# Problem Solving and Modelling Task

Alexander Arthur

February 18, 2022

## Contents



Word count excluding QCAA exclusions (e.g. title pages, data tables and equations): 1683

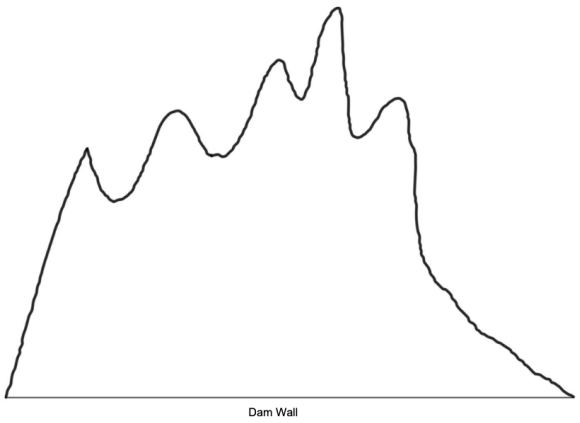



Figure 1: Diagram of outline of dam

# 1 Introduction

The given task is to predict the date on which a population of bacteria will completely cover the surface of the water in a dam (Figure 1). In order to do this, the area of the lake must first be calculated from the diagram provided. Once this area is known, the growth of the bacteria can then be projected and used to calculate the point at which the lake will be covered.

# 2 Assumptions and observations

## 2.1 Observations

**Observation 1.** When the diagram of the dam is oriented with the straight wall colinear with the $x$-axis, the rest of the outline passes the vertical line test, meaning that it can be modelled with a mathematical function.

**Observation 2.** The size of the bacterial population was recorded at the end of each day, meaning that the measurements are consistent. However, measurements were not taken daily; there are only four data points provided. This means that a method must be used to extrapolate this data foward to a point in time where the area covered exceeds that of the lake.

**Observation 3.** The area covered by the bacteria increases each day, suggesting that its gradient will be positive and thus that it will eventually increase to be larger than that of the dam. This means that the dam will eventually be covered in bacteria by a certain date, which this investigation will find.

## 2.2 Assumptions

In order to simplify the problem, certain assumptions must be made either due to a lack of information provided or to bring the scope of the problem to a reasonable scale.

**Assumption 1.** If the water level of the dam were to change the water's surface area would also change, since it is unlikely the walls of the dam are perfectly vertical. Changing the surface area of the dam would in turn affect the point at which the bacteria would cover the entire lake, affecting the result of the investigation. Since no data is provided on the water level, it is reasonable to assume that it is static, and will not have any effect on the surface area of the lake.

**Assumption 2.** In a similar vein, there are many factors which can influence the growth rate of a population of bacteria, including access to a food source, temperature and competition for space. Similarly to the dam's water level, no data is provided around this, meaning that it must be assumed that the none of these factors will affect the bacteria as the population grows.

# 3 Translation of mathematical concepts

As per the requirements of the task, a mathematical approach must be used to calculate the area of the dam. To satisfy this requirement, the definite integral of a piecewise function will be used to calculate the area of the dam.

In order to create this piecewise function, the data was broken up into eight sections so as to reduce the complexity of the model needed for each section. In order to select these functions, regression analysis was used.

The Weierstrass approximation theorem states that for any continuous function $f$ over the real interval $[a, b]$ and any $\epsilon > 0$ there exists a polynomial $p$ such that $|f(x) - p(x)| < \epsilon$, where $a \leq x \leq b$. This means that since the only metric used to select a function is its accuracy within the specified domain, regression was used to select a function within each defined section, and the regression models used were strictly polynomial; the degree of the polynomial was simply increased until the function fit the data to a sufficient degree of accuracy ($R^2 > 0.97$).

To generate the points to be used for these regressions, a Python script was written to list the coordinates of the topmost black pixel in each column of pixels (Appendix 6.1, page, 11). This generated a list of 1262 pixel coordinates, which were scaled to the dimensions of the dam such that one unit on the $x$ and $y$ axes were each equal to 1 meter and subsequently used to generate regression models (Figure 2).

