Analysis of Statistical Models to Predict Attainment of Safe Pathogen Levels in Biosolids Treatments in Jordan

Final Report to the Sustainable Development of Drylands Project

Prepared by John Bear and Dean Billheimer, PhD
Statistical Consulting Laboratory, University of Arizona

September 27, 2010

Summary

In our analysis of wastewater pathogens we looked at statistical models to predict attainment of safe levels. We had data for two pairs of drying beds and a pair of aging ponds in Jordan. We explored models with various weather predictors and time. The most plausible prediction model was one in which log counts of pathogens decrease linearly in time. For the drying beds, in late spring/summer weather, the upper bound of a 99% prediction interval reaches the fecal coliform threshold of 1000 by 40 days. This result indicates that future drying beds should reach fecal coliform counts of less than 1000 by 40 days (under similar starting conditions and weather conditions).

The ‘time series’ nature of the data suggested modeling incremental changes in (log) pathogen load as a function of weather conditions. Unfortunately, we saw little correlation between pathogen reduction and weather conditions. The lack of correlation may be attributable to the relatively homogeneous weather conditions during the study periods. Different weather conditions, for example in a different season, might result in a different association. However, this result suggests that weather data may not be reliable predictors of decreases in pathogen load.

Following Akrum’s earlier work, we also explored the use of cumulative weather data to predict reduction in pathogen load. However, there were two complications to this approach.

1. Poor correlation between incremental changes in pathogens and weather data suggest that ‘cumulative’ weather may not be a reliable predictor.

2. The available weather data contain missing values for two of the drying beds.

We estimated the missing values from weather data from three years with complete records. Statistical models using the imputed weather data fit substantailly worse than those using only attested weather data. This indicates that imputed weather data may not be useful for this modeling task. For these reasons, we are not enthusiastic about using cumulative weather data for modeling decreases in pathogen load.

For the aging ponds we conducted similar analyses. Again total days provided a reliable indicator of pathogen load, over the observed weather conditions. This time though, microbial testing stopped when the microbial counts first attained the safe levels. Our statistical model identified these stopping observations as possible low outliers. Although the observations indicated that fecal coliforms were below 1000, the mean response from the statistical model did not. There is not enough information to know the actual coliform levels for these ponds. This illustrates that statistical modeling confers a benefit for evaluating noisy observations in conjunction with entire data set history.

Introduction

Our goal is to model the decrease in wastewater pathogens in drying beds and aging ponds so as to be able to predict when the pathogen levels have most likely decreased below the safe threshold. Dr. Akrum Tamimi obtained the data, which are for some sites in Jordan.

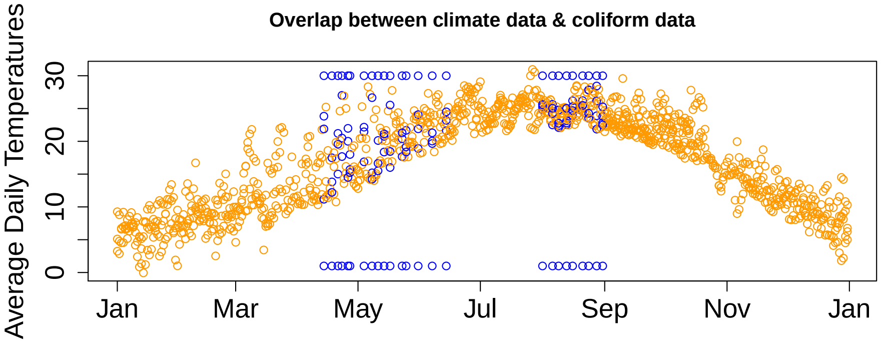

Figure 1: Overlap between climate data & coliform data

Before we begin with the analysis, we show a graph illustrating the parts of the year for which we have data, essentially spring and summer.

The gold circles in the graph above represent average daily temperatures for the years 2006 to 2009 overlaid for each day of the year. Blue circles denote days for which we have recorded data about fecal coliform and salmonella. For visibility, blue circles are also shown along the bottom and top of the graph for those days. We have no data for days when the average temperature is below 10 degrees C.

