very similar and little to no benefit is attained from supplying historical data for the wells to be predicted within this model framework, unlike Chapter 8. As with other models, we run the algorithm from 4 different initial values to produce 4 chains with $5,000$ burn-in samples discarded and $10,000$ saved posterior samples. Our prior information as specified in Section 7.2.5 allows us to use the same prior information as Section 5.5 for all regression coefficients and precision parameters although an argument could be made to assign different prior information per component. Given our prior assumption that the weighting coefficients are an order of magnitude smaller than the regression coefficients, we assume $\omega_{jk} \sim N(0, 10^{-1})$ for all non-pivot components $k = 2, \ldots, K$.

7.6.1 Choosing K

Candidate models for several appropriate choices of number of components up to $K = 4$ are fit, including the special case $K = 1$ that was shown in Chapter 5. No label switching occurred in these fits and so no relabelling algorithm was employed. Each model is fit with 4 chains, as described in Appendix B, with different initial values drawn from the prior to verify convergence and highlight any issues such as algorithm sensitivity to starting values.

For both prediction scenarios, LMWO and hold-out future, when we assume $K = 2$ we find that the MCMC algorithm converges on two distinct posterior modes sug-

gesting a potential problem with multimodality. Interestingly, we found no evidence of the sampler switching between these modes and instead found convergence was dependent on the starting values of the model parameters. We can rule out these issues as label switching, since the log likelihood is different between chains even though it is invariant to relabelling. The reasoning for poor consistency could be due to the suspected low signal-to-noise ratio of these groundwater data from site A. As the posterior predictive draws are indistinguishable between the two posterior modes, we opt for the pragmatic approach of presenting the model fit with the best log score given by calculating the LPD on the out-of-sample data. Multimodality becomes an increasingly prevalent problem as the assumed number of components increases with evidence of potential multimodality seen in these analyses for $K \geq 3$; various modes produce similar posterior predictive samples and are likely to be a product of the non-identifiability of the MoE models. As such, we find sufficient evidence that assuming more than $K = 4$ components leads to poor parameter estimation with no benefit to prediction and so these are not considered as viable models.

Qualitative methods to choose the optimal number of components such as visualising component allocations and identifying "empty" components that are similar to other components or contain very few observations are difficult to apply to many models. Calculating posterior allocation probabilities, as done in Section 7.5.1, can also help identify such "empty" components quantitatively. Further complicating this process is the seemingly random relabelling between chains that are caused by different starting values. Instead, we initially apply the quantitative method of comparing the model metrics when applied to the in-sample and out-of-sample datasets.

	Δ LPD (SE)	Δ WAIC (SE)	Δ PSIS (SE)
K = 1	-868.6 (38.8)	-802.1 (39.9)	-801.7 (39.9)
K = 2	-276.7 (22.2)	-228.4 (23.3)	-228.0 (23.2)
K = 3	-94.5 (14.4)	-69.2 (15.3)	-68.9 (15.2)
K = 4	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)

Table 7.4: Model comparison metrics based on in-sample data. LMWO, MoE, $K = 1, ..., 4$.

	Δ LPD (SE)
K = 1	-34.7 (8.6)
K = 2	0.0 (0.0)
K = 3	-23.5 (7.3)
K = 4	-6.8 (10.3)

Table 7.5: Model comparison metrics based on out-of-sample data. LMWO, MoE, $K = 1, ..., 4$.

Table 7.4 and Table 7.5 show model comparison metrics log pointwise predictive density (LPD), widely applicable information criterion (WAIC) and Pareto-smoothed importance sampling (PSIS) based on the pointwise log likelihood for both analytes. In-sample metrics show that including more components will improve the fit with diminishing returns as the assumed number of components K increases. LPD is expected to increase as the number of total components increases because the model is more overfitted each time; it is surprising that WAIC and PSIS follow a similar trend as these metrics take into account the number of parameters in the model. We hypothesise that there is an optimal in-sample number of components that leads to

the 'best' WAIC and PSIS where adding more components will not increase LPD enough to offset the increase in model complexity.

However, we opt to not fit more than $K = 4$ components based on the out-of-sample LPD scores and increasingly poor chain mixing to be discussed. The combination of which suggests that adding more components to these models may improve in-sample model comparison metrics without improving the models' predictive performance on previously unseen data. In the calculation of the PSIS metric, a diagnostic corresponding to an estimated shape parameter, $\hat{k}$, as defined in Section 1.7 is produced (Vehtari *et al.*, 2017). All values of $\hat{k}$ are reliable for the censored regression special case where $K = 1$ and no diagnostic values exceed 1 for any candidate models. Fewer than 5 PSIS pointwise estimates are spurious for the model with $K = 4$ assumed number of components as the diagnostic values falls into $[0.5, 1)$. PSIS approximates leave-one-out (LOO) cross-validation and so problematic $\hat{k}_i$ values are indicative of very different marginal posteriors when all data is used and all data except observation i is used. It is expected that the PSIS estimate will only become less reliable as more components are included.

In summary, choosing the assumed number of components K can be viewed as a multi-attribute decision problem with no objective best answer. Moreover, the qualitative appearance of our predictions and quantitative LPD scores indicate that each models' predictive performance on previously unseen data is fairly insensitive to this choice. Therefore, we opt for the simpler model in terms of computational speed and number of parameters with $K = 2$ assumed components over $K = 4$ assumed components; we avoid the model where $K = 3$ components are assumed due to poor mixing and relatively low LPD (OOS) score. Comparison to the model with $K = 1$ will be included in Chapter 9 as part of a thesis-wide model comparison.

7.6.2 Regression Parameters

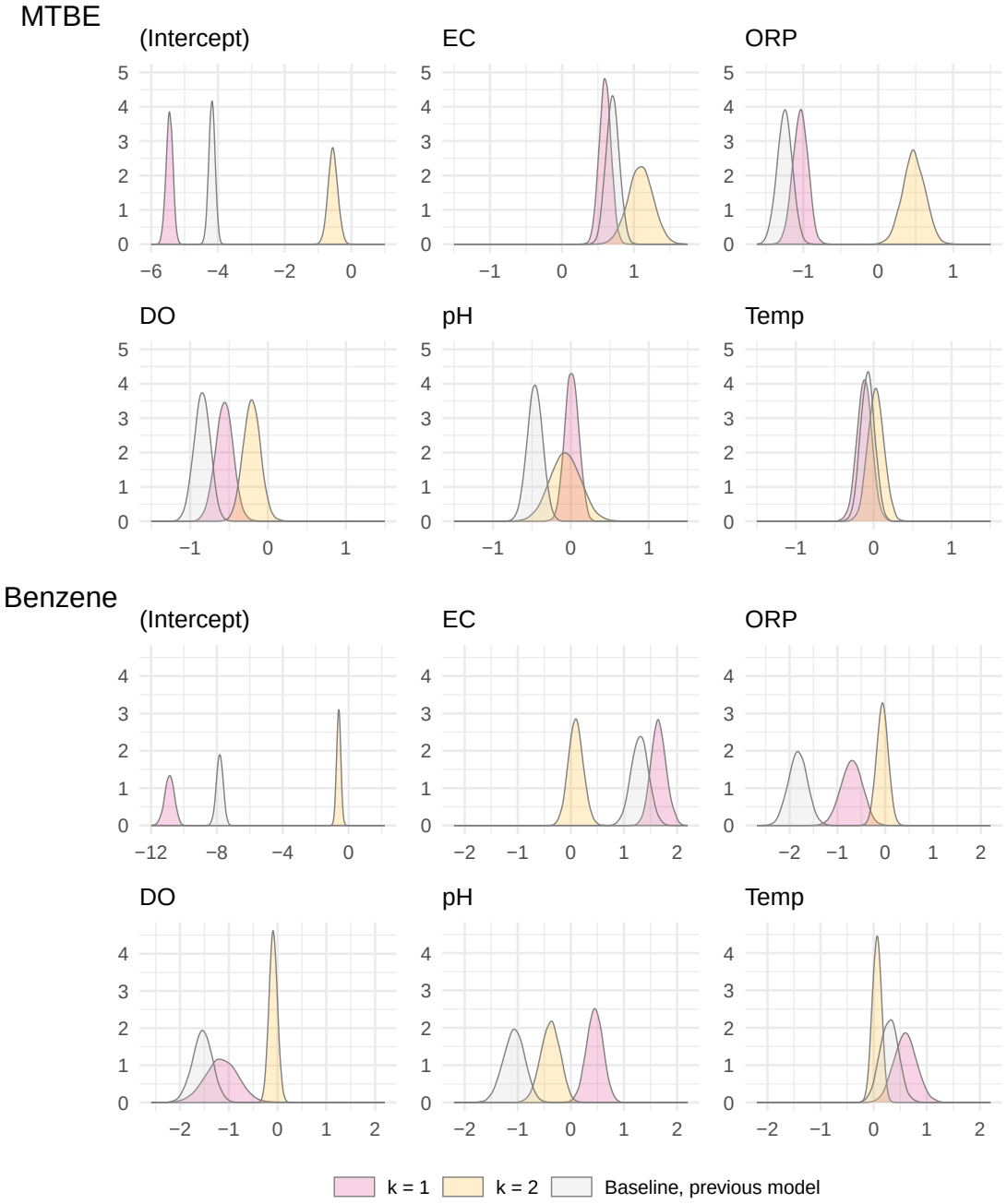

Figure 7.6: Marginal posterior density of regression coefficients, censored regression in grey for comparison. LMWO, MoE, $K = 2$.

Since we assume only two components, we visualise regression coefficients for each component with baseline results from Section 5.5.1 in the same panel for comparison, as shown in Figure 7.6. One might expect that the baseline posterior densities would tend to lie between the posterior densities where $K = 2$ was assumed but this is not the case as shown by the impact of pH on both MTBE and benzene concentrations. This is the case for the intercept coefficients where this model appears to favour placing extreme observations in an intercept only sub-model that would be the case if all regression coefficients for that component were 0 valued. Further work could investigate if a prior distribution that is not centred on 0 would produce alternative posterior densities or agree with our results that show no evidence that the predictors have an effect on benzene concentrations, conditional on being in component 2.

Electrical conductivity (EC) appears to have a positive impact on either analyte concentration with a posterior mean greater than 0.5 for all densities shown except the second component associated with benzene concentrations.

7.6.3 Weighting Parameters

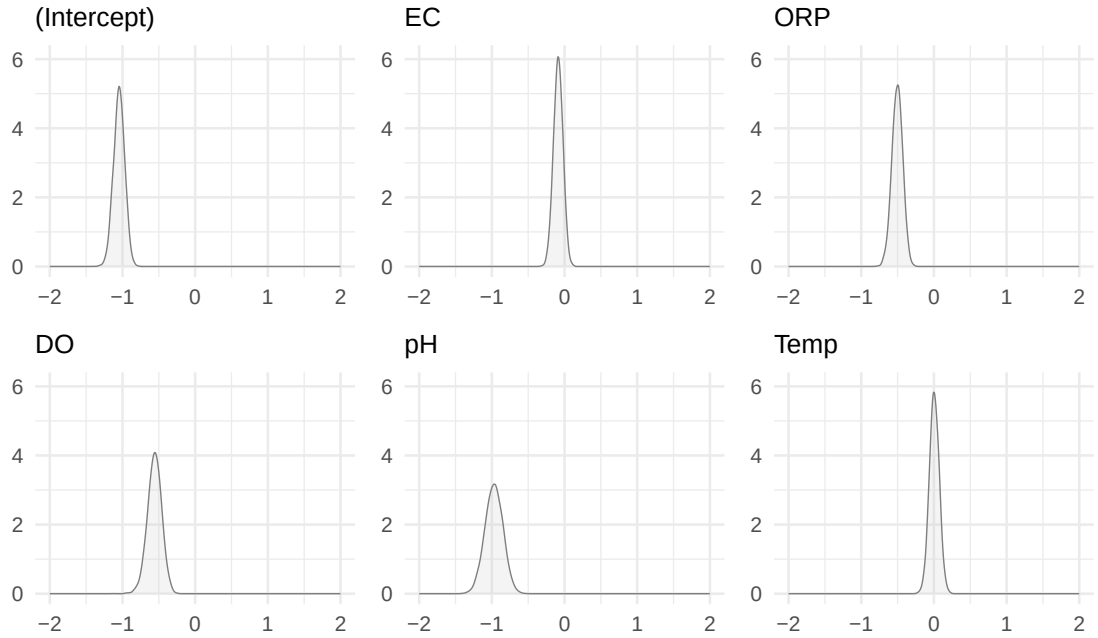

Figure 7.7: Marginal posterior density of weighting coefficients. LMWO, MoE, $K = 2$.

For each of the candidate models fit, we have taken the first label, $k = 1$ to be our pivot component to maintain identifiability. Therefore, only posterior draws associated with the $k = 2$ component are shown in Figure 7.7. The posterior mean of the intercept weighting coefficient is negative which suggests that when all normalised predictors are at the observed average and take the value of 0, it is more likely that observation will be in the first component. Combining this inference with Figure 7.6, where component 1 is shown to have a more negative intercept for both MTBE and benzene concentrations, reveals that average predictor values lead to low analyte concentrations. This is consistent with our initial assumption that many perimeter wells deflate the sample mean of each analyte concentration.

We observe that even though the regression parameter for electrical conductivity (EC) had a significant impact, the corresponding weighting coefficient shows very little prior to posterior update. Similarly, temperature presents a lack of a posterior effect in either set of predictors corroborating its low importance in the random forest models from Section 4.4.2.

Weighting estimates appear to stay within the $[-2, 2]$ range which is reasonable when comparing to the simulation study in Section 7.5. The posterior uncertainty of these weighting parameters appear to be consistent with EC and temperature showing more precise posterior distributions than the other predictors shown.

Interpretability of weighting coefficients is more difficult than the regression coefficients since the impact of a unit increase in ORP on the allocation probabilities is more difficult to parse. On the other hand, there is information in the sign of each weighting coefficient and we observe that oxidation reduction potential (ORP), dissolved oxygen (DO) and pH all have greater density for negative values and do not contain zero in their 95% credible interval. Hence, an increase in ORP, DO or pH decreases the estimated posterior probability of the corresponding observation being in the second component.

7.6.4 Precision Parameters

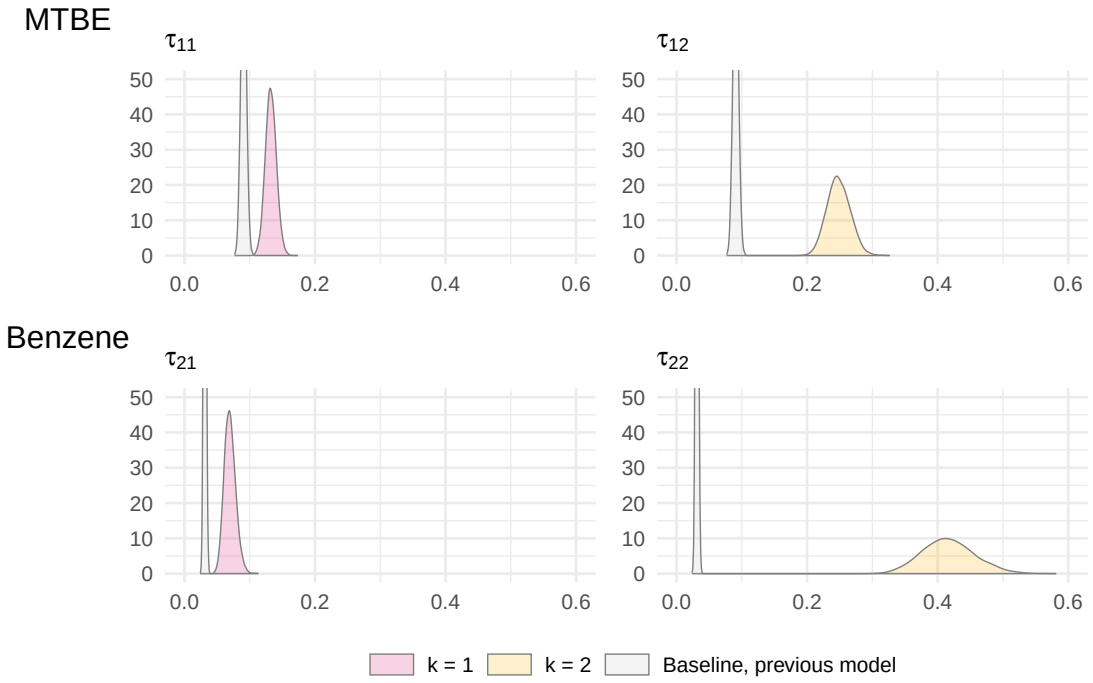

Figure 7.8: Marginal posterior density of precision parameters, censored regression in grey for comparison. LMWO, MoE, $K = 2$.

For the precision parameters, we observe posterior draws that are shown as estimated densities in Figure 7.8. Again, we see similarity in component behaviour due to the shared component allocations and high correlation of these dependent variables. In particular, the component $k = 1$ reveals an estimated precision that is greater than the baseline parameter from Chapter 5 but with greater posterior uncertainty. The component $k = 2$ shows an even bigger increase in posterior mean and also an increase in posterior uncertainty. While this may be due to a better fit to these data, we have anecdotally found a tendency to group censored observations when censoring is explicitly modelled in a mixture of experts model. In the extreme case where one component contains all left-censored observations, that component would have a smaller variance and therefore a larger precision.

7.6.5 Latent Variables

To better understand the latent structure we have observed affecting model parameters, we visualise posterior allocations by combining two types of plots:

1. stacked histogram representing estimated posterior allocation probabilities, per observation and ordered by analyte concentration;
2. ordered quantile plot of the data with censoring shown.

We composite these plots into a single visualisation per analyte by ordering the y-axis by observed concentrations or detection limits. Figure 7.9 for MTBE and Figure 7.10 for benzene, show that analyte concentrations are a key driver in the posterior distributions of the latent variables with many censored observations corresponding to the first component. That is, a censored observation is more likely to be in the first component, *a posteriori*, whereas a higher than average concentration is more likely to be in the second component. For uncensored concentrations closer to the median value, we see uncertainty and changes between components during the MCMC algorithm as evidenced by the probabilities closer to 0.5 than 0 or 1. While a trend is clear in Figure 7.9, it is not a perfect relationship with analyte concentrations implying that predictors are having a greater influence on component allocation probabilities for some high-leverage observations.

Posterior allocation probabilities, by definition, are conditional on all supplied data Y and X; a greater signal in weighting parameters could lead to model fits where the component allocations are more dependent on the predictors. Recreating these visualisations for models with more components assumed, say $K = 4$, reveals a similar trend of allocation probabilities depending on analyte concentrations. This is not too surprising as the posterior distributions for each components' intercept term are clearly separated, as shown in Figure 7.6.

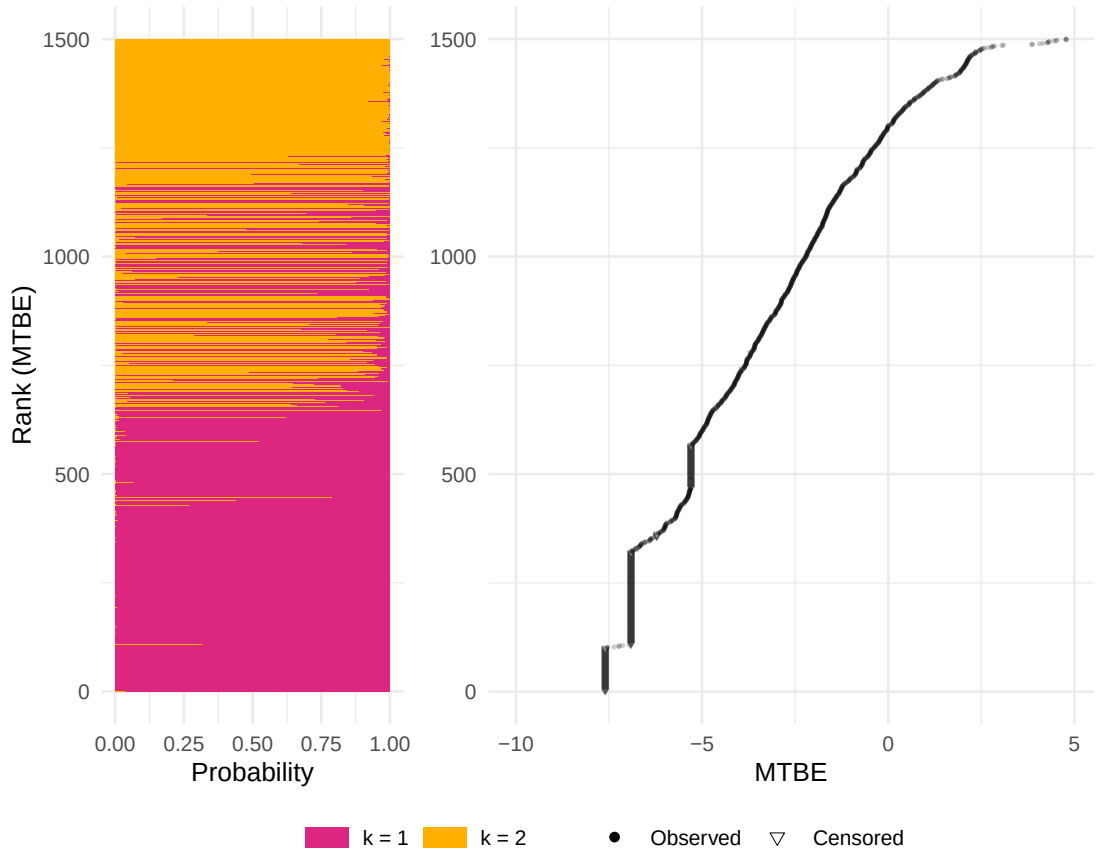

Figure 7.9: Estimated posterior probabilities (left) and quantile plots of MTBE data (right). MTBE, LMWO, MoE, $K = 2$.

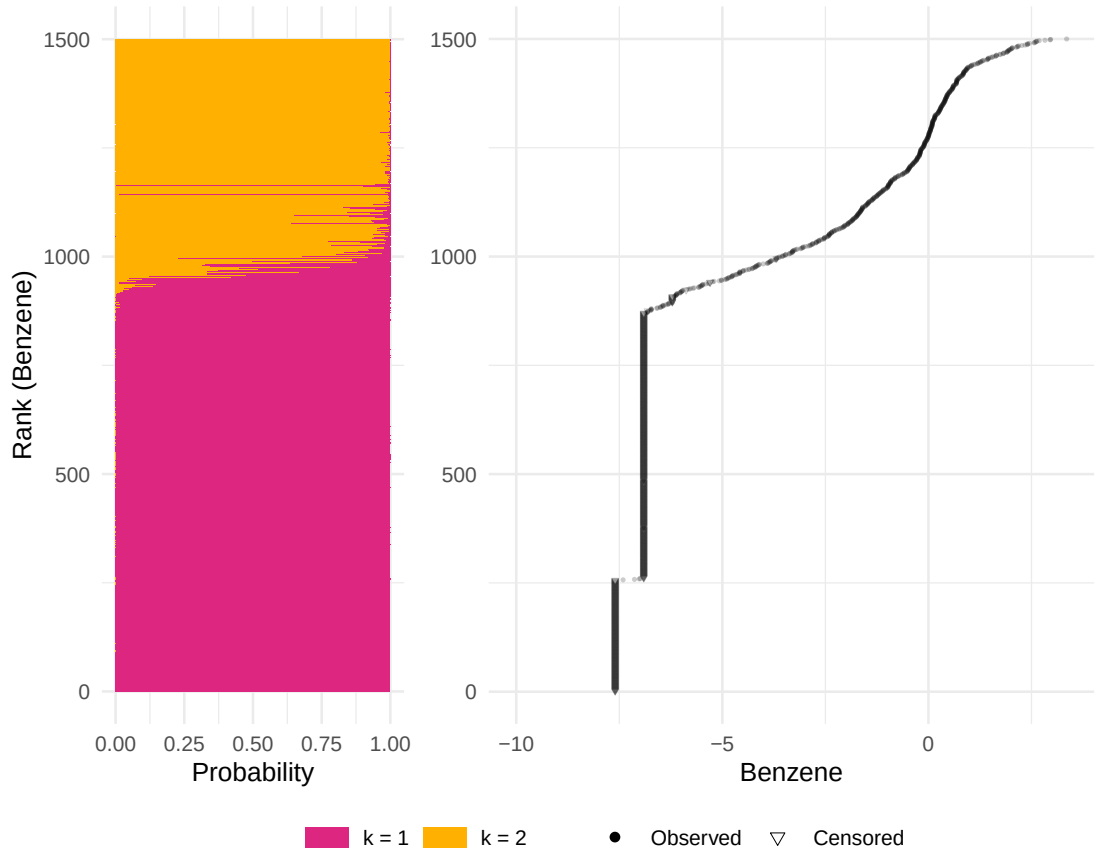

Figure 7.10: Estimated posterior probabilities (left) and quantile plots of benzene data (right). Benzene, LMWO, MoE, $K = 2$.

When applied to highly censored data, it is common to observe the role of at least one component to contain many censored observations leading to a potentially better model than $K = 1$. This is most noticeable in Figure 7.10, most censored data can be identified due to sorting by benzene concentration and corresponding allocation probabilities overwhelmingly favour $k = 1$. For $K = 2$, similar results were found in Terry *et al.* (2019) where a linear regression was improved upon by classifying the data into plume and non-plume observations. This explanation may not necessarily hold true for all data. For instance, we observe a mix in allocation variable posterior modes used in prediction of the "Focus" holdout well, shown in Figure 7.11.

7.6.6 Prediction

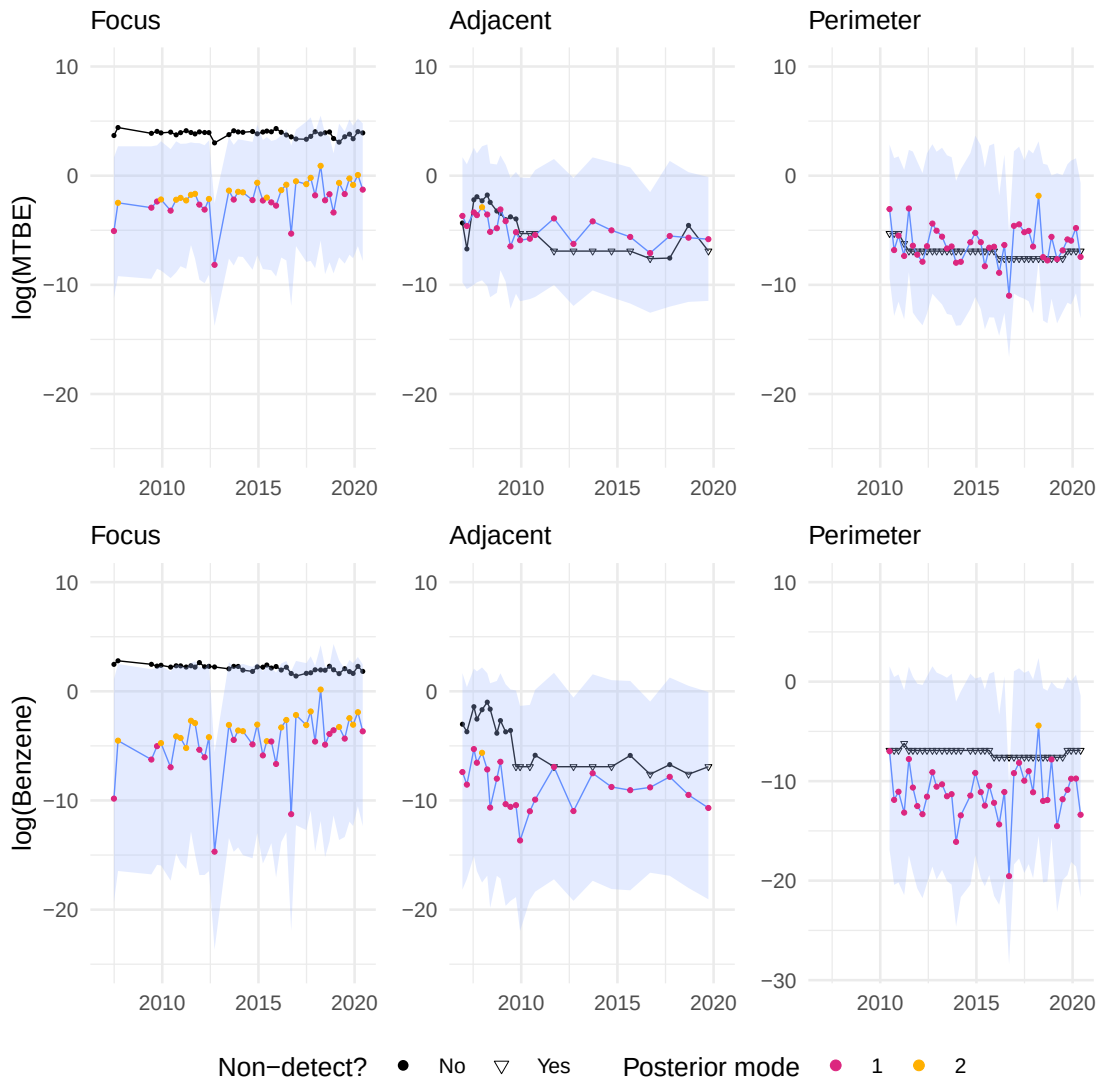

Figure 7.11: Predictions with comparison to truth in black. LMWO, MoE, $K = 2$.

Prediction of MTBE, as shown in Figure 7.11, is fairly accurate for the adjacent well and underestimates the edge-case focus well as expected; predictions for the perimeter well appear to be much more volatile than the measured values maintaining an average at the common detection limit but not below it. The MoE extension has somewhat improved the outlier prediction for the observation circa 2012 Q2 from

the "Focus" holdout well, which was previously predicted to be around −10 (log scale) in Figure 5.6. However, for extreme ORP and DO values we observe a much wider prediction interval, tending toward extreme negative values, as seen for the 2016 observation from the adjacent well.

Prediction of benzene follows a similar trend to MTBE with a slight improvement on the perimeter well as most predictions are below the observed detection limit. However, Figure 7.11 highlights an idiosyncrasy of the censored MoE model that is made more prominent when too many components are assumed; for components with few observations or too few uncensored observations, the prediction intervals become much wider. This is due to the regression coefficient estimates having wide credible intervals. Wide prediction intervals are desired for a component with minimal data but when combined with extreme outliers in the predictors, as shown in Figure 2.4, the model will produce extreme point estimates with similarly extreme prediction intervals.

7.7 Conclusion

Mixture of Experts (MoE) describes a whole suite of models capable of clustering data into some chosen number of components to deal with non-linearity. The process of splitting the observed data into partitions is shared with random forest models and we see evidence in Chapter 4 and this chapter of a low-signal process. In these Bayesian models, further issues such as potential multimodality, label switching and poor mixing highlight a problematic posterior distribution which is only made worse by the presence of censoring. These issues can be dealt with by an experienced practitioner with domain knowledge of the data to tease out interesting relationships.

Our results highlight the dual purpose of hydrocarbon groundwater monitoring sites: monitoring analyte concentrations within

1. a general area (including perimeter wells);
2. a specific area around a base of operations (excluding perimeter wells).

We include all perimeter well observations into the training data to better understand the underlying correlation between analytes and predictors but in our case study, there is little information in these perimeter wells. A more expert-driven model would incorporate these beliefs into each low-concentration well through the use of a more granular per-well prior distribution.

Within a MoE model, one may want to jointly model the analytes as we have done in Chapter 6, using data augmentation as described in Section 1.5.4 or defining a component-specific version of the bivariate likelihood (2.2). However, since we have defined a model where all analytes are only independent conditional on the latent variables then they are unconditionally dependent by definition. That is, through the mechanism of the estimating component membership based on the predictors we have captured some of the high correlation between the analytes.

In summary, results shown have revealed evidence that the relationships between analyte concentrations and our defined predictors can be heterogeneous within a site. In particular, non-detect observations where analyte concentrations are known to be an order of magnitude lower are better modelled separately due to different relationships, evidenced by different regression coefficients in Section 7.6.2. These results concur with the mechanistic reactive transport model (RTM) introduced in Chapter 3 where each component can be viewed as a different "phase" with disparate dominating reactions such as aerobic and anaerobic degradation.

Chapter 8

Varying Intercept

8.1 Introduction

For censored regression models from Chapter 5, and mixture of experts (MoE) from Chapter 7, there is very little difference between our training data within each prediction scenario that was introduced in Section 1.6, LMWO and hold-out future. This chapter aims to reconcile this by leveraging key metadata from the observations using the corresponding wells as a natural grouping to help explain more of the residual variation.

Linear mixed effects models can be more suitable than other models when such a natural grouping of the data exists, unlike MoE models where groups are not observed and must be inferred by some clustering method. For example, suppose we are interested in modelling academic performance using a test score as the dependent variable and our dataset contains multiple schools, each student would correspond to a single observation that would be inherently associated with one of the schools (Gelman *et al.*, 1995). A linear mixed effects model would be appropriate if we believed the parameters to vary between groups.

In a hydrocarbon groundwater monitoring context, a natural grouping we implement is to use the well identifiers that each observation is associated with. We could reduce the number of groups by further classifying these wells into categories, for example,

- off-site wells;
- boundary wells;
- downgradient pumping wells.

Well classifications were provided for some, not all, groundwater network datasets made available to us, and any potential grouping is highly dependent on the specific operations of each site. Therefore, we will not pursue this possibility but there is a clear opportunity for further work.

In general, suppose y_{ik} is the i^{th} observation from the k^{th} group, then a linear mixed effects regression takes the form

$$y_{ik} \sim N(\alpha_k + \mathbf{x}_{ik}^T \boldsymbol{\beta}_k, \tau_k^{-1}),$$

for $i = 1, \dots, n_s$ and $k = 1, \dots K$ where K is the total number of groups and parameters α_k, $\boldsymbol{\beta}_k$ and τ_k are allowed to vary between groups. Parameters are

- the same across groups, which we refer to as a *common* effect;
- or group-specific, which we refer to as a *varying* effect.

Other literature may refer to these terms as fixed and random effects respectively, however we find these nomenclature inappropriate for a Bayesian context and not having a single unambiguous definition (Gelman, 2005).

We can more compactly express these models as an extension to (5.2) for all analytes $j = 1, \dots, n_y$ as

$$\mathbf{y}_j = X\boldsymbol{\beta}_j + Z\boldsymbol{\gamma}_j + \boldsymbol{\epsilon}_j$$

where all predictors collected in the X matrix are multiplied by common effects $\boldsymbol{\beta}$, all predictors collected in the Z matrix are multiplied by varying effects $\boldsymbol{\gamma}$ and $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \tau_j^{-1})$ for all $i = 1, \dots, n_s$ as before.

The decision to make a parameter common or varying within each groups is completely dependent on the context. We consider two special cases,

1. varying intercept: varying β_{0j} but common $\beta_{1j}, \dots, \beta_{n_x j}$;
2. varying slopes model: varying all regression coefficients.

While the varying slopes model is appealing, our main interest in the regression coefficients is to assist in understanding a relationship between each analyte and predictor that may be generalisable to other groundwater sites. Varying the regression coefficients will not help us achieve this goal as interpretation will rely on the groundwater well from which data was sampled making it difficult to extend inference to a new well within the same site and impossible to compare analyses at different sites. Instead, we are motivated by Revie *et al.* (2017) to use a varying intercept model to "sweep up" inter-well variation by modelling it directly.

Due to irregular sampling schedules described in Section 2.1, it is more natural to describe our varying intercept model by

$$y_{ij} = \gamma_{w_i j} + \mathbf{x}_i^T \boldsymbol{\beta}_{-0\,j} + \epsilon_{ij} \tag{8.1}$$

where w_i indexes the well that produced the water sample from which the log analyte concentration y_{ij} and predictors $\mathbf{x}_i$ were measured from. We denote $\boldsymbol{\beta}_{-0\,j} = (\beta_{1j}, \dots, \beta_{n_x j})^T$ to be all regression coefficients corresponding to analyte j except

the intercept β_{0j}; exclusion of the intercept is to help with interpretation. If an intercept was also included then $\gamma_{w_i\,j}$ would represent the deviation from a common intercept at well w_i, without the common intercept we interpret the variable $\gamma_{w_i\,j}$ to be the expected concentration of analyte j at well w_i when all normalised predictors take their mean value, $\mathbf{x}_i = \mathbf{0}_{n_x}$. We will refer to these varying parameters, $\boldsymbol{\gamma}_j = (\gamma_{1j}, \ldots, \gamma_{n_w\,j})$, as *well effects* where n_w denotes the number of wells in the observed training data.

8.2 Prior

Since this model is an extension of the censored regression described in (5.2), we assume the same prior for $\beta_{-0\,j}$ and τ_j, that is,

$$\begin{aligned} \boldsymbol{\beta}_{-0,j} &\sim N_{n_x}(\mathbf{m}_\beta, V_\beta), \\ \tau_j &\sim Ga(a_\tau, b_\tau). \end{aligned} \tag{8.2}$$

For the regression coefficients we typically set $V_\beta = c\, I_{n_x}$ for some scalar c as we lack information *a priori* that these regression coefficients would be correlated. In contrast, we do have prior information suggesting that the well effects are correlated and even spatially correlated. Therefore, we present two models with different prior assumptions for the well effects.

1. A hierarchical normal prior that will induce correlated effects, to be described in Section 8.2.1;
2. Spatial prior that leverages Gaussian processes (GP) as introduced in Section 6.2.1 applied to the spatial data associated with each groundwater well, to be described in Section 8.2.2.

8.2.1 Hierarchical

The hierarchical prior for these models takes the form

$$\gamma_{wj} \sim N(\gamma_{\mu j}, \gamma_{\tau j}^{-1}),$$
$$\gamma_{\mu j} \sim N(m_{\gamma_\mu}, p_{\gamma_\mu}^{-1}),$$
$$\gamma_{\tau j} \sim Ga(a_{\gamma_\tau}, b_{\gamma_\tau}),$$

where $m_{\gamma_\mu}, p_{\gamma_\mu}^{-1}, a_{\gamma_\tau}$ and b_{γ_τ} are prior hyperparameters to be chosen for each analyte $j = 1, \ldots, n_s$. The upshot of this prior specification is that the multivariate marginal prior distribution of $\boldsymbol{\gamma}_j = (\gamma_{1j}, \ldots, \gamma_{n_w j})^T$ will have identical marginal expectations for each element and a covariance matrix where the main diagonal variances exceed the off-diagonal positive covariances inducing positive covariance between each effect. That is, suppose $\gamma_{wj} = \gamma_{\mu j} + \epsilon_{wj}$ where $\epsilon_{wj} \sim N(0, \gamma_{\tau j}^{-1})$ are independent to $\gamma_{\mu j}$ and we assert $w \neq w'$, then

$$E(\gamma_{wj}) = E(\gamma_{\mu j}) + E(\epsilon_{wj}) = m_{\gamma_{\mu j}},$$
$$\mathrm{Var}(\gamma_{wj}) = \mathrm{Var}(\gamma_{\mu j}) + \mathrm{Var}(\epsilon_{wj}) = p_{\gamma_\mu}^{-1} + \frac{b_{\gamma_\tau}}{a_{\gamma_\tau} - 1},$$
$$\mathrm{Cov}(\gamma_{wj}, \gamma_{w'j}) = \mathrm{Cov}(\gamma_{\mu j} + \epsilon_{wj}, \gamma_{\mu j} + \epsilon_{w'j}) = p_{\gamma_\mu}^{-1}.$$

Note that the independence assumption ensured $\mathrm{Cov}(\gamma_{\mu j}, \epsilon_{wj})$ and $\mathrm{Cov}(\epsilon_{wj}, \epsilon_{w'j})$ were 0 valued and the law of total variance with properties of the inverse gamma distribution are needed to derive $\mathrm{Var}(\epsilon_{wj})$.

The important takeaway for our application in these models is that the correlation of wells w and w' must be between 0 and 1 with a higher correlation corresponding to smaller values of $\mathrm{Var}(\epsilon_{wj}) = \frac{b_{\gamma_\tau}}{a_{\gamma_\tau} - 1}$, that is determined *a priori*.

In fact,

$$\text{Cor}(\gamma_{wj}, \gamma_{w'j}) = \frac{\text{Cov}(\gamma_{wj}, \gamma_{w'j})}{\sqrt{\text{Var}(\gamma_{wj})}\sqrt{\text{Var}(\gamma_{w'j})}} = \frac{p_{\gamma_\mu}^{-1}}{p_{\gamma_\mu}^{-1} + b_{\gamma_\tau}(a_{\gamma_\tau} - 1)^{-1}}.$$

8.2.2 Spatial

The grouping in (8.1) relates to wells with corresponding spatial coordinates and we want to assume, *a priori*, that wells that are closer together are more correlated (Tobler, 1970). A Gaussian process (GP) with Matérn kernel function is assumed since the grouping is spatially motivated, but a similar prior can be recreated with any metadata about the observations. At hydrocarbon groundwater monitoring sites, this approach is especially appealing since most wells are clustered together around an operational base as shown in Figure 2.2 and prediction of more sparsely located perimeter wells, that typically report lower concentration and more non-detect measurements, can be aided by stronger prior beliefs.

We construct a subset of the per-observation spatiotemporal coordinates, described in Section 1.4, that correspond to the unique spatial locations of all wells in the training data. That is, $\tilde{S} = (\tilde{\mathbf{s}}_1, ..., \tilde{\mathbf{s}}_{n_w})^T$ where $\tilde{\mathbf{s}}_w = (\tilde{s}_{wx}, \tilde{s}_{wy})$ denotes the spatial coordinates of well w. The well effects are then assumed, *a priori*,

$$\boldsymbol{\gamma}_j \sim GP(\mathbf{0}, k(\tilde{\mathbf{s}}, \tilde{\mathbf{s}}'|\Theta_k)),$$

for arbitrary spatial coordinates $\tilde{\mathbf{s}}$ and $\tilde{\mathbf{s}}'$, choice of covariance function k and hyperparameters Θ_k. We assume the mean of the Gaussian process is zero-valued everywhere for notational simplicity but this need not be the case (Rasmussen *et al.*, 2006). We choose the Matérn covariance function with the ν parameter fixed at 3/2 using the same justifications used in Section 6.2.1, that is,

$$K_{\text{Matérn}}(\tilde{\mathbf{s}}, \tilde{\mathbf{s}}') = \gamma_\alpha^2 \left(1 + \sqrt{3}d\right) \exp\left(-\sqrt{3}d\right),$$

where γ_α^2 is the amplitude hyperparameter and d is the weighted distance between two wells, after weighting by characteristic length-scales $\boldsymbol{\gamma} = (\gamma_x, \gamma_y)^T$, explicitly defined as

$$d = \left\| \frac{\tilde{\mathbf{s}} - \tilde{\mathbf{s}}'}{\boldsymbol{\gamma}} \right\| = \sqrt{\frac{(\tilde{s}_x - \tilde{s}'_x)^2}{\gamma_x^2} + \frac{(\tilde{s}_y - \tilde{s}'_y)^2}{\gamma_y^2}}.$$

We considered a squared exponential kernel function with automatic relevance determination (ARD) in preliminary simulation studies (Beckers, 2021). That is,

$$K_{\text{ARD}}(\tilde{\mathbf{s}}, \tilde{\mathbf{s}}') = \gamma_\alpha^2 \exp\left(-\frac{d^2}{2}\right).$$

However, we found the Matérn 3/2 to be consistently more computationally stable and able to recover true characteristic length-scales with equal or less difficulty than the ARD kernel in almost all simulations. Future iterations of this model could look into more complex choices of kernel functions including non-isotropic kernels motivated by a groundwater monitoring application.

For the prior to be fully specified, it remains to define prior distributions for the GP hyperparameters $\Gamma_K = (\gamma_\alpha, \gamma_x, \gamma_y)$. Both characteristic length-scales have a positive support and are assumed to follow gamma distributions *a priori*. We take the same approach applied to precision parameters and fix the prior mean and investigate sensitivity by changing the prior variance. Suitability of this prior mean can be verified by comparing expert opinion to the correlation matrix produced when γ_α is fixed at 1. Even though γ_α has a vastly different interpretation to the other GP hyperparameters, it is still a strictly positive parameter and is only ever evaluated in the chosen kernel as γ_α^2. We have found that results are somewhat insensitive to choice of prior with truncated normal and gamma prior distributions producing similar posterior draws for all model parameters and no change to posterior predictive draws. Hence, we opt for the arguably more interpretable prior $\gamma_\alpha \sim Ga(a_{\gamma_\alpha}, b_{\gamma_\alpha})$.

8.3 Novel Well Effects

To produce draws from the posterior predictive distribution which quantify our predictions and uncertainty, we make use of draws from the posterior distribution of the model parameters with justification given in Appendix B.1. For these varying intercept models, the situation may arise where a groundwater well, not observed in the original training data, exists in the new data to be predicted. In fact, the LMWO prediction scenario guarantees this very situation. Therefore, there exists two scenarios where analyte concentrations to be predicted either correspond to a

- well *observed* in the training dataset with an already defined well effect;
- *novel* well, with no historical data observed in the training data, and so the well effect has not been defined.

This distinction motivates the nomenclature of **observed** well effects that are dependent on prior choice and observed data, and **novel** well effects that are dependent on the prior specification and posterior observed well effects. Posterior draws of observed well effects are generated during the model fitting algorithm so it only remains to describe how to obtain posterior predictive draws for the novel well effects.

Suppose we denote ${n_w}^*$ novel well effects as $\boldsymbol{\gamma}^* = (\gamma_1^*, \dots, \gamma_{{n_w}^*}^*)^T$ and observed well effects as $\boldsymbol{\gamma}^o = (\gamma_1^o, \dots, \gamma_{{n_w}^o}^o)^T$, where we have made the analyte-specific subscript j implicit as all statements hold for all $j \in \{1, \dots, n_y\}$. For both choice of priors the joint distribution of all well effects is assumed normal, *a priori*,

$$\begin{pmatrix} \boldsymbol{\gamma}_j^* \\ \boldsymbol{\gamma}_j^o \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_* \\ \boldsymbol{\mu}_o \end{pmatrix}, \begin{pmatrix} \Sigma_* & \Sigma_{*o} \\ \Sigma_{o*} & \Sigma_o \end{pmatrix} \right),$$

where $\boldsymbol{\mu}_*$ and $\boldsymbol{\mu}_o$ are the marginal expectations of the respective well effects and the covariance matrix is as shown.

It follows from properties of the multivariate normal distribution that the conditional distribution of our novel well effects is also normal, specifically,

$$\begin{aligned}\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^o &\sim N(\boldsymbol{\mu}_{*|o}, \Sigma_{*|o}),\\ \boldsymbol{\mu}_{*|o} &= \boldsymbol{\mu}_* - \Sigma_{*o}\Sigma_o^{-1}(\boldsymbol{\gamma}_j^o - \boldsymbol{\mu}_o),\\ \Sigma_{*|o} &= \Sigma_* - \Sigma_{*o}\Sigma_o^{-1}\Sigma_{o*}.\end{aligned}$$

Posterior draws for the novel well effects are then obtained by drawing from the conditional distribution

$$\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^o = \boldsymbol{\gamma}^{o\,(m)} \sim N(\boldsymbol{\mu}_{*|o}, \Sigma_{*|o}),$$

where $\boldsymbol{\gamma}^{o\,(m)}$ denotes the $m = 1, \ldots M$ posterior draws of the observed well effects.

8.4 Groundwater Application

There are a total of 4 models for both hierarchical and spatial priors applied to both prediction scenarios, LMWO and holdout future. When there is more information in the form of historical data, we notice that our inferences and predictions are less dependent on the choice of prior. In the holdout future models where all well effects are observed and thus based on historical data, we may present a single model to act as representative for the near identical model outputs.

We assert the same prior hyperparameters described in (5.7) for the parameters shared with the censored regression model and repeat prior choices for both analytes. For the hierarchical prior, we assume $m_{\gamma_\mu} = 0$, $p_{\gamma_\mu}^{-1} = 1.5$, $a_{\gamma_\tau} = 2$ and $b_{\gamma_\tau} = 1$. These choices correspond to a sufficiently vague prior on γ_τ while inducing a specified prior well-to-well correlation of 0.6, appropriate for site A. A better prior could be elicited from experts, but we found each analysis is fairly insensitive to these choices.

While the prior distributions of the GP hyperparameters are all gamma distributions, it is imperative to specify different parametrisations for the amplitude and length-scale hyperparameters due to their differing interpretation. By viewing the resultant correlation matrix for different values of γ_x and γ_y, we have found that values of 0.2 for both is consistent with expectations and "range" values we can obtain from the variograms shown in Figure 2.8. However, asserting a $Ga(0.2, 1)$ distribution leads to poor mixing and unrealistic length-scales in the y direction, this appears to be indicative of the eastwards movement of analyte solute where our data has very little information about correlation in the y direction. To rectify this, we reduce the prior density on large length-scale values by decreasing prior variance of γ_x and γ_y. As the most difficult parameter to estimate, it should not be surprising that our analyses are highly sensitive to the choice of characteristic length-scale prior. Hence,

$$\gamma_\alpha \sim Ga(2, 1),$$
$$\gamma_x, \gamma_y \sim Ga(2, 10).$$

8.4.1 Regression Parameters

We expect the marginal posteriors of the regression coefficients to be similar for either choice of well effect prior, hierarchical or spatial, as the correlation between wells should not affect the impact of some predictor, say electrical conductivity (EC), on an analyte such as MTBE. Similarly, since the training data for each prediction scenario, LMWO and holdout future, have a substantial overlap this choice should have little impact on the regression coefficients. Our results are in line with these expectations and we only show Figure 8.1 for effects on MTBE and Figure 8.2 for effects on benzene, with models using a hierarchical prior producing very similar plots (not shown) highlighting the aforementioned prior insensitivity.

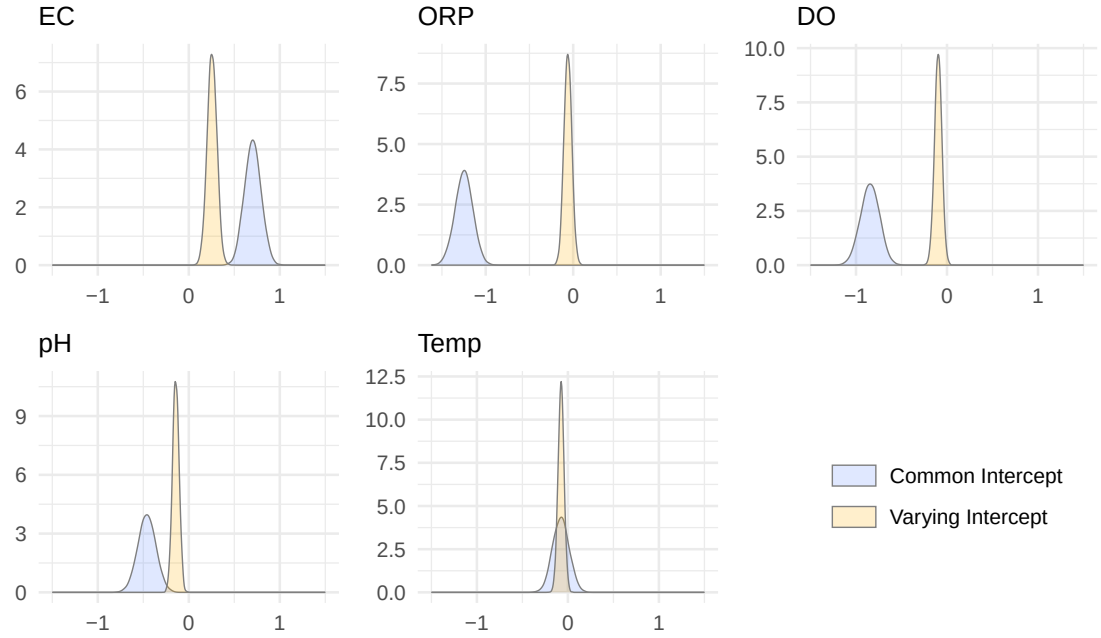

Figure 8.1: Marginal posterior density of regression coefficients. MTBE, LMWO, varying intercept (spatial prior) compared to censored regression.

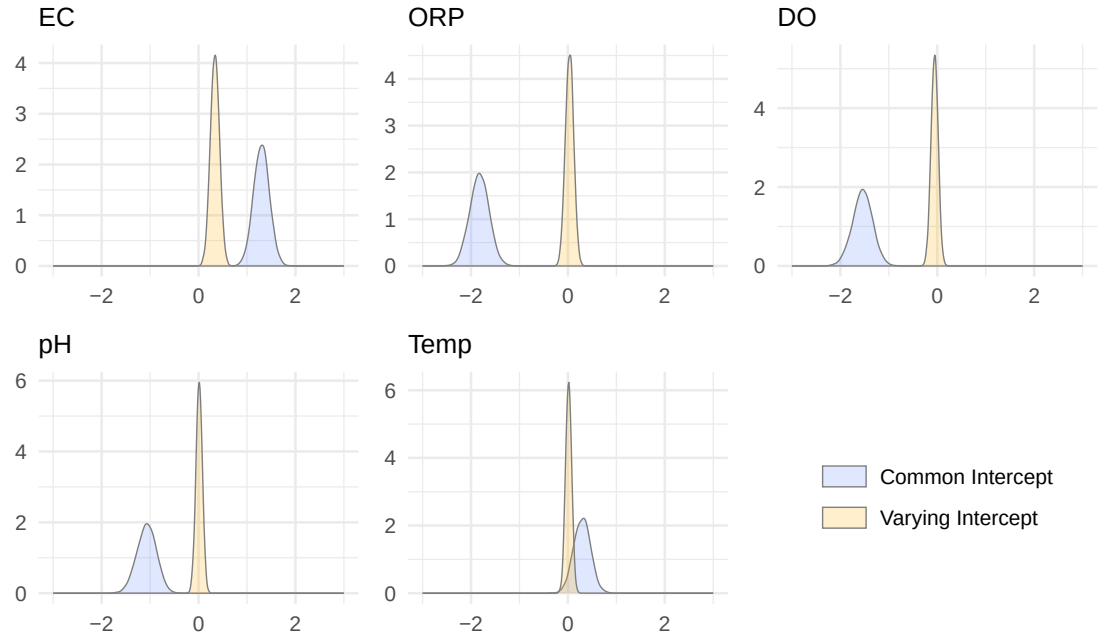

Figure 8.2: Marginal posterior density of regression coefficients. Benzene, LMWO, varying intercept (spatial prior) compared to censored regression.

By including a varying intercept or well effect in the analysis of the site A data, we observe that the sign of each regression coefficient is typically unchanged but the magnitude of all parameters is reduced. The impact of ORP on MTBE log concentrations shows a density with a posterior mean of roughly -1.3, with common intercept, reduced to approximately -0.05 when a varying intercept is used. The coefficients corresponding to benzene concentrations shown in Figure 8.2 show even more reduction than coefficients corresponding to MTBE concentrations; parameters for all predictors except conductivity (EC) contain a 0 in the 95% credible interval implying that only EC has a significant impact on benzene within site A, directly contradicting Figure 5.2.

Due to decreases in effect magnitude, it may be the case that some predictors that were significant in the censored regression described in Chapter 5 are no longer significant in the corresponding varying intercept model. The outcome of these models suggest that the impact of our predictors on analyte concentrations such as MTBE and benzene may be overestimated in a censored regression where inter-well variation is not adequately taken into account.

8.4.2 Precision Parameters

The varying intercept models were motivated by a desire to "sweep up" the inter-well variation and reduce the overall residual variation. We quantify how well these models succeed in this goal by comparing the precision of the varying intercept models to the counterpart model with a common intercept. Again, we notice little difference between the posterior draws for each proposed prediction scenario and choice of well effect prior.

Our expectations are met again as the precision of both analytes, MTBE and benzene increase by a order of magnitude. We also observe the posterior uncertainty on

the precision increase substantially for the varying intercept models which is not too surprising as the more complex model splits the variation into measurement error and inter-well variation leading to a more difficult to estimate parameter.

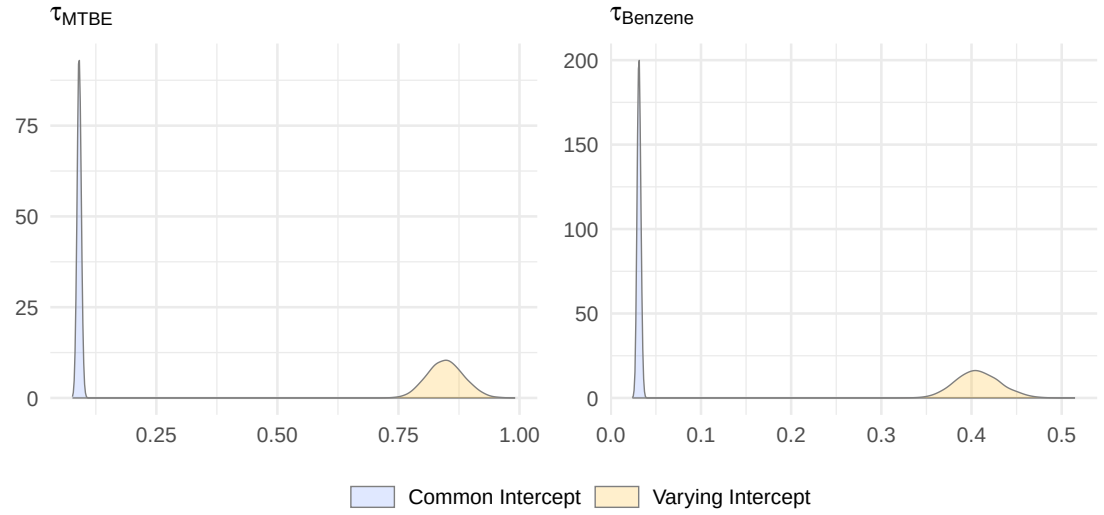

Figure 8.3: Marginal posterior density of precision for both analytes. LMWO, varying intercept (spatial prior) compared to censored regression.

8.4.3 Well Effects

Our inferences of the well effects naturally fall into the two categories defined in Section 8.3, observed well effects and novel well effects. We find that the marginal posteriors of the observed well effects are mainly based on data and so the results are not affected by choice of prior or prediction scenario; this is the same behaviour we have noticed in the regression and precision parameters. On the other hand, the novel well effect parameters are based on the posterior draws of the observed well effects and the correlation between each novel well and observed well is specified differently, *a priori*, by design.

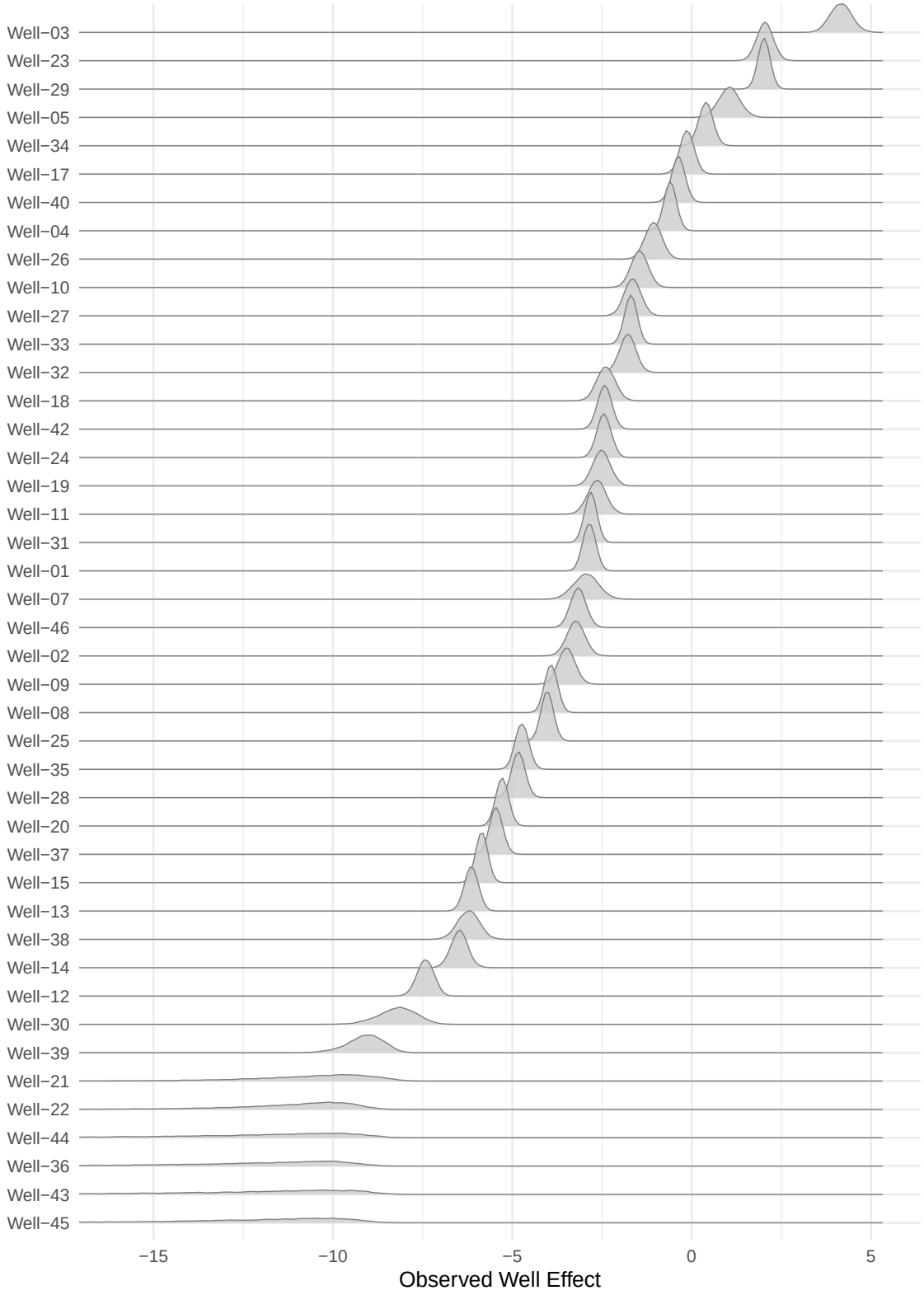

Figure 8.4: Marginal posterior densities of observed well effects. MTBE, LMWO, varying intercept with spatial prior.

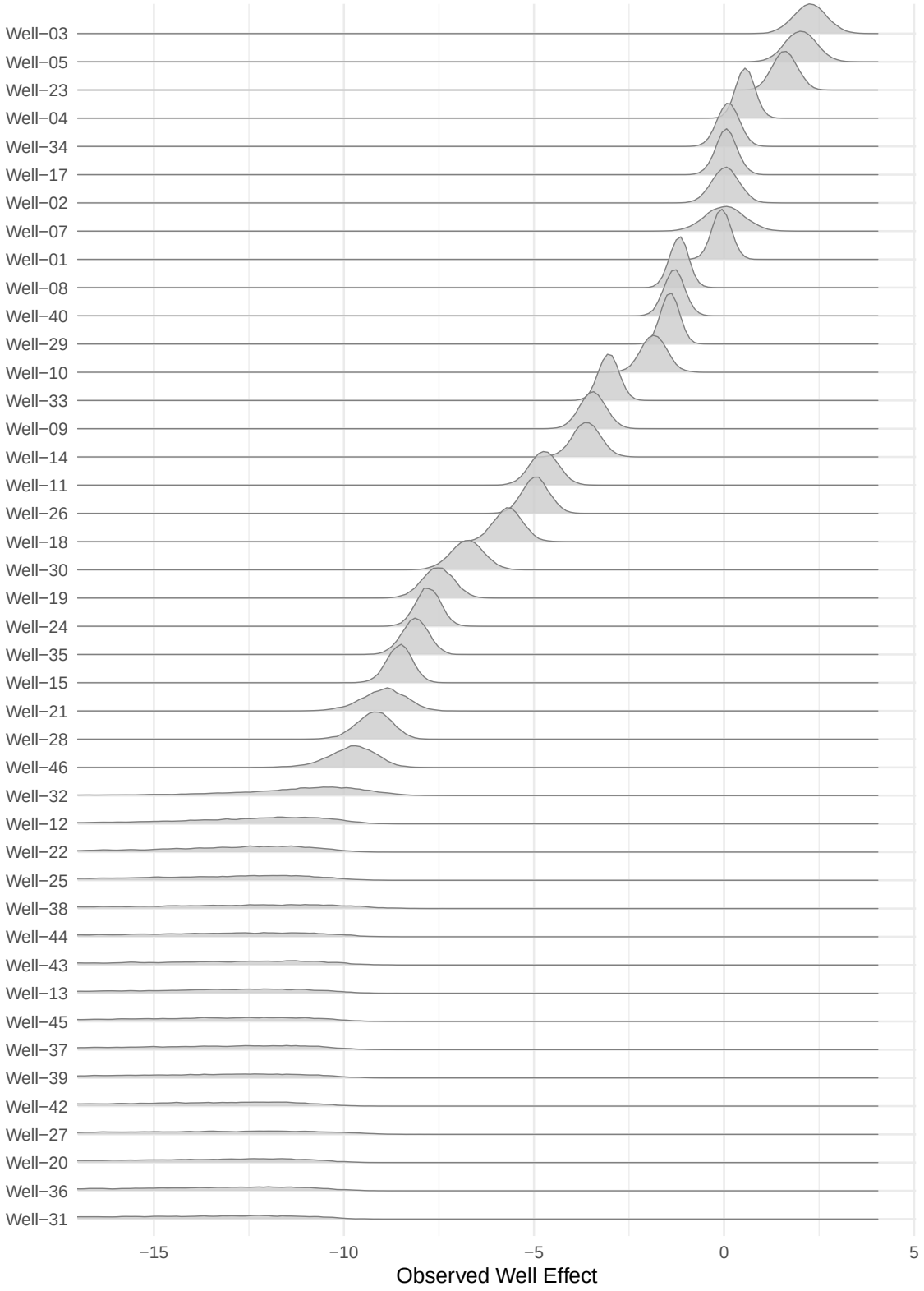

Figure 8.5: Marginal posterior densities of observed well effects. Benzene, LMWO, varying intercept with spatial prior.

If the marginal posterior distributions of the observed well effects were similar, the varying intercept model would not be suitable as it would hold little benefit over a fixed intercept. On the contrary, we notice clear heterogeneity of the wells for both MTBE, Figure 8.4, and benzene, Figure 8.5 with a clear ranking of each well. Of particular interest is "Well-03" that was highlighted as a key location of interest due to the historically high concentrations, relative to the other wells in the site.

The height of the peaks in these posterior densities correspond to the posterior certainty of each well effect; well effects with the lowest posterior means tend to also be the most uncertain. Perimeter wells, for example "Well-43", "Well-44" and "Well-45", typically report non-detect observations between 80% and 100% of the time where each concentration is bounded above by a relatively low detection limit. Therefore, the observed well effects with a posterior mean below average and posterior uncertainty above average are in line with what we expect when investigating hydrocarbon groundwater monitoring networks.

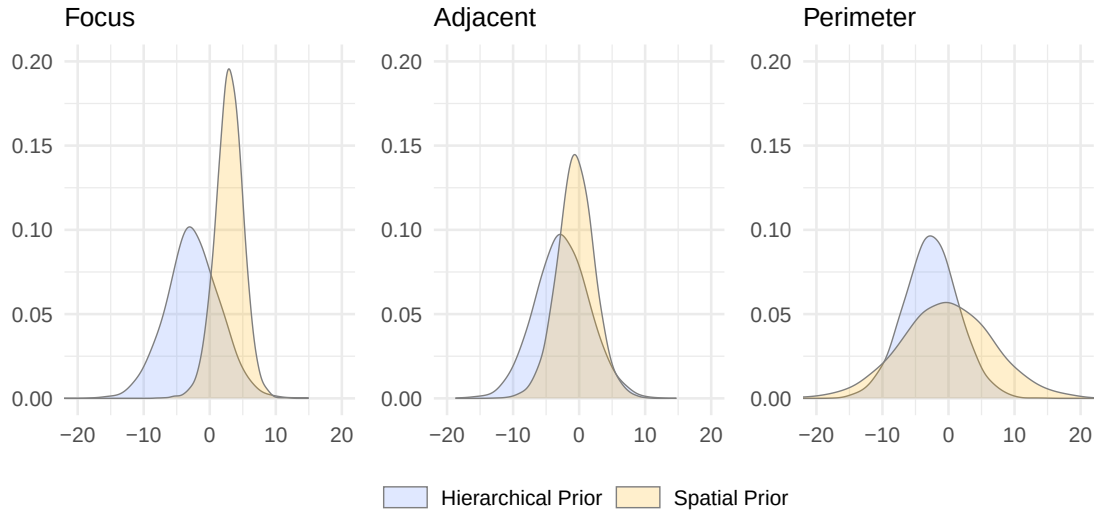

Figure 8.6: Marginal posterior predictive densities of novel well effects. MTBE, LMWO, varying intercept.

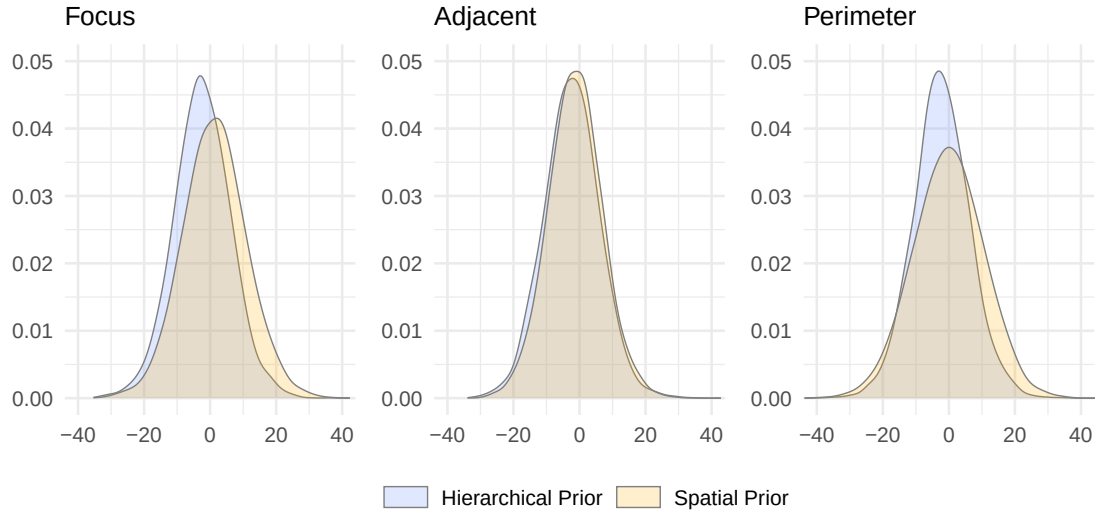

Figure 8.7: Marginal posterior predictive densities of novel well effects. Benzene, LMWO, varying intercept.

For the holdout wells, we either obtain observed well effects based on partial data in the holdout future scenario (not shown) or novel well effects in the LMWO scenario. Figure 8.6 highlights a situation where the spatial prior could yield some benefit over the non-spatial hierarchical prior; the focus well effect is increased based on its proximity to high concentration reporting "Well-03". The variation of the posterior distributions for the other two novel well effects is shown to decrease for the relatively close "Adjacent" well and increase for the remote "Perimeter" well. A similar pattern can be discerned for benzene but the magnitude of the described changes is much smaller and could be due to randomness in the data. When there is little to no information in the data, as we believe may be the case for benzene, the posterior distributions shown in Figure 8.7 are much more insensitive to the choice of prior.

8.4.4 Hyperparameters

Hyperparameters of the hierarchical prior, $(\gamma_\mu, \gamma_\tau)^T$, and spatial prior, $(\gamma_\alpha, \gamma_x, \gamma_y)^T$, are estimated during the MCMC algorithm and can provide us with valuable insight.

One should be careful to understand the nuance of the differing interpretations between γ_μ and the fixed intercept β_0 described in Chapter 5. That is, γ_μ represents the expected mean of several analyte concentrations arising from a novel well when no spatial information is used and the predictors are measured to be at their respective averages. On the other hand, β_0 represents the expectation of a single analyte observation, given average predictor observations. We remark that both of these parameters have similar marginal posterior means that are both similar to estimates of the mean concentration of each respective analyte for the observed training data. However, due to the aforementioned differences we also see a greater posterior uncertainty for γ_μ in Figure 8.8.

Due to the higher degree of censoring, benzene is expected to be the more difficult analyte to model and we do observe a less precise well effect mean and lower expected well effect precision, γ_τ, as shown in Figure 8.8. Another reason for this heightened uncertainty of the well effects could be that benzene is able to be transported quicker than MTBE which is consistent with the fact that benzene has a smaller molecular mass. However, expert opinion from data providers and an independent analysis with GWSDAT (Jones *et al.*, 2014), as introduced in Section 1.2.1, contradicts this possibility showing MTBE as the more mobile species within the site.

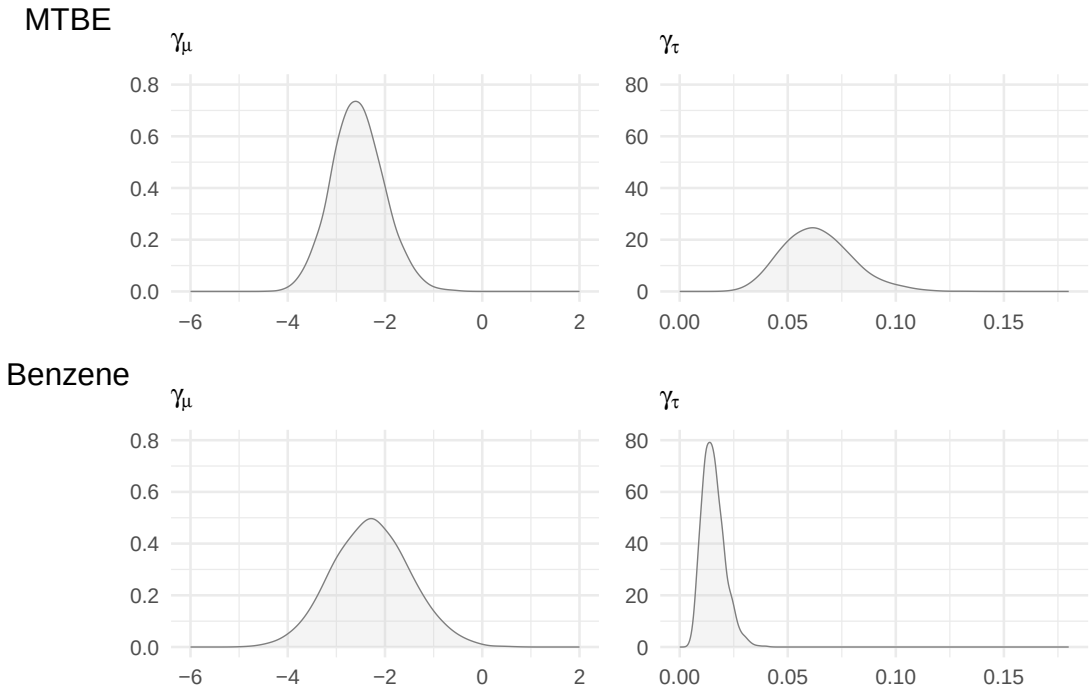

Figure 8.8: Marginal posterior densities of hierarchical hyperparameters. LMWO, varying intercept with hierarchical prior.

For the spatial prior, we are interested in the posterior distributions of the amplitude parameter, γ_α, and characteristic length-scales, γ_x, γ_y, as presented in Figure 8.9, where a smaller length-scale corresponds to smaller correlation over the same distance. We see a similar inference for the amplitude parameter for both MTBE and benzene, a high variance for the well effects. That is, consider the variance of a well effect which is equal to the square of the amplitude parameter as all other terms in the kernel are redundant when the distance is zero. For example, a value of $\gamma_\alpha = 5$ corresponds to marginal variances all equal to 25 for each well effect.

Our analysis of each analyte differs in the length-scale inferences. For MTBE, we note a prior-to-posterior update in the form a reduction from prior mean of 0.2 to a smaller posterior mean. Relative similarity between the marginal posteriors implies a simplified kernel with common length-scales would be adequate. Figure 8.9 shows

the greatest prior to posterior update for the length-scale in the x direction for well effects corresponding to benzene where reasonable values are an order of magnitude smaller than the y direction. We infer from these results that the spatial correlation diminishes much quicker in the x direction than the y direction and future work could confirm this empirically using a bi-directional variogram as discussed in Section 2.5.

Key differences in the MTBE-based parameters and benzene-based parameters shown in Figure 8.8 and Figure 8.9 remind us that while a high correlation is almost guaranteed for these analytes it is not a perfect correlation. As such, it is important to remember that each of the chemical species have different intrinsic properties that can affect the data motivating our decision to present results from both.

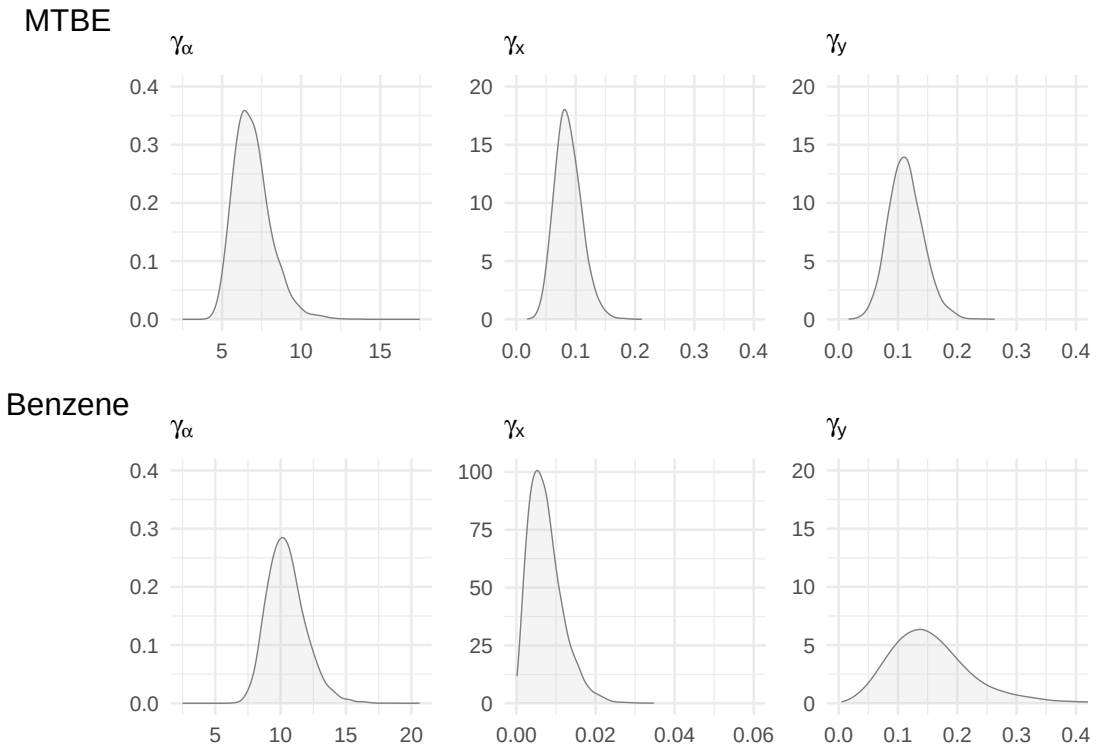

Figure 8.9: Marginal posterior densities of Gaussian process hyperparameters. LMWO, varying intercept with spatial prior.

8.4.5 Prediction

For the holdout future prediction scenario, the linear predictor which will be used in the posterior predictive distribution is based on observed well effects, not novel well effects. Hence, we would not expect the choice of prior to impact our predictions since it has not had a substantial impact on the observed well effects and so we again use the spatial prior as a representative of model output that is identical up to the randomness of the MCMC algorithm.

Considering Figure 8.10 and Figure 8.11, predictions for holdout wells close to the base of operations, that is "Focus" and "Adjacent", are extremely similar and insensitive to prior choice and prior hyperparameters. On the other hand, predictions for the holdout well "Perimeter" are only pragmatically similar since all predictions and intervals lie below the true non-detect observations. In fact, we notice a more uncertain posterior predictive distribution for the spatial prior than the hierarchical prior. This highlights a key issue with the Gaussian process (GP) approach, where a length-scale that results in an appropriate prior covariance between "Focus" and "Adjacent" may lead to an inappropriate covariance between "Focus" and "Perimeter". Our choice of prior mean for the characteristic length-scales were based on the cluster of wells in Figure 2.2, but we would ideally have a prior covariance equal to that of the hierarchical prior for distances exceeding 50% of the site. Future work could look into a GP kernel that tends to a lower limit instead of 0 for "large" distances to emulate a site-wide correlation and spatial correlation that we would expect from these data.

8.4.5.1 Holdout Future

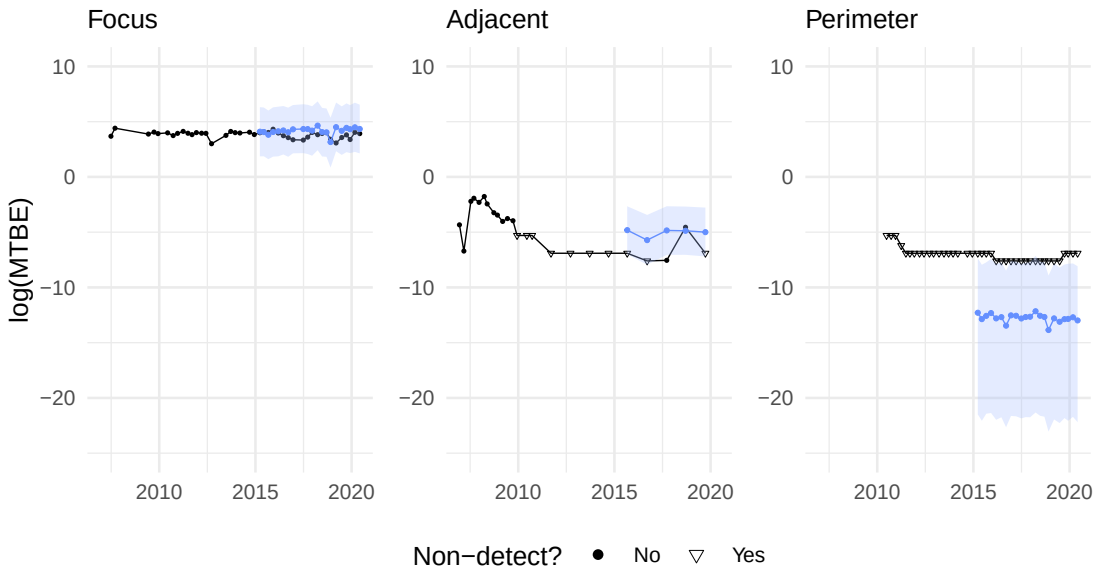

Figure 8.10: Predictions with comparison to truth in black. MTBE, holdout future, varying intercept with spatial prior.

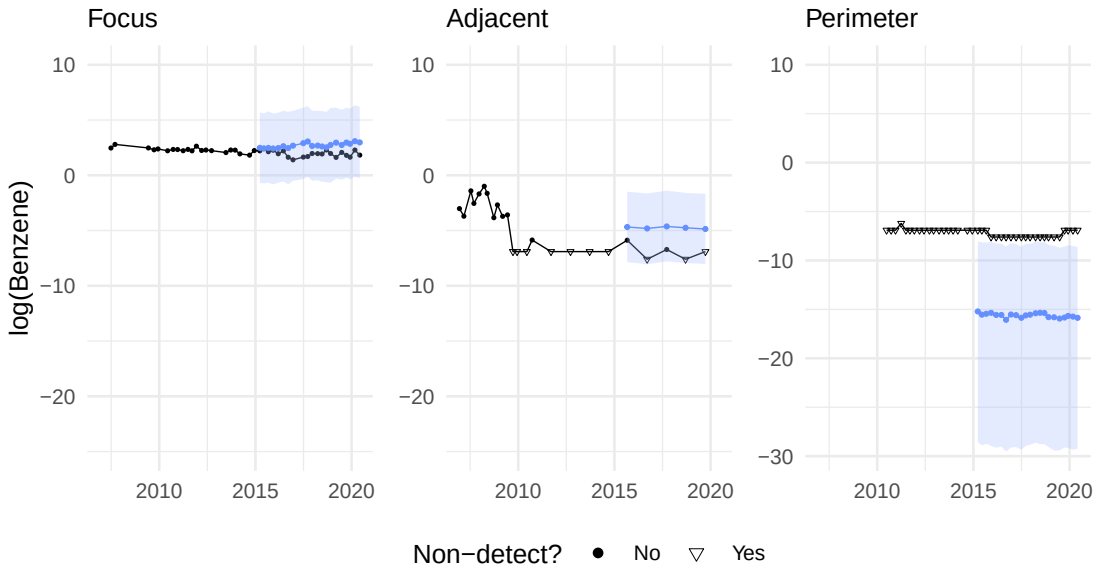

Figure 8.11: Predictions with comparison to truth in black. Benzene, holdout future, varying intercept with spatial prior.

In the holdout future case, we see similarities for MTBE, Figure 8.10, and benzene, Figure 8.11. By using historical data, prediction for the "Focus" well is much better than a common intercept model and this is impressive given the edge-case nature of this well that reports the second highest average concentrations. Prediction is not as good with the "Adjacent" well, likely due to a concentration that decreases over time leading to consistent overprediction in this case.

The combination of these two results highlights the power and shortcomings of the varying intercept models. We have reframed the problem of predicting analyte concentrations to predicting deviations from the mean within analyte concentrations. In either case, if there is a low signal-to-noise ratio in these data as we expect based on the results in Chapter 4, prediction is going to be a difficult task. Moreover, for time series with a downward trend that is common in groundwater monitoring there is a material risk of overestimation, as we see with "Adjacent". The ramifications for upward trends in analyte concentrations are much worse as these models may underestimate concentrations leading to worse decision making.

We observe little difference in the quantitative predictive metrics for in-sample data shown in Table 8.1 and out-of-sample data shown in Table 8.2 with near identical results omitted when the modelled analyte was benzene. These results are consistent with our previous assertion that these models are less sensitive to prior choice, spatial or hierarchical, when there exists historical data for all wells to be predicted.

	Δ LPD (SE)	Δ WAIC (SE)	Δ PSIS (SE)
MTBE, Spatial	-0.1 (1.2)	0.0 (0.0)	0.0 (0.0)
MTBE, Hierarchical	0.0 (0.0)	-0.5 (1.4)	-0.5 (1.5)

Table 8.1: Model comparison metrics based on in-sample data. MTBE, holdout future, varying intercept.

	Δ LPD (SE)
MTBE, Spatial	0.0 (0.0)
MTBE, Hierarchical	-0.2 (0.1)

Table 8.2: Model comparison metrics based on out-of-sample data. MTBE, holdout future, varying intercept.

8.4.5.2 LMWO, Hierarchical Prior

All hierarchical prior predictions appear to be very similar across the wells. This is overtly expected as the model has no method of distinguishing each holdout well and so all novel well effects are drawn from the same distribution. Our observed values to be predicted still fall within the prediction intervals shown but since all of the predictive power lies with the predictors, this model does not appear to have a benefit over the common intercept models from a qualitative standpoint.

Figure 8.12 and Figure 8.13 show the LMWO predictions of MTBE and benzene concentrations respectively when a hierarchical prior is assumed. Trivially, each prediction is similar because this model does not use any information about each well and identical data for all predictors would lead to identical posterior predictive distributions to be drawn from. Comparing the prediction intervals of Figure 8.12 to Figure 5.6 shows that the uncertainty is only slightly increased for MTBE concentrations. Whereas, comparing Figure 8.13 to Figure 5.7 shows a greater increase in uncertainty. This could be due to heterogeneity of the observed well effects as evidenced by the smaller estimated precision γ_τ visualised in Figure 8.8.

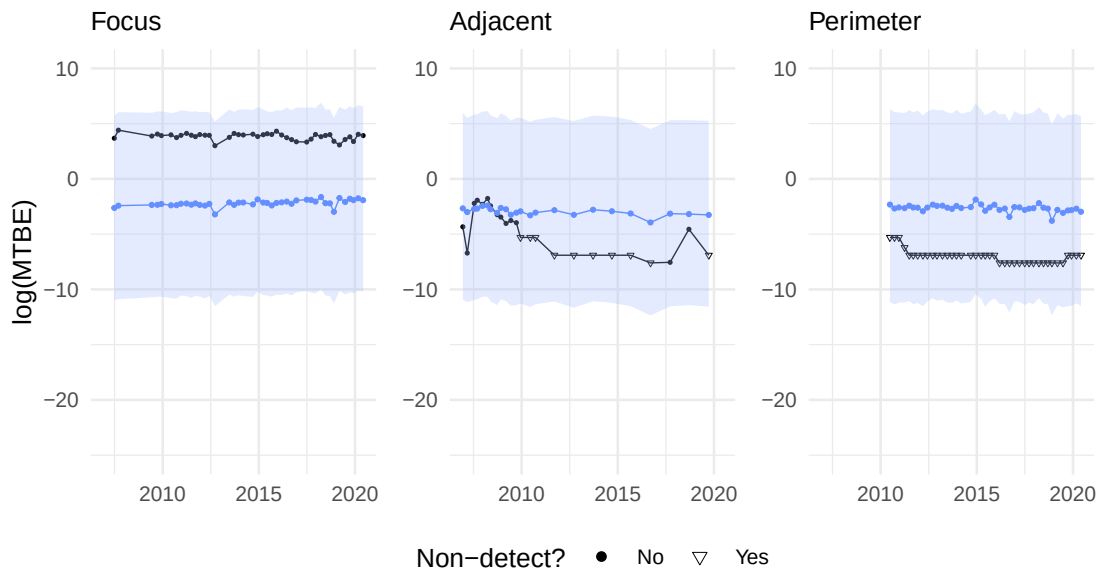

Figure 8.12: Predictions with comparison to truth in black. MTBE, LMWO, varying intercept with hierarchical prior.

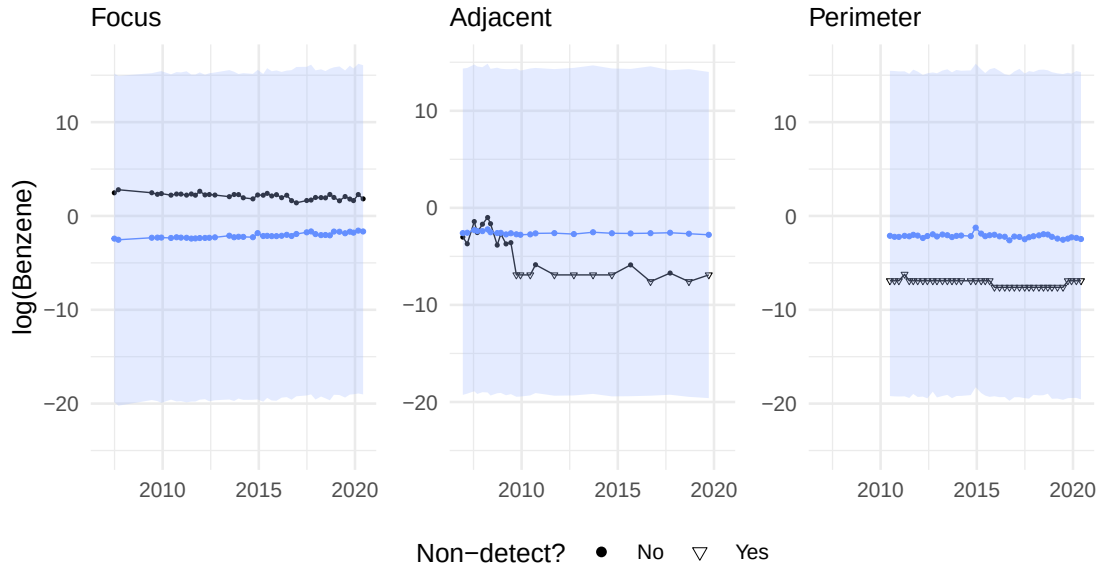

Figure 8.13: Predictions with comparison to truth in black. Benzene, LMWO, varying intercept with hierarchical prior.

8.4.5.3 LMWO, Spatial Prior

When using the spatial prior to predict MTBE as shown in Figure 8.14, we present good prediction of the "Focus" well for these models and even mirror the downward change halfway through 2012. The more spatially remote wells "Adjacent" and "Perimeter" get progressively worse with overprediction posing a potential problem. The true values shown for both these wells fall within our prediction intervals highlighting reasonable predictions qualitatively.

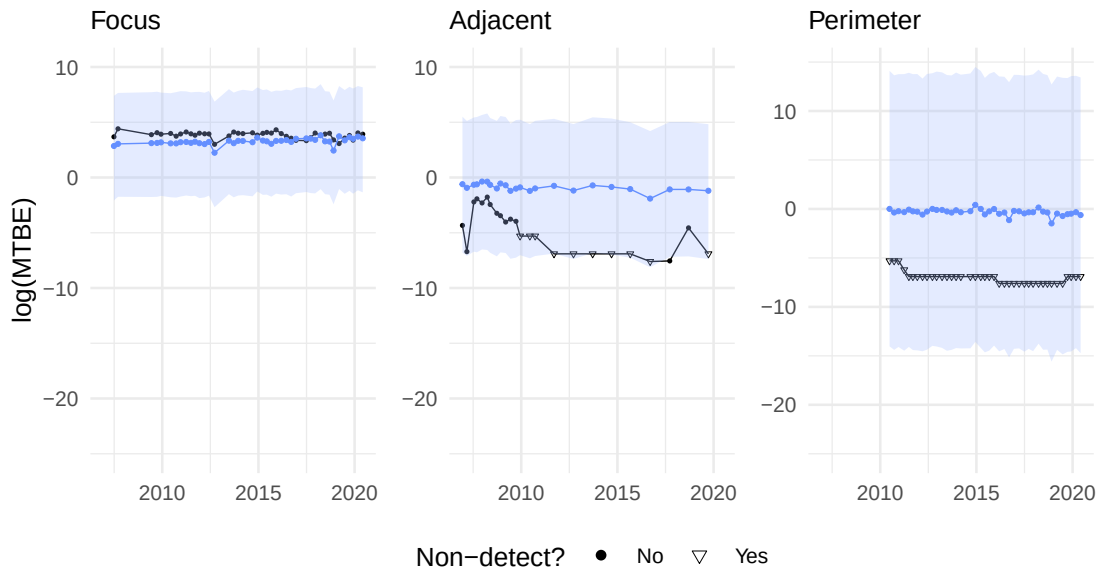

Figure 8.14: Predictions with comparison to truth in black. MTBE, LMWO, varying intercept with spatial prior.

We see that these models struggle with the excessively censored analyte benzene, showing more uncertain predictions in the spatial case, Figure 8.15, than in the hierarchical case, Figure 8.13. These predictions may be improved if we chose a GP prior that was specific to benzene instead of choosing a GP prior based on all analytes. However, it is likely the case that there is too little information in these data due to high censoring.

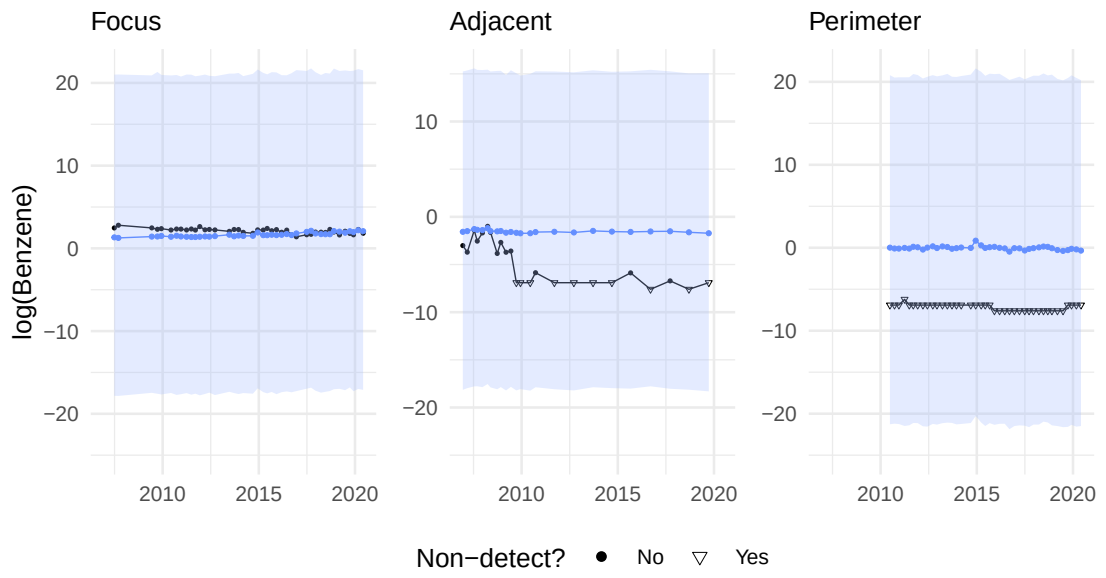

Figure 8.15: Predictions with comparison to truth in black. Benzene, LMWO, varying intercept with spatial prior.

8.4.5.4 LMWO, Comparison

To make quantitative comparisons between the spatially motivated and hierarchical prior choices we construct four tables for each unique combination of modelled analyte, MTBE or benzene, and prior choice. Table 8.3 contains the model comparison metrics based on in-sample data for MTBE and Table 8.4 reports the same information for benzene. Both tables show very little difference in prior choice, which is to be expected as the in-sample data must only contain observed well effects with no novel well effects by definition. As such, we see a similar comparison to the holdout future scenario and conclude there is no evidence that the choice of prior impacts predictive performance of the observed data.

	Δ LPD (SE)	Δ WAIC (SE)	Δ PSIS (SE)
MTBE, Spatial	-0.0 (1.1)	0.0 (0.0)	0.0 (0.0)
MTBE, Hierarchical	0.0 (0.0)	-0.7 (1.3)	-0.4 (1.4)

Table 8.3: Model comparison metrics based on in-sample data. MTBE, LMWO, varying intercept.

	Δ LPD (SE)	Δ WAIC (SE)	Δ PSIS (SE)
Benzene, Spatial	-0.3 (0.4)	0.0 (0.0)	-0.2 (0.4)
Benzene, Hierarchical	0.0 (0.0)	-0.1 (0.4)	0.0 (0.0)

Table 8.4: Model comparison metrics based on in-sample data. Benzene, LMWO, varying intercept.

The most pertinent scenario for these varying intercept models corresponds to when we assess the models' ability to not only predict completely new data but to predict new data from a novel well. To this end, we consider model comparison metrics based on out-of-sample data for the model trained in the LMWO scenario in Table 8.5 for MTBE and Table 8.6 for benzene. The quantitative results agree with the qualitative results for MTBE corroborating our conclusions that better prediction of the wells close to the site, "Focus" and "Adjacent", is worth the trade-off of a wider prediction interval for the remote "Perimeter" well. On the other hand, benzene shows how the GP prior may perform worse than a hierarchical prior in some cases; different inferences for the GP hyperparameters for MTBE and benzene may indicate that the spatial information is less useful to prediction of the benzene concentrations.

	Δ LPD (SE)
MTBE, Spatial	0.0 (0.0)
MTBE, Hierarchical	-51.4 (10.9)

Table 8.5: Model comparison metrics based on out-of-sample data. MTBE, LMWO, varying intercept.

	Δ LPD (SE)
Benzene, Spatial	-3.9 (0.7)
Benzene, Hierarchical	0.0 (0.0)

Table 8.6: Model comparison metrics based on out-of-sample data. Benzene, LMWO, varying intercept.

8.5 Conclusion

In this chapter we have explored a modification from a site-wide intercept to a well-specific intercept as an improvement to model fit, parameter estimation and model prediction. We are able to directly model these well effects within a Bayesian linear mixed effects model and have provided two prior distributions that allow for different types of positive correlation between these wells. Within a hydrocarbon groundwater monitoring site, the Gaussian process (GP) prior is well motivated with a clear rationale whereas the hierarchical prior may be better suited to applications where the spatial information is less pertinent.

A clear advantage to these models is the ability to quantify the impact of observed wells and still be able to predict never before seen novel wells that may be purely hypothetical. When the spatial component of these well effects are leveraged, we

see a substantial yet expected increase in accuracy of novel well effects which directly impact the models' predictive performance on new data. Such an increase is dependent on the spatial information contained within the data and may negatively impact the prediction of perimeter wells.

Moreover, prediction within these models appear to rephrase the problem of prediction. Instead of relying on the predictors to estimate unknown analyte concentrations, we are now tasked with estimating a deviation from the sample mean obtained from historical data. That is, even with better predictive performance we are no closer to being able to use predictor measurements to make statements about the direction of future analyte concentrations, a key statistic required by groundwater site managers.

Similarly, the varying intercept models have revealed how estimated regression coefficients from previous models may be spurious and biased, not only due to the presence of left censoring but also due to a tangible well effect. More accurate well effects should lead to more accurate estimates of other parameters including the measurement precision τ_j and regression coefficients $\boldsymbol{\beta}_{-0\,j}$ by "sweeping up" excess residual variation not explained by the model (Revie *et al.*, 2017). Results from our case study, site A, highlight the potential of false-positive errors within our baseline censored regression models. Repeating these analyses at other hydrocarbon groundwater monitoring sites would be required to increase our confidence in the estimated general impact of each predictor on analyte concentrations.

Further work could take lessons learned from McLean *et al.* (2019) and extend these models to also include a temporal effect and leverage the entire spatiotemporal data instead of just the spatial aspect. Previous attempts at modelling the spatiotemporal effect jointly in Section 6.2 have proven difficult to fit, so a separated well effect and temporal effect would be of great interest.

Chapter 9

Conclusion

Throughout this thesis, we have been single-mindedly focused on a single hydrocarbon groundwater monitoring site, site A, in our attempts to elicit a relationship between the log concentrations of analytes, MTBE and benzene, and predictors, electrical conductivity (EC), oxidation reduction potential (ORP), dissolved oxygen (DO), pH and temperature. Another key interest is the prediction of some influential holdout wells in multiple prediction scenarios based on hypotheticals that are expected to occur at a groundwater monitoring site. Given more data availability we could have applied all techniques to multiple sites to demonstrate where these models and inferences would generalise to other hydrocarbon groundwater monitoring sites and maybe even groundwater monitoring sites interested in other analytes of particular concern.

Chapter 1 established the background and motivated the problem. This was followed by an in-depth look into these data in Chapter 2. Chapter 3 set expectations using models that are not common in the statistics literature including mechanistic reactive transport models (RTM) and tree-based random forest models were described in Chapter 4. Using the univariate censored regression model described in Chapter 5

as a baseline model, we have presented new models as possible extensions throughout the subsequent chapters, each dealing with a problematic assumption such as analyte independence, linearity and spatial independence. Specifically, Chapter 6 allowed our analytes to be modelled using multivariate or even matrix-variate distributions; Chapter 7 investigated phases motivated by the RTM model in a mixture modelling framework known as mixture of experts and Chapter 8 directly modelled a well effect by using a varying intercept model, a special case of a linear mixed effects model.

Each model extension can potentially improve prediction in some areas, but at the expense of others; we favour the pragmatic approach of focusing on wells at key locations as opposed to the best average prediction because of the presence of perimeter wells within these data. Regression coefficient estimates unfortunately agree with our initial assessment of a low signal-to-noise ratio since each time a model is improved, any significant effect is seen to be biased and presenting a spurious effect. This is definitely the case for predictors like temperature but may not be the case for more promising predictors such as EC and ORP. Harbingers of this difficult to detect effect existed in the low R^2 values of the random forest models that have a tendency to overfit and in the simulated RTM data based on Bemidji (Ng *et al.*, 2015) where noiseless predictor observations have smaller changes than the expected measurement error that would exist in the real data.

With sufficient motivation, one could exploit the generality of each extension and combine several key ideas, for example a varying intercept model with concomitant variables in a mixture of experts framework would be possible if one was to answer key questions including if the varying intercepts should further vary by component. In a reality where computational resources were much larger, maybe due to the introduction of quantum computing, one could even fit a matrix-variate regression with

varying intercepts and concomitant variables. However, such a model is definitely computationally infeasible currently and we would like to remind the reader that the simpler model is preferred and commonly just as good, if not better, than the overly complex model in terms of our comparison metrics.

9.1 Final Comparison

Our original intent was to compare the best model from each of the modelling chapters. However, for reasons discussed in Section 6.5, the multivariate models described in Chapter 6 do not deal with censoring which may lead to incomparable models. Therefore, we present our final model comparison tables for the univariate models only including the censored regression from Chapter 5, the mixture of experts model with $K = 2$ components from Chapter 7 and both versions of the varying intercept models from Chapter 8.

9.1.1 Leave-Multiple-Well-Out (LMWO)

	Δ LPD (SE)	Δ WAIC (SE)	Δ PSIS (SE)
Censored Regression	-2611.2 (71.8)	-2549.7 (75.9)	-2547.5 (76.1)
MoE, 2 Components	-2019.3 (62.7)	-1976.0 (66.8)	-1973.9 (66.9)
Varying Intercept, Spatial	-0.3 (1.3)	0.0 (0.0)	0.0 (0.0)
Varying Intercept, Hierarchical	0.0 (0.0)	-0.6 (1.5)	-0.3 (1.6)

Table 9.1: Model comparison metrics based on in-sample data. Both analytes, LMWO.

Again, we use the LMWO prediction scenario to calculate metrics for the in-sample data with near identical tables resulting from the holdout future scenario. We see a continuation of the variable performance of the varying intercept models as Table 9.1 shows the 'best' model based on in-sample data is the varying intercept model. Another conclusion from Table 9.1 is that the mixture of experts model shows considerable increase over the single component censored regression. Finally, we are reminded that for in-sample data, the choice of prior for the observed well effects is of little consequence.

	Δ LPD (SE)
Censored Regression	-34.7 (8.6)
MoE, 2 Components	0.0 (0.0)
Varying Intercept, Spatial	-47.5 (24.2)
Varying Intercept, Hierarchical	-94.2 (16.1)

Table 9.2: Model comparison metrics based on out-of-sample data. Both analytes, LMWO.

When interested in our holdout wells in the out-of-sample case, we observe that with the introduction of novel well effects into the varying intercept models, concentrations are better predicted using the spatial prior as evidenced by Table 9.2. However, we present results slightly favouring the 2 component MoE model. As mentioned in Section 8.4.5.4, this could be due to poor estimation of the novel well effects for benzene concentrations regardless of prior choice, spatial or hierarchical. Since the increase in log pointwise predictive density (LPD) is small relative to the other models, we assume all models within 3 standard errors could be viable models.

9.1.2 Holdout Future

Changing to the prediction scenario where historical data for each holdout well, up to 2015, is also observed in model training and focusing on the task of predicting observations post-2015 reveals expected results. Table 9.3 appears contradictory to the LMWO case since varying intercept models are favoured by approximately 5 standard errors over the next best model. This is expected since the well effects are better estimated from data observed at that well and the sensitivity of choice of prior, spatial or hierarchical, is lowered drastically.

	Δ LPD (SE)
Censored Regression	-98.8 (17.1)
MoE, 2 Components	-65.1 (13.3)
Varying Intercept, Spatial	0.0 (0.0)
Varying Intercept, Hierarchical	-0.2 (0.1)

Table 9.3: Model comparison metrics based on out-of-sample data. Both analytes, future.

9.2 Further Work

Several opportunities for extending upon this body of work have been described in the relevant section and we will summarise those ideas here. We avoid discussing the optimal well placement or removing wells in this section as we are aware of such work ongoing in the University of Glasgow (Radvanyi *et al.*, 2023) and want to avoid overlap.

Anecdotally, random forest models appear to have become very popular with data scientists in recent times and we have recognised the key requirements that are asked of the dependent variable data. That is, a random forest can be built to handle censoring if one can define a measure of location, for example a mean, and a measure of spread to define impurity. Some methods of defining a mean from left-censored data including MLE would be inappropriate as the nodes most distant from the root will have very few observations, by design. However, if further work could compose a sample mean function and sample variance function for left censored data, it would be straightforward to construct a left-censored random forest model.

Some models may be improved by existing methods and supplementary investigations would reveal the feasibility of each. One could revisit the matrix-variate regression model with solutions to computational infeasibility using gradient-free methods, sub-sampling of the training data before model fitting or emulation techniques. Similarly, we could alleviate the multimodality from Chapter 7 using reversible jump Markov chain Monte Carlo (RJMCMC) (Richardson & Green, 1997) or tempering to fully explore each posterior mode (Neal, 1996, Jasra *et al.* (2005)).

When there are only 2 analytes of interest, we have a bivariate left-censored regression that uses a correlation parameter to be estimated (Newton & Rudel, 2007). Future work looking into the impact this would have on each of the univariate mod-

els would be very interesting and require less work than other suggestions since we have already used likelihood-based approaches such as slice sampling (Neal, 2003) and Hamiltonian Monte Carlo (Stan Development Team, 2023). A more challenging approach would be to create a generalised multivariate distribution for several potentially censored analytes that is capable of defining a density for any possible combination of censoring indicators.

In the interest of predicting analyte concentrations with previously defined predictors, we have avoided time series approaches to these data. The benefit of this decision is that we are able to predict analyte concentrations without requiring the previous observation of that variable. Moreover, missing data and irregular sampling patterns, common in groundwater monitoring, has little to no impact on other observations for these models. However, there is clearly a temporal dependence in both analyte and predictor data as evidenced by the use of time-series plots throughout the thesis and noticeable autocorrelation. Further work could investigate including a temporal aspect into the model such as drift, removing any potential seasonality or autoregressive parameters. While these opportunities are not the focus of our research, more effectively modelling the residual variance by removing temporal dependence could improve estimation and prediction as we have shown in Chapter 8.

While all aforementioned suggestions would be of great interest to the statistical community, the best approach for research impact would be to investigate repeatability and reproducibility at other groundwater monitoring sites. That is, organisations investigating hydrocarbon concentrations in addition to sites with a focus on some other water quality indicator, for example lead pollution, nitrate pollution or radioactivity within the water sample.

Appendix A

Glossary

A.1 Analyte Collections

In hydrocarbon groundwater monitoring data it is common to see grouping of variables as a representative of the highly correlated analytes.

- **BTEX**: benzene, toluene, ethylbenzene and all xylenes;
- **Total Petroleum Hydrocarbons (TPH)**: total crude oil hydrocarbons with 6 to 35 carbon atoms, sometimes split into various ranges such as
  - **Gasoline Range Organics (TPH-GRO)**: C_6 to C_{10} alkanes, lower boiling point than TPH-DRO (Environmental Protection Agency, 2015);
  - **Diesel Range Organics (TPH-DRO)**: C_{10} to C_{28} alkanes;
- Other groups may classify these hydrocarbons with more groups such as
  - **Low Range Hydrocarbons (LRH)** for C_5 to C_8 alkanes;
  - **Mid Range Hydrocarbons (MRH)** for C_9 to C_{18} alkanes;
  - **High Range Hydrocarbons (HRH)** for C_{19} to C_{35} alkanes.

A.2 Analytes

- **Benzene** (C_6H_6): petrochemical with a rigid hexagonal shape formed from 6 carbon atoms with a hydrogen bond for each. Has been confirmed to be carcinogenic and a major health concern (World Health Organization, 2010).
- **Toluene** (C_7H_8): alternatively $C_6H_5\,CH_3$, a methyl group (CH_3) attached to a phenyl group, that is a benzene ring minus one of the hydrogen atoms.
- **Ethylbenzene** (C_8H_{10}): alternatively $C_6H_5\,CH_2\,CH_3$, a further deviation from benzene where one bond is replaced by phenyl group (C_6H_5) and another by methylene [CH_2].
- **Xylene** (C_8H_{10}): alternatively $(CH_3)_2\,C_6H_4$, structurally is a benzene ring with a methyl group replacing two hydrogen bonds. Data is provided as total xylenes or by a further classification depending on the locations of the two methyl groups as shown in Figure A.1:
  - ortho-xylene neighbours or 1-2;
  - meta-xylene neighbours once removed or 1-3;
  - para-xylene: opposite vertices of hexagon or 1-4.

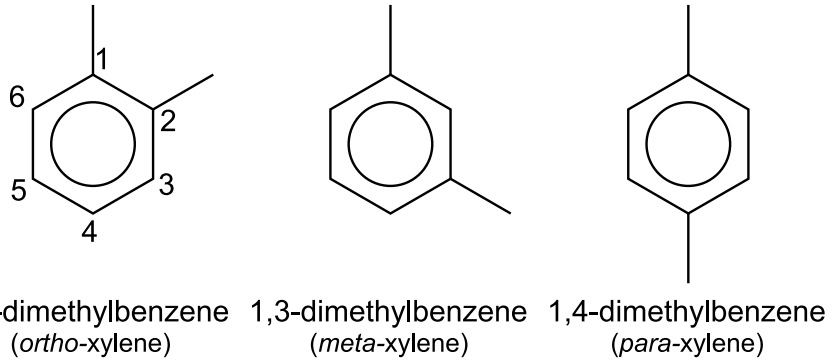

Figure A.1: Illustrative schematic of IUPAC nomenclature of alicyclic compounds.

- **Methyl tert-butyl ether (MTBE)** ($(CH_3)_3COCH_3$): volatile flammable liquid that is a fuel additive as opposed to a crude oil compound that has multiple uses including raising the oxygen content of gasoline.
- **Naphthalene** ($C_{10}H_8$): has the appearance of two fused benzene rings sharing two carbon atoms, for this reason it is the simplest polycyclic aromatic hydrocarbon.

A.3 Predictors

- **pH**: measure of acidity of the water sample; since values are already on the log scale one could also use the activity of the hydrogen ions in the solution, say a_{H^+}, where $\text{pH} = -\log(a_{H^+})$.
- **Conductivity (EC)**: direct measurement of how well the sample can conduct electricity per volume.
- **Temperature**: the temperature of the water during analysis.
- **Dissolved Oxygen (DO)**: the concentration of oxygen gas that has been incorporated into the water
- **Oxidation Reduction Potential (ORP)**: water quality parameter, reported in volts or millivolts, that conveys the presence of an oxidising agent from a high ORP measurement or the presence of a reducing agent from a low ORP measurement.

Appendix B

Markov chain Monte Carlo (MCMC) Methods

The nature of MCMC algorithms involve drawing samples from the desired posterior distribution, or stationary distribution, which is extremely appealing when the target is intractable due to a complex normalising constant (Gamerman & Lopes, 2006). To further validate these methods, we often start the algorithm at a variety of starting values and assess multiple outputs, known as chains. This thesis utilises several MCMC-based algorithms including

- Gibbs sampling, (Section 5.2.2);
- Slice sampling, (Neal, 2003);
- Hamiltonian Monte Carlo, (Stan Development Team, 2023).

Output of any of these MCMC methods is said to have *good mixing* if the target distribution is well explored; it follows that algorithms producing draws with low autocorrelation are likely to have better mixing than one with many autocorrelated draws. Mixing is impacted by several factors such as model complexity, data used,

prior information or parameter starting values. For each model fit we have assessed the mixing of each chain using several diagnostics including density, trace and autocorrelation plots and numerical quantities such as effective sample size.

Convergence is also a desirable property where we want the output to converge to the target stationary distribution in a relatively short time and to converge to the same solution regardless of the algorithm's initialisation state. Multiple chains converging on the same distribution improves our confidence convergence. Further discussion around assessing convergence can be found in Gelman *et al.* (1995).

Initial samples from MCMC output, known as "burn-in", are typically discarded as it can take time for the algorithm to reach the stationary distribution and these samples may not be representative of the stationary distribution. An alternate method of discarding samples involves "thinning" the output where only every k^{th} sample is kept; while thinning can reduce autocorrelation yielding more information per sample, many argue that it is often unnecessary and inefficient (Link & Eaton, 2012), although there is much debate around this topic. Any output we have thinned has been for pragmatic reasons such as computational feasibility or to obtain less autocorrelated samples without increasing output storage requirements.

B.1 Posterior Predictive Densities

Consider the general case where we observe some *training* data as a collection of dependent variables Y, independent variables X and potentially some metadata S that often takes the form of spatiotemporal coordinates in our application. All are assumed to be matrices with n_s rows or *observations* and a number of columns specified in the corresponding model description.

By performing Bayesian inference with some chosen model on these data, we estimate the posterior distribution of the model parameters Θ. That is, by Bayes' theorem

$$\pi(\Theta|Y, X, S) = \frac{\pi(Y|X, S, \Theta)\,\pi(\Theta)}{\pi(Y, X, S)} \propto \pi(Y|X, S, \Theta)\,\pi(\Theta).$$

When new data is observed, say Y', X', S', our objective is to make predictions with quantifiable uncertainty on the response variables Y'. This is achieved through the use of the posterior predictive distribution,

$$\pi(Y'|Y, X, S, X', S') = \int \pi(Y'|X, S, X', S', \Theta)\,\pi(\Theta|Y, X, S)\ \mathrm{d}\Theta,$$

which is the joint posterior density $\pi(Y', \Theta|Y, X, S, X', S')$ with the model parameters marginalised out. In practice, these distributions are realised in the forms of finite draws that are based on the Monte Carlo estimate,

$$\pi(Y'|Y, X', S') \approx \frac{1}{M} \sum_{m=1}^{M} \pi(Y'|\Theta^{(m)}, X', S'),$$

where $\Theta^{(m)}$ denotes the m^{th} posterior draw of the model parameters.

Appendix C

Model Code

One of the primary motivations behind this project included transparency and contributing to open-source projects where possible. To this end, several R packages were produced and made publicly available on code repository services like GitHub and GitLab.

- Chapter 5 uses the mixture of experts code with $K = 1$ components assumed;
- Chapter 6 uses the `bmnr` package, https://github.com/nclJoshCowley/bmnr;
- Chapter 7 uses the `bmoe` package, https://github.com/nclJoshCowley/bmoe;
- Chapter 8 uses the `visp` package, https://github.com/nclJoshCowley/visp.

All packages are open source and will be made public in due course and will improve with new users and feedback. Hence, we include the original model code here for reproducibility.

C.1 Bayesian Mixture of Experts

JAGS code used to fit Bayesian mixture of experts model (Chapter 7).

```
var regr[p_regr, n_y, k], prec[n_y, k], unnormalised_probs[n_s, k];
data {
  # Ones trick
  for (i in 1:n_s) { for (yi in 1:n_y) { ones[i, yi] = 1 }}
  C = 10000
}

model {
  # Transformed parameters
  for (i in 1:n_s) {
    for (ki in 1:k) {
      unnormalised_probs[i, ki] = exp(x_wt[i, ] %*% wt[, ki])
    }
  }

  # Likelihood
  for (i in 1:n_s) {
    z[i] ~ dcat(unnormalised_probs[i, 1:k])

    for (yi in 1:n_y) {
      ones[i, yi] ~ dbern(L[i, yi] / C)
      L[i, yi] = ifelse(is_nd[i, yi], y_cdf[i, yi], y_pdf[i, yi])

      y_cdf[i, yi] = pnorm(y[i, yi], mean[i, yi], prec[yi, z[i]])
      y_pdf[i, yi] = dnorm(y[i, yi], mean[i, yi], prec[yi, z[i]])

      mean[i, yi] = x_regr[i, ] %*% regr[, yi, z[i]]
    }
  }

  # Prior
  for (ki in 1:k) { for (yi in 1:n_y) {
    prec[yi, ki] ~ dgamma(prec_shape, prec_rate)
  }}

  for (j in 1:p_regr) { for (yi in 1:n_y) { for (ki in 1:k) {
    regr[j, yi, ki] ~ dnorm(0, regr_prec)
  }}}

  for (j in 1:p_wt) {
    for (ki in 2:k) { wt[j, ki] ~ dnorm(0, wt_prec) }
    wt[j, 1] = 0
  }
}
```

C.2 Bayesian Multivariate Normal Regression

Stan code for Bayesian multivariate normal regression models (Chapter 6).

```
data {
  // Dimensions
  int<lower=1> n_s;
  int<lower=1> n_y;
  int<lower=1> n_x;

  // Data
  matrix[n_s, n_y] y;
  matrix[n_s, n_x] x;

  // Prior Hyperparameters
  real<lower=0> regr_prec;
  real<lower=0> covar_y_df;
  cov_matrix[n_y] covar_y_scale;
}

parameters {
  matrix[n_x, n_y] regr;
  cov_matrix[n_y] covar_y;
}

model {
  matrix[n_y, n_y] L_covar_y = cholesky_decompose(covar_y);
  matrix[n_s, n_y] mean_y = x * regr;

  // Prior
  for (yi in 1:n_y) regr[, yi] ~ normal(0, 1 / sqrt(regr_prec));
  covar_y ~ inv_wishart(covar_y_df, covar_y_scale);

  // Likelihood
  for (ii in 1:n_s) y[ii, ] ~ multi_normal_cholesky(mean_y[ii, ], L_covar_y);
}
```

C.3 Varying Intercept

Stan code for varying intercept models (Chapter 8).

Spatial Prior

```
functions {
  #include /functions/lcens_norm.stan
  #include /functions/gp_matern32_cov_ard.stan
}

data {
  // Dimensions
  int<lower=1> n_s;
  int<lower=1> n_x;
  int<lower=1> n_groups;
  int<lower=1> n_gp_dims;

  // Optional Left-Censoring
  int<lower=0, upper=1> is_left_cens;

  // Data
  vector[n_s] y;
  int<lower=0, upper=1> is_nd[is_left_cens ? n_s : 0];
  matrix[n_s, n_x] x;
  int<lower=0, upper=n_groups> groups[n_s];
  vector[n_gp_dims] coords[n_groups];

  // Prior Hyperparameters
  real<lower=0> regr_prec;
  real<lower=0> prec_shape;
  real<lower=0> prec_rate;

  // GP Prior Hyperparameters
  real<lower=0> gp_scale_shape;
  real<lower=0> gp_scale_rate;
  real<lower=0> gp_length_shape;
  real<lower=0> gp_length_rate;

  // GP Parameters (Assumed known)
  real<lower=0> gp_nugget;
}

parameters {
  vector[n_x] regr;
  real<lower=0> prec;
  vector[n_groups] vary_eff;
  real<lower=0> gp_scale;
  vector[n_gp_dims] gp_length;
}

model {
  // Priors
  regr ~ normal(0, sqrt(1 / regr_prec));
  prec ~ gamma(prec_shape, prec_rate);
```

```
42    // GP Prior
43    gp_scale ~ gamma(gp_scale_shape, gp_scale_rate);
44    for (k in 1:size(gp_length)) {
45      gp_length[k] ~ gamma(gp_length_shape, gp_length_rate);
46    }

47    matrix[n_groups, n_groups] gp_covar;
48    gp_covar = gp_matern32_cov_ard(coords, gp_scale, gp_length);

49    for (g in 1:n_groups) gp_covar[g, g] = gp_covar[g, g] + gp_nugget;

50    vary_eff ~ multi_normal_cholesky(
51      rep_vector(0, n_groups),
52      cholesky_decompose(gp_covar)
53    );

54    // Likelihood
55    vector[n_s] y_mean = vary_eff[groups] + (x * regr);
56    if (is_left_cens) {
57      y ~ lcens_norm(is_nd, y_mean, sqrt(1 / prec));
58    } else {
59      y ~ normal(y_mean, sqrt(1 / prec));
60    }
61  }
```

Hierarchical Prior

```
functions {
  #include /functions/lcens_norm.stan
}

data {
  // Dimensions
  int<lower=1> n_s;
  int<lower=1> n_x;
  int<lower=1> n_groups;

  // Optional Left-Censoring
  int<lower=0, upper=1> is_left_cens;

  // Data
  vector[n_s] y;
  int<lower=0, upper=1> is_nd[is_left_cens ? n_s : 0];
  matrix[n_s, n_x] x;
  int<lower=0, upper=n_groups> groups[n_s];

  // Prior Hyperparameters
  real<lower=0> regr_prec;
  real<lower=0> prec_shape;
  real<lower=0> prec_rate;

  // Hierarchical Prior Hyperparameters
  real<lower=0> vary_eff_mean_prec;
  real<lower=0> vary_eff_prec_shape;
  real<lower=0> vary_eff_prec_rate;
}

parameters {
  vector[n_x] regr;
  real<lower=0> prec;
  vector[n_groups] vary_eff;
  real vary_eff_mean;
  real<lower=0> vary_eff_prec;
}

model {
  // Priors
  regr ~ normal(0, sqrt(1 / regr_prec));
  prec ~ gamma(prec_shape, prec_rate);

  // Hierarchical Prior
  vary_eff_mean ~ normal(0, sqrt(1 / vary_eff_mean_prec));
  vary_eff_prec ~ gamma(vary_eff_prec_shape, vary_eff_prec_rate);
  vary_eff ~ normal(vary_eff_mean, sqrt(1 / vary_eff_prec));

  // Likelihood
  vector[n_s] y_mean = vary_eff[groups] + (x * regr);
  if (is_left_cens) {
    y ~ lcens_norm(is_nd, y_mean, sqrt(1 / prec));
  } else {
    y ~ normal(y_mean, sqrt(1 / prec));
  }
}
```

C.4 Stan Functions

Matérn 3/2 Kernel with Characteristic Length Scales

This Stan function extends the built-in `gp_matern32_cov` to have characteristic length-scales.

```
/*
 * Matern 3/2 Kernel with Characterstic Length Scales
 *
 * @param x array of vector. Distance between these two vectors are needed.
 * @param gp_scale real. Multiplicative amplitude of the kernel function.
 * @param gp_length vector. Multiple length scales, one for each dimension.
 *
 * @return Covariance matrix with 'size(x)' rows and columns
*/
matrix gp_matern32_cov_ard(vector[] x, real gp_scale, vector gp_length) {
    int n_r = size(x);
    matrix[n_r, n_r] out;

    real gp_scale_sq = pow(gp_scale, 2);
    real dist;

    for (i in 1:(n_r - 1)) {
      out[i, i] = gp_scale_sq;

      for (j in (i + 1):n_r) {
        dist = sqrt(dot_self((x[i] - x[j]) ./ gp_length));
        out[i, j] = gp_scale_sq * (1 + sqrt(3) * dist) * exp(-1 * sqrt(3) * dist);
        out[j, i] = out[i, j];
      }
    }

    out[n_r, n_r] = gp_scale_sq;

    return out;
  }
```

Left-censored Normal Log Likelihood

We define a custom distribution for the univariate left-censored response variable using the `_lpdf` syntax.

```
/*
 * Increment with Left-censored Normal Log-likelihood
 *
 * @param y Vector, either observed value (is_nd = T), otherwise detection limit.
 * @param is_nd Array of T/F, TRUE implies censoring for that observation.
 * @param mu Vector, linear predictor.
 * @param sigma Real, standard deviation.
 *
 * @return lp__
*/
real lcens_norm_lpdf(vector y, array[] int is_nd, vector mu, real sigma) {
  real result = 0;

  int n_cens = sum(is_nd);
  int n_obs = rows(y) - n_cens;

  int which_cens[n_cens];
  int which_obs[n_obs];

  int i_cens = 1;
  int i_obs = 1;

  for (i in 1:rows(y)) {
    if (is_nd[i] == 1) {
      which_cens[i_cens] = i;
      i_cens += 1;
    } else {
      which_obs[i_obs] = i;
      i_obs += 1;
    }
  }

  result += normal_lpdf(y[which_obs] | mu[which_obs], sigma);
  result += normal_lcdf(y[which_cens] | mu[which_cens], sigma);

  return result;
}
```

Bibliography

Adamson, D. T., McHugh, T. E., Rysz, M. W., Landazuri, R. & Newell, C. J. 2012 Field investigation of vapor-phase-based groundwater monitoring. *Groundwater Monitoring & Remediation* **32** (1), 59–72.

Allman, E. S., Matias, C. & Rhodes, J. A. 2009 Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics* **37** (6A), 3099–3132.

Anastasopoulos, P. C., Shankar, V. N., Haddock, J. E. & Mannering, F. L. 2012 A multivariate tobit analysis of highway accident-injury-severity rates. *Accident Analysis & Prevention* **45**, 110–119.

Appelo, C. A. J. & Postma, D. 2005 *Geochemistry, groundwater and pollution.* CRC Press.

Baize, D., Bellanger, L. & Tomassone, R. 2009 Relationships between concentrations of trace metals in wheat grains and soil. *Agronomy for Sustainable Development* **29** (2), 297–312.

Beckers, T. 2021 An Introduction to Gaussian Process Models. *arXiv preprint arXiv:2102.05497* .

Boys, R. J. & Henderson, D. A. 2002 On determining the order of Markov

dependence of an observed process governed by a hidden Markov model. *Scientific Programming* **10** (3), 241–251.

Breiman, L. 2001 Random forests. *Machine Learning* **45** (1), 5–32.

Breiman, L., Friedman, J., Olshen, R. A. & Stone, C. J. 1984 *Classification and regression trees*. Routledge.

Brown Jr, B. W., Hollander, M. & Korwar, R. M. 1973 Nonparametric tests of independence for censored data with application to heart transplant studies. Florida State University.

Celeux, G. 1998 Bayesian inference for mixture: the label switching problem. In *Compstat* (ed. R. Payne & P. Green), 227–232. Springer, Physica, Heidelberg.

Chamroukhi, F. 2016 Unsupervised learning of regression mixture models with unknown number of components. *Journal of Statistical Computation and Simulation* **86** (12), 2308–2334.

CL:AIRE 2017 Petroleum Hydrocarbons in Groundwater: Guidance on assessing petroleum hydrocarbons using existing hydrogeological risk assessment methodologies. CL:AIRE, ISBN: 978-1-905046-31-7.

Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977 Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39** (1), 1–22.

Eilers, P. H. & Marx, B. D. 1996 Flexible smoothing with b-splines and penalties. *Statistical Science* **11** (2), 89–121.

Environmental Protection Agency 2015 Method 8015c: Nonhalogenated organics by gas chromatography. https://www.epa.gov/sites/default/files/2015-12/documents/8015c.pdf, accessed: 2023-12-06.

Essaid, H. I., Bekins, B. A., Herkelrath, W. N. & Delin, G. N. 2011 Crude oil at the Bemidji site: 25 years of monitoring, modeling, and understanding. *Groundwater* **49** (5), 706–726.

Eubank, R. L. 1999 *Nonparametric Regression and Spline Smoothing*. CRC Press.

Evers, L., Molinari, D., Bowman, A., Jones, W. & Spence, M. 2015 Efficient and automatic methods for flexible regression on spatiotemporal data, with applications to groundwater monitoring. *Environmetrics* **26** (6), 431–441.

Ferguson, T. S. 1973 A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.

Fridley, B. L. & Dixon, P. 2007 Data augmentation for a Bayesian spatial model involving censored observations. *Environmetrics: The official journal of the International Environmetrics Society* **18** (2), 107–123.

Friedman, J. H. 2001 Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29** (5), 1189–1232.

Frühwirth-Schnatter, S. 2006 *Finite Mixture and Markov Switching Models*. Springer.

Frühwirth-Schnatter, S., Celeux, G. & Robert, C. P. 2019 *Handbook of Mixture Analysis*. CRC Press.

Gamerman, D. & Lopes, H. F. 2006 *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.

Gelman, A. 2005 Analysis of variance - why it is more important than ever. *The Annals of Statistics* **33** (1), 1–53.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. 1995 *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A., Goodrich, B., Gabry, J. & Vehtari, A. 2019 R-squared for Bayesian regression models. *The American Statistician* **73** (3), 307–309.

Genuer, R., Poggi, J.-M., Genuer, R. & Poggi, J.-M. 2020 *Random forests with R*. Springer.

George, B. J., Gains-Germain, L., Broms, K., Black, K., Furman, M., Hays, M. D., Thomas, K. W. & Simmons, J. E. 2021 Censoring trace-level environmental data: statistical analysis considerations to limit bias. *Environmental Science & Technology* **55** (6), 3786–3795.

Gneiting, T. & Raftery, A. E. 2007 Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association* **102** (477), 359–378.

Gong, W., Reich, B. J. & Chang, H. H. 2021 Multivariate spatial prediction of air pollutant concentrations with INLA. *Environmental research communications* **3** (10), 101002.

Goodfellow, I., Bengio, Y. & Courville, A. 2016 *Deep Learning*. MIT Press.

Gormley, I. C. & Frühwirth-Schnatter, S. 2019 Mixtures of experts models. In *Fruhwirth-Schnatter, S. Celeux, G., Robert, CP (eds.). Handbook of Mixture Analysis*. CRC Press.

Greenwell, B. M. 2017 pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal* **9** (1), 421–436.

Gregorutti, B., Michel, B. & Saint-Pierre, P. 2015 Grouped variable importance with random forests and application to multiple functional data analysis.

Computational Statistics & Data Analysis **90**, 15–35.

Helsel, D., Hirsch, R., Ryberg, K., Archfield, S. & Gilroy, E. 2020 Statistical methods in water resources: U.S. Geological Survey Techniques and Methods, book 4, chap. A3. U.S. Geological Survey.

Helsel, D. R. 2005 More than obvious: better methods for interpreting nondetect data. *Environmental Science & Technology* **39** (20), 419A–423A.

Helsel, D. R. 2011 *Statistics for Censored Environmental Data using Minitab and R*, 2nd edn. John Wiley & Sons.

Helsel, D. R. & Cohn, T. A. 1988 Estimation of descriptive statistics for multiply censored water quality data. *Water Resources Research* **24** (12), 1997–2004.

Hennig, C. 2000 Identifiability of models for clusterwise linear regression. *Journal of classification* **17** (2).

Hoff, P. D. 2009 *A First Course in Bayesian Statistical Methods*. Springer.

Hornung, R. W. & Reed, L. D. 1990 Estimation of average concentration in the presence of nondetectable values. *Applied Occupational and Environmental Hygiene* **5** (1), 46–51.

Huynh, B.-T. 2019 Estimation and feature selection in high-dimensional mixtures-of-experts models. PhD thesis, LMNO Lab CNRS, UMR 6139, University of Caen.

Interstate Technology & Regulatory Council 2021 Sustainable Resilient Remediation Training (SRR-1). https://clu-in.org/conf/itrc/SRR/, accessed: 2022-09-30.

Iranmanesh, A., Arashi, M. & Tabatabaey, S. 2010 On conditional applications of matrix variate normal distribution. *Iranian Journal of Mathematical*

Sciences and Informatics **5** (2), 33–43.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. & Hinton, G. E. 1991 Adaptive mixtures of local experts. *Neural Computation* **3** (1), 79–87.

Jaeger, B. C., Welden, S., Lenoir, K., Speiser, J. L., Segar, M. W., Pandey, A. & Pajewski, N. M. 2023 Accelerated and interpretable oblique random survival forests. *Journal of Computational and Graphical Statistics* 1–16.

Jasra, A., Holmes, C. C. & Stephens, D. A. 2005 Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science* **20** (1), 50–67.

Jones, W. R., Spence, M. J. & Bonte, M. 2015 Analyzing groundwater quality data and contamination plumes with GWSDAT. *Groundwater* 513–514.

Jones, W. R., Spence, M. J., Bowman, A. W., Evers, L. & Molinari, D. A. 2014 A software tool for the spatiotemporal analysis and reporting of groundwater monitoring data. *Environmental Modelling and Software* **55**, 242–249.

Kleinbaum, D. G., Klein, M. *et al.* 2012 *Survival Analysis: A Self-Learning Text*. Springer.

Kruschke, J. 2014 *Doing Bayesian Data Analysis: A tutorial with R, JAGS, and Stan*, 2nd edn. Academic Press.

Lewandowski, D., Kurowicka, D. & Joe, H. 2009 Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* **100** (9), 1989–2001.

Li, L. 2023 Reactive transport in the subsurface. The Pennsylvania State University. https://www.e-education.psu.edu/png550/node/698, accessed: 2023-11-13.

Li, L., Maher, K., Navarre-Sitchler, A. *et al.* 2017 Expanding the role of reactive transport models in critical zone processes. *Earth-science reviews* **165**, 280–301.

Liaw, A. & Wiener, M. 2002 Classification and Regression by randomForest. *R News* **2** (3), 18–22.

Link, W. A. & Eaton, M. J. 2012 On thinning of chains in MCMC. *Methods in ecology and evolution* **3** (1), 112–115.

Lockwood, J., Schervish, M. J., Gurian, P. L. & Small, M. J. 2004 Analysis of contaminant co-occurrence in community water systems. *Journal of the American Statistical Association* **99** (465), 45–56.

MacQueen, J. *et al.* 1967 Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 281–297. Oakland, CA, USA.

Malander, M. E. 2016 Optimization of Groundwater Monitoring at a Research Facility in New Jersey (GWSDAT). https://gro-1.itrcweb.org/optimization-of-groundwater-monitoring-at-a-research-facility-in-new-jersey-gwsdat/?print=pdf, accessed: 2023-08-12.

Matheron, G. 1963 Principles of Geostatistics. *Economic Geology* **58** (8), 1246–1266.

McLachlan, G. J. & Peel, D. 2000 *Finite mixture models*. Wiley.

McLean, M., Evers, L., Bowman, A., Bonte, M. & Jones, W. 2019 Statistical modelling of groundwater contamination monitoring data: a comparison of spatial and spatiotemporal methods. *Science of The Total Environment* **652**, 1339–1346.

Mclean, M. I. 2018 Spatio-temporal models for the analysis and optimisation of groundwater quality monitoring networks. PhD thesis, University of Glasgow.

Meinshausen, N. & Ridgeway, G. 2006 Quantile regression forests. *Journal of Machine Learning Research* **7** (6).

Molinari, D. A. 2014 Spatiotemporal modelling of groundwater contaminants. PhD thesis, University of Glasgow.

Morris, D. E., Oakley, J. E. & Crowe, J. A. 2014 A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software* **52**, 1–4.

Neal, R. M. 1996 Sampling from multimodal distributions using tempered transitions. *Statistics and Computing* **6**, 353–366.

Neal, R. M. 2003 Slice sampling. *The Annals of Statistics* **31** (3), 705–767.

Newton, E. & Rudel, R. 2007 Estimating correlation with multiply censored data arising from the adjustment of singly censored data. *Environmental Science & Technology* **41** (1), 221–228.

Ng, G.-H. C., Bekins, B. A., Cozzarelli, I. M. *et al.* 2015 Reactive transport modeling of geochemical controls on secondary water quality impacts at a crude oil spill site near Bemidji, MN. *Water Resources Research* **51** (6), 4156–4183.

Oliver, M. A., Webster, R. *et al.* 2015 Basic steps in geostatistics: the variogram and Kriging. Springer.

Papastamoulis, P. 2016 label.switching: An R package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software, Code Snippets* **69** (1).

Pebesma, E. J. 2004 Multivariable geostatistics in S: the gstat package. *Computers & Geosciences* **30**, 683–691.

Plummer, M. *et al.* 2003 JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 1–10. Vienna, Austria.

Pörtner, H.-O., Roberts, D. C., Adams, H., Adler, C., Aldunce, P., Ali, E., Begum, R. A., Betts, R., Kerr, R. B. & Biesbroek, R. 2022 Climate change 2022: impacts, adaptation and vulnerability. IPCC Geneva, The Netherlands.

Radvanyi, P., Miller, C., Alexander, C., Low, M., Jones, W. R. & Rock, L. 2023 Computationally efficient ranking of groundwater monitoring locations. *Proceedings of the 37th International Workshop on Statistical Modelling* .

Rasmussen, C. E., Williams, C. K. *et al.* 2006 *Gaussian Processes for Machine Learning*. Springer.

Redner, R. A. & Walker, H. F. 1984 Mixture densities, maximum likelihood and the EM algorithm. *SIAM review* **26** (2), 195–239.

Revie, M., Wilson, K. J., Holdsworth, R. & Yule, S. 2017 On modeling player fitness in training for team sports with application to professional rugby. *International Journal of Sports Science & Coaching* **12** (2), 183–193.

Richardson, S. & Green, P. J. 1997 On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (Statistical Methodology)* **59** (4), 731–792.

Rossi, P. E., Allenby, G. M. & McCulloch, R. 2012 *Bayesian Statistics and Marketing*. John Wiley & Sons.

Sahoo, I. & Hazra, A. 2021 Contamination mapping in Bangladesh using a multivariate spatial Bayesian model for left-censored data. *arXiv preprint arXiv:2106.15730* .

Schmidt, F., Wainwright, H. M., Faybishenko, B., Denham, M. & Eddy-Dilek, C. 2018 In situ monitoring of groundwater contamination using the Kalman filter. *Environmental Science & Technology* **52** (13), 7418–7425.

Schwarz, G. 1978 Estimating the dimension of a model. *The Annals of Statistics* **6** (2), 461–464.

Silva, M. M. V. G., Gomes, E. M., Isaías, M., Azevedo, J. M. M. & Zeferino, B. 2017 Spatial and seasonal variations of surface and groundwater quality in a fast-growing city: Lubango, Angola. *Environmental Earth Sciences* **76**, 1–17.

Singh, A. & Nocerino, J. 2002 Robust estimation of mean and variance using environmental data sets with below detection limit observations. *Chemometrics and Intelligent Laboratory Systems* **60** (1-2), 69–86.

Sreekanth, J., Lau, H. & Pagendam, D. 2017 Design of optimal groundwater monitoring well network using stochastic modeling and reduced-rank spatial prediction. *Water Resources Research* **53** (8), 6821–6840.

Stan Development Team 2023 Stan modeling language users guide and reference manual. v2.26.1. https://mc-stan.org.

Steefel, C. I., DePaolo, D. J. & Lichtner, P. C. 2005 Reactive transport modeling: An essential tool and a new research approach for the earth sciences. *Earth and Planetary Science Letters* **240** (3-4), 539–558.

Stephens, M. 2000 Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62** (4), 795–809.

Summers, J. K. 2020 *Water Quality: Science, Assessments and Policy*. BoD.

Taffahi, H., Bensouda, N. & Salih-Alj, Y. 2013 Automated groundwater monitoring using telemetry. In *2013 4th International Conference on Intelligent Systems, Modelling and Simulation*, 596–600. IEEE.

Terry, N., Day-Lewis, F. D., Lane Jr, J. W., Trost, J. J. & Bekins, B. A. 2019 Geophysical mapping of plume discharge to surface water at a crude oil spill site: Inversion versus machine learning. *Geophysics* **84** (5), EN67–EN80.

Therneau, T. & Atkinson, B. 2022 *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1.19.

Tobin, J. 1958 Estimation of relationships for limited dependent variables. *Econometrica* **26** (1), 24–36.

Tobler, W. R. 1970 A computer movie simulating urban growth in the detroit region. *Economic geography* **46** (sup1), 234–240.

Tomlinson, D., Thornton, S., Thomas, A., Leharne, S. & Wealthall, G. 2014 *An Illustrated Handbook of LNAPL Transport and Fate in the Subsurface*. CL:AIRE.

Vehtari, A., Gabry, J., Magnusson, M. *et al.* 2023 loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.6.0.

Vehtari, A., Gelman, A. & Gabry, J. 2017 Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27** (5), 1413–1432.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y. & Gabry, J. 2015 Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646* .

Watanabe, S. & Opper, M. 2010 Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* **11** (12), 3571–3594.

Wedel, M. 2002 Concomitant variables in finite mixture models. *Statistica Neerlandica* **56** (3), 362–375.

Williams, C. J., Wilson, K. J. & Wilson, N. 2021 A comparison of prior elicitation aggregation using the classical method and SHELF. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **184** (3), 920–940.

World Health Organization 2010 Exposure to benzene: a major public health concern. *World Health Organization* 1–5.

Zeng, Q., Wen, H., Huang, H., Pei, X. & Wong, S. 2017 A multivariate random-parameters tobit model for analyzing highway crash rates by injury severity. *Accident Analysis & Prevention* **99**, 184–191.

Zens, G. 2019 Bayesian shrinkage in mixture-of-experts models: identifying robust determinants of class membership. *Advances in Data Analysis and Classification* **13** (4), 1019–1051.

Zhang, H., Zimmerman, J., Nettleton, D. & Nordman, D. J. 2019 Random forest prediction intervals. *The American Statistician* .

Zoffoli, H. J. O., Varella, C. A. A., do Amaral-Sobrinho, N. M. B., Zonta, E. & Tolón-Becerra, A. 2013 Method of median semi-variance for the analysis of left-censored data: comparison with other techniques using environmental data. *Chemosphere* **93** (9), 1701–1709.