Investigation into the Concentration of a Contaminant at a City's Main Water Supply (UQ Project 2)

MATH4074 - Uncertainty Quantification
Group project report - Spring semester
2021/22

School of Mathematical Sciences
University of Nottingham

Group 2:
Jake Denton
Edwina Haran
Thomas Kenyon
Charles Lamb

We have read and understood the School and University guidelines on plagiarism. We confirm that this work is our own, apart from the acknowledged references.

Abstract

Accurate knowledge of the concentration of contaminants in the water supply of a city is essential to identify health risks associated with providing water to residents. In this report, we consider the uncertainty in the concentration of a contaminant given 100 observations on a square grid at time 0.015. We have been informed that if the concentration of the contaminant measured at the water supply exceeds the critical value $c_{shut} = 1.8$, the water supply must be shut down. This report finds estimates for the initial concentration of the contaminant (i.e. at time 0), then investigates how the concentration of this contaminant evolves for later times. This is done in order to help the local government decide whether to shut down the water supply and if so, when and for how long.

We find that the initial concentration is centred at a peak within the aquifer not far from the main city water supply. Time series data of the concentration at the city water supply shows us that many trajectories do exist with concentration values exceeding this critical value between the times of $t = 0$ and $t \approx 0.18$, although on average most of the trajectories do not exceed this value. Consequently, we do recommend that the city shuts down the water supply between these times, as we cannot be certain that the water is safe.

Contents

1 Introduction

There has been a contamination incident in the local area, and there is evidence that the contaminant has reached the point within the aquifer that provides water to the city. We have been tasked firstly with determining the initial concentration of the contaminant and secondly with advising the local authorities on whether or not to shut the main city water supply down. We must also provide rigorous estimates of uncertainty about our predictions.

This report is structured as follows. We will begin by looking at the background and assumptions given in the problem which are necessary in order to complete the tasks at hand. Following this, we will look at the data provided in order to check its quality and get an idea of what the solution might look like. We will then outline our mathematical approach to the problem, before presenting the outcomes of this analysis as well as checking the sensitivity of these results. Finally, we will use these results to make a recommendation to the local authorities on the appropriate action to take, as well as acknowledging the limitations of our investigation.

1.1 Defining the Problem Mathematically

We are informed that we can make the following assumptions relating to the problem:

- The aquifer consists of one thin layer and the area of interest is the unit square. The main water supply is located at $(x^*, y^*) = (0.495, 0.495)$.
- The concentration of the contaminant, $c(x, y, t)$, satisfies the diffusion equation:

$$\frac{\partial c}{\partial t} - \kappa \Delta c = 0 \text{ in } [0,1]^2,\ t > 0 \tag{1.1}$$

  with homogeneous (zero) Dirichlet boundary conditions. Previous geological investigations tell us that we can be fairly confident that the diffusivity constant of the aquifer is $\kappa = 0.05$.
- We are told that the water supply of the city must be shut down if the concentration of the contaminant exceeds $c_{shut} = 1.8$.
- The measurements given in the data are contaminated with Gaussian errors of zero mean and standard deviation $\sigma = 0.2$.

Note that for the majority of this report, we fix the constant $\kappa = 0.05$. In the further work section, we test the sensitivity of our results with respect to this value.

1.2 The Data

We have been given 100 measurements of the concentration of the contaminant at given locations on the unit square. We denote the locations and measurements of the concentration at these points by

$$\mathcal{D} = \big\{(x_m, y_m, c_m) : m = 1, \ldots, 100\big\}. \tag{1.2}$$

Here, (x_m, y_m) are the measurement locations and c_m denotes the corresponding observed concentration of the contaminant, which has been collected at time $t = 0.015$. All quantities involved were transformed to be dimensionless before they were given to us, so we are not aware of what this actually represents in real-time (see discussion for more details). A visualisation of the data is presented below.