Analysis with Time (Days) as a Predictor

Fecal Coliform

The data on fecal coliform come from 6 sites, really 3 pairs of sites. Sites db3 and db13 are paired, and have similar rates. Sites db1 and db2 are also paired. In addition, there are some aging ponds which are deeper, and take longer. These are aging33 and aging34. Fecal coliform counts from these experiments are shown in the graph below. Pairs of sites have similar symbols, e.g., squares for db3 and db13, one hollow, and one filled. Similarly, the aging ponds are both indicated by triangles, a filled triangle for aging33, and a hollow triangle for aging34.

The generally accepted standard is that counts below 1000/g are considered safe, so we have drawn a horizontal line corresponding to that level. Since the vertical axis is in units of the log (base 10) of fecal coliform counts, the horizontal line is drawn at log(1000) which is 3.

This next graph, Figure 3, is very similar, but the independent variable is now a measure of cumulative solar radiation. We don't have solar radiation data for drying beds db1 and db2, so there are only lines for db3, db13, aging33, and aging34.

Modeling Fecal Coliform Decay vs. Time

From the previous graphs it is clear that the times for the aging ponds are longer than the times for the drying beds, and that the drying beds seem to perform similarly to one

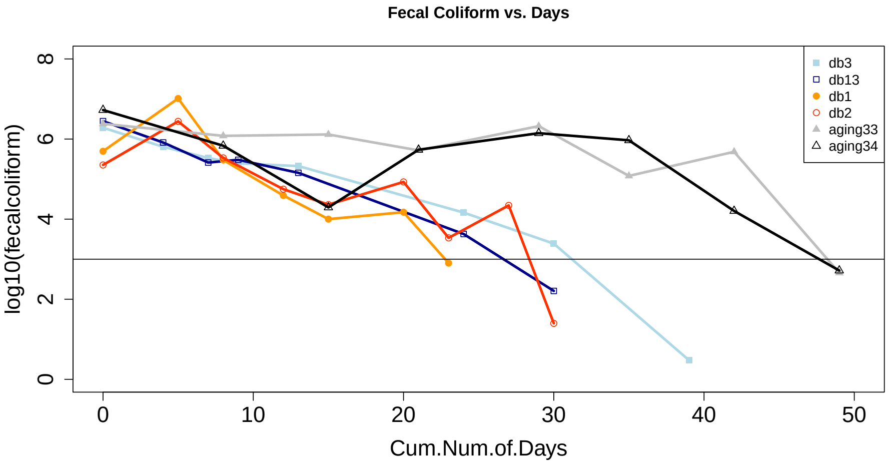

Figure 2: Fecal Coliform vs. Days

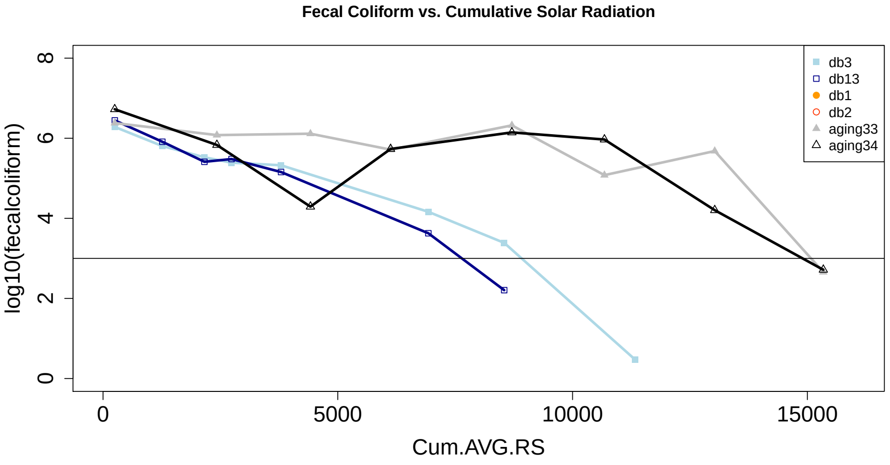

Figure 3: Fecal Coliform vs. Cumulative Solar Radiation

another. So we will start with a regression on the data from the drying beds. We will start with the assumption that the appropriate model for the amount of fecal coliform is one of exponential decay. In subsequent sections we investigate the various weather factors, especially cumulative solar radiation, but for now, a good first step is to regress the log of the fecal coliform against the cumulative number of days.

