# Time Series Analysis of Bovine TB Herd Incidence in Ireland (2011-2023).

Rory O' Sullivan
20232721

April 28, 2024

## 1 Introduction

This project will explore herd incidence % of Bovine Tuberculosis (bTB) in the Republic of Ireland at a state level, gathering data from 26 counties. The data was sourced from the Central Statistics Office (CSO) [1]. The data was collected quarterly from the beginning of 2011 to the end of 2023 [2]. Herd incidence % can be defined as the percentage of herds that experience a new TB breakdown in a given period, in this case, a quarter.

There are three main ways cattle obtain TB - purchasing infected cattle, presence of residual (undetected) infection within herds and spread from wildlife, mainly badgers. The Eradicate TB Programme aims to eradicate bTB by 2030, though this has been noted as unlikely (More, 2023) [3]. The programme employs various measures to attempt to reduce the spread of bTB including spreading awareness of bTB and how to combat it, vaccinating badgers and more bTB tests being carried out.

There are 52 entries in this dataset. For this project, the last 10% will be removed to train the model, leaving us with 47 entries. From Figure 1, we can see a seasonal pattern with what seems to be maybe a quadratic trend. We notice straight away that the herd incidence grows from Q1-Q4 and sharply drops down to Q1 again every year. The values are increasing towards the end of the graph quickly, especially in the recent unseen data shown in teal, which the model may not be able to foresee. Notably, there is one key outlier in this dataset at 2021 Q1 whereby the incidence rate drops to a very low level. There is also a repeated value of 2.59 in 2018Q1 and 2018Q2. Exploring other data collected by the CSO in this dataset not examined in this project such as the number of animals in the area similarly showed repeated values for these quarters. The cause of this is unknown.

**Note:** There was no clear definition of the obtainment of these numbers from the CSO, however, I suspect that these incidence rates are not showing on a quarter-to-quarter basis, rather starting from Q1 and rolling until the end of the year and then resetting, explaining how the rates strictly only increase within the year. This point is outside the scope of this project as we are only interested in making a model.

## 2 Identifying the Model

From our introduction, we will likely have a SARIMA model due to the clear seasonality of the data. To examine what components influence the data the most, we can decompose the dataset. Figure 2 decomposes the series into trend, seasonal and random components, from which we obtain the following:

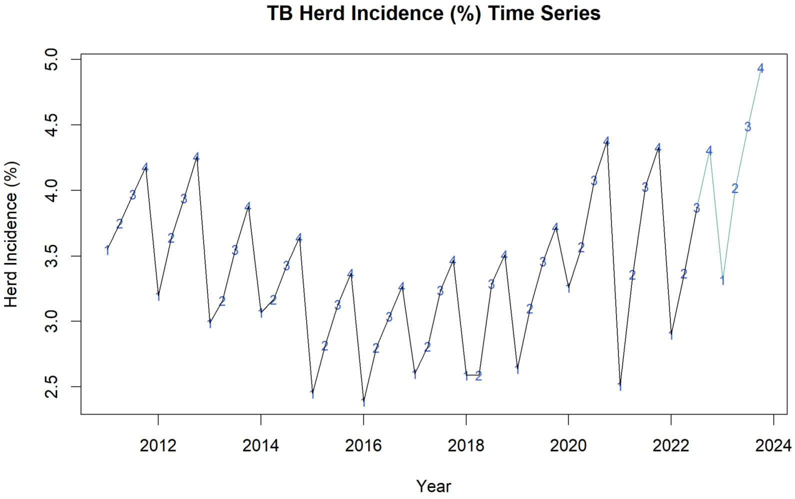


Figure 1: Plot of the time series data. The remaining 10% of the data is fitted in teal.

- **Trend:** We can see what seems to be a quadratic trend in the data, with the range of the trend being 0.93.
- **Seasonal:** Here we note the cycle we saw earlier whereby the data will increase within the year followed by a large drop and the cycle will repeat. The range for the seasonal values here is 1.03
- **Random:** This plot does seem to have a random pattern except for the sharp dip where we noted the outlier in 2021 Q1. The range here is 0.81

Interpreting the different ranges associated with these components, we can see that there does seem to be a strong trend and seasonal component within the data, with the randomness having less of an impact than the former components.

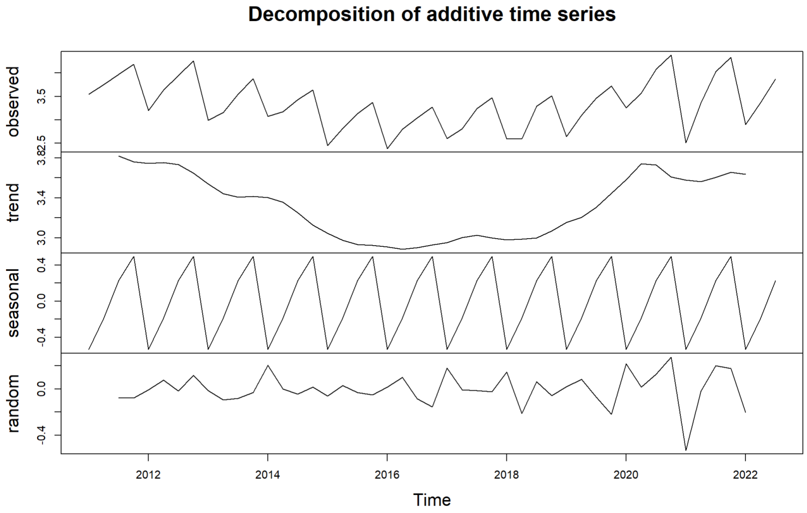


Figure 2: Decomposing the time series into trend, seasonal and random components.

# 3 Transforming the Series

As we noted what seemed a quadratic trend in the data, we can perform a second-order difference on our data to eliminate this trend. One could argue, for say a further 10 years into the data, the quadratic trend would likely not continue and over time, the process may seem linear. Therefore we will fit models with both a first and second-order difference and we can compare how each of these models perform.

Other methods were also considered to make the data stationary. These methods included using the Box-Cox transformation and taking the log of the data.

- **Box-Cox:** The Box-Cox transformation yielded an MLE of 0.6 with a confidence interval of (0,2). This interval is very wide and the associated plot from the transformation didn't bring much certainty with this answer. Transforming the data by taking it all to the power of 0.6, didn't transform the data too much at all.
- **Log:** Similarly to the Box-Cox transform, the natural log transform did very little to the data.
- **Note:** Though the differences here were very little, both methods were tried until predictions. The log and linear trend Box-Cox yielded nearly identical results to the linear model, therefore I saw it as redundant. Similarly the Box-Cox transform with assuming a quadratic trend proved similar to the ordinary quadratic trend model.