To predict the point at which the bacteria will cover the lake, regression will again be used, since it is an efficient way of finding a mathematical function that fits as closely as possible a set of discrete data points.

In an ideal scenario, a single function could be used to model the entire length of the dam's outline. However, due to the computational limits of the available software a polynomial regression of sufficiently high degree is not possible. This means that the data was instead divided into several sections (Figure 3), each of which were then used to run a polynomial regression. Since these sections were selected to have more simple shapes than the entire dataset, polynomials of much lower degree ($< 6$) were sufficient to replicate their shape to a sufficient degree of accuracy.

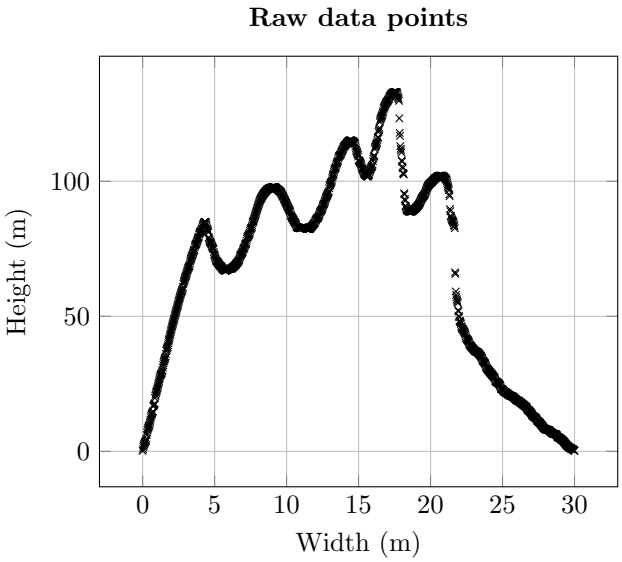



Figure 2: Raw data from Python script

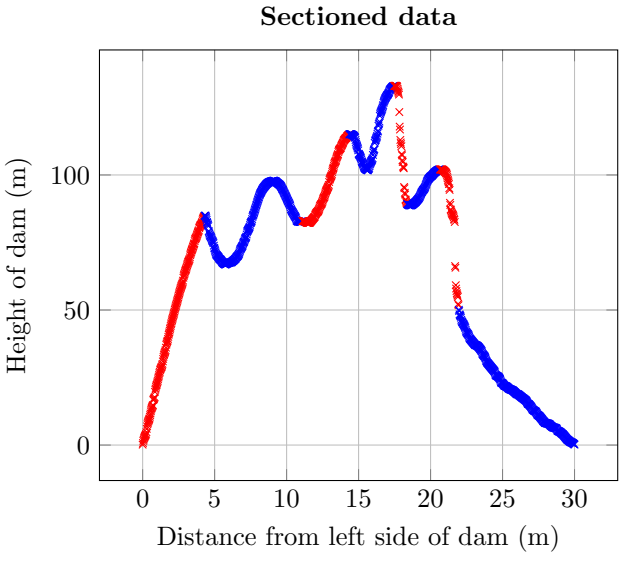



Figure 3: Data shown in coloured sections

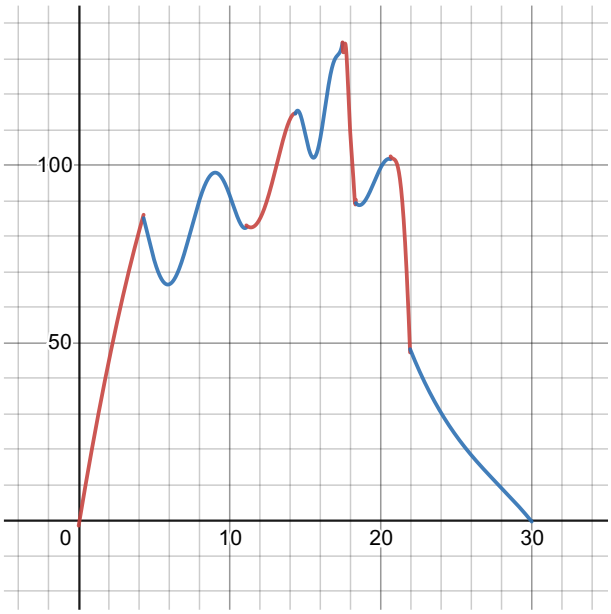



Figure 4: Piecewise function modelling outline of dam

## 3.1 Calculation of area

Let us define a variable $b$ for the boundaries between the sections, such that $b_0 = 0$, $b_1 = 4.3 \ldots b_8 = 30$. Let us also define the piecewise function $f$ as follows (Figure 4):