A model for decay of pathogen load is described as:

$$Y = \alpha e^{-\beta t}$$
$$logY = log(\alpha) - \beta t.$$

Using the data from the four drying beds, and not the aging ponds, we get coefficients, standard errors, and t-statistics using a robust regression method (Tukey's biweight, Ripley, http://www.stats.ox.ac.uk/pub/StatMeth/Robust.pdf). The p-values are less than 10^{-10} and were rounded to 0.

The R^2 value for this model fit is approximately 0.8. Note that we are using cumulative number of days, and the log of the fecal coliform count.

```
                  Value Std. Error  t value
(Intercept)      6.4350     0.2019  31.8718
Cum.Num.of.Days -0.1249     0.0115 -10.8359
```

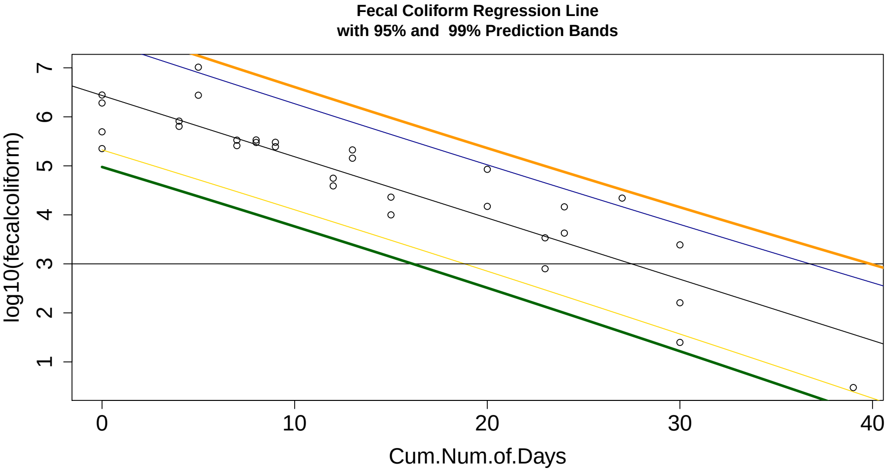

Figure 4: Fecal Coliform Regression Line with 95% and 99% Prediction Bands

In Figure 4 we show a plot of the data with 95% and 99% prediction bands. The outer bands are for the 99% prediction interval. The upper prediction band intersects the horizontal line for $log_{10}(1000)$ at around 40 days. Thus our preliminary results suggest that during the time of year these experiments were conducted, we have 99% confidence that new drying bed experiments will reach less than 1000 fecal coliforms by 40 days.

Modeling Salmonella Decay vs. Time

For Salmonella we repeated the anaysis. Figure 5 is a graph of the salmonella counts for the drying beds and the aging ponds, versus the cumulative number of days. The horizontal line at 0.48 corresponds to a salmonella count of 3.

Using robust regression as before, with method="MM" for Tukey's biweight, we get the results below.

```
                  Value Std. Error t value
(Intercept)      1.1085     0.1252  8.8576
Cum.Num.of.Days -0.0572     0.0113 -5.0609
```