From these conclusions, I will choose to analyse two models in this project. They are the one that assumes a linear trend, and the one that assumes a quadratic trend.

## 3.1 The Linear Trend Model

Assuming the trend will stay linear in the data, the data was differenced once and also differenced at lag = 4 to remove the seasonal trend from the data.

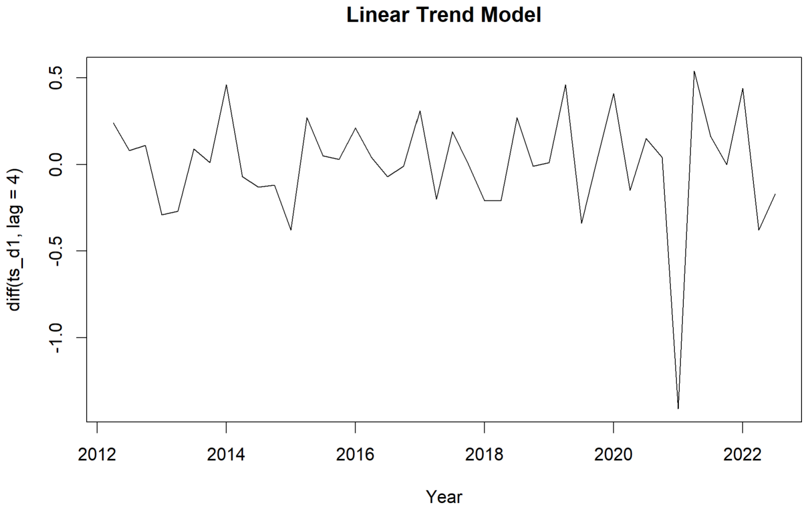


Figure 3: Plot of the once differenced time series.

The data appears reasonably stationary with a seemingly constant mean and near-constant variance. A glaring outlier corresponding to 2021 Q1 can be seen here. We can further investigate the stationarity of the data by running an Augmented Dicky-Fuller test. Table 1 returns the result of this test, showing stationarity in the data.

| Dickey-Fuller | Lag order | p-value |
|---|---|---|
| $-6.05$ | 3 | 0.01 |

Table 1: Augmented Dickey-Fuller Test for stationarity.

We will now examine the ACF and PACF of the series so that we may identify a suitable model for the data.

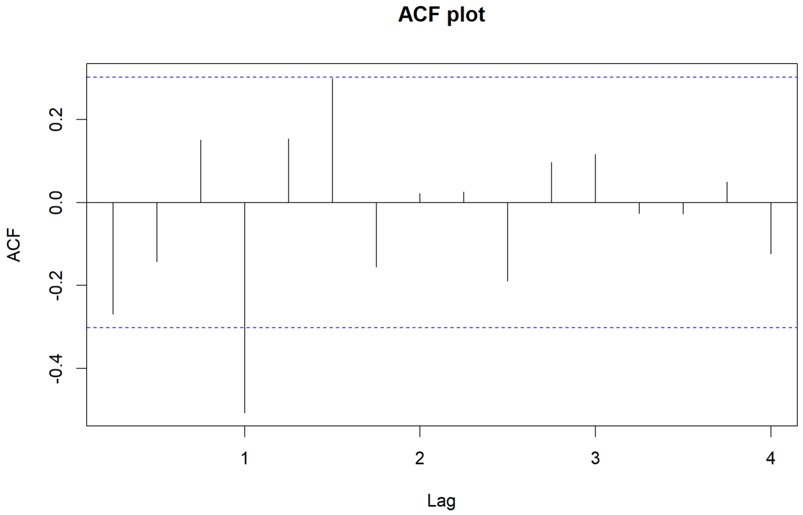


Figure 4: ACF of the series.

Examining the ACF, we note a significant autocorrelation at lag = 1, indicating a seasonal MA(1) term. There are no other apparent significant lags to indicate any non-seasonal MA terms.

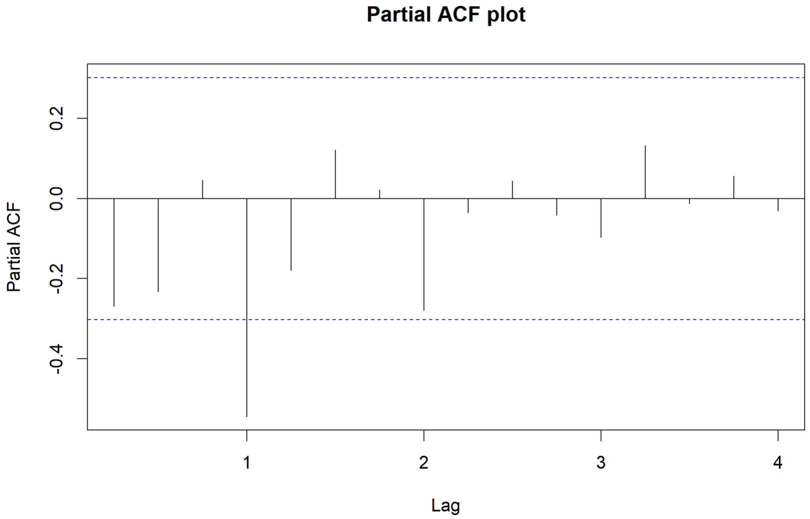


Figure 5: Partial ACF of the series.

Reviewing the partial ACF of the series, we again note a significant partial autocorrelation at lag = 1 however this decays on the seasonal lags, indicating an MA seasonal component. For the non-seasonal component, again there seems to be no significant correlations.

Using the Extended Autocorrelation Function (EACF) here for model selection isn't suitable as we have seasonality, however, the grid does have a column of X's on MA(3), indicating the seasonality