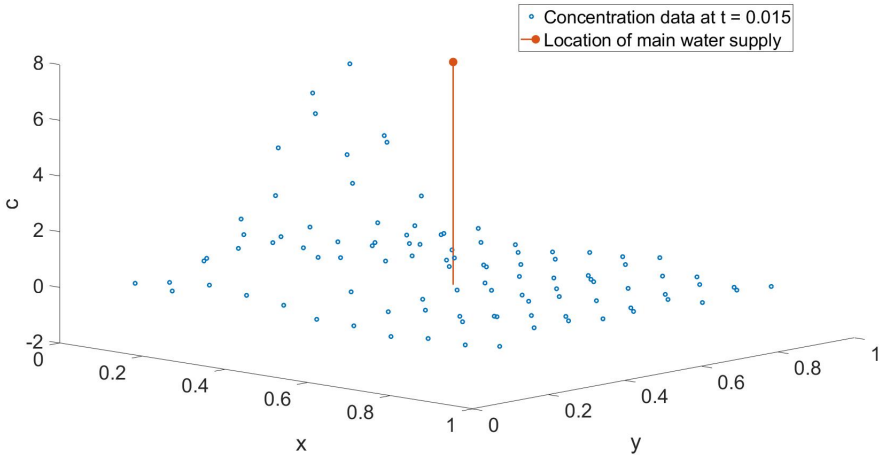
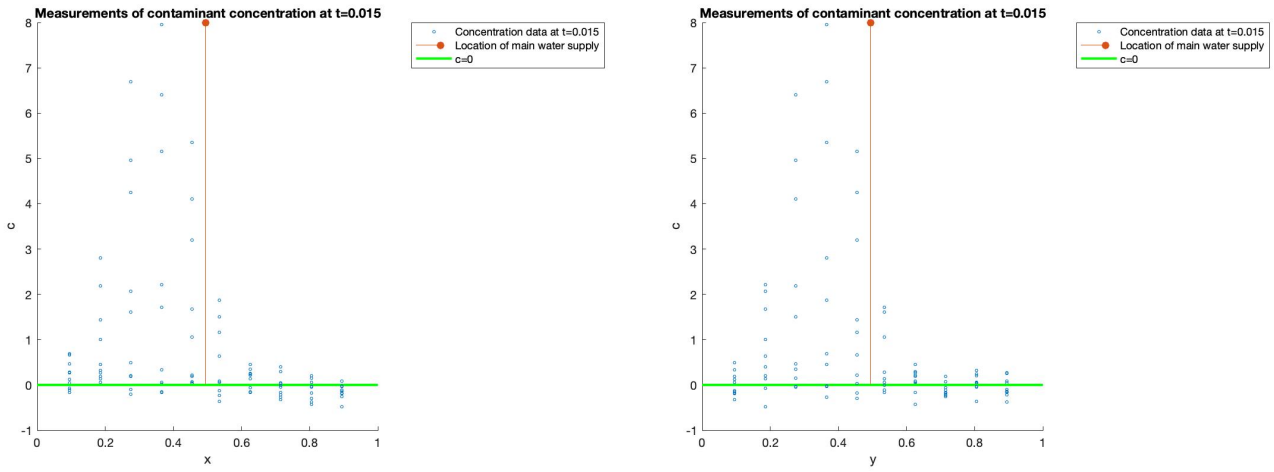


Figure 1: Concentrations of the contaminant at t = 0.015

From this plot, we can clearly see a broad peak in the concentration located in the lower quadrant (where $x, y < 0.5$) of the grid. In each of the other quadrants the concentrations are settled around zero. The location of the maximum in the data, and so also the maximum of the broad peak, is at $(0.365, 0.365)$. This lies close to the main water supply, just a Euclidean distance of 0.1838 away. Also worth noting is the substantial (36) number of points with a negative concentration measurement (see figures below). This does not make logical sense, and thus will need to be kept in mind when interpreting results. We assume that this is due to the Gaussian errors and not human error in measuring the concentration. The true measurements are likely zero or very near to zero, so as long as our results reflect this, the negative results can be overlooked.



(a) x_m measurements

(b) y_m measurements

Figure 2: Figures showing measurements with negative concentrations.

2 Methodology

Our overall objective can be split into two separate tasks. Firstly, this constitutes using the data supplied to find the initial concentration of the contaminant across the aquifer, an **inverse** problem. The second task is as follows: using the solution of the inverse problem, forward propagate and find the concentration of the contaminant in the water supply (and importantly, near the city water supply) at later times. This is a **forward** problem. In both cases, we consider the uncertainty present in our solutions.

To do any of this, we first need to solve the bounded 2D diffusion equation and find the operator matrix.

2.1 Solution to Bounded 2D Diffusion Equation and Operator Matrix

We aim to show that the solution to the diffusion equation (1.1) takes the form:

$$c(x,y,t) = \sum_{n,m=1}^{\infty} A_{mn} \sin(n\pi x) \sin(m\pi y) e^{-\lambda_{mn} t}. \tag{2.1}$$