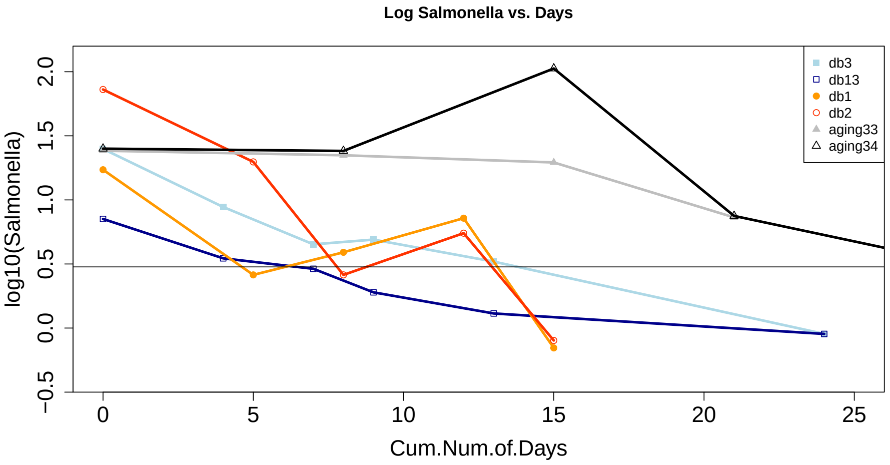

Figure 5: Log Salmonella vs. Days

Here the R^2 value is about 0.6. The graph in Figure 6 shows a plot of the data with the regression line through it, and a horizontal line representing a safety threshold at $log_{10}(3) = 0.477$. The plot also shows 95% and 99% prediction bands. The outer bands are the 99% bands.

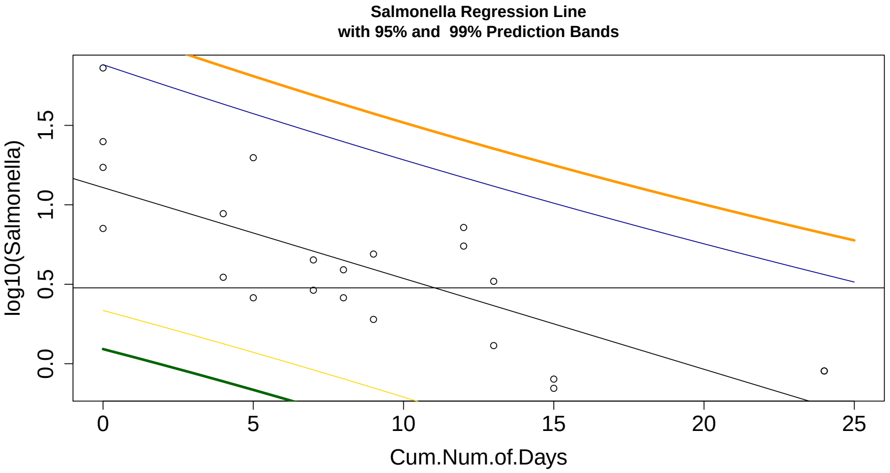

Figure 6: Salmonella Regression Line with 95% and 99% Prediction Bands

Summary of First Analysis

For our initial analysis we have used the data from the drying beds, not the aging ponds. As a first pass, we used the single predictor variable, cumulative number of days. The model was that the amount of fecal coliform or salmonella decays exponentially with passage of time. Since the time for decay of fecal coliform appears longer than for salmonella, this first analysis suggests a time of around 40 days corresponds to 99% confidence that the pathogens have dissipated to the safe threshold. The caveat here is that this is based on only four drying beds observed in the dry time of year. The model says nothing about what would happen at other times of the year, e.g., fall and winter.

Incremental Decay

The second way in which we analyzed the data involved looking at the log of the pathogen counts for the beginnings and endings of the measurement time intervals, and taking the difference. So if a measurement was taken on one day, and then another measurement was taken three days later, we computed the decrease in the logs of the two measurements. To incorporate weather data, we use cumulative values (e.g., cumulative relative humidity), but now those numbers represent the sum of the weather data on the three days between the two measurements.

Fecal Coliform Incremental With Weather

The table below summarizes the ability of weather variables to predict incremental reduction in fecal coliform load. R-squared is one standard way of assessing how well a model fits the data. The AIC column is for the Akaike Information Criterion, which is roughly the log likelihood of the model, minus a penalty based on the number of predictors. For the AIC criterion, smaller values indicate a better model fit. Also for AIC, differences of less than 2 might be attributable to random variation. These weather predictors did not substantially improve a model using days. R^2 for the best fecal coliform predictor in this way, relative humidity, was 0.12. However, the p-value (0.08) suggests that cumulative relative humidity (accumulated over the days between bacterial sampling) is not a strong predictor of incremental decrease in fecal coliform counts.