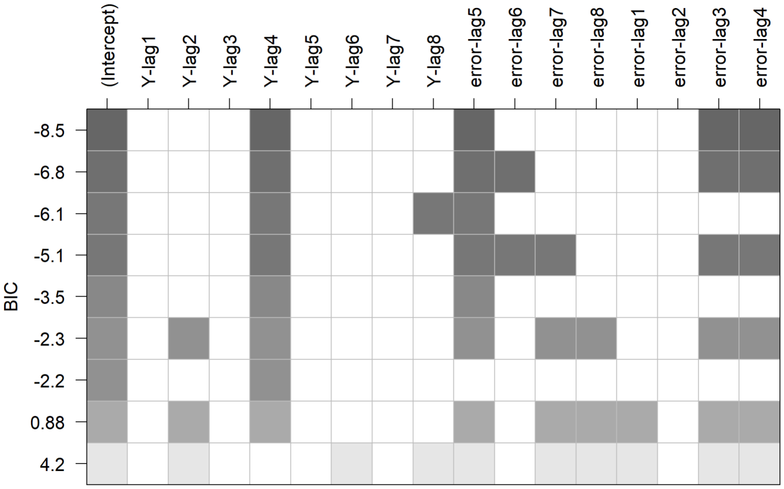
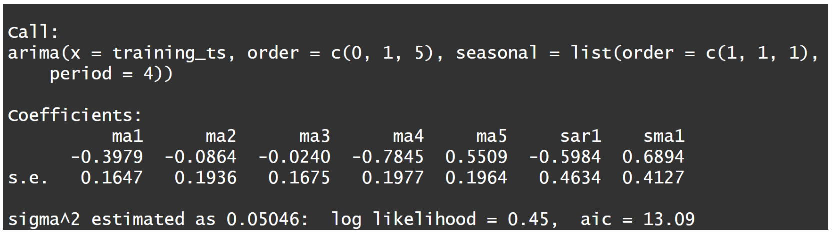


Figure 6: Using armasubsets() on the data.

is indeed there in the data. Turning to using the armasubsets() function, the plot outlines different significance to what I spotted. The plot here is picking up on a seasonal AR model, with a seasonal MA and a non-seasonal MA(5) with a significant MA(3) also, indicating here that an MA(1) may be significant due to it being one from the seasonal level.

Now that we will plot our initial SARIMA model, the parameters I am choosing to go off on my first run are (0,1,5),(1,1,1) with period 4. The R output from this call can be seen in Figure 7 below. The model outputs an AIC of 13.09, picking up significance on just the MA(4) and MA(1) terms. It's not picking up on the significance of the seasonal MA term, though this may just be due to all the other terms in the model now as I believe, from our earlier analysis that there should be a seasonal MA term in the model.



```
Call:
arima(x = training_ts, order = c(0, 1, 5), seasonal = list(order = c(1, 1, 1),
    period = 4))

Coefficients:
          ma1      ma2      ma3      ma4     ma5     sar1    sma1
      -0.3979  -0.0864  -0.0240  -0.7845  0.5509  -0.5984  0.6894
s.e.   0.1647   0.1936   0.1675   0.1977  0.1964   0.4634  0.4127

sigma^2 estimated as 0.05046:  log likelihood = 0.45,  aic = 13.09
```

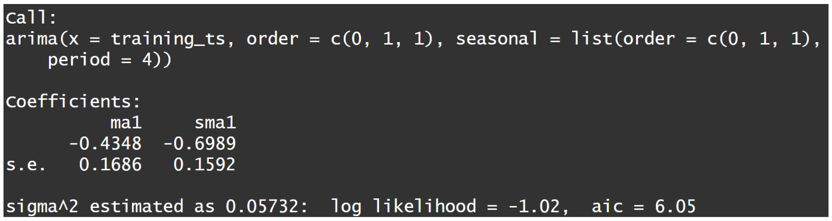
Figure 7: Initial ARIMA output for the model assuming a linear trend.

To trim the model, I remove the MA terms one by one until all are significant. This leaves me with just one MA non-seasonal term. Following this, I removed the AR seasonal term as it is insignificant and we also identified prior that this term shouldn't be in the model from the PACF. In the end, the final model we are left with is a seasonal ARIMA $(0, 1, 1) \times (0, 1, 1)_4$. The output of the model can be seen in Figure 8 below.

Both terms in the model are significant with the AIC now being much lower at 6.05, indicating a better-performing, more parsimonious model. Mathematically, we can write this model as:

$$(1 - B)(1 - B^4)Y_t = (1 - 0.43B)(1 - 0.7B^4)e_t \tag{1}$$

with $\sigma_e^2 \approx 0.057$.



```
Call:
arima(x = training_ts, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1),
    period = 4))

Coefficients:
          ma1     sma1
      -0.4348  -0.6989
s.e.   0.1686   0.1592

sigma^2 estimated as 0.05732:  log likelihood = -1.02,  aic = 6.05
```

Figure 8: Final ARIMA output for the model assuming a linear trend.

### 3.1.1 Overfitting the Model

Though we started with a model with many terms and reduced it to a more parsimonious model, I also aimed to find any room for improvement in this current model via overfitting. Changing the number of non-seasonal and seasonal terms in the models yielded models with higher AICs and insignificant terms. We conclude that this model is our final model assuming a linear trend.

## 3.2 The Quadratic Trend Model

In this section, we explore the model whereby we applied a second-order difference the start, assuming a quadratic trend followed by a seasonal difference to remove seasonality.

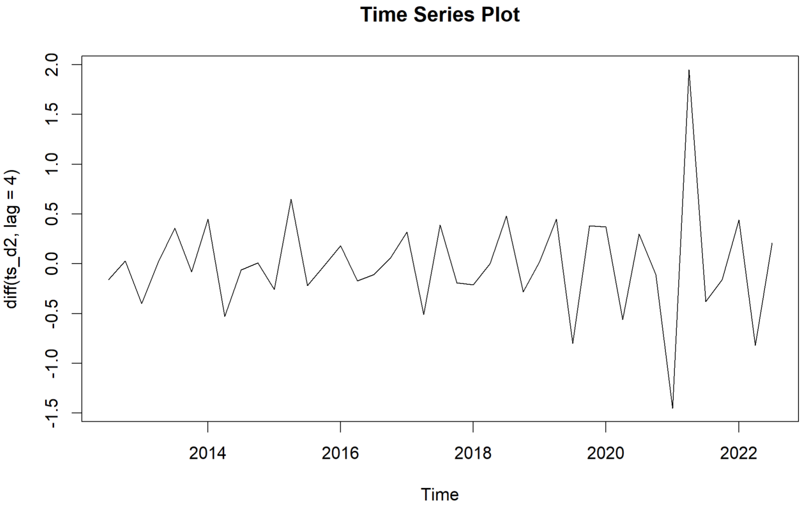


Figure 9: Plot of the twice differenced time series.

The data appears approximately stationary with seemingly constant mean and variance. Comparing this to our linear trend model, the outlier is still apparent with it naturally going in the opposite direction from an added order difference. Running an Augmented Dicky-Fuller test further indicates stationarity in the data.

| Dickey-Fuller | Lag order | p-value |
|---|---|---|
| $-5.57$ | 3 | 0.01 |

Table 2: Augmented Dickey-Fuller Test for stationarity.

We will now examine the ACF and PACF of the series so that we may identify a suitable model for the data.