$$f(x) = \begin{cases} f_1(x) & b_0 \leq x < b_1 \\ f_2(x) & b_1 \leq x < b_2 \\ \vdots & \vdots \\ f_7(x) & b_6 \leq x < b_7 \\ f_8(x) & b_7 \leq x \leq b_8 \end{cases} \tag{1}$$

The values of $b_n$ were selected to create sections of data that were simple to create regression models for and are as follows:

$$\begin{aligned} b_0 &= 0 \\ b_1 &= 4.3 \\ b_2 &= 11.1 \\ b_3 &= 14.4 \\ b_4 &= 17.5 \\ b_5 &= 18.4 \\ b_6 &= 20.7 \\ b_7 &= 22.0 \\ b_8 &= 30 \end{aligned}$$

The subfunctions $f_1 \cdots f_8$ were found using a graphing calculator using polynomial regression models, repeatedly increasing the degree by one until the model's $R^2$ was greater than 0.97 and the model was visually close to the shape of the data. The functions found are as follows:

$$
\begin{aligned}
f_1(x) &= -1.2364x^2 + 25.667x - 1.4657 \\
f_2(x) &= 0.0922x^5 - 3.4033x^4 + 47.748x^3 - 315.07x^2 + 968.69x - 1022.9 \\
f_3(x) &= -0.4898x^4 + 22.736x^3 - 388.82x^2 + 2903.4x - 7897 \\
f_4(x) &= 1.1895x^6 - 110.38x^5 + 4255.2x^4 - 8.7230 \cdot 10^0 4x^3 + 1.0027 \cdot 10^6 x^2 \\
&\quad - 6.1276 \cdot 10^6 x + 1.5551 \cdot 10^7 \\
f_5(x) &= 6046.8x^6 - 6.5135 \cdot 10^5 x^5 + 2.9233 \cdot 10^7 x^4 - 6.9967 \cdot 10^8 x^3 \\
&\quad + 9.4192 \cdot 10^9 x^2 - 6.7624 \cdot 10^{10} x + 2.0228 \cdot 10^{11} \\
f_6(x) &= -3.4000x^3 + 199.63x^2 - 3897.0x + 25388 \\
f_7(x) &= 253.31x^5 - 26978x^4 + 1.1491 \cdot 10^6 x^3 - 2.4471 \cdot 10^7 x^2 + 2.6052 \cdot 10^8 x \\
&\quad - 1.1093 \cdot 10^9 \\
f_8(x) &= -0.0556x^3 + 4.6621x^2 - 134.73x + 1347.2
\end{aligned}
$$

Using the subfunctions of $f$, the area $A$ inside the lake can be calculated using the sums of the definite integrals of each function $f_n(x)$ within its applicable domain $b_{n-1} \leq x < b_n$, otherwise expressed as:

$$A = \sum_{n=1}^{8} \int_{b_{n-1}}^{b_n} f_n(x)\, dx \tag{2}$$

Equation 2 was then evaluated using a graphics calculator for each $n$.

| $n$ | Area ($m^2$) |
|---|---|
| 1 | 198.22 |
| 2 | 561.52 |
| 3 | 311.77 |
| 4 | 316.16 |
| 5 | 101.00 |
| 6 | 218.10 |
| 7 | 113.73 |
| 8 | 163.43 |
| Total | 2027.93 |

This process will now be demonstrated for $n = 1$:

$$\begin{aligned}
A_1 &= \int_{b_0}^{b_1} f_1(x)dx \\
&= \int_0^{4.3} \left(-1.2364x^2 + 25.667x - 1.4657\right) dx \\
&= \left[-0.41213x^3 + 12.8335x^2 - 1.4657x\right]_0^{4.3} \\
&= (-0.41213 \cdot 4.3^3 + 12.8335 \cdot 4.3^2 - 1.4657 \cdot 4.3) \\
&\quad - (-0.41213 \cdot 0^3 + 12.8335 \cdot 0^2 - 1.4657 \cdot 0) \\
&= 198.2217
\end{aligned}$$

A graphing calculator was used to evaluate Equation 2 and the area of the lake found to be 2027.93m$^2$.