	Coefficient	P-value	R-squared	AIC
Days	0.11	0.19	0.07	80.33
Tair	0.00	0.38	0.03	81.34
RH	0.00	0.08	0.12	78.82
RS	0.00	0.24	0.05	80.69
Press	0.00	0.19	0.07	80.32
WindSpd	0.04	0.16	0.08	80.02
Rain	0.38	0.26	0.05	80.81

Table 1: Predictors for incremental change in fecal coliform

Next there's a scatterplot for the predictor, days, (Figure 7) followed by a scatter plot for the predictor, relative humidity (Figure 8). Both plots suggest a weak relationship between decreasing counts and day (or relative humidity). Note, however, that the correlation between incremental days and incremental relative humidity is 0.91. Thus, it is unclear that relative humidity adds substantially to the prediction model.

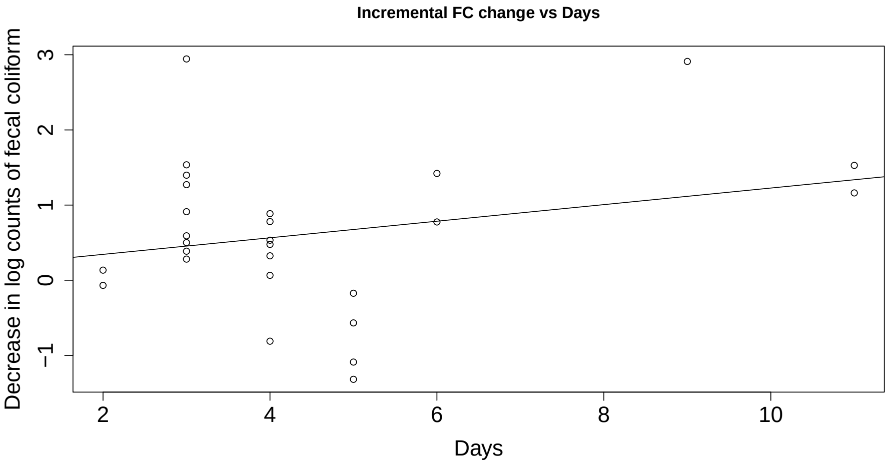

Figure 7: Incremental FC change vs Days

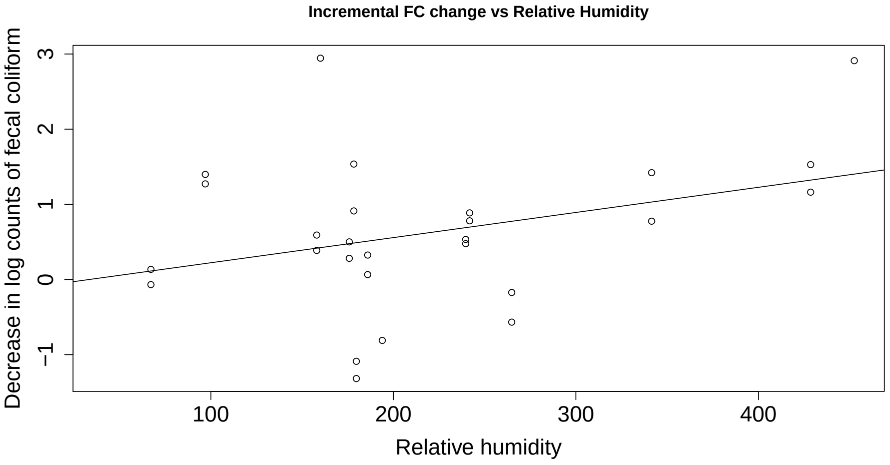

Figure 8: Incremental FC change vs Relative Humidity

Salmonella - Incremental Decay

In the table below we have omitted rain this time because for all the relevant days, the value for rain was 0. None of the models fit well at all. This can be seen both from the table of R^2 values and also the from the scatter plots in Figures 9 and 10. The plots show the decrease in salmonella counts vs. days, and separately, vs. relative humidity.