Examining the ACF, we note a significant autocorrelation at lag = 1, indicating a seasonal MA(1)

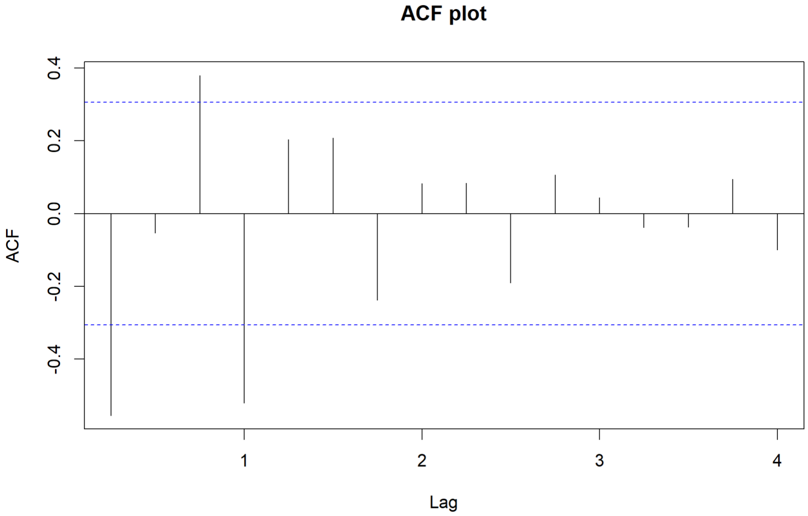


Figure 10: ACF of the series.

term. We also note a significant autocorrelation at lag 0.25 and 0.75, indicating that a non-seasonal MA(1) term may be suitable.

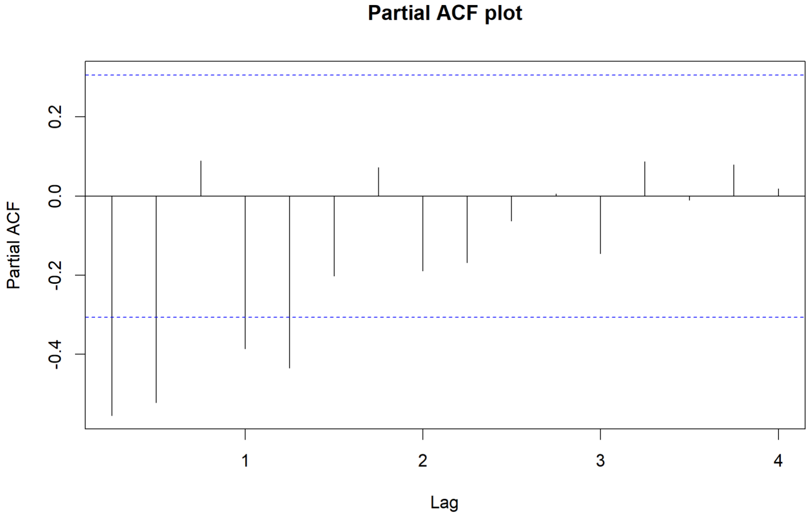


Figure 11: Partial ACF of the series.

Examining the partial ACF of the series, we again note a significant partial autocorrelation at lag = 1, however, this decays on the seasonal lags, indicating an MA seasonal component. For the non-seasonal component, it is less clear here, indicating perhaps an AR(1) or AR(2) although the argument could be made here that the first two non-seasonal lags are tapering.

As with the linear trend model, examining the EACF for model selection isn't suitable as we have seasonality, however, the grid does have a column of X's at MA(3), indicating the seasonality is indeed there in the data. Turning to using the armasubsets() function, the plot outlines different significance to what I spotted.

The plot here is picking up on a non-seasonal AR(6) with significant terms on 1 and 6. It is also highlighting a non-seasonal MA(6) with significant terms of 1,5 and 6. The function does not pick up on any seasonality unlike what we saw in the ACF and PACF plots.

For the initial model, the parameters I am choosing to go off on my first run are $(2, 2, 2) \times (0, 1, 1)_4$.

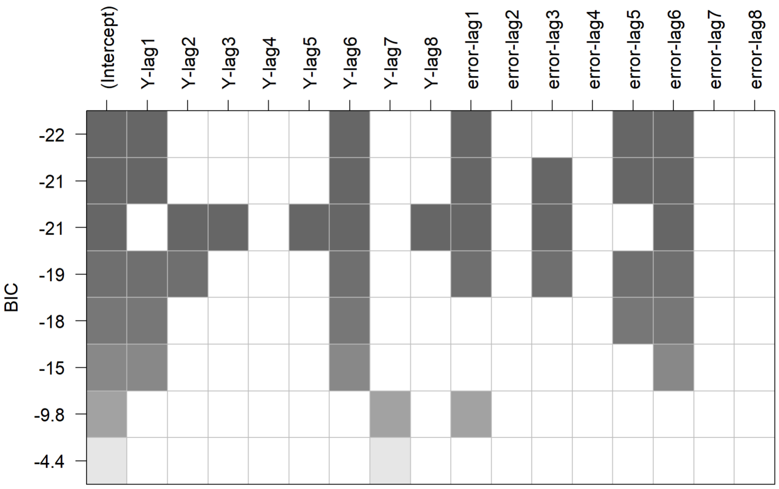
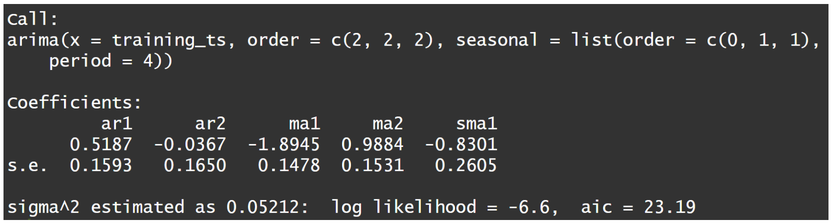


Figure 12: Using armasubsets() on the data.

I'm not choosing to agree with armasubsets() entirely here as when using these terms, they are mostly insignificant. It is of course important to highlight that armasubsets() does not include insignificant terms in its' model. The R output from this call can be seen in Figure 13 below.