Stepping back a moment, this ansatz (initial guess) could be found by using separation of variables, before forming an infinite sum of the solutions using the superposition principle. We need to check that this does indeed satisfy the diffusion equation, and find the associated constants A_{mn} and λ_{mn}. Starting with the coefficients, taking the partial derivatives of c twice with respect to x and y and once with respect to t, before substituting into equation (1.1) gives:

$$\sum_{n,m=1}^{\infty} -\lambda_{mn} A_{mn} \sin(n\pi x) \sin(m\pi y) e^{-\lambda_{mn} t} = \kappa \sum_{n,m=1}^{\infty} A_{mn}(-\pi^2(n^2+m^2)) \sin(n\pi x) \sin(m\pi y) e^{-\lambda_{mn} t} \tag{2.2}$$

Comparing coefficients on both sides, this holds when $\lambda_{mn} = \kappa\pi^2(n^2 + m^2)$. As for A_{mn}, let us simplify the notation by introducing $v_{mn}(x,y) = sin(n\pi x)sin(m\pi y)$, which represent the two-dimensional basis functions of the Laplace operator (see UQ lecture notes which show this). Now, setting $t = 0$ in equation (2.1) and denoting the initial condition (quantity of interest for inverse problem) by $m(x,y)$, we have:

$$c(x,y,0) = m(x,y) = \sum_{n,m=1}^{\infty} A_{mn} v_{mn}(x,y) \tag{2.3}$$

Taking the inner product of both sides with the basis function $v_{\hat{m}\hat{n}}(x,y)$, we have:

$$\int_0^1 \int_0^1 f(x,y) v_{\hat{m}\hat{n}}(x,y) dxdy = \sum_{n,m=1}^{\infty} A_{mn} \int_0^1 \int_0^1 v_{mn}(x,y) v_{\hat{m}\hat{n}}(x,y) dxdy$$

$$= \sum_{n,m=1}^{\infty} A_{mn} (\tfrac{1}{4} \delta_{m\hat{m}} \delta_{n\hat{n}}) = \tfrac{1}{4} A_{\hat{m}\hat{n}}.$$

On the second line, the Kronecker deltas come from the fact that unless $m = \hat{m}$ and $n = \hat{n}$, the integral evaluates to zero (using standard trigonometric results). The factor $\frac{1}{4}$ comes from the fact that when you set $m = \hat{m}$ and $n = \hat{n}$, you have the integral of the

product of two squared sine terms $\sin^2(n\pi x)\sin^2(m\pi y)$ (see definition of v_{mn}). This can be separated into the product of the integral of $\sin^2(n\pi x)$ with the integral of $\sin^2(m\pi y)$ over the same interval. Each of these evaluates to $\frac{1}{2}$ (again, using standard trigonometric results), so you get the factor $\frac{1}{2}^2 = \frac{1}{4}$ from the integration overall. Given that we get zero in this sum unless $m = \hat{m}$ and $n = \hat{n}$, the sum simplifies to the final term on the right. Relabelling (for consistency) and multiplying both sides by 4 gives:

$$A_{mn} = 4\int_0^1 \int_0^1 m(x,y)\sin(n\pi x)\sin(m\pi y)dydx. \tag{2.4}$$

We have found the values of the coefficients in (2.1). Incorporating these into the equation gives:

$$c(x,y,t) = \int_0^1 \int_0^1 m(\hat{x},\hat{y})\left(\sum_{n,m=1}^{\infty} 4e^{-\kappa\pi^2\left(n^2+m^2\right)t}\sin(n\pi\hat{x})\sin(m\pi\hat{y})\sin(n\pi x)\sin(m\pi y)\right)d\hat{x}d\hat{y}. \tag{2.5}$$

Here, the sum in brackets is the Green's function for the diffusion problem, which we denote by $G(\hat{x},\hat{y},x,y,t)$. It is repeated below as it will be incredibly useful in the formation of the operator matrix:

$$G(\hat{x},\hat{y},x,y,t) = \sum_{n,m=1}^{\infty} 4e^{-\kappa\pi^2\left(n^2+m^2\right)t}\sin(n\pi\hat{x})\sin(m\pi\hat{y})\sin(n\pi x)\sin(m\pi y). \tag{2.6}$$

We now check that the form (2.5) satisfies the conditions of the problem, which are:

$$c(0,y,t) = 0 \tag{2.7}$$

$$c(x,0,t) = 0 \tag{2.8}$$

$$c(1,y,t) = 0 \tag{2.9}$$

$$c(x,1,t) = 0 \tag{2.10}$$

Indeed, setting $x = 0, 1$ or $y = 0, 1$ makes $c(x,y,t)$ vanish for all times t, since $sin(i\pi) = 0$ for all integers i. Furthermore, it is worth noting that when t is set equal to zero, $m(x,y)$ is returned. This is not completely obvious, since we still have an integral and infinite sum on the right-hand side, but since we have used the basis functions on the $L^2(0,1)$ space, any function can be expanded in terms of this basis so they are equivalent forms.