## 3.2 Projection of bacteria growth

The data provided for the area covered by the bacteria is as follows.

| Date | Area ($m^2$) |
|---|---|
| 1 Jan | 1 |
| 9 Jan | 20 |
| 15 Jan | 119 |
| 19 Jan | 368 |

In order to project the size of the bacterial population, multiple types of function were tested using regression analysis. The types of function tested were exponential, quadratic and cubic. Regression analyses for each of these functions were performed using a graphics calculator and the results summarised below (Figure 5.

| Type | Form | $R^2$ |
|---|---|---|
| Exponential | $re^{ax} - b$ | 1 |
| Quadratic | $ax^2 + bx + c$ | 0.9701 |
| Cubic | $ax^3 + bx^2 + cx + d$ | 1 |

Both the exponential and cubic functions have an $R^2$ value of 1, indicating that they both exactly fit the data provided. However, the cubic model predicts that the area covered by the bacteria is negative in the 24 hours before the first measurement was taken. This is not physically possible, meaning that this model does not accurately model the size of the population and that therefore the exponential model will be used to predict the growth of the bacteria.

The exponential function found was

$$f(x) = 1.80065e^{0.280261x} - 1.80409 \tag{3}$$

Using this regression model, the point at which the area covered by the bacteria can be found as follows:

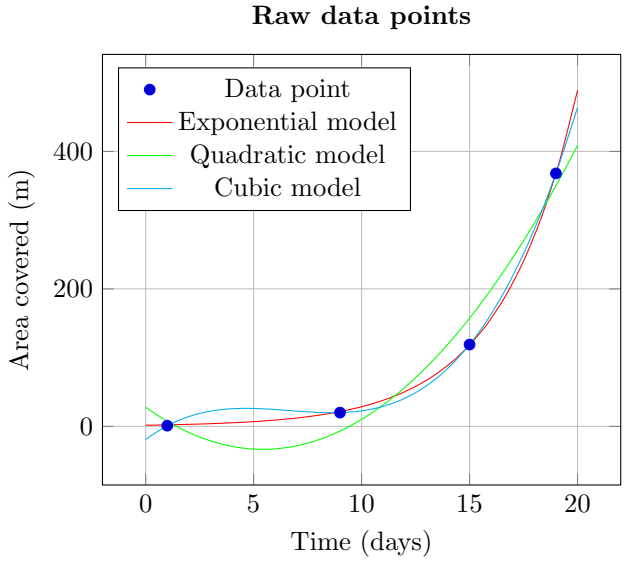



Figure 5: Raw data and regression models of bacteria population

$$
\begin{aligned}
2027.9283 &= 1.80065e^{0.280261x} - 1.80409 \\
2026.1197 &= 1.80065e^{0.280261x} \\
\frac{2026.1197}{1.80065} &= e^{0.280261x} \\
\ln 1125.21572765 &= 0.280261x \\
x &= \frac{\ln 1125.21572765}{0.280261} \\
x &= 25.03636597
\end{aligned}
$$

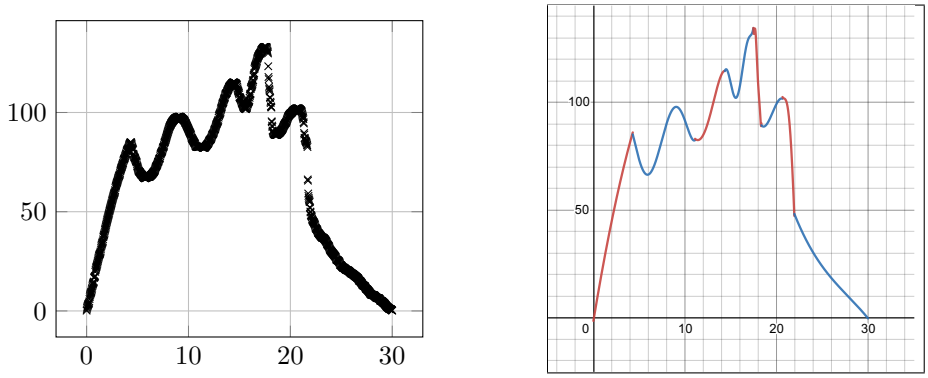
This means that the dam will be entirely covered by bacteria on the $25^{\text{th}}$ of January.

# 4 Evaluation

## 4.1 Reasonableness

To determine whether the predicted date of complete coverage is reasonable, we must first determine the reasonableness of the estimation of the area of the dam. In order to do this, the image of the dam was scaled such that 1 pixel represented a 10 by 10cm area. A Python script was then written to count how many pixels were beneath the dam outline in each column of pixels, using the rectangle method of estimating the area underneath a curve. The script returned a value of 204 020 pixels, or 2040.2m$^2$. This is an error of 0.64% from the predicted area of 2027.93, an amount which is small enough to be negligible.

In addition, the area of a triangle with a base of 30m and a height of 133m is 1995m$^2$, which is also very close to 2027.93. Since both of thes methods of




(a) Raw data

(b) Piecewise function

Figure 6: Comparison of raw data and piecewise function

approximating the area of the dam are close to the actual value calculated, the solution for the area of the dam is reasonable.

The prediction of when the bacteria will cover the dam is also reasonable, since it is based on valid assumptions. Specifically, the assumption that no outside factors would influence the growth of the bacteria was both necessary and reasonable, because the amount of resources consumed by a population of bacteria in 25 days would not exceed that which is available in the entire body of water of a dam. Similarly, the assumption that the water level of the dam was valid, since dams will actively regulate their level by increasing or decreasing output to match that which is flowing in, resulting in no net change in quantity of water and thus in surface area.

## 4.2 Evaluation of model and result

### 4.2.1 Strengths

The model had a very close fit with the data in each section ($R^2 > 0.97$), and was also very visually similar to the plot of the raw data (Figure 6). This indicates that the shape whose area was calculated matched that of the physical dam closely, and since the process of integration bears no inherent inaccuracies it is likely that the area calculated is highly accurate.

Similarly, the model for the size of the bacteria population also had a very high $R^2$ value of 0.9999, indicating that it is also a very close fit for the data collected. An exponential function is also theoretically the correct type of function to model the size of a population of bacteria, since each generation the population multiplies by a certain amount when ignoring outside factors as per Assumption 2.

Another strength of the model is the extent to which it makes use of computational resources to increase its accuracy. Using a Python script to generate the set of points used for regression greatly increases the number of points available without introducing any inaccuracies due to human error. This in turn makes the regression models generated more accurate due to this elimination of human error.

#### 4.2.2 Limitations

The model is limited by the fact that only one set of section boundaries were used. Another set of boundaries may have yielded more accurate regression models, and thus a more accurate calculation of the area of the dam. The model is also limited by the fact that the Python scripts used are highly specialised to this task. While the approach they use is widely applicable, the specific implementation used is not highly versatile, limiting their usefulness outside the context of this task.

# 5 Conclusion

# 6 Appendix

## 6.1 Python script

```
if __name__ == "__main__":
    img = Image.open('dam cropped.png')

    [width, height] = img.size
    rowslist = []
    xvals = []
    yvals = []
    regressionDegree = 10

    # make image binary (each pixel is either white or black)
    img = img.point(mappoint)



    # find coordinates of black pixels
    for x in range(width):
        for y in range(height):
            if img.getpixel((x, y)) == (0, 0, 0):
                # print(x, y, img.getpixel((x, y)))
                rowslist.append([x, height - 1 - y])
                xvals.append(x)
                yvals.append(height - 1 - y)


    with open('points.csv', 'w') as file:
        writer = csv.writer(file)
        writer.writerows(rowslist)
```