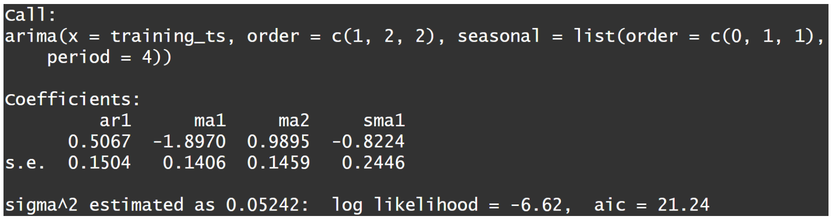


```
Call:
arima(x = training_ts, order = c(2, 2, 2), seasonal = list(order = c(0, 1, 1),
    period = 4))

Coefficients:
         ar1      ar2      ma1     ma2     sma1
      0.5187  -0.0367  -1.8945  0.9884  -0.8301
s.e.  0.1593   0.1650   0.1478  0.1531   0.2605

sigma^2 estimated as 0.05212:  log likelihood = -6.6,  aic = 23.19
```

Figure 13: Initial ARIMA output for the model assuming a quadratic trend.

The model outputs an AIC of 23.19, picking up significance on all terms bar the second term of the AR(2) non-seasonal.

To trim the model, I remove the second AR non-seasonal term due to insignificance. The final model we are left with is a seasonal ARIMA $(1, 2, 2) \times (0, 1, 1)_4$. The output of the model can be seen below.



```
Call:
arima(x = training_ts, order = c(1, 2, 2), seasonal = list(order = c(0, 1, 1),
    period = 4))

Coefficients:
         ar1      ma1     ma2     sma1
      0.5067  -1.8970  0.9895  -0.8224
s.e.  0.1504   0.1406  0.1459   0.2446

sigma^2 estimated as 0.05242:  log likelihood = -6.62,  aic = 21.24
```

Figure 14: Final ARIMA output for the model assuming a quadratic trend.

All terms in the model are now significant with the AIC now being much reasonably lower at 21.24, indicating a weak-moderate argument that this model is a better-performing, more parsimonious model. This model can be written mathematically as:

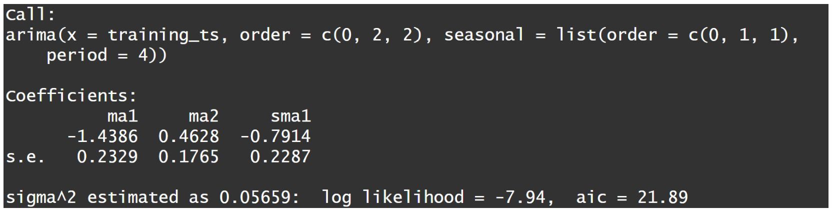
$$(1 - 0.5B)^2(1 - B^4)Y_t = (1 - 1.89B + 0.99B^2)(1 - 0.82B^4)e_t \tag{2}$$

with $\sigma_e^2 \approx 0.052$.

### 3.2.1 Overfitting the Model

Though we started with a model with many terms and reduced it to a more parsimonious model, I will still aim to find any room for improvement in this current model via overfitting. Changing the number of non-seasonal and seasonal terms in the models yielded models with higher AIC's and insignificant terms.

Interestingly, if the AR(1) non-seasonal term is removed (though it is significant), the AIC just increases marginally by 0.65, indicating no real argument for a difference in performance by the models. In Section 4 where we will compare the models' forecasting abilities, both of these quadratic trend models will be included. The models' R output can be seen in Figure 15 below. This alternate, reduced model can be written mathematically as:



```
Call:
arima(x = training_ts, order = c(0, 2, 2), seasonal = list(order = c(0, 1, 1),
    period = 4))

Coefficients:
          ma1     ma2     sma1
      -1.4386  0.4628  -0.7914
s.e.   0.2329  0.1765   0.2287

sigma^2 estimated as 0.05659:  log likelihood = -7.94,  aic = 21.89
```

Figure 15: Alternate ARIMA output for the model assuming a quadratic trend.

$$(1 - B)^2(1 - B^4)Y_t = (1 - 1.44B + 0.46B^2)(1 - 0.79B^4)e_t \tag{3}$$

with $\sigma_e^2 \approx 0.057$.

# 4 Forecasting the Models

Now that we have acquired three different models, we shall test their forecasting abilities in this section. For a general picture, I plotted all three models against the full dataset (including the testing data). Figure 16 below illustrates their performance against the actual values in the dashed black line.

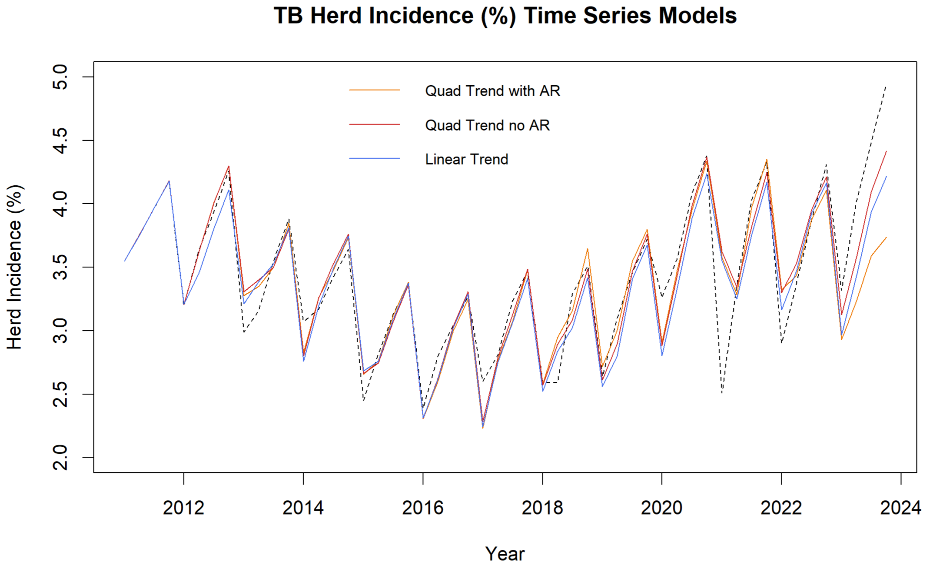


Figure 16: All models plotted against the complete dataset.

From the figure, we can see that the models perform reasonably the same. The models seem to be all reasonably accurate at predicting the Q4 values, however, they are not as consistent with the Q1 values. In particular, 2021 Q1 differs wildly from the models and even perhaps there is a good difference for 2017 Q1 also.

On the predicted values for the models, the model assuming a quadratic trend with no AR terms hits closest to the mark. The model assuming a linear trend is closely behind this model. Interestingly, the quadratic trend model including the significant non-seasonal AR(1) term, predicts values much lower than the other quadratic trend model.

From this figure and the similarity in AICs of both the models assuming a quadratic trend, this project will not report on residual analysis for the quadratic trend model with the non-seasonal AR(1) model. We note that the only difference between the structuring of our two final models now, is the order of the non-seasonal differencing.

# 5 Residual Analysis of the Final Models

Now that we have narrowed down our selection to just two models, we will examine each model's residuals using a variety of techniques such as examining their ACF, using a Q-Q plot, performing a Ljung-Box test and more.

## 5.1 Linear Trend Model

Looking at the linear models residuals vs fitted values in Figure 17, we can see that the residuals seem to have fairly constant variance here with one notable outlier located at the bottom of the plot, corresponding to 2021Q1.

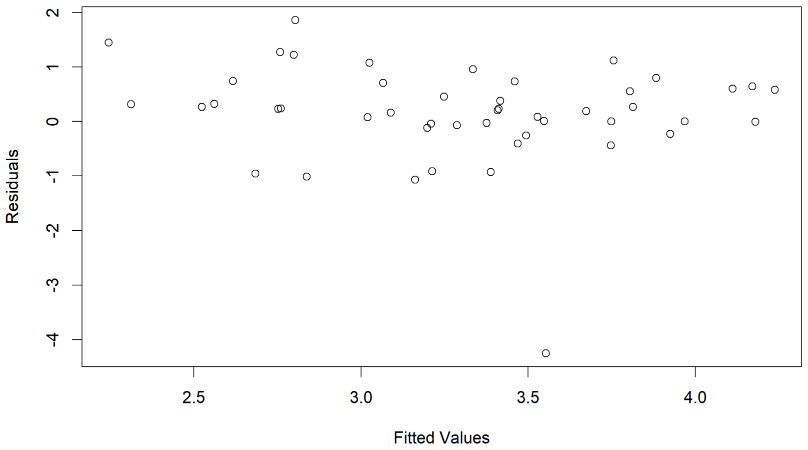


Figure 17: Fitted vs residual values for the linear trend model.

Now examining a Q-Q plot of the residuals, the points do fray from the line towards the end and we again have the one large outlier on the bottom left of the plot.

Further examining the residuals by looking at its histogram, the general shape isn't completely normal, it is fairly narrow and has large drop-offs from the right middle. The outlier at -5 is also apparent here. As expected, a Shapiro-Wilk test rejects the assumption of normality.

Plotting the ACF of the residuals in Figure 20 shows no significant autocorrelations at any of the lags. There may be some tapering at the beginning however this is fairly minor.

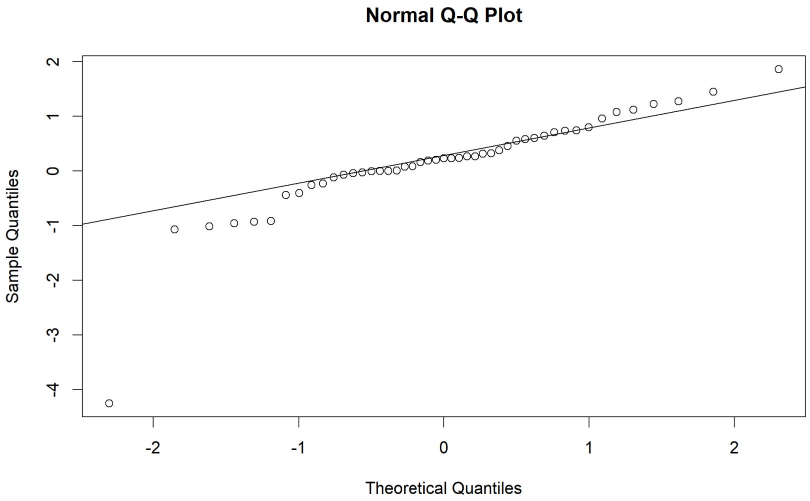


Figure 18: Q-Q Plot for the linear trend model.

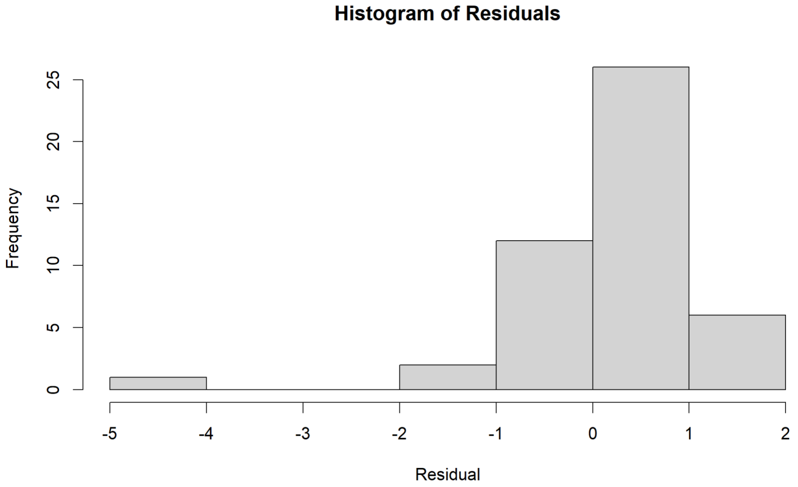


Figure 19: Histogram of residuals for the linear trend model.

Finally, examining the residuals by plotting the different p-values for Ljung-Box tests on a range of lags can be seen in Figure 21. All values here are above the significance line implying there is no autocorrelation in the residuals.

## 5.2 Quadratic Trend Model

Looking at the models residuals vs fitted values in Figure 22, we can see that the residuals display a reasonably similar pattern to the linear trend model. There is more of what seems maybe a funneling pattern found in the plot here with higher value terms being predicted more accurately. This corresponds to what we saw earlier in Section where we noted that the models seemed to better predict the final quarters of the year compared to Q1.

Now examining a Q-Q plot of the residuals, the points fray even more than the linear trends model from the line towards the ends and we again have the one large outlier on the bottom left of the plot.

Further examining the residuals by looking at a histogram in Figure 24, the general shape isn't completely normal, however, it is more evenly distributed than the linear trend model residuals. The

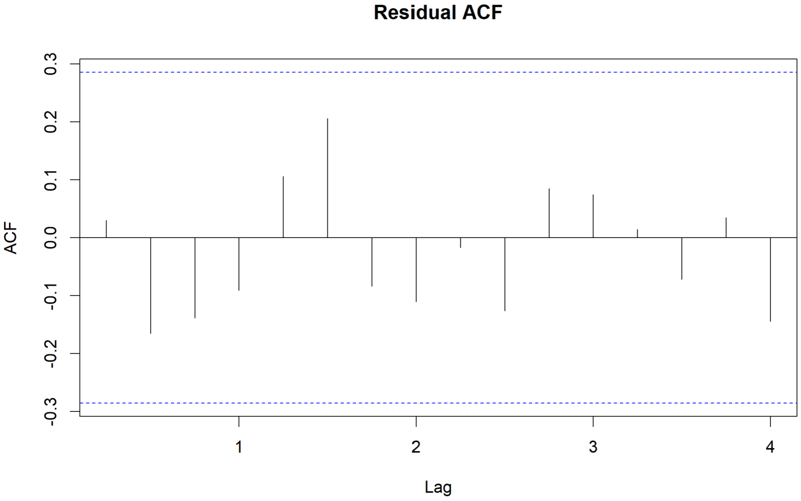


Figure 20: ACF plot of residuals for the linear trend model.

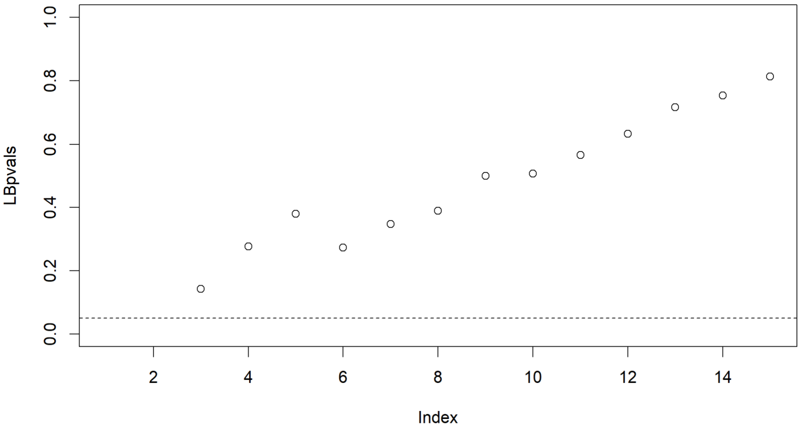


Figure 21: Ljung-Box p-values for residual autocorrelation plotted for the linear trend model.

middle left bar of the main area has a much higher frequency here, indicating this model fits values with more of an equal distribution between those points above and below their actual values. Again, the outlier at -5 is also apparent here. As expected, a Shapiro-Wilk test rejects the assumption of normality.

Plotting the ACF of the residuals in Figure 25 shows no significant autocorrelations at any of the lags. The plot is almost identical to the ACF for the residuals in the linear trend model. This is naturally due to these models having the same parameters bar the order of non-seasonal differencing.

Finally, examining the residuals by plotting the different p-values for Ljung-Box tests on a range of lags yields the following figure. Again, all values here are above the significance line implying there is no autocorrelation in the residuals. Interestingly, p-values here are much further from the significance line here compared to the linear trend model, where they are a lot closer for the lower lags.

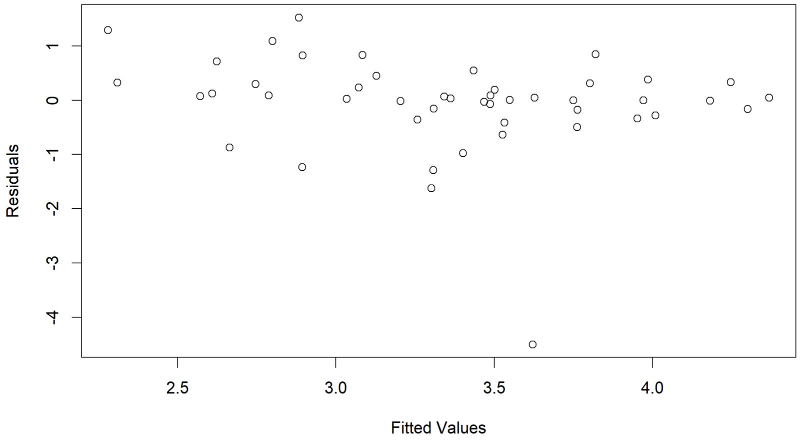


Figure 22: Fitted vs residual values for the quadratic trend model.

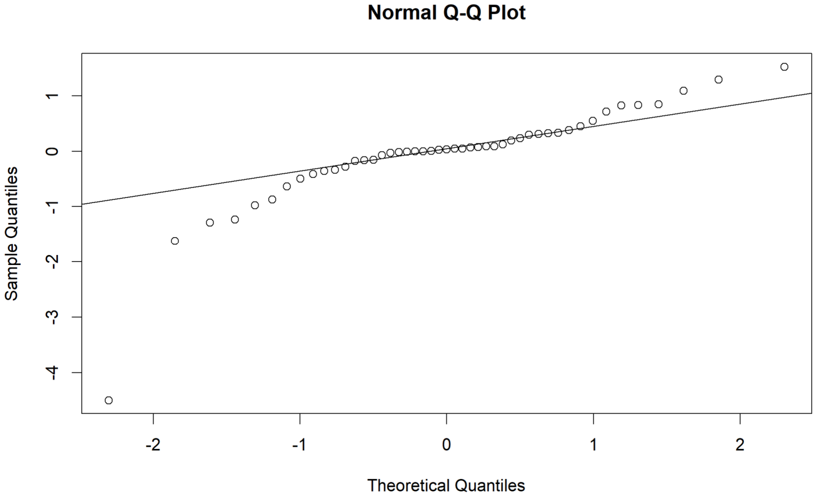


Figure 23: Q-Q Plot for the quadratic trend model.

# 6 Comparing Predictions and Final Model

As we saw in Section 4, the quadratic trend model with no AR term hits closest to the target out of all our models. Comparing the standard error on predictions of the next five quarters in both models, yields the following table.

| **Model** | **22Q4** | **23Q1** | **23Q2** | **23Q3** | **23Q4** |
|---|---|---|---|---|---|
| Linear | 0.25 | 0.28 | 0.31 | 0.34 | 0.40 |
| Quadratic | 0.25 | 0.29 | 0.33 | 0.36 | 0.42 |

Table 3: Comparing the standard errors of both models for their predictions

The standard errors for both models are very similar, indicating similar predictive accuracy.

When it comes to picking which model will be the final model, an argument could be made for both of these final two. Their residuals behave reasonably the same with some minor differences between the two. They have almost identical errors for their predictions and overall they model the existing data quite well. While the quadratic trend model predictions do hit closer to the mark, it is not hugely

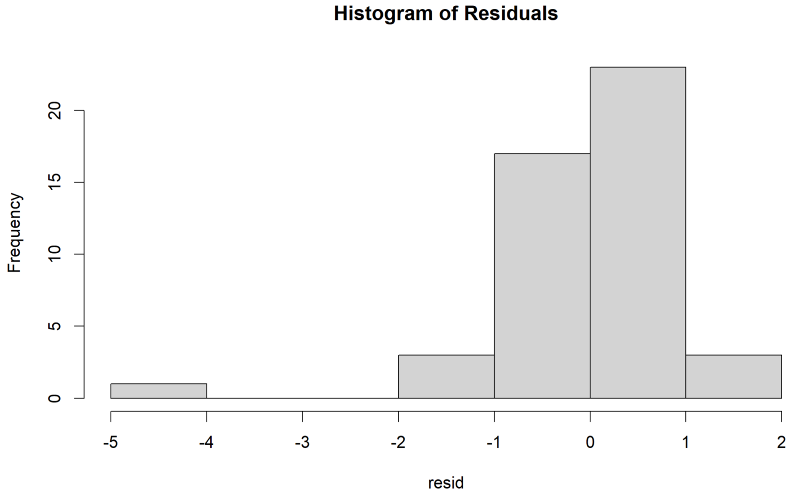


Figure 24: Histogram of residuals for the quadratic trend model.

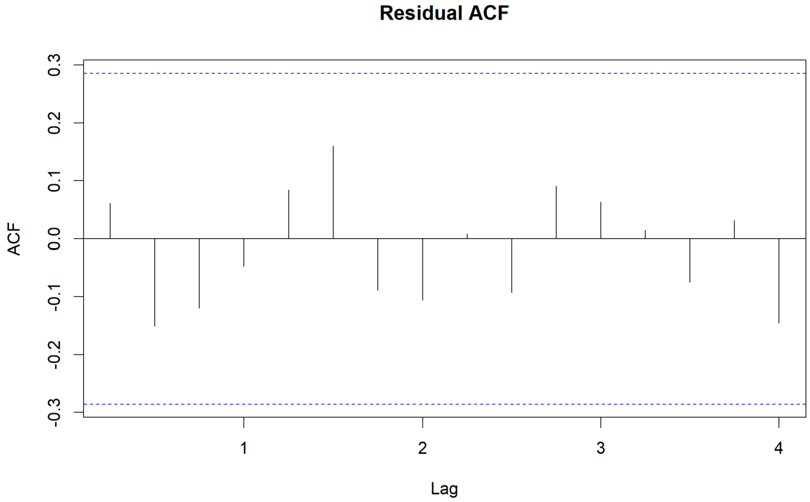


Figure 25: ACF plot of residuals for the quadratic trend model.

more accurate than the linear trend model.

Overall, if the aim of this project was to get the most accurate predictions on the final 10% of the dataset, I would choose the quadratic trend. However, long term, I imagine and would hope that the bTB herd incidence % levels will not continue to increase and will likely begin to come down again. Therefore, for a longer-term model, the single differencing model may be more appropriate. Taking into consideration that the quadratic model only guesses slightly better than the linear, the linear trend model is likely the superior model here and will be considered as my final model.

# 7 Forecasting the Final Model

This section aims to give a more in-depth view of Section 4, where we are now just considering our final model. Figure 27 below shows the predicted values given from our final model with an associated 80% and 95% prediction interval alongside the true values in the dashed line.

The model predicts all values less than their true values. We noted this may happen at the start of this

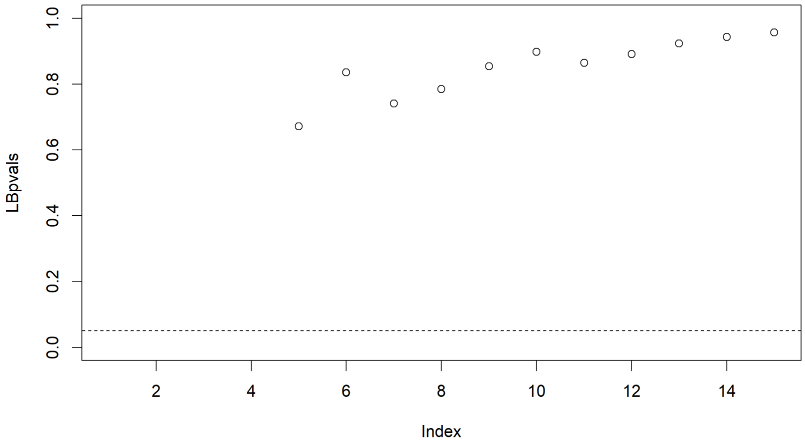


Figure 26: Ljung-Box p-values for residual autocorrelation plotted for the quadratic trend model.

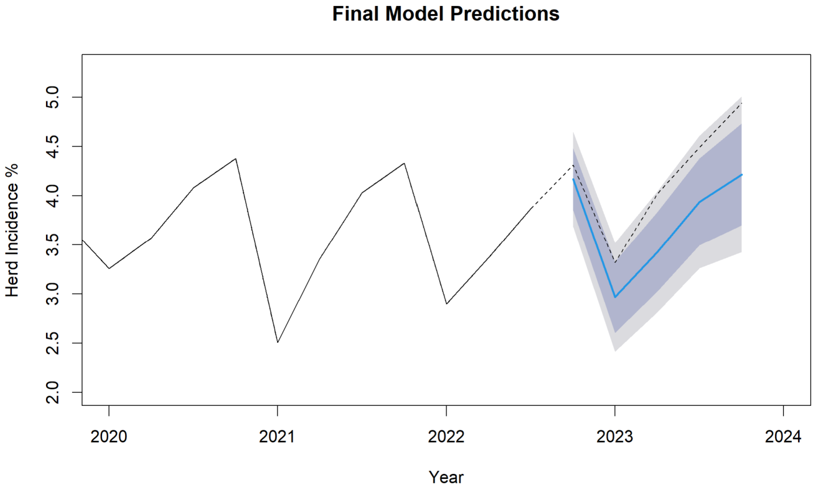


Figure 27: Plot showcasing the predictions of the final model with 80% and 95% prediction intervals.

report due to the incidence rates increasing in recent times. The true values are located just within the 95% prediction interval of our model. Overall this model performs well with reasonable prediction, and may prove better if and when the incidence rates begin to flatten out.

# References

[1] Ireland, Department of Agriculture, Food and the Marine, Central Statistics Office (2023), Bovine Tuberculosis DAQ01, available https://data.cso.ie/

[2] Ireland, Department of Agriculture, Food and the Marine (2024) National Bovine TB Statistics, available: https://www.gov.ie/en/publication/5986c-national-bovine-tb-statistics-2020/
**Note:** This poster notes the 2023Q4 herd incidence % which was not given in the CSO dataset.

[3] More, S.J. 'bTB eradication in Ireland: where to from here?' Ir Vet J 76, 11 (2023). available: https://doi.org/10.1186/s13620-023-00239-8