2.2 Forming the Operator Matrix

In order to apply Bayesian methods, we need to first define the action of the operator, which when applied to the initial concentration $m(x,y)$ (the function we wish to find) returns the measurements d. Written mathematically, we have:

$$\mathcal{F}(m) = d. \tag{2.11}$$

Note that the function $m(x,y)$ belongs to the class $L^2(0,1)$ whilst the vector of measurements d belongs to $\mathbb{R}^{100}$ (since we have 100 measurements). Therefore we define

$\mathcal{F} : L^2(0,1) \to \mathbb{R}^{100}$ as:

$$\mathcal{F}(m) = (\mathcal{F}_1(m), ..., \mathcal{F}_{100}(m)), \tag{2.12}$$

where

$$\mathcal{F}_j(m) = \int_0^1 \int_0^1 m(\hat{x}, \hat{y}) G(x_m, y_m, \hat{x}, \hat{y}, 0.015) d\hat{x} d\hat{y}, \text{ where } j = 1, ..., 100. \tag{2.13}$$

Time is set equal to 0.015 within the Green's function due to the fact that the measurements are given at this time. To get to these measurements, the initial concentration needs to be propagated for this length of time. Let us define a spatial discretisation of this, which will allow us to express this operator in matrix form. Splitting the unit square N times in each direction (let step-size $h = \frac{1}{N}$), we get N^2 equally-sized cells and using the midpoint rule to evaluate the integral gives:

$$\mathcal{F}_j(m) \approx \sum_{i=1}^{N^2} h^2 m(x_i, y_i) G(x_m, y_m, x_i, y_i, 0.015), \tag{2.14}$$

where (x_i, y_i) represent the centres of each cell in the discretisation. We can rewrite this sum as the product of a matrix $\mathbf{F}$ and a vector $\mathbf{m}$, where the elements of the matrix are given by:

$$\mathbf{F}_{m,i} = h^2 G(x_m, y_m, x_i, y_i, 0.015). \tag{2.15}$$

The indices (m, i) tell us which measurement and cell we are focusing on respectively. In this way, we obtain an MxN^2 matrix which represents the action of the operator. For our investigation, we used 50^2 cells and a step-size of $h = 0.02$ in each spatial direction. Note that use of the midpoint rule adds an additional numerical error. This is a second-order method, so this error should be proportional to h^3. We do not consider how this numerical error affects our results. However, we do note that a smaller step-size could be used and the error could be investigated, for example by using MATLAB to perform the integration exactly and seeing how large the error is for one individual cell.

2.3 The Inverse Problem

Once we have our operator matrix $\mathbf{F}$ representing the action of the diffusion equation on the concentration as described in the previous section, we can perform techniques such as Tikhonov regularisation and Bayesian methods to estimate the initial concentration.

2.3.1 Tikhonov Regularisation

Note that the operator involved in the diffusion equation is compact (see references for more information), and the problem is ill-posed in the sense of Hadamard. Tikhonov regularisation is a method to address the lack of continuity with respect to the measurements that arise from the compactness of $\mathcal{F}$. Having performed a discretisation, it is simple to find a regularised solution to our problem. The initial concentration with Tikhonov constant δ can be shown to be equal to:

$$\mathbf{m}^{h,\delta} = \mathbf{F}^T (\mathbf{F}\mathbf{F}^T + \delta \mathbf{I}_M)^{-1} \mathbf{d}, \tag{2.16}$$

where $\mathbf{F}^T$ is the transpose of the operator matrix $\mathbf{F}$ and I_M is the identity matrix with dimension $M = 100$. Setting $\delta = 10^{-8}$ and finding the Tikhonov solution above before

plotting gives the following figure:

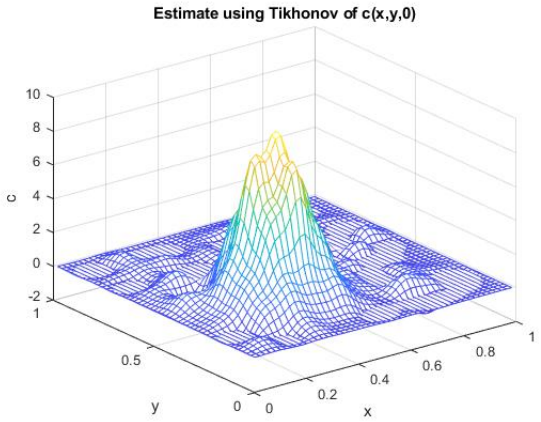


Figure 3: Estimate using Tikhonov of c(x,y,0)

We can see that there is a clear large peak on the lower left (observing from the x-axis), and the rest of the initial concentration is settled around zero as we expected from the data. This motivates us as to what the initial concentration might plausibly look like. However, Tikhonov regularisation does not provide us with any measure of uncertainty, so we need to consider Bayesian methods instead, which we proceed with now.

2.3.2 Bayesian Approach

In the Bayesian approach, we assume an initial guess for the probability distribution of the unknown function (in our case the initial concentration), which we call the prior, and then we use Bayes' theorem to find the posterior distribution. The data $\mathbf{d}$ is assumed to be of the form (conditioning upon the initial condition):

$$\mathbf{d}|\mathbf{m} = \mathcal{F}(\mathbf{m}) + \eta \sim N(\mathcal{F}(m), \Gamma), \tag{2.17}$$

where η is the vector of measurement errors, assumed to be independent centred Gaussian variables with variance 0.2^2. Γ is the covariance matrix of these measurement errors, which under the assumption of independence is the identity matrix multiplied by the variance 0.2^2. Note that this could be adapted for our problem, it seems unlikely that the variance of the errors in the measurements is equal across the whole aquifer. In terms of further work, one could use a higher variance near the peak and a lower variance away from the peak, and even consider adding dependencies between the measurements.

Our prior assumption is that the function m representing the initial concentration is a Gaussian process. The mean of this prior is set equal to 1 (we used 0 first and found more reliable results with the prior equal to 1) for all points, whilst the Matern covariance matrix is used for the prior covariance. We tested a large number of hyperparameters and decided a good choice giving reliable results is $(\sigma^2, \nu, \tau) = (1, 1, 1)$. One option for further work is to take these hyperparameters as random variables. This strong dependence on parameter choice was a difficulty in our analysis.

Since our operator is linear, the posterior is also a Gaussian process with mean and

covariance given by:

$$f_{pos} = f_{prior} + C_{prior}\mathbf{F}^T(\mathbf{F}C_{prior}\mathbf{F}^T + \Gamma)^{-1}(\mathbf{d} - \mathbf{F}f_{prior}), \tag{2.18}$$

$$C_{pos} = C_{prior} - C_{prior}\mathbf{F}^T(\mathbf{F}C_{prior}\mathbf{F}^T + \Gamma)^{-1}\mathbf{F}C_{prior}. \tag{2.19}$$

Using all of this information as well as the operator matrix found from the previous section, we can obtain samples of the posterior Gaussian. The plot below shows a sample from the prior process as well as the corresponding sample from the posterior.

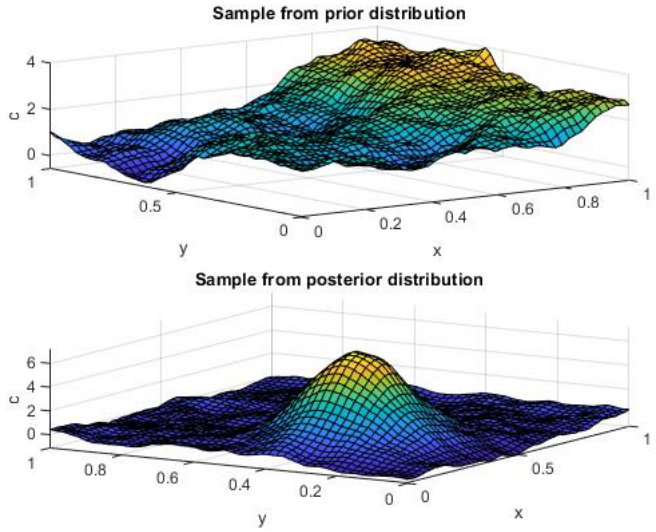
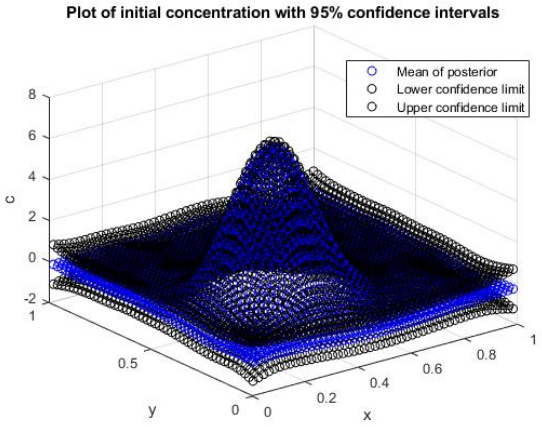


Figure 4: A plot of samples from initial prior and posterior distributions.

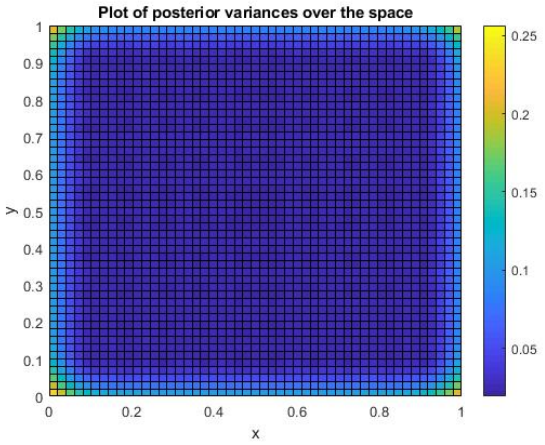
Comparing with the Tikhonov solution, we again have a broad peak and observations near zero away from the peak. In contrast with the Tikhonov solution however, we have a way to obtain confidence intervals. The main diagonal of the posterior covariance matrix gives us the variance of the points in the discretisation, so using the posterior mean, a 95% confidence interval for the initial concentrations in the discretisation is given by:

$$f_{pos} \pm 1.96 * diag(C_{pos}). \tag{2.20}$$

The plot below (left) displays the posterior mean of the initial concentration and the confidence intervals around it:



(a) Posterior mean estimate with 95% confidence interval bands



(b) Posterior variances over the unit square

Figure 5: Posterior mean and variance estimates of the initial concentration

The colorplot on the right of the figure displays the posterior variances of the points on the grid. Note the regular pattern of dark patches where the variance is much lower. These are at the measurement locations, where you would expect lower variance.

2.4 The Forward Problem

We begin the analysis of the forward problem by considering the value of the initial concentration at the main water supply. Using the posterior mean and variance from the Bayesian approach, we have the following histogram:

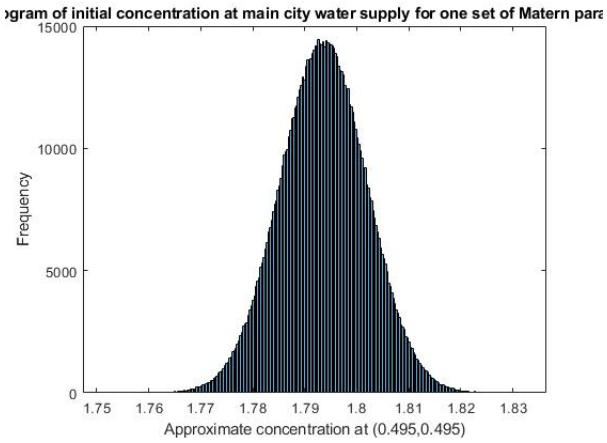


Figure 6: Histogram of concentrations at the main water supply, t = 0

This shows us that it is possible for the initial concentration at the main water supply to be above the threshold value of 1.8.

Recall the solution to the diffusion problem from section 2.1 (equation 2.5). We can first check that the solution to the inverse problem is sensible by using the Green's function to propagate forward the initial concentration and obtain the concentration at time 0.015,

from which we can compare with the data. We have the following:

$$c(x, y, 0.015) = 4 \sum_{n,m=1}^{\infty} e^{-\kappa\pi^2(n^2+m^2)*0.015} sin(n\pi x) sin(m\pi y) \int_0^1 \int_0^1 m(\hat{x}, \hat{y}) sin(n\pi\hat{x}) sin(m\pi\hat{y}) d\hat{x} d\hat{y} \tag{2.21}$$

For the integral, we apply the midpoint rule again with the same step-size ($h = 0.02$). As for the sum, we truncate after n=m=30 terms, which introduces another source of uncertainty. The outcome of this is the following surface, shown alongside the data measurements:

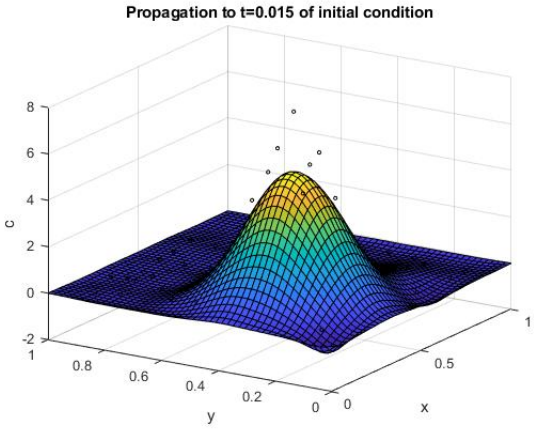


Figure 7: Checking inverse solution with data at t = 0.015

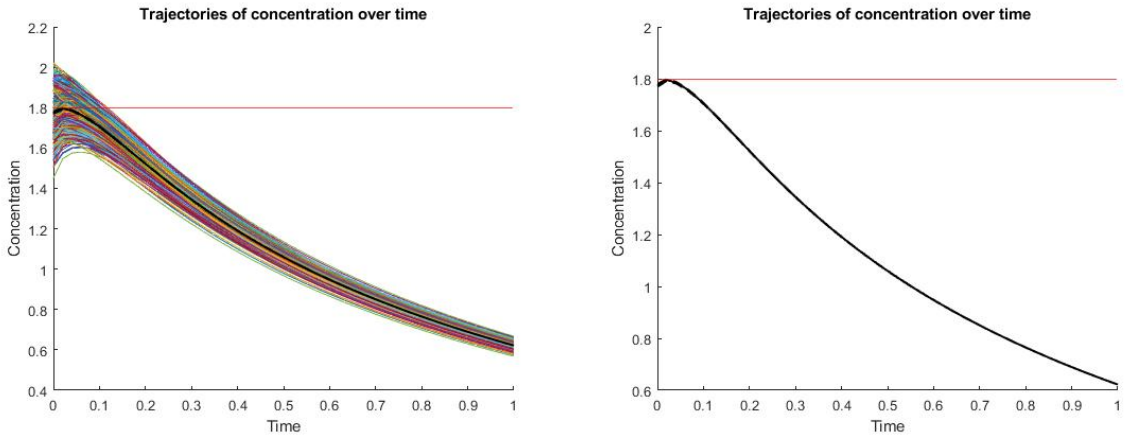
This surface fits the data away from the peak very well, and indeed follows the behaviour near to the peak. However, there are clearly some points above the surface at the peak. This could be due to the truncation of the infinite series, or a combination of this with the numerical error of two uses of the midpoint rule at each cell. Nevertheless, the shape of this plot is not unreasonable compared with the data.

Moving on, we shift our focus to the location of the main city water supply at $(x^*, y^*) = (0.495, 0.495)$. Consider $c(0.495, 0.495, t)$:

$$c(.495, .495, t) = 4 \sum_{n,m=1}^{\infty} e^{-\kappa\pi^2\left(n^2+m^2\right)t} \sin\left(.495n\pi\right) \sin\left(.495m\pi\right) \int_0^1 \int_0^1 m(\hat{x}, \hat{y}) \sin\left(n\pi\hat{x}\right) \sin\left(m\pi\hat{y}\right) d\hat{x} d\hat{y} \tag{2.22}$$

We can evaluate the integral via the midpoint rule. We fix a sample from our posterior of the Gaussian process to obtain the function $m(\hat{x}, \hat{y})$ then get results for the concentration for many later times.

Each sample we do this for gives a separate trajectory of the PDE over time at the location (0.495,0.495). We can then use Monte Carlo techniques to find the average as a function of time, as well as a 95% confidence interval. This should give us a good idea of the times at which there is a risk of exceeding the threshold value $c_{shut} = 1.8$. We provide plots of the trajectories (left) and the Monte Carlo average/confidence intervals (right) below.



(a) Trajectories from 1000 samples (b) Mean and 95% confidence band

Figure 8: Results from performing Monte Carlo

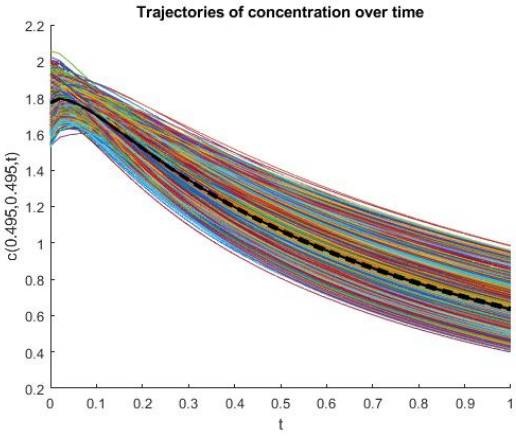
We can see immediately that the mean value never exceeds the critical threshold, however we do see a huge number of trajectories above the critical value (the red line on the plot) at earlier times. This suggests that there is some risk of the contamination exceeding the value at which the water supply must be shut down, between times of $t = 0$ and $t \approx 0.18$. We can see that the peak of most of the trajectories (and the mean) is roughly $t = 0.05$, before the measurements are even taken, and even at this point the average is still below the critical value of 1.8, so we can be fairly confident that on average this is not exceeded. Despite this, a large number of trajectories do exceed the critical value, so we should reference this and side with caution in our recommendation.

3 Further Work and Discussion

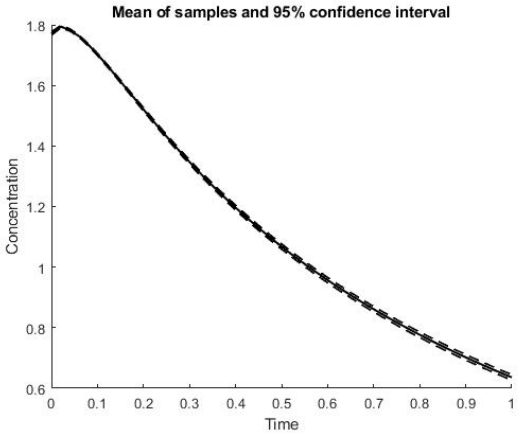
3.0.1 Assumption on Kappa

At the beginning of the investigation, we were told that κ is confidently around the value 0.05. Since this is not definitive, we decided to investigate the sensitivity of our results around this value. In regards to this, we instead suppose that kappa is some random value from the uniform distribution, $\kappa \sim Unif(0.03, 0.07)$. We can once again make use of Monte Carlo by fixing everything, then sampling many different κ and performing the same analysis as described earlier.

Taking 1000 samples of κ from this uniform distribution, and once again calculating the mean and 95% confidence interval from the trajectories generated using Monte Carlo, we obtain the following:



(a) Trajectories from 1000 samples



(b) Mean and 95% confidence band

Figure 9: Results from performing Monte Carlo with variable kappa

These figures show that allowing κ to be random results in more trajectories being further away from the mean at later times, however if we focus on the left side of the graph we see a similar scenario to before, with the mean being slightly below $c_{shut} = 1.8$, however we see a large proportion of trajectories being above this value at early times. More trajectories stay above the critical value for slightly longer in this scenario, although these results suggest to us that we should not significantly change our recommendation based on this, and that loosening our assumption on the value of κ does not have a detrimental effect on our results.

4 Concluding Remarks

4.1 Limitations & Discussion

One of the main limitations to our investigation is that evaluating (2.5) is very computationally expensive, as it is an infinite sum and requires the sampling of a two-dimensional Gaussian process. On top of this, we have to evaluate the integral numerically at each term in the sum. This means that, especially when performing the forwards propagation where this has to be evaluated at 50 time points for many samples, the code can take a significant amount of time to run. This in itself limited the number of samples we could feasibly take for Monte Carlo for example. One possible solution to this problem would be to investigate and implement a fast Fourier transform to make this computation faster.

As mentioned previously, there is further uncertainty prevalent in the choice of hyperparameters in the Matern prior. Further analysis of this choice could be considered, as well as the effect of changing the covariance of the measurement errors, for example by considering non-uniform variance. The last thing to mention is that the time we consider must have been scaled somewhat, it could well represent years, days or centuries. This has an affect on our recommendation, as for example it is not plausible to shut down the water supply for a small amount of time (if it represents days) as we imagine this would take a long time.

4.2 Conclusion

In conclusion, Bayesian techniques have allowed us to quantify the uncertainty around the initial concentration of the contaminant in the aquifer. Using samples of a posterior Gaussian process, we have managed to propagate this uncertainty forward, with focus on the location of the city water supply. We have found that it is possible (in fact, many trajectories exhibit this feature) for the concentration to exceed the critical value of $c_{shut} = 1.8$. Since contamination is a massive health risk to residents in the city, we recommend being cautious, and shutting down the water supply between the times of $t = 0$ and $t = 0.2$. This will ensure that the water has a concentration below the critical value upon reopening. This recommendation is despite the fact that Monte Carlo showed us that on average, concentrations found by sampling at this point do not exceed this value, since we only had 1000 trajectories there is simply not enough evidence to rule out the concentration having exceeded the shutdown value. More trajectories exceeding this value were also found when we tested the sensitivity of our results with respect to the diffusion constant κ, supporting this conclusion.

References

[1] M. Hancock. *The heat and wave equations in 2D and 3D.* Link: https://ocw.mit.edu/courses/18-303-linear-partial-differential-equations-fall-2006/pages/lecture-notes/. Last accessed: 04/04/22. This reference provides more information on the solution of the heat/diffusion equation.

[2] G. Holzegel. *Functional Analysis.* Link: https://www.imperial.ac.uk/people/g.holzegel/document/2311/notes1/?notes1.pdf. Last accessed: 04/04/22. This reference gives information on functional analysis and compact operators.

[3] C. Constanda. *Solution Techniques for Elementary Partial Differential Equations.* Chapman and Hall/CRC, 2016.

[4] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems.* Springer, Dordrecht, 2005.