	Coefficient	P-value	R-squared	AIC
Days	0.01	0.76	0.01	23.04
Tair	0.00	0.59	0.02	22.82
RH	-0.00	0.91	0.00	23.13
RS	0.00	0.66	0.01	22.92
Press	0.00	0.77	0.01	23.04
WindSpd	-0.00	0.75	0.01	23.02

Table 2: Predictors for incremental change in salmonella

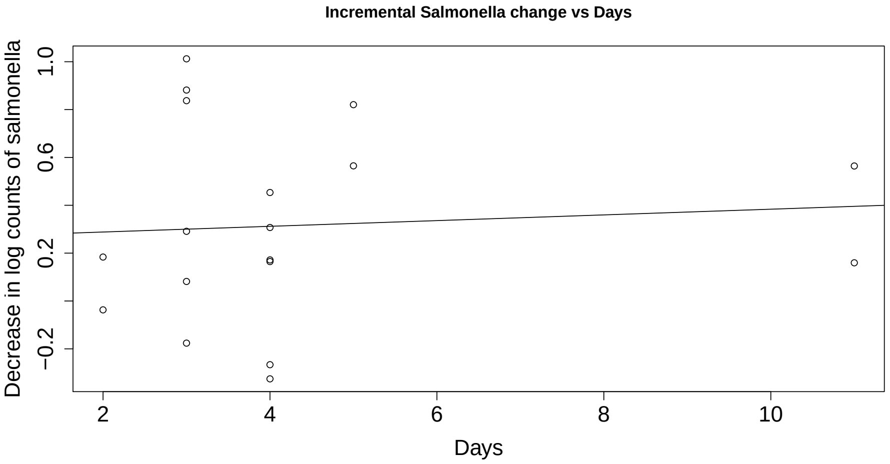

Figure 9: Incremental Salmonella change vs Days

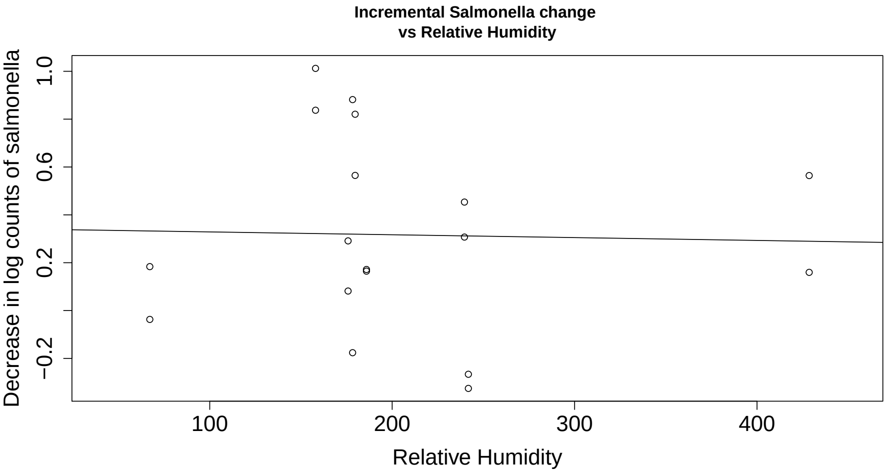

Figure 10: Incremental Salmonella change vs Relative Humidity

Weather Predictors in Exponential Decay Model

Following Akrum's earlier work, we also explored the use of cumulative weather data to predict reduction in pathogen load. However, the results from incremental changes (summarized in the previous section) suggest that weather may not be a reliable predictor. Because cumulative weather data are highly correlated with the number of days, their "apparent" predictive ability is likely artifactual. In statistics jargon, a cumulative weather variable and number of days are said to be "confounded." We include these analyses here for comleteness, but do not consider them reliable.

Imputing Weather Data

Since a great deal of the time, we are missing weather data for the days where we have salmonella and fecal coliform readings, