Arsenic Contamination in Nova Scotia's Private Well Water: A Spatial-Temporal Statistical Analysis

Zirui Dong
Supervised by: Dr. Cindy Feng, Dr. Edward Susko
Department of Mathematics and Statistics, Dalhousie University

April 23, 2024

1 Abstract

Arsenic contamination in drinking water is a critical public health concern in Nova Scotia's well water. Long-term exposure to arsenic in drinking water exceeding the safety threshold could lead to a wide range of public health issues, including cardiovascular problems, skin conditions, and various types of cancer. While existing knowledge outlines connections between precipitation and arsenic levels, understanding of the intricacies of this relationship, and more broadly, the influence of climate change on arsenic in drinking water, remains limited. Factors affecting arsenic levels, including climate variation, exhibit gradual spatial-temporal variations. Despite the inherent time and space attributes in drinking water monitoring data, few studies have adopted a spatial-temporal modelling approach to analyze high-resolution spatial and temporal datasets.

Multiple datasets on climate variables arsenic, precipitation, and temperature data from 2000 to 2021 in Nova Scotia, Canada, were cleaned and linked. To model arsenic exceedance in Nova Scotia's well water, a generalized additive model with logistic regression is employed to predict the proportion of well water in Nova Scotia exceeding the threshold value of 5 microgrammes per litre (ug/L) for arsenic contamination. Regions with a high probability of arsenic exceedances were identified, likely due to environmental factors, with geographical location and groundwater region being the most significant.

2 Introduction

Arsenic is a naturally occurring, poisonous substance in air, water, and soil, which is present at an exceeding level against its safety threshold of 5 ug/L in the well water of Nova Scotia. Over the past decades, arsenic contamination in drinking water has become a significant public health concern.

The long-term effects of consuming drinking water with arsenic contamination over its safety threshold significantly harm the public. As such, exceeding a safety threshold is associated with an increased bladder and skin cancer risk. Also, exposure to arsenic could lead to lung, digestive tract, liver, kidney, and lymphatic and hematopoietic cancers.

The arsenic level corresponds to a wide range of environmental factors. Precipitation levels, temperature, and geographic disparities significantly impact the arsenic level. As a result, a multivariate statistical analysis is essential to model the arsenic level precisely, given environmental variables. Thus, this study utilized a generalized additive model with logistic regression to model the proportion of well water samples exceeding the safety threshold of 5 ug/L for arsenic contamination in Nova Scotia, Canada, from 2000 to 2021.

Arsenic modelling has been a crucial yet overlooked subject in public health and biostatistics. The growing recognition of the public health risks underscores a critical need for precise statistical modelling and prediction of arsenic levels in Nova Scotia's well water, as such an endeavour is significantly crucial to the safety of Nova Scotia's residents.

According to Kenndy and Drage (2017), a significant portion of Nova Scotia's private well users, approximately 37%, are situated in high-risk zones where arsenic levels exceed Health Canada's Maximum Allowable Concentration (MAC). Moreover, the overall proportion of private wells with arsenic levels surpassing the Health Canada MAC could be as high as 20%, potentially impacting around 90,000 individuals (Kenndy & Drage, 2017). This underscores the critical importance of implementing robust arsenic contamination protocols, particularly as many people are exposed to arsenic contamination risks in Nova Scotia.

This research explores the spatial effects of arsenic levels, finding the significant environmental factors, pinpointing the regions with substantial exceedances, and providing insights into the ecological factors causing arsenic exceedances. This research aims to serve as a vital resource for analyzing arsenic contamination in the public health sector of Nova Scotia, thus enhancing public health protection in Nova Scotia.

3 Data Collection and Preparation

3.1 Arsenic Concentration Data Collection

To model arsenic contamination in well water, daily data on arsenic concentrations from domestic wells in Nova Scotia, sourced from the NRR Geological Survey is obtained. To ensure the accuracy of the modelling, data collected between 2000 and 2021, excluding information from the islands in the region, is utilized.

Arsenic concentrations are in microgrammes per litre (ug/L), and a 5 ug/L safety threshold is applied. The arsenic concentrations were categorized into 0 (indicating safe) and 1 (indicating unsafe/exceeding the safety threshold) based on the safety threshold to facilitate logistic regression modelling.

To incorporate temporal effects into the analysis, a seasonal variable was introduced to account for the timing of each measurement.

The arsenic data collection contains several significant factors, including:

- Arsenic Concentration: Arsenic concentration has been categorized based on the 5 ug/L safety threshold. Concentrations within the threshold are 0, while those exceeding the threshold are 1.
- Location (Eastings, Northings): This factor represents the geographic coordinates of each domestic well, using the NAD83, zone 20 coordinate system.
- Time: Time-related information for each data record, including the date, month, and year.
- Season: The season in each data record.

3.2 Precipitation and Temperature Data Collection

Data from Climate Data Canada were utilized, providing mean monthly precipitation and temperature values corresponding to the time column in the arsenic dataset. Initially recorded as longitude and latitude, the location information in the precipitation and temperature data was converted to eastings and northings in NAD83, zone 20.

The climate data considered in this study includes the following information:

- Location (Eastings, Northings): Spatial coordinates of each weather station in the NAD83, zone 20 coordinate system.
- Time: Time-related information for each record, including the month and year.
- Precipitation: Representing the 50th percentile of precipitation total under the Representative Concentration Pathway (RCP) 2.6 scenario.
- Temperature: Indicating the 50th percentile of mean temperature (tg) under the Representative Concentration Pathway (RCP) 2.6 scenario.

3.3 Merging Arsenic Concentration, Precipitation, and Temperature Data

The arsenic concentration, precipitation, and temperature data were merged using Euclidean distance, resulting in a new dataset for subsequent analysis.

To combine the arsenic concentration dataset with the climate data effectively, we implemented a matching process. Initially, we systematically looped through each entry in the arsenic concentration dataset. For each entry, we compared its temporal and spatial attributes with those in the climate dataset. This comparison enabled us to identify the most closely matched point in the climate data corresponding to each entry in the arsenic dataset.

4 Methods

This section introduces two key components of the statistical analysis: generalized additive models (GAMs) and logistic regression within the framework of GAMs. These statistical methods are crucial for understanding the relationships between arsenic levels in well water and their predictor variables.

4.1 Logistic Regression for Binary Responses

Logistic regression is a statistical model specifically designed to predict the probability of binary outcomes.

Logistic regression uses the logit function to linearize the relationship between the dependent and independent variables, making it suitable for regression analysis.

The logit function is written as:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

The logit function is the inverse of the logistic function, which transforms the probability p of the dependent variable to the log-odds $logit(p)$ to map probabilities from the range (0,1) to the entire real number line.

The logistic regression equation in the form of GAMs could be presented as follows:

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+x_1\beta_1+x_2\beta_2+x_3\beta_3+x_4\beta_4\ +\dots)}}$$

where

- $P(Y=1)$ represents the probability that the binary outcome variable of Y equals 1.
- β_0 is the intercept, representing log-odds of binary outcome variable equals 11 while all predictor variables equal 0.
- $x_1, x_2, \dots, x_k$ are predictor variables in logistic regression.
- $\beta_1, \beta_2, \dots, \beta_k$ are coefficients of each predictor variables $x_i, x_2, \dots, x_k$.

In logistic regression, the predicted probability ($\hat{p}_i$) of the event (e.g., $Y = 1$) occurring for the i-th observation is denoted as $P(Y = 1|X_i)$, where X_i represents the predictor variables for the i-th observation.

The predicted probability $\hat{p}_i$ represents the model's estimate of the probability that the event (e.g., success) will occur for the i-th observation, given its predictor variables X_i. In other words, it is the model's estimate of the expected value of the binary outcome variable Y for the i-th observation.

For example:

- If $\hat{p}_i = 0.7$, the model predicts a 70% probability that the event (e.g., success) will occur for the i-th observation, given its predictor variables.
- If $\hat{p}_i = 0.3$, the model predicts a 30% probability that the event (e.g., success) will occur for the i-th observation, given its predictor variables.

The logistic regression model transforms the linear combination of predictor variables and their coefficients using the exponential function into a calculated probability between 0 and 1. This probability is later utilized to predict the binary outcome. The threshold of logistic regression is often 0.5, but in practice, the appropriate threshold is determined and adjusted based on the specific context of statistical research.

In this research, logistic regression is a valuable tool for determining whether arsenic contamination in Nova Scotia's well water will likely exceed the safety threshold of 5 microgrammes per litre (ug/L). The arsenic levels are classified as either high or low, where high indicates a level exceeding the safety threshold of 5 microgrammes per litre (ug/L), while low indicates a level within the safety limit.

Additionally, to enhance the predictive modelling, the Generalized Additive Model (GAM) is merged with logistic regression. The GAM allows for the incorporation of smooth functions, accommodating non-linear relationships between predictor variables and the response variable, which is crucial in capturing the complexity of arsenic contamination patterns. This integration of GAM and logistic regression offers a more comprehensive and nuanced understanding of the factors influencing arsenic levels in Nova Scotia's well water.

4.2 Modelling with GAMs

Generalized additive models, known as GAMs, are statistical modelling techniques wherein the linear response variable is presumed to be a function of linear combinations of unknown smooth functions of one or more predictor variables. GAMs employ smoothing techniques to capture nonlinear relationships between predictors and the response variable. GAMs are exceptionally useful when a nonlinear relationship exists between predictor and outcome variables. GAMs extend the traditional linear regression framework by allowing nonlinear relationships between the response and predictor variables. The application of smooth functions serves to linearize nonlinear effects between the response variable and predictor variables.

In the context of generalized additive models, the usual way the link functions without specification is $g(\mu_i) = \mu_i$, where μ_i is the mean response. In this study, the link function is specified as logit, $g(\mu_i) = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$, where i is the mean response for the i -th individual.

A generic way of defining the parameters of the GAM model with spatial-temporal terms is

$$g(\mu_i) = X_i^T \theta + \sum_j f_j(Z_{ij}) \tag{1}$$

. Where X_i stands for the predictor variables without smoothing functions, and θ is the corresponding parameter vector. $Z_{i,j}$ stands for i -th observation for j -th predictor variable with smoothing function applied. Since the effects of space and time are included, spatial location terms i and time t are added to the formula, where i refers to the spatial location of each private well, and t relates to the year and month of each observation. In this model, the equation of the GAM is written as:

$$g(\mu_{i,t}) = \alpha_0 + Source_{i,t} + f_1(Eastings_{i,t}, Northings_{i,t}) + Season_{i,t} +$$
$$GroundwaterRegion_{i,t} + f_2(Dates_{i,t}) + f_3(Precipitation_{i,t}) +$$
$$f_4(Temperature_{i,t})$$

where

- α_0: The intercept term.
- $g(\mu_{i,t}) = \text{logit}(\mu_{i,t}) = \log\left(\frac{\mu_{i,t}}{1-\mu_{i,t}}\right)$: The logit link function transforms the probabilities from 0 to 1 to the entire real number line.
- i represents the spatial location of the wells.
- t represents the date when arsenic is measured.
- $Precipitation_{i,t}$ and $temperature_{i,t}$ represent the average monthly precipitation and temperature values that t is considered at location i.
- $GroundwaterRegion_{i,t}$ represent the bedrock geology of groundwater region at location i with time t.
- $\mu_{i,t} = P(Y = 1|X_i, Z_i) = \frac{\exp[g(\mu_{i,t})]}{1+\exp[g(\mu_{i,t})]}$: The expected value of the response variable for the i-th observation.
- $f_1, \ldots, f_4$: Thin-plate regression spline (TPRS) functions of the corresponding predictor variables.

The choice of the GAM model was due to its versatility in accommodating smooth functions, such as splines. This feature proves invaluable in detecting nonlinear impacts of predictor variables on arsenic levels and spatial and temporal variables. It makes the GAM model particularly suitable for analyzing data with varying geographical locations, providing a comprehensive understanding of the factors influencing arsenic contamination across diverse locations in Nova Scotia.

Thin-plate Regression Spline (TPRS) Smooth functions in the context of GAMs refer to flexible functions used to model the relationship between predictor variables and response variables nonlinearly. These functions capture complex patterns and relationships that linear relationships cannot describe. Smooth functions are applied to individual predictor variables and are often represented using spline functions, smoothing splines, or penalized splines. Splines are considered an improvement over piecewise polynomials as they use multiple polynomial segments connected smoothly at specific points called knots.

Thin-plate splines are smoothing splines used by default in the mgcv package in R. The thin plate spline function, denoted as $f_j(Z)$, is defined as:

$$f_j(Z) = \sum_k \delta_k |Z - Z'_{k,j}|^3 + b_1 + b_2 Z$$

where

- $f_j(Z)$ represents the smooth function of the predictor variable Z.
- $\sum_k$ denotes a summation of all the basis functions in the smooth term.
- δ_k represents a coefficient to be measured.
- $|Z - Z'_{k,j}|$ represents the absolute difference between the predictor variable Z and the knots $Z'_{k,j}$ are the distinct observed values of $Z_{i,j}$.
- 3 represents the degree of the local polynomial used in the smoothing function. In this case, a cubic polynomial is used.
- b_1 and b_2 are coefficients associated with linear terms (e.g., linear trend) included in the smoothing function.

The penalty term of thin-plate regression spline functions is defined as:

$$U_j(f_j) = \int \left[\left(\frac{d^2 f_j(z)}{dz^2} \right) \right]^2 dz = \boldsymbol{\alpha^T S_j \alpha}$$

where

- $U_j(f_j)$ represents the objective function, which calculates the squared integral of the second derivative of the j -th smooth function.
- f_j represents the j -th smooth function.
- $\frac{d^2 f_j(z)}{dz^2}$ denotes the second derivative of the smooth function $f_j(z)$ with respect to z.
- $\int$ denotes the integral sign, indicating the integration over the entire range of the predictor variable Z_j.
- α represents a vector of coefficients of the marginal smooths.

- S_j represents the penalty matrix or penalty function, which does not involve the unknown parameters for the basis functions α.

The expression $U_j(f_j)$ is used in penalized regression methods, such as penalized splines or generalized additive models (GAMs), to impose smoothness on the fitted function. By penalizing the second derivative, it encourages a smoother fit to the data.

Penalized Log-Likelihood Maximization In the context of TPRS, the model could be written as $g(\mu) = X\theta$, where $g(\mu)$ means an n-vector $\mu = (\mu_1, \ldots, \mu_n)$ with link function $g()$ applied. To estimate the parameter θ, which contains all the parameters needed to be estimated, the aim is to maximize the penalized log-likelihood

$$l_p(\theta|y) = l(\theta|y) - \frac{1}{2}\sum_{j=1}^{U} \lambda_j U_j(f_j)$$

where

- λ_j is the j -th smoothing parameter, which provides a tradeoff between the model's goodness of fit and smoothness.
- U_j are the penalty terms with U denoting the total number of penalty terms.
- $l(\theta|y) = \sum_{i=1}^{n} \{y_i \log[p(y_i = 1|\theta)] + (1 - y_i)\log[1 - p(y_i = 1|\theta)]\}$ is the log-likelihood associated with Bernoulli response (Wood, 2006).

Note that X here is not the X matrix associated with the fixed predictors, but a matrix that includes terms like $|Z - Z'_{k,j}|$. and $Z_{i,j}$. The vector θ includes β all the δ parameters and the b_1 and b_2 parameters.

The penalized log-likelihood is maximized by a penalized iteratively reweighted least squares (P-IRLS) algorithm, which minimizes the penalized sum of squares of

$$\sum_{i=1}^{n} \left\{ w_i^{[m]}(z_i^{[m]} - g_i) \right\}^2 + \sum_{j=1}^{J} \lambda_j U_j(f_j)$$

where

- m, denoting the m -th penalized iteratively reweighted least squares (P-IRLS) iteration
- $z_i^{[m]} = g_i^{[m]} + g'(\mu_i^{[m]})(y_i - \mu_i^{[m]})$, the pseudo data vector in fitting TPRS
- $w_i^{[m]} = \frac{1}{\sqrt{V(\mu_i^{[m]})g'(\mu_i^{[m]})^2}}$, the weight matrix taking account of variances
- $V_i^{[m]}$ is proportional to the variance of $Y_i^{[m]}$ according to the current estimate $\mu_i^{[m]}$ (Feng 2022).

Knots Placement According to Simon Wood (2023), the GAMs fitting algorithm treats all the unique observations as knots, and "thin plate regression splines are constructed by starting with the basis and penalty for a full thin plate spline and then optimally truncating this basis, to obtain a low rank smoother" (Wood, 2023).

Regularization Parameter Since the iterative process of leave-one-out cross-validation is computation intensive with data sets containing large numbers of observations, λ_j, the smoothing parameters, "are estimated by minimizing the cross-validation score for each working penalized linear model of the P-IRLS iteration." The score is formulated as:

$$\mathrm{GCV} = \frac{n \left\| \sqrt{W}(z - Z\theta) \right\|^2}{[n - \mathrm{tr}(A)]^2}$$

where:

- A is the influence matrix and $tr(A)$ is the effective degrees of freedom or the effective number of parameters.
- $\mathbf{z}$ is a vector of pseudodata z_i. z_i are pseudodata in the form of $z_i = g(\mu_i) + g'(\mu_i)(y_i - \mu_i)$.
- W is the weight matrix that accounts for the weighting for each observation for variance and curvature, with $w_i = \frac{1}{\sqrt{V(\mu_i)g'(\mu_i)^2}}$ variances.

4.3 Effective Degrees of Freedom

The Effective Degrees of Freedom (EDF) in GAM measures the linearity of the smooth terms in the model. An EDF of 1 represents a linear relationship, and an EDF greater than 2 means a solid non-linear relationship.

The EDF is calculated by $trace(A)$,

$$A = X(X^T W X + S)^{-1} X^T W$$

where:

- A is the influence matrix.
- X is the redefined design matrix containing all the predictor variables. It includes terms like the original $X_{i,j}$ and terms from the smooth functions like $|Z_{i,j} - Z'_{k,j}|^3$.
- S is the smoothing matrix representing penalties on the coefficients, it depends on the S_j coming from the penalty terms.

The equation calculates the effective degrees of freedom matrix by applying penalties to the model's smoothing or regularization terms. The penalties control the flexibility or non-linearity in the model, and the resulting A matrix quantifies the effective degrees of freedom associated with each smoothing or regularization term. This matrix is essential for understanding the complexity of the model and for model selection or regularization purposes. The EDF is calculated by summing the diagonal components of the degrees of freedom matrix A.

4.4 Software and Tools

Data analysis was conducted using a combination of statistical software and packages. The following software and tools played a significant role in facilitating the research:

- R: R, an open-source programming language renowned for its rich repository of libraries and packages, was chosen as the primary platform for this statistical data analysis.
- dplyr: The 'dplyr' package facilitates efficient data manipulation, enabling essential data cleaning tasks to be executed on raw and unstructured datasets.
- ggplot2: For data visualization. With 'ggplot2,' informative heat maps are created to visually represent arsenic levels and other predictor variables with geographical locations across Nova Scotia. These visualizations enhanced the capacity to convey data insights intuitively.
- mgcv: The 'mgcv' package in R proved indispensable in the modelling process. This package enabled users to fit a GAM with logistic regression, offering a robust framework for modelling the intricate relationship between arsenic levels and the selected predictor variables.

5 Exploratory Data Analysis

In this exploratory data analysis, we focus on investigating the distribution of arsenic level relationship between arsenic levels and key predictor variables Additionally, we examine seasonal and spatial variations in arsenic levels to gain a comprehensive understanding of the temporal and geographic patterns of arsenic contamination.

5.1 Distribution of Arsenic Levels

5.1.1 Histogram of Arsenic Levels

Figure 1 gives a histogram depicting the arsenic levels found in well water across Nova Scotia during the study period spanning from 2000 to 2021. This histogram aids in comprehending the distribution pattern of arsenic levels within the region. The graph illustrates that the predominant range for arsenic levels spans from 0 ug/L to 5 ug/L, encompassing the majority of recorded values. Conversely, only 371 out of 2444 arsenic measurements surpassed the safety threshold of 5 ug/L.

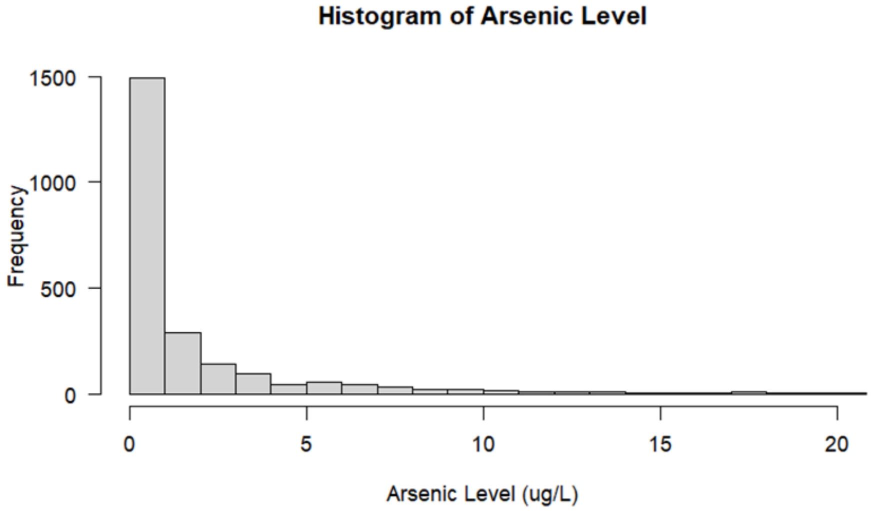

Figure 1: The distribution of arsenic level in ug/L using histogram.

5.2 Distribution of Arsenic Levels by Predictor Variables:

5.2.1 Proportion of Arsenic Exceedances by Well Type

Figure 2 gives a bar plot to visualize the proportion of arsenic exceedances categorized by well source. Drilled wells have the highest proportion of arsenic exceedance of 17.9%, suggesting a potential vulnerability in water quality associated with this source. Wells of unknown sources have slightly lower arsenic exceedances of 17.6%. Dug well has the lowest proportion of arsenic exceedance of 2.5%.

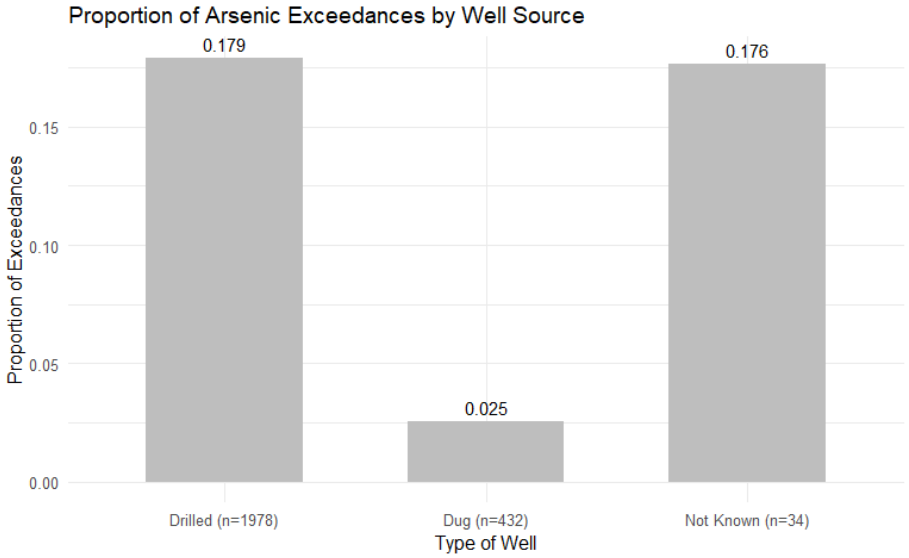

Figure 2: The proportion of arsenic exceedances by well type with bar plot.

5.2.2 Counts of Arsenic Exceedances by Well Type

Figure 3 provides a bar plot to illustrate the counts of arsenic exceedances categorized by well source. Drilled wells exhibit the highest count of arsenic exceedances, totalling 354. In contrast, dug wells and unknown sources demonstrate lower counts of arsenic exceedances, with 11 and 6 instances, respectively.

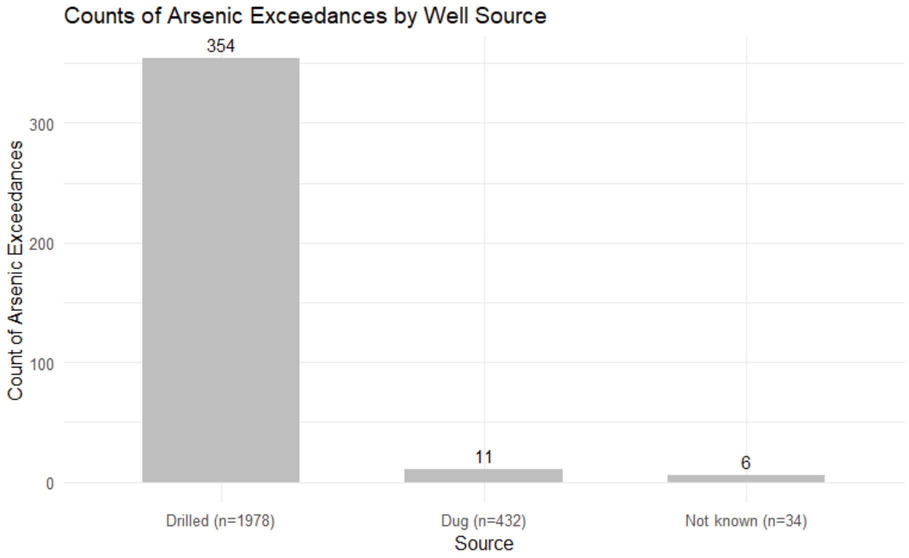

Figure 3: The counts of arsenic exceedances are categorized by well type using a bar plot.

5.2.3 Proportion of Arsenic Exceedances by Groundwater Region

Figure 4 gives a bar plot to visualize the proportion of arsenic exceedances categorized by groundwater region. Metamorphic and plutonic exhibit a higher proportion of arsenic exceedances of 29 % and 25 % compared to other regions, while the surficial groundwater region has the lowest proportion of arsenic exceedances of 2%.

5.2.4 Counts of Arsenic Exceedances by Groundwater Region

Figure 5 provides a bar plot to illustrate the counts of arsenic exceedances categorized by groundwater region. Sedimentary has the highest count of arsenic exceedances of 133, followed by metamorphic of 129. The volcanic region has the lowest exceedance count of 5, and the surficial has the second-lowest exceedance count of 13.

5.3 Seasonal Variation of Arsenic Levels

5.3.1 Proportion of Arsenic Levels Exceeding Threshold in Each Month and Each Season

Understanding the monthly and seasonal distribution of arsenic levels is crucial for detecting temporal variations in contamination trends. This study computed the proportion of arsenic levels surpassing the safety threshold of 5 µg/L

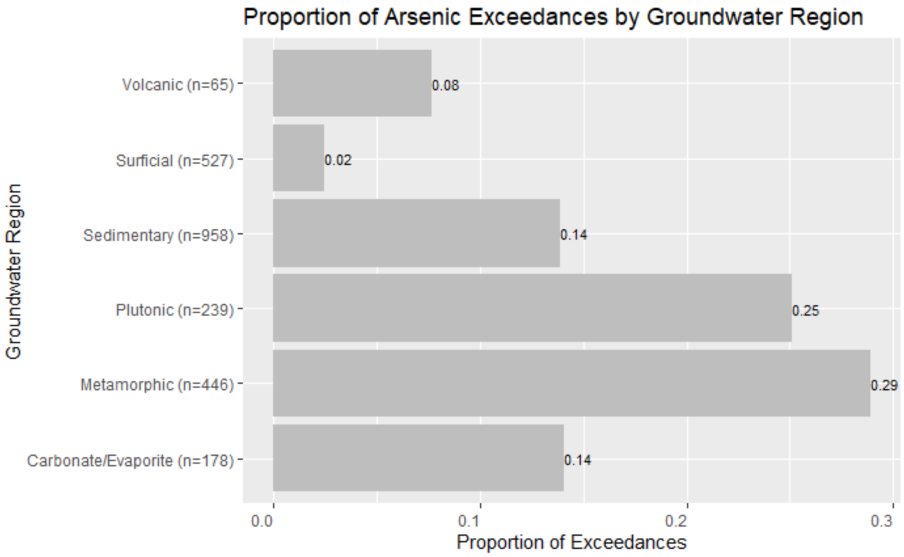

Figure 4: The proportion of arsenic exceedances by groundwater region using a bar plot.

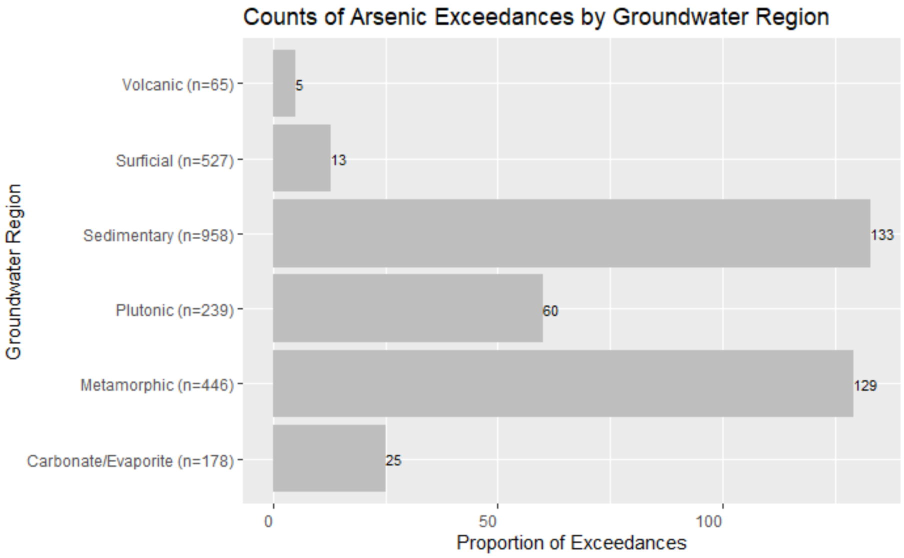

Figure 5: The counts of arsenic exceedances are categorized by groundwater region using a bar plot.

for each month. This graphical depiction allows the discerning of any observable monthly and seasonal patterns and fluctuations in arsenic contamination, offering insights into potential seasonal trends. Figure 6 gives a bar plot to examine the monthly and seasonal distribution of arsenic levels is crucial for identifying temporal variations in contamination. The proportion of arsenic levels exceeding the safety threshold of 5 µg/L is calculated for each month. This graphical representation allows the discovery of any discernible monthly and seasonal patterns and fluctuations in arsenic contamination, providing insights into potential seasonal trends.

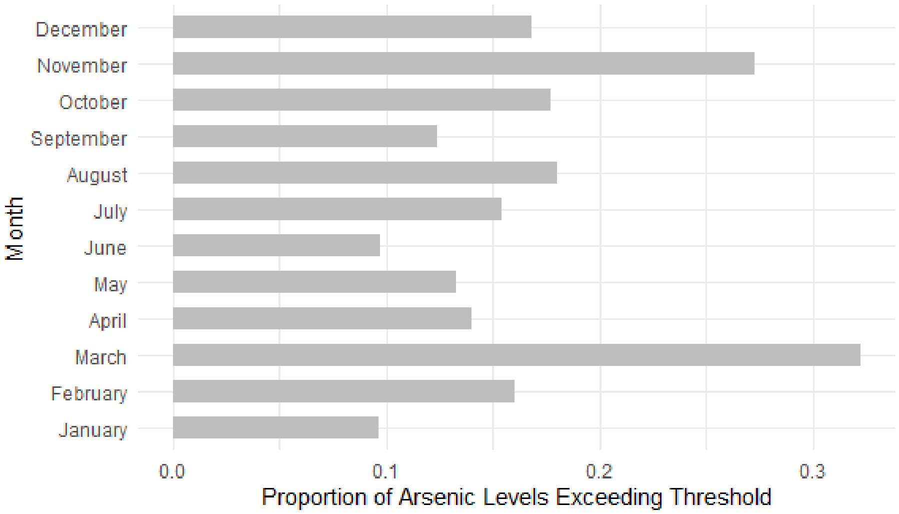

Figure 6: The proportion of arsenic levels exceeding the safety threshold of 5 ug/L in each month.

Figure 7 gives a bar plot to visualize the proportion of arsenic exceedances across seasons. Elevated proportions are observed in March and November. This observation suggests a potential seasonal trend in arsenic contamination, with higher occurrences during these specific months. The graph highlights spring as having the highest proportion, followed by fall, winter, and summer, revealing a distinct seasonal pattern in arsenic contamination. Further analysis and exploration of contributing factors during periods with a higher proportion of arsenic contamination may provide valuable insights into the temporal dynamics of arsenic levels in Nova Scotia's well water.

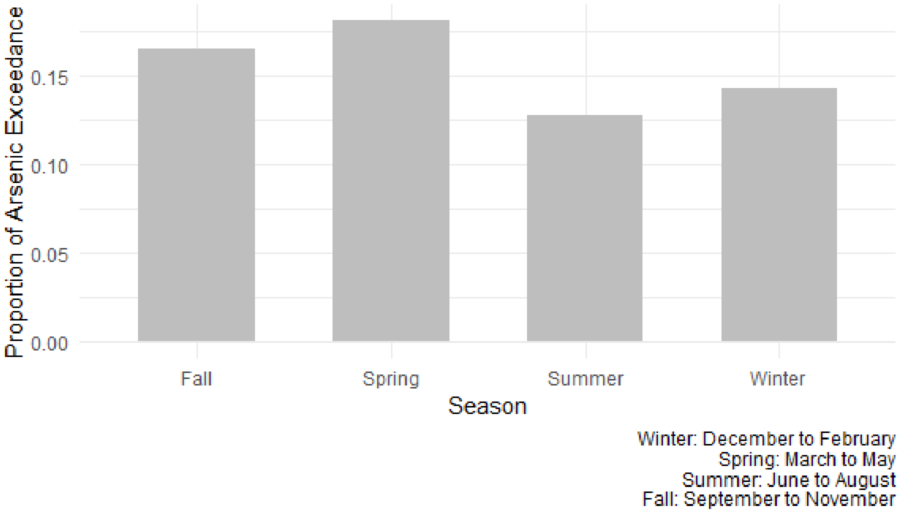

Figure 7: The proportion of arsenic levels exceeded the safety threshold of 5 ug/L in each season.

5.4 Distribution of Precipitation Levels

5.4.1 Histogram of Precipitation Levels by Seasons

Figure 8 presents a histogram depicting the distribution of precipitation levels across the study period (2000 to 2021). This graphical representation allows for a visual assessment of the variability and frequency of precipitation, providing insights into potential seasonal trends.

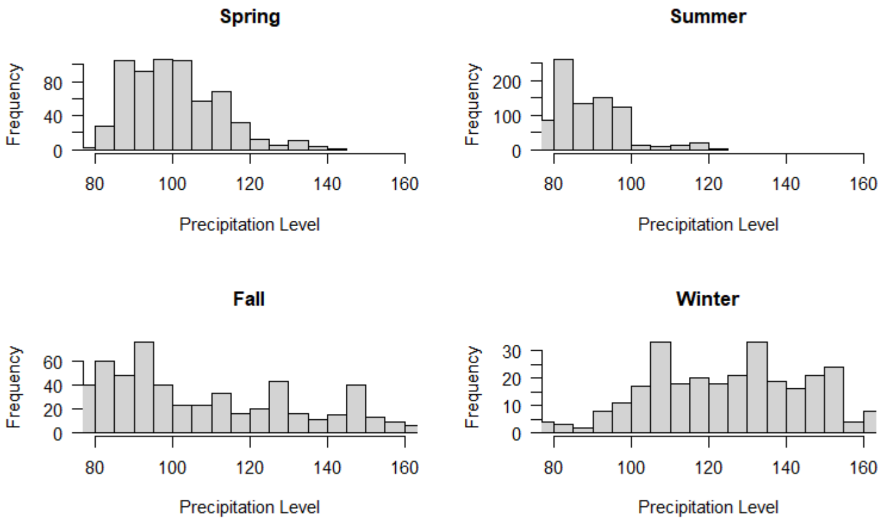

Figure 8: The histogram of precipitation levels is categorized by seasons in a histogram.

The histogram depicting precipitation levels in Nova Scotia reveals a predominant concentration of records within the range of 80 to 100. The distribution exhibits a notable right skewness, indicating that most precipitation observations are clustered toward the lower end of the scale. However, a few outliers extend beyond 140, suggesting exceptionally high precipitation. This right-skewed pattern implies that while most of the recorded precipitation values fall within a specific range, there are occasional more intense and extreme precipitation events in the dataset.

5.4.2 Bar Plot of Precipitation Levels by Months

Figure 9 provides a bar plot of precipitation levels by each month to visualize the distribution of precipitation levels across the year.

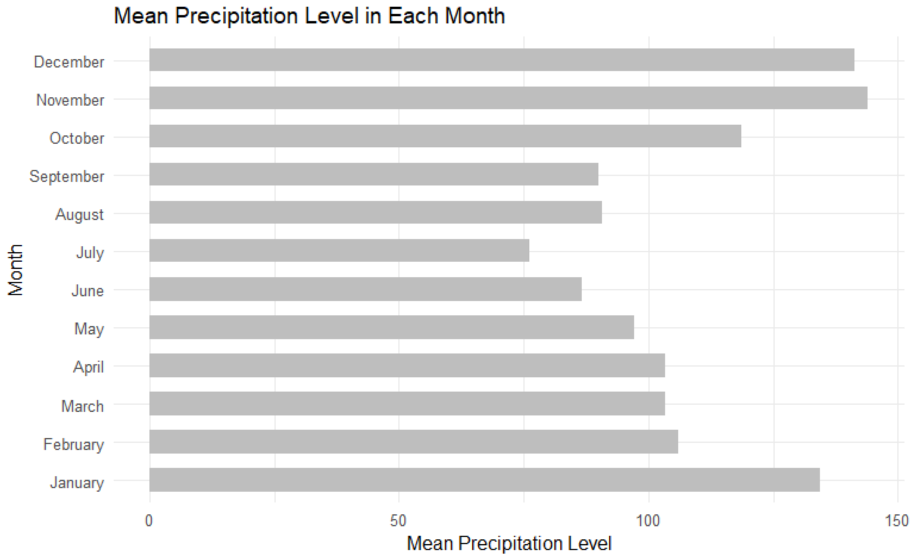

Figure 9: The mean precipitation level in each month in a bar plot.

5.5 Spatial Pattern of Arsenic Levels

5.5.1 Distribution of Arsenic Levels Exceeding Threshold in Nova Scotia

To gain insights into the spatial distribution of arsenic contamination across Nova Scotia, we conducted a spatial examination by mapping the distribution of arsenic levels exceeding the safety threshold. The dataset was partitioned into 20 by 20 grids to create the heat map, computing the average proportion of arsenic exceedances within each grid. Subsequently, this information was depicted through a heatmap using ggplot. Figure 10 displays a heat map that maps the distribution of arsenic levels surpassing the safety threshold across Nova Scotia, providing a spatial examination of arsenic contamination. This visualization aids in identifying geographic hotspots of contamination. Incorporating these visual representations into the results allows for a detailed exploration of the distribution of arsenic levels. Yarmouth and Halifax exhibit substantially higher arsenic levels than other Nova Scotia regions.

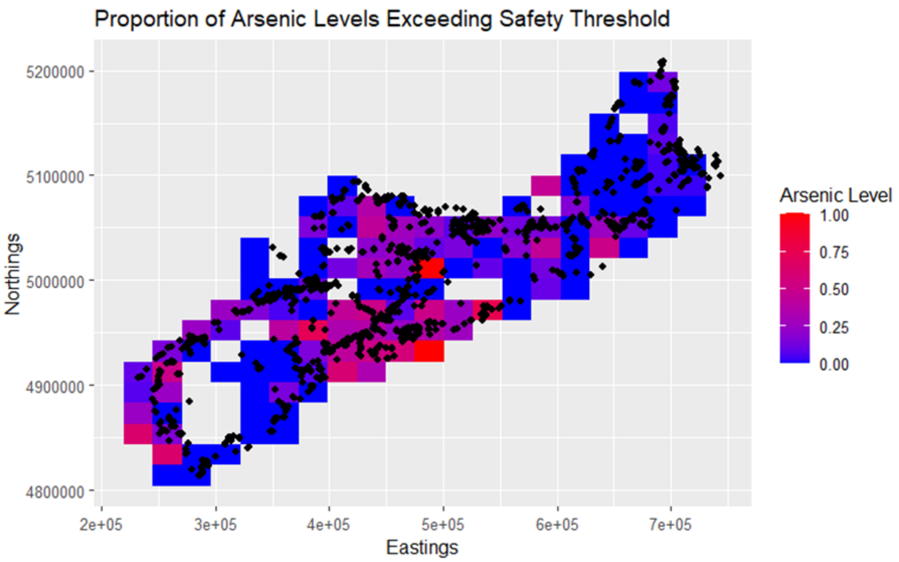

Figure 10: The proportion of private wells with arsenic levels exceeding the safety threshold of 5 ug/L in Nova Scotia.

5.5.2 Relationship between Arsenic Levels and Precipitation Levels

This study discovered a significant positive correlation between arsenic and precipitation levels in Nova Scotia's well water. According to a generalized additive model with precipitation levels being a predictor variable and arsenic levels as the response variable, the estimated approximate significance of smooth term of precipitation levels provided an adequate degree of freedom of 1.001, with a significant p-value of 1.25e-05, suggesting a positive, near-linear effect between the precipitation levels and the arsenic levels.

Spatial and Temporal Variation of Precipitation Levels and Arsenic Levels Figure 11 to Figure 18 provide stratified heat maps to visualize the spatial and temporal variation of precipitation levels and arsenic levels across Nova Scotia every five years to visualize the correlation between precipitation and arsenic levels. The findings highlight a substantial correlation between precipitation and arsenic levels, revealing that regions experiencing elevated precipitation, particularly Yarmouth and Halifax also exhibit higher arsenic contamination.

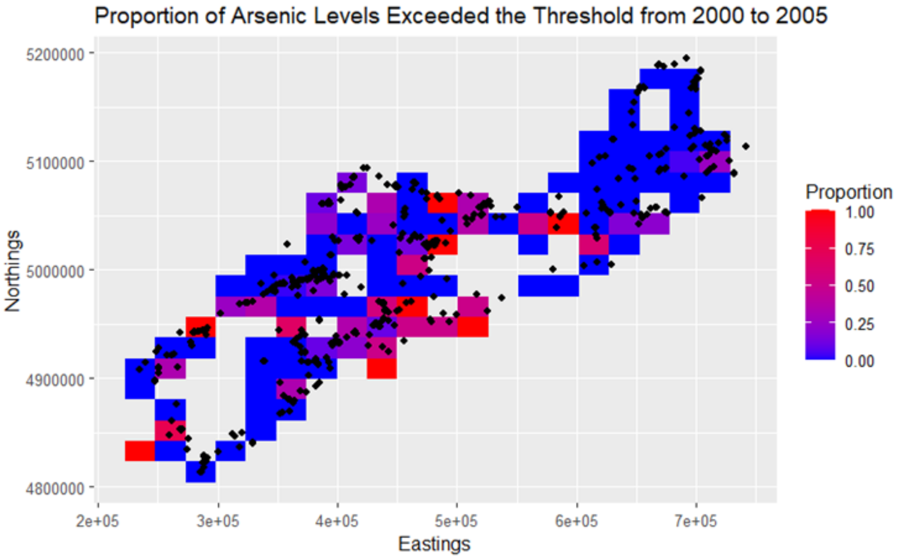

Figure 11: The proportion of private wells with arsenic levels exceeding the safety threshold of 5 ug/L from 2000 to 2005.

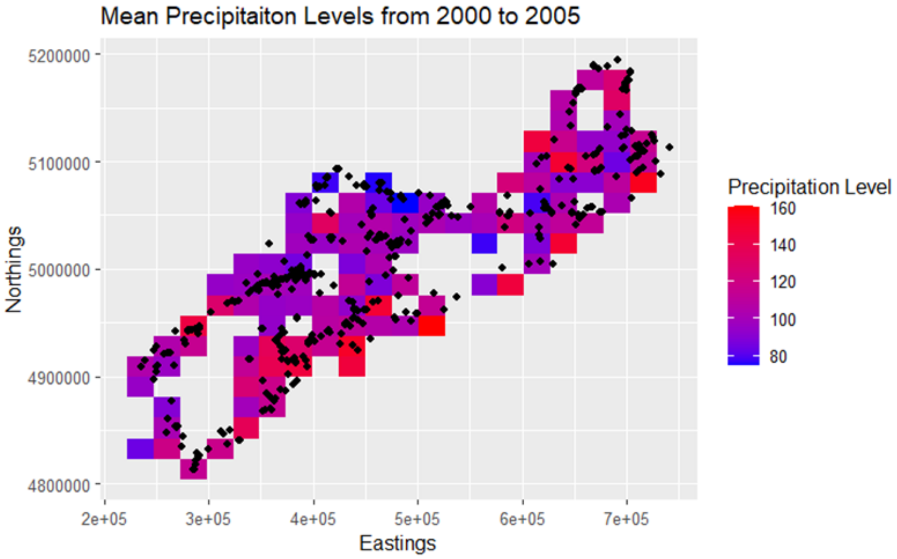

Figure 12: The mean precipitation levels from 2000 to 2005.

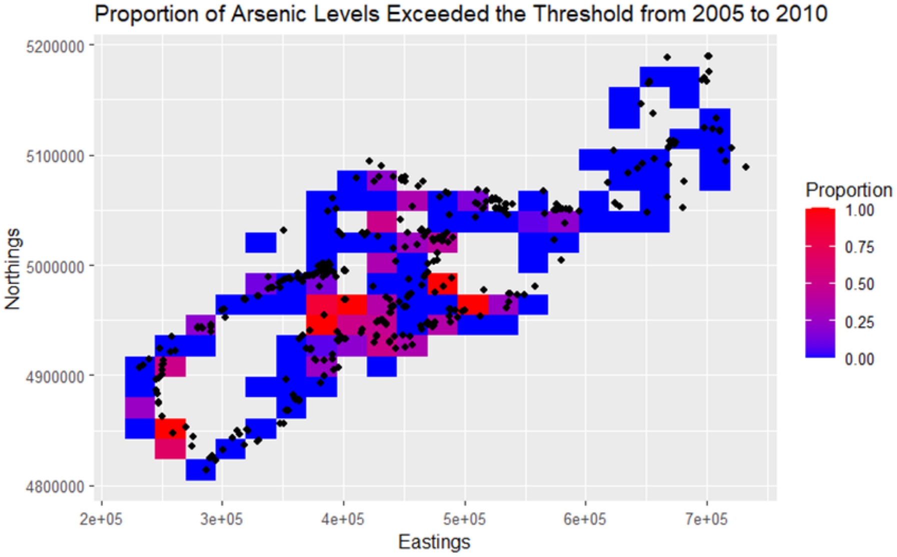

Figure 13: The proportion of private wells with arsenic levels exceeding the safety threshold of 5 ug/L from 2005 to 2010.

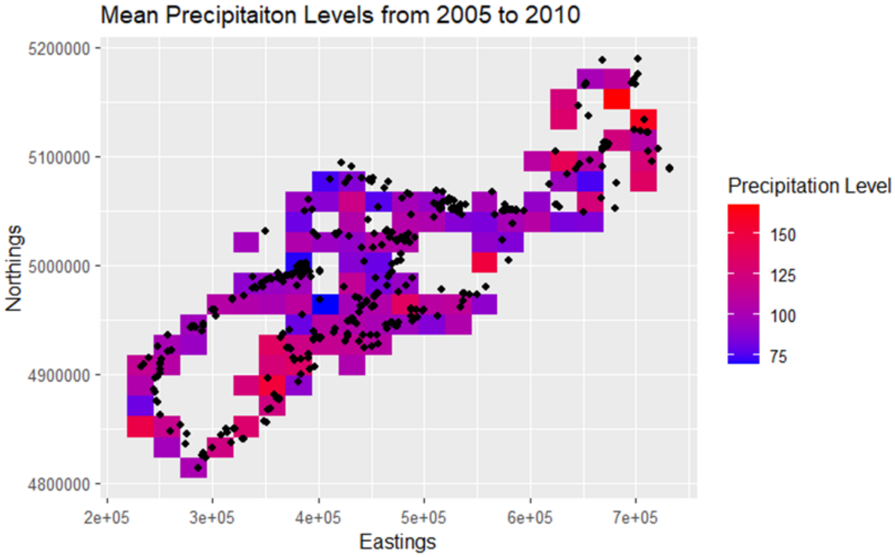

Figure 14: The mean precipitation levels from 2005 to 2010.

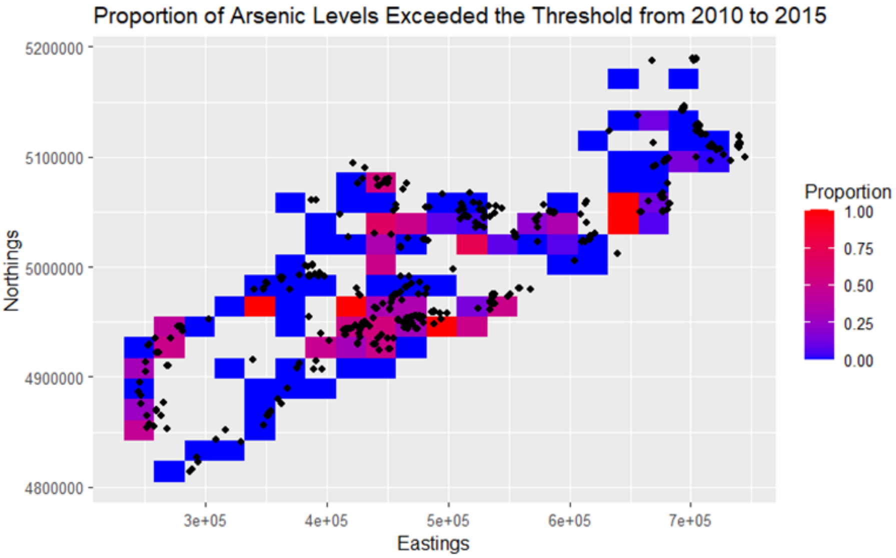

Figure 15: The proportion of private wells with arsenic levels exceeding the safety threshold of 5 ug/L from 2010 to 2015.

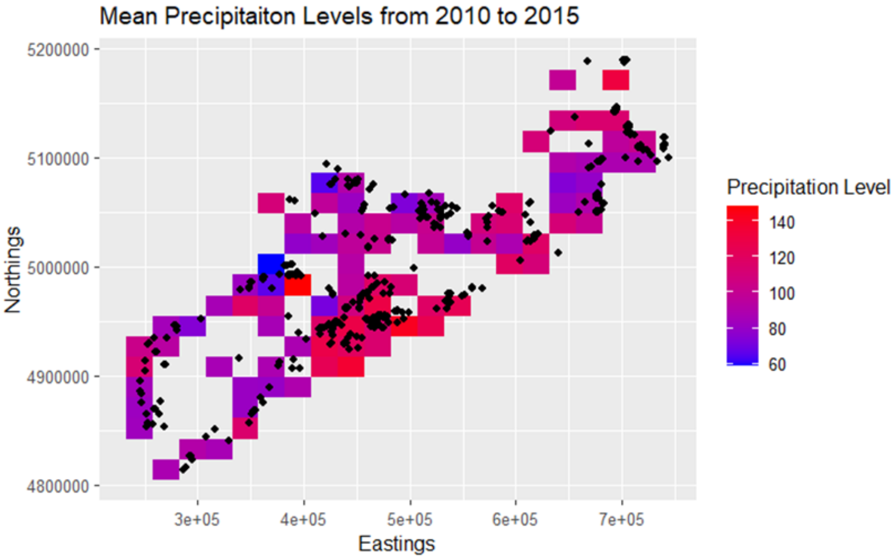

Figure 16: The mean precipitation levels from 2010 to 2015.

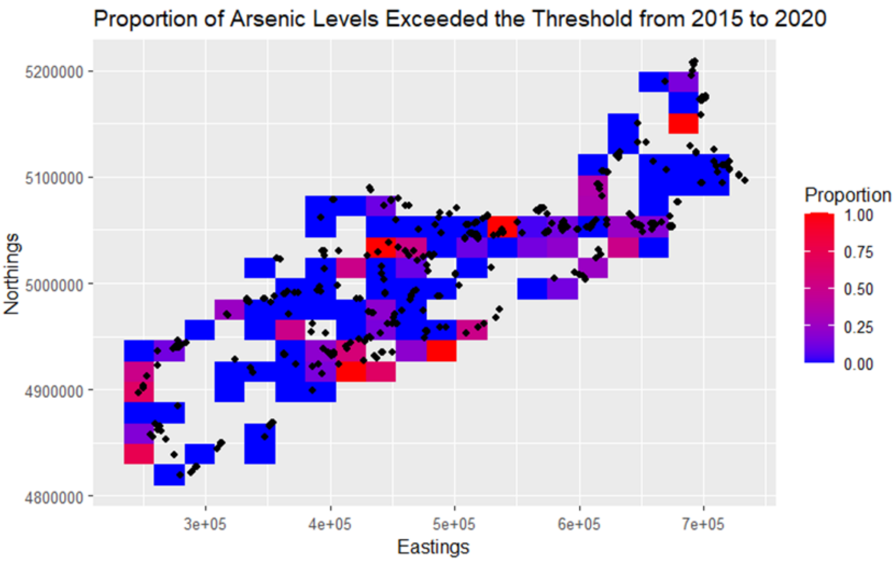

Figure 17: The proportion of private wells with arsenic levels exceeding the safety threshold of 5 ug/L from 2015 to 2020.

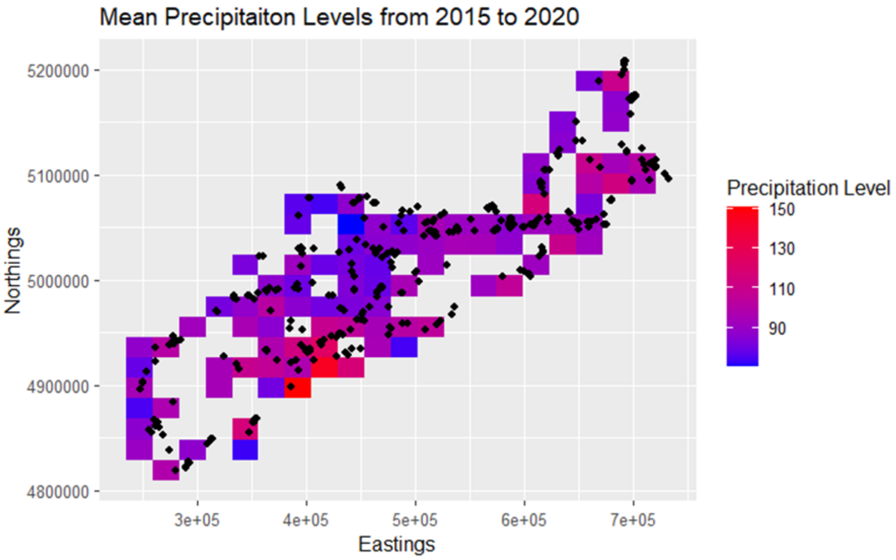

Figure 18: The mean precipitation levels from 2015 to 2020.

5.6 Interaction Between Well Type and Groundwater Region

Table 1 shows the interactions between well type and groundwater region variables, providing a breakdown of groundwater types categorized by the source of the well, offering insights into how different types of wells are distributed across various groundwater regions.

Table 1: The counts of observations illustrate the interaction between well type and groundwater regions.

	Surficial	Carbonate /Evaporite	Metamorphic	Not Known	Plutonic	Sedimentary	Volcanic
Dug	432	0	0	0	0	0	0
Drilled	92	178	446	0	239	958	65
Not Known	3	0	0	31	0	0	0

6 Results

6.1 Results of multivariate regression analysis

In this section, we present the results of several competing GAMs developed to investigate the relationship between predictor variables and arsenic contamination in well water. We began by considering a full model that included all available covariates. Subsequently, we developed a series of reduced models by excluding certain covariates based on theoretical considerations and model fit criteria. We present the results of each model, starting with the full model and then progressing to the reduced models, to provide insight into the relative importance of different predictor variables in predicting arsenic contamination. This approach allows us to assess the robustness of our findings and identify key predictors associated with elevated arsenic levels in well water.

6.1.1 Results of the full model analysis

Table 2 shows the parametric coefficients representing the estimated relationships between the predictor and response variables. The odds of arsenic exceedance are significantly higher in drilled wells, spring season, and metamorphic and sedimentary groundwater regions. Here are the findings:

Notably, the odds of arsenic contamination are significantly higher in drilled wells, compared to other well types. Spring seasons have higher odds compared to other seasons. Additionally, the type of groundwater region also plays a role, with metamorphic and sedimentary regions showing higher odds of contamination. As shown in Table 2, the odds of arsenic exceeding the safety threshold for drilled wells are 10% higher than for dug wells (OR=1.109, 95% CI: 0.227, 5.426). Season data indicates that spring exhibits the highest odds of arsenic exceedances, with an odds ratio of 3.186 (95% CI: 1.543, 6.578), followed by fall, showing an odds ratio of 2.347 (95% CI: 1.055, 5.220), and summer with an odds ratio of 2.134 (95% CI: 0.859, 5.301). Groundwater region analysis reveals that carbonate/evaporite areas have notably higher odds of arsenic exceedances, with an odds ratio of 14.580 (95% CI: 3.157, 67.327). Metamorphic regions follow, displaying an odds ratio of 10.462 (95% CI: 2.360, 46.383), while plutonic and sedimentary areas show odds ratios of 8.347 (95% CI: 1.719, 40.523) and 8.763 (95% CI: 2.008, 38.246) respectively. Volcanic regions exhibit lower but still notable odds, with an odds ratio of 4.642 (95% CI: 0.805, 26.768).

Table 2: Summary of multivariable logistic GAM results for parametric terms. Exceedances represent the proportion of arsenic exceeding the threshold. (n=2444)

Predictor Variable	**Odds Ratio**	**95% CI for Odds Ratio**	***P*-value**	n	**Exceedances**
Well Type					
- Drilled Well	1.109	(0.227, 5.426)	0.898	1978	0.179
- Dug Well	1			432	0.025
Season					
- Spring	3.186	(1.543, 6.578)	0.002	629	0.181
- Fall	2.347	(1.055, 5.220)	0.036	557	0.165
- Summer	2.134	(0.859, 5.301)	0.102	964	0.127
- Winter	1			294	0.143
Groundwater Region					
- Carbonate / Evaporite	14.580	(3.157, 67.327)	< 0.001	178	0.140
- Metamorphic	10.462	(2.360, 46.383)	0.002	446	0.289
- Plutonic	8.347	(1.719, 40.523)	0.008	239	0.251
- Sedimentary	8.763	(2.008, 38.246)	0.004	958	0.139
- Volcanic	4.642	(0.805, 26.768)	0.086	65	0.077
- Surficial	1			527	0.025

For the smoothing spline terms included in the full model, Table 3 presents the Effective Degrees of Freedom (EDF) and corresponding p-values. Additionally, Figure 19 illustrates the partial effect plots, illustrating how changes in each smooth term affect the log odds of arsenic exceedance probability while holding other variables constant. The smoothing spline terms include:

- Eastings and Northings: These represent the geographical coordinates of each domestic well. The effective degrees of freedom (EDF) for this smooth term is estimated at 23.975, with a highly significant p-value (p ¡ 2e-16). This indicates a complex, non-linear relationship between spatial location and arsenic contamination levels, suggesting a significant influence of spatial coordinates on arsenic contamination.
- Date Sampled: This smooth term captures the relationship between the date of sampling and arsenic levels. It has an estimated EDF of 2.943,

with a p-value of 0.3543. The non-significant p-value suggests that the relationship between sampling date and arsenic contamination is not statistically significant.

- Monthly precipitation: With an estimated EDF of 5.081 and a p-value of 0.1437, the relationship between precipitation and arsenic contamination does not appear to be statistically significant.

- Monthly temperature: This smooth term exhibits an estimated EDF of 2.286 and a p-value of 0.0451. The significant p-value suggests a nonlinear relationship between mean temperature and arsenic contamination, indicating statistical significance.

Table 3: The approximate significance of smooth terms in the GAM model for modelling the arsenic levels. EDF stands for effective degrees of freedom.

Predictor Variable	EDF	P-value
Eastings and Northings	23.975	< 0.001
Date Sampled	2.943	0.3543
Monthly Precipitation Level	5.081	0.1437
Monthly Temperature Level	2.286	0.0451

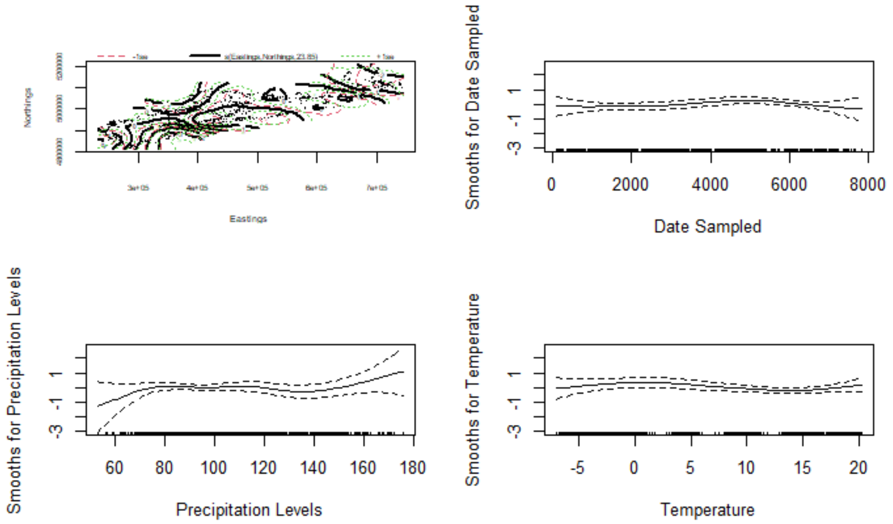

Figure 19: The partial effect plots for the smooth terms including spatial terms, date sampled, precipitation levels, and temperature in the GAM model.

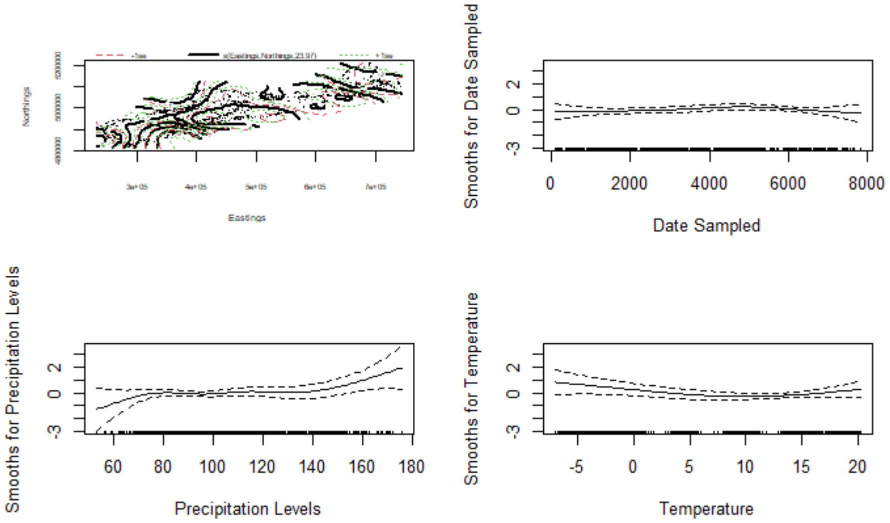

Figure 20: ROC curve of the GAM model modelling arsenic levels.

Figure 20 shows the ROC curve, a graph visualizing the performance of the binary classifier model at all classification thresholds, providing insights regarding model selection. The AUC, the area under the ROC curve, ranges from 0 to 1 and evaluates the binary classification model's performance as 82.0%, and the 95% confidence interval of AUC is 0.799-0.839. In the ROC curve, each point corresponds to a threshold in the model. For example, as depicted in the plot, each point on the curve represents a specific threshold value. If the predicted probability from the model exceeds the threshold value, it classifies as a positive outcome.

In practical terms, consider a scenario where a threshold of 0.2 is set. If the predicted probability of a positive outcome exceeds 0.2, it will be classified as a positive prediction. However, this comparatively low threshold leads to a high false positive rate, resulting in instances of being incorrectly classified as positive. On the other hand, the true positive rate remains high, indicating that the model correctly identifies a significant proportion of actual positives.

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) could be used to compare model fit. AIC is an estimator of prediction error, which measures the relative quality of models and is formulated as $\text{AIC} = 2k - 2\ln(\hat{L})$. While BIC, another model selection criterion, has a different penalty term compared to AIC, formulated as $\text{BIC} = -2\ln(\hat{L}) + k\ln(n)$. where:

- $\hat{L}$ represents the maximized value of the likelihood function for the model.
- k represents the number of independent variables used.

- n represents the number of observations.

The model has an AIC of 1716.423 and a BIC of 1984.919. Although they do not provide any insights alone without comparing them to other models, they could be used for further research regarding model selection.

The adjusted R-squared value for the model is 0.172, which means that it can explain around 17.2% of the variation in the response variable. The deviance explained by the model is 22%. The UBRE is an unbiased risk estimator, which estimates the mean-squared error (MSE) of an estimator. It considers both bias and variance to provide an unbiased estimate of the true prediction error associated with the estimator. A lower UBRE indicates that the estimator has a lower expected prediction error (risk), and a UBRE close to 0 indicates that the estimator performs well without significant bias or variance. The UBRE for this model is estimated as -0.29205, which suggests that the model has a small amount of bias in estimating the parameters. These outputs provide valuable insights into the factors influencing arsenic contamination in well water, emphasizing the significance of spatial coordinates, specific sources, and temperature as essential determinants.

6.2 Examine the Predictability of Variables by Variable Removal

6.2.1 Results of the Multivariable Analysis Excluding Groundwater Region from the Full Model

Since the well type is highly predictive, additionally, Table1 shows a substantial correlation between well type and the groundwater region variables, the groundwater region variable was removed from the model to investigate the effects of well types. Table 4 and Table 5 show the model's estimation and approximate significance of smooth terms when the groundwater region variable is removed. Similar to Table 2, the drilled well has a significant impact on arsenic levels. However, it has an odds ratio of 10.420 (95% CI: 5.511, 19.702), which is ten times higher than the odds ratio of dug wells observed previously. According to Table 5, the approximate significance of smooth terms remained relatively unchanged.

Table 4: Model estimates of the GAM model without groundwater region variable for modelling the arsenic levels. Exceedances represent the proportion of arsenic exceeding the threshold. (n=2444)

Predictor Variable	**Odds Ratio**	**95% CI for Odds Ratio**	P**-value**	n	**Exceedances**
Well Type					
- Drilled Well	10.420	(5.511, 19.702)	< 0.001	1978	0.179
- Dug Well	1			432	0.025
Season					
- Spring	3.216	(1.556, 6.649)	0.002	629	0.181
- Fall	2.428	(1.106, 5.331)	0.027	557	0.165
- Summer	2.211	(0.902, 5.422)	0.082939	964	0.127
- Winter	1			294	0.143

Table 5: The approximate significance of Smooth Terms in the GAM model without groundwater region variable for modelling the arsenic levels. EDF stands for effective degrees of freedom.

Predictor Variable	EDF	P-value
Eastings and Northings	24.507	< 0.001
Date Sampled	3.331	0.372
Monthly Precipitation Level	4.674	0.189
Monthly Temperature Level	2.382	0.026

6.2.2 Model after Removal of the Spatial Terms

Since there could potentially exist a confounding effect of spatial terms and climate variables, a model after the removal of spatial terms is fit to take out the spatial effects in the model. According to the findings, significant factors affecting arsenic levels include seasonal variations (spring, fall, and summer), groundwater types (carbonate/evaporite, metamorphic, plutonic, and sedimentary), as well as date and temperature variables.

Table 6: Model estimates of the GAM model without spatial terms for modelling the arsenic levels, Exceedances represent the proportion of arsenic exceeding the threshold. (n=2444)

Predictor Variable	**Odds Ratio**	**95% CI for Odds Ratio**	**P-value**	n	**Exceedances**
Well Type					
- Drilled Well	0.741	(0.159, 3.447)	0.703	1978	0.179
- Dug Well	1			432	0.025
Season					
- Spring	4.176	(1.983, 8.795)	< 0.001	629	0.181
- Fall	4.301	(1.978, 9.354)	< 0.001	557	0.165
- Summer	3.613	(1.488, 8.774)	0.005	964	0.127
- Winter	1			294	0.143
Groundwater Region					
- Carbonate / Evaporite	10.411	(2.377, 45.593)	0.001	178	0.140
- Metamorphic	19.126	(4.580, 79.877)	< 0.001	446	0.289
- Plutonic	15941	(3.756, 67.654)	< 0.001	239	0.251
- Sedimentary	8.686	(2.096, 35.992)	0.003	958	0.139
- Volcanic	4.701	(0.872, 25.335)	0.072	65	0.077
- Surficial	1			527	0.025

Table 7: The approximate significance of Smooth Terms in the GAM model without spatial terms for modelling the arsenic levels. EDF stands for effective degrees of freedom.

Predictor Variable	EDF	P-value
Date Sampled	2.361	0.030
Monthly Precipitation Level	1.001	0.051
Monthly Temperature Level	7.499	< 0.001

6.2.3 Model after Removal of the Season Variable

Since season could potentially exhibit collinear effects with variables such as temperature and precipitation, it is removed from the model to investigate the predictability of other smooth terms. Table 8 shows the model's approximate significance of smooth terms when the season variable is removed. As a result, the temperature variable became insignificant, and the significance of the date sampled and monthly precipitation level remained unchanged. Figure 21 shows the partial effect plots for the smooth terms. The notable difference between Figure 19 and Figure 21, the original and new partial effect plots for temperature, is the shift in the log odds of temperature. Initially starting from 1 and ending around 0.5, it has now moved towards 0, indicating its loss of significance.

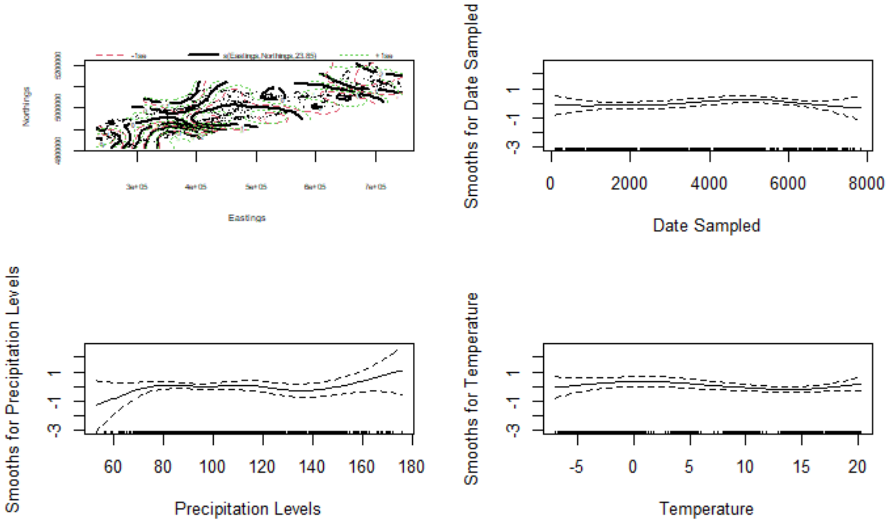

Figure 21: The partial effect plots for the smooth terms including spatial terms, date sampled, precipitation levels, and temperature in the GAM model without the season variable.

Table 8: The approximate significance of Smooth Terms in the GAM model without season variable for modelling the arsenic levels. EDF stands for effective degrees of freedom.

Predictor Variable	EDF	P-value
Eastings and Northings	23.851	< 0.0001
Date Sampled	3.368	0.212
Monthly Precipitation Level	5.159	0.382
Monthly Temperature Level	3.358	0.170

6.2.4 Investigating Model Predictability Effects by Removal of Variables

The effort to enhance model predictability involved systematically removing each variable and comparing the resulting Area Under the Curve (AUC). Table 9 resulting AUC values along with their corresponding 95% confidence intervals (CI) and the Akaike Information Criterion (AIC) scores. According to the results, removing spatial terms yielded the lowest AUC of 73.3%, suggesting the spatial terms had the most significant impact on the model's predictability.

Table 9: The predictability of the GAM model for modelling arsenic levels after removing each variable in terms of AUC.

	AUC	95% CI of AUC	AIC
Full Model	82.0%	(79.9%-84.0%)	1716.42
Variable Removed			
Well Type	82.0%	(80.0%-83.9%)	1712.61
Eastings and Northings	73.3%	(70.7%-75.8%)	1884.81
Date Sampled	82.0%	(79.8%-84.0%)	1716.32
Monthly Precipitation Level	81.7%	(79.6%-83.6%)	1719.55
Monthly Temperature Level	81.9%	(79.8%-83.8%)	1721.58
Groundwater Region	81.1%	(78.9%-83.1%)	1729.52
Season	82.0%	(80.2%-84.1%)	1720.91

6.2.5 Investigating Model Predictability Effects by Removing Each Variable and Spatial Terms

Since Table 9 suggests that spatial terms contribute the most to the model's predictability, this section systematically removes each variable and spatial term to assess their impact on the model's predictability for each variable. Table 10 presents the resulting AUC values along with their corresponding 95% confidence intervals (CI) and the Akaike Information Criterion (AIC) scores. Noticeably, removing the groundwater region variable from the model led to the lowest AUC of 70.1%, suggesting groundwater region has the second most impact on the model's predictability besides spatial terms.

Table 10: The predictability of the original GAM model for modelling arsenic levels without spatial terms and removal of each variable in terms of AUC

	AUC	95% CI of AUC	AIC
Full Model without Spatial Terms	73.3%	(70.4%-75.8%)	1884.81
Variable Removed			
Well Type	73.3%	(70.5%-75.8%)	1881.13
Date Sampled	73.2%	(70.3%-75.6%)	1890.93
Monthly Precipitation Level	73.3%	(70.7%-75.9%)	1886.35
Monthly Temperature Level	72.2%	(69.5%-75.0%)	1901.19
Groundwater Region	70.1%	(0.673%-0.728%)	1927.54
Season	73.9%	(71.3%-76.6%)	1892.24

7 Discussions

7.1 Regional Variation of Arsenic Levels

Understanding regional variations in arsenic contamination is vital for effectively regulating drinking water. The findings indicate significant regional differences in arsenic levels, particularly in Yarmouth and Halifax. The elevated arsenic levels in these areas underscore the pressing need for targeted interventions and regulatory measures. Mitigating risks and enhancing community safety can be achieved by implementing tailored water quality monitoring and interventions in high-risk regions.

7.2 Seasonal Variation of Arsenic Levels

Examining the seasonal patterns in arsenic levels is another essential of water safety regulations. This research reveals spring and fall seasons, particularly in March and November, have the highest arsenic levels, surpassing the 5 ug/L safety threshold. For both spring and fall, the odds ratio is positive: 3.186 for Spring and 2.347 for Fall. This indicates a positive association between these seasons and arsenic levels in well water. The significance of this association varies: Spring shows a highly statistically significant relationship with a p-value of 0.00172, while Fall's association is marginally significant with a p-value of 0.036. The results point to a distinct seasonal variation in arsenic levels. To optimize arsenic monitoring, prioritizing efforts during months with higher levels can enhance the efficiency and effectiveness of water monitoring strategies.

7.3 Other Significant Factors of Arsenic Levels

This research indicates that spatial variables like eastings and northings (longitude and latitude) contribute the most to the model's overall predictability in predicting arsenic levels across Nova Scotia. This aligns with previous observations of regional variations in arsenic levels across the province. Drilled wells exhibit a notably higher proportion of arsenic levels, surpassing the safety threshold of 5 ug/L. To ensure the safety of drinking water, it's imperative to implement targeted measures specifically for drilled wells throughout Nova Scotia.

A strong correlation exists between precipitation and arsenic levels across Nova Scotia. Regions with higher precipitation levels, such as Yarmouth, Halifax, and Cape Breton, also exhibit elevated arsenic levels. This correlation underscores the importance of tailored monitoring efforts in regions experiencing higher precipitation levels and during seasons characterized by increased precipitation.

Furthermore, the number of sampling dates beyond 2000 demonstrates a significant positive association with arsenic levels. This suggests a concerning trend of escalating arsenic contamination across Nova Scotia from 2000 to 2020. Implementing proactive water quality regulations is crucial to mitigate health risks associated with arsenic contamination.

7.4 Spatial Confounding

Spatial confounding is a challenge in this research, as a collinear effect exists between spatial terms and other environmental variables such as precipitation and temperature. A model without spatial terms is fit to discover the impact of climate variables without the effects of spatial terms. Comparing the results from the model output with and without spatial terms, notable changes emerged upon removing spatial terms from the model. Specifically, the summer and fall seasons transitioned from negative to favourable log odds with significance, and the dates and precipitation became significant with positive log odds. Meanwhile, the effects of other variables remained unchanged.

7.5 Predictability of Variables by Removing Each Variable and Spatial Terms

Spatial terms were found to have the greatest impact on the overall predictability of the model. Consequently, the individual effects of each climate variable on predictability were assessed by removing spatial terms from the model without spatial terms. Removing the groundwater region variable from the model without spatial terms resulted in the lowest AUC and the highest increase in AIC. These findings indicate that besides the spatial term, the groundwater region variable contributes to the model's predictability and overall quality the most. Removing other variables slightly decreased the model's overall predictability and quality, although the changes were insignificant.

7.6 The Effects of Climate Change on Arsenic Levels

Climate change poses significant implications for arsenic levels, as this study suggests that essential climate factors, such as temperature, are a substantial factor in arsenic levels. This study indicates that when spatial terms are removed from the model, temperature becomes significant in the model, and temperature is the third most significant variable besides spatial terms and groundwater region. This finding underscores the potential benefits of including temperature variables in arsenic modelling.

A simulation study by Melissa A. Lombard et al. (2021) suggests that decreased precipitation levels are associated with an increased probability of high arsenic exposure from private domestic wells. The results of this simulation study contradict the findings, as according to the visualizations, the regions with lower precipitation tend to have lower arsenic levels.

A study by Craig T. Connolly et al.(2022) suggests that in the context of surface flooding, the frequency, and duration of flooding play crucial roles in explaining the heterogeneity of arsenic concentrations in groundwater. Where environments experiencing sustained surface flooding tend to have higher levels of groundwater arsenic, while areas with high interannual flooding frequencies but shorter durations each year may exhibit lower arsenic concentrations. Flooding with longer durations tends to have higher levels of arsenic. Yet, shorter-duration flooding may result in lower groundwater arsenic levels due to the absence of strongly reducing conditions favourable for arsenic mobilization.

Understanding the impact of climate change events such as drought and flooding on arsenic exposure is critical for developing strategies to mitigate the risks associated with arsenic contamination in drinking water. By recognizing the influence of climate change on arsenic levels, policymakers, and researchers can implement proactive measures to monitor arsenic exposure and ensure drinking water safety.

7.7 The Effects of Bedrock Geology towards Arsenic Levels

The groundwater region, representing bedrock geology, emerged as a pivotal factor contributing to the predictability and quality of the model. Its removal from the spatially unadjusted model led to a notable decline in the Area Under the Curve (AUC), dropping from 73.3% to 70.1%, accompanied by a substantial increase in the Akaike Information Criterion (AIC) from 1881.81 to 1927.54. This underscores the significant role of bedrock geology as the second most influential variable impacting the model's predictability.

According to Kennedy and Drage (2017), arsenic levels in Nova Scotia's well water are closely associated with the underlying bedrock geology. Specifically, regions characterized by metamorphic and plutonic bedrock in southern Nova Scotia exhibit notably higher concentrations and rates of arsenic exceedance (Kennedy & Drage, 2017). Conversely, groundwater regions associated with sedimentary and carbonate/evaporite bedrock tend to demonstrate lower to moderate concentrations of arsenic, as indicated by the study's findings. Based on the results, it appears that the carbonate/evaporite groundwater region exhibits the highest odds ratio towards arsenic levels among the bedrock geology categories considered in this analysis. This contrasts with the expectation based on the report, which suggests that carbonate/evaporite groundwater typically has lower arsenic concentrations. Additionally, this analysis highlights other significant variables associated with bedrock geology. Specifically, the metamorphic, sedimentary, and plutonic groundwater regions demonstrate substantial odds ratios towards arsenic levels, with decreasing magnitudes from metamorphic to plutonic. This suggests that these geological formations may contribute to elevated arsenic levels in groundwater.

The study used publicly available and privately sourced data spanning 40 years from the Nova Scotia Groundwater Chemistry Database. Despite Health Canada's threshold for arsenic being set at 10 ug/L, the study adopted a laboratory threshold of 5 ug/L, which is consistent with the research methodology. Additionally, the study highlighted that arsenic levels in Nova Scotia's well water correlate with aquifer geochemistry factors, including Eh (redox potential or oxidation) and pH levels. Specifically, the research suggests higher pH levels are associated with elevated arsenic levels in the well water samples analyzed.

Similar to this research, the study encountered challenges from significant spatial heterogeneity. Certain areas of Nova Scotia exhibited high-density well-water chemistry data, while data from other locations were unavailable. This highlights the importance of addressing spatial heterogeneity in arsenic modelling efforts within Nova Scotia to ensure a comprehensive and representative understanding of regional arsenic levels.

7.8 Strengths and Limitations of This Research

This research provides insights into arsenic modelling and policymaking regarding arsenic contamination, particularly within Nova Scotia. The studies delve into the complex relationship between arsenic contamination and environmental factors in the region. Factors such as precipitation and geographical location are investigated for their influence on arsenic concentrations. This understanding enhances the knowledge of arsenic contamination dynamics in Nova Scotia.

In the field of arsenic modelling, one of the main challenges is the diverse formats of environmental datasets available in Nova Scotia, which can be heterogeneous. To overcome this issue, these datasets are integrated into a standardized format. This approach enables a more comprehensive analysis and interpretation of the data. The research highlights the importance of environmental factors such as well type and precipitation, offering valuable guidance for monitoring efforts and resource allocation. This aids in developing more effective water safety interventions to minimize the risks of arsenic exposure and protect public health. Additionally, gaining insights into the spatial-temporal variations of arsenic levels is useful for policymakers who manage arsenic contaminants. This study identifies the regions and seasons in Nova Scotia with elevated arsenic levels, providing policymakers with valuable information. This information can aid in implementing targeted water treatment protocols to reduce arsenic exposure risks in drinking water sources.

One limitation of this study is that daily climate data was not available, which restricted the analysis from incorporating lagged effects of climate variables. As climate variables often take time to materialize, the study faced limitations in capturing these temporal dynamics, such as the delayed impact of rainfall on precipitation exceedances. The study could not determine whether the influence of a climate variable on arsenic levels is immediate or requires several days or months to unfold. In a similar study by Caroline M. Andy et al. (2017), logistic regression was used to model arsenic levels using 374 of the most significant variables. The accuracy of the model could be improved by adding other climate variables.

The study also faces limitations due to non-uniform samplings. Urban areas of Nova Scotia are more frequently sampled than rural areas. Variations in the sampling frequency of locations across different regions of Nova Scotia may introduce biases and negatively affect the accuracy of the analyses. Climate models generate precipitation records and are subject to certain limitations. These climate models utilize smoothing methods that may smooth out extraordinarily high or low precipitation values. Using raw precipitation data could enhance the accuracy of modelling.

7.9 Methodological Innovations for Future Studies

Future studies of arsenic modelling, particularly in Nova Scotia, can benefit significantly from future methodological innovations. Addressing the bias in arsenic sampling across different regions requires either the advancement of sampling techniques to achieve more uniform representation or a statistical approach of utilizing advanced modelling methods such as spatial-temporal modelling to minimize biases. The datasets of this study came from various formats. For example, in this study, the arsenic data uses Eastings and Northings for geological locations, and the precipitation and temperature data use Longitude and Latitude. Merging datasets manually by calculating Euclidean distance could be time-consuming. Therefore, statistical packages could be developed to ensure efficient data merging. Additionally, implementing a real-time precipitation monitoring system can ensure the accuracy and precision of precipitation recordings. Real-time data collection can provide timely and accurate information on precipitation patterns, allowing for more reliable modelling of arsenic contamination dynamics.

As this research mainly focuses on analyzing existing data, transitioning from descriptive analysis to predictive modelling can enhance the proactive management of arsenic contamination risks. By implementing predictive analytics techniques, researchers can forecast future trends in arsenic levels and anticipate potential contamination events, enabling preemptive measures to safeguard drinking water safety in Nova Scotia. The AUC levels evaluated in this study serve as reference points for comparing the quality of future models. Additionally, this study encountered the problem of spatial confounding, which could be addressed by implementing methods to eliminate spatial confounding effects by methods such as spatial+. Furthermore, other factors, such as PH level, may potentially contribute to the quality of the model since arsenic levels correspond to many environmental factors besides the variables used in this study. According to Joel E. Podgorski et al. (2017), "High soil pH can drive arsenic desorption and is indicative of an evaporative environment, which further raises arsenic concentrations" (Podgorski, 2017). Therefore, adding other climate factors might benefit future modelling, since arsenic levels correspond to different climate variables.

8 Conclusion

This study examined the complex dynamics of arsenic contamination throughout Nova Scotia by employing a generalized linear regression model with logistic regression techniques. Notable correlations between arsenic levels and various environmental factors are discovered by integrating diverse datasets and conducting thorough statistical analyses. These factors included geographical location, precipitation levels, well types, dates, seasons, and temperature variations, with geographical location and groundwater region emerging as the most influential. Additionally, high-risk regions and seasons are identified while investigating spatial-temporal patterns of arsenic levels.

The study encountered several limitations, including the unavailability of daily climate data, which precluded the ability to explore the lagged effects of climate variables, non-uniform samplings, and smoothed-out precipitation records. These constraints hindered the ability to investigate the lagged impact of climate variables effectively.

Despite these limitations, the findings of the study provide valuable insights for policymakers addressing arsenic contamination in Nova Scotia. The research also catalyzes future endeavours to refine arsenic contamination models and enhance understanding of the intricate dynamics of arsenic levels within the region.

9 Appendix

Table 11 summarizes the divergence in precipitation levels and occurrences of arsenic exceedances among different regions, considering their regional disparities.

Table 11: Mean arsenic and precipitation levels are summarized by county.

County	Mean Arsenic	Mean Precipitation
Annapolis	0.1163	97.2891
Antigonish	0.1381	91.2827
Cape Breton	0.0316	90.3675
Colchester	0.2047	95.4403
Cumberland	0.1748	90.5292
Digby	0.1452	99.8048
Guysborough	0.1316	100.0447
Halifax	0.3595	115.6557
Hants	0.3860	91.4307
Inverness	0.0690	NA
Kings	0.0737	95.0007
Lunenburg	0.1771	105.2624
Pictou	0.1195	93.0550
Queens	0.0159	112.2513
Richmond	0.1688	94.2692
Shelburne	0.0000	107.5214
Victoria	0.0209	101.4880
Yarmouth	0.3514	93.7846

10 References

1. Andy, Caroline & Fahnestock, Maria & Lombard, Melissa & Hayes, Laura & Bryce, Julie & Ayotte, Joseph. (2017). Assessing Models of Arsenic Occurrence in Drinking Water from Bedrock Aquifers in New Hampshire. Journal of Contemporary Water Research & Education. 160. 23. 10.1111/j.1936-704X.2017.03238.x.

2. Connolly, C. T., Stahl, M. O., DeYoung, B. A., & Bostick, B. C. (2022). Surface Flooding as a Key Driver of Groundwater Arsenic Contamination

in Southeast Asia. *Environmental Science & Technology, 56*(2), 928-937. DOI: 10.1021/acs.est.1c05955

3. Farzan, S. F., Karagas, M. R., & Chen, Y. (2013). In utero and early life arsenic exposure in relation to long-term health and disease. Toxicology and applied pharmacology, 272(2), 384–390. https://doi.org/10.1016/j.taap.2013.06.030

4. Feng C. Spatial-temporal generalized additive model for modeling COVID-19 mortality risk in Toronto, Canada. Spat Stat. 2022 Jun;49:100526. doi: 10.1016/j.spasta.2021.100526. Epub 2021 Jul 6. PMID: 34249608; PMCID: PMC8257405.

5. Kennedy, G. W., & Drage, J. M. (*2017*). An Arsenic in Well Water Risk Map for Nova Scotia based on Observed Patterns of Well Water Concentrations of Arsenic in Bedrock Aquifers. *Open File Report ME 2017-003.*

6. Lombard, M. A., Daniel, J., Jeddy, Z., Hay, L. E., & Ayotte, J. D. (2021). Assessing the Impact of Drought on Arsenic Exposure from Private Domestic Wells in the Conterminous United States. *Environmental Science & Technology, 55*(3), 1822-1831. DOI: 10.1021/acs.est.9b05835

7. Podgorski, J., & Berg, M. (2020). Global threat of arsenic in groundwater. Science (New York, N.Y.), 368(6493), 845–850. https://doi.org/10.1126/science.aba1510

8. Podgorski, J. E., Eqani, S. A. M. A. S., Khanam, T., Ullah, R., Shen, H., & Berg, M. (2017). Extensive arsenic contamination in high-pH unconfined aquifers in the Indus Valley. *Science advances*, *3*(8), e1700935. https://doi.org/10.1126/sciadv.1700935

9. Quansah, R., Armah, F. A., Essumang, D. K., Luginaah, I., Clarke, E., Marfoh, K., Cobbina, S. J., Nketiah-Amponsah, E., Namujju, P. B., Obiri, S., & Dzodzomenyo, M. (2015). Association of arsenic with adverse pregnancy outcomes/infant mortality: a systematic review and meta-analysis. Environmental health perspectives, 123(5), 412–421. https://doi.org/10.1289/ehp.1307894

10. Ravenscroft, Peter & Brammer, Hugh & Richards, Keith. (2009). Arsenic Pollution: A Global Synthesis. RGS-IBG Book Series. 1. 10.1002/9781444308785.

11. Tolins, M., Ruchirawat, M., & Landrigan, P. (2014). The developmental neurotoxicity of arsenic: cognitive and behavioral consequences of early life exposure. Annals of global health, 80(4), 303–314. https://doi.org/10.1016/j.aogh.2014.09.005

12. United Nations Children's Fund & World Health Organization. (2018). Arsenic primer: Guidance on the investigation and mitigation of arsenic contamination. Retrieved from https://www.who.int/publications/m/item/arsenic-primer

13. Wood, S. N. (2006). Generalized Additive Models: An Introduction with R. CRC Press.

14. Wood, S. (2023). Package 'mgcv' (Version 1.9-1) [Computer software]. Comprehensive R Archive Network. https://cran.r-project.org/web/packages/mgcv/index.html

15. Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1), 95-114. https://doi.org/10.1111/1467-9868.00374

16. World Health Organization. (2022, December 7). Arsenic [Fact sheet]. https://www.who.int/publications/m/item/arsenic

17. Urdangarin, A., Goicoa, T., & Ugarte, M. D. (2023). Evaluating recent methods to overcome spatial confounding. *Rev Mat Complut, 36*, 333–360. DOI: 10.1007/s13163-022-00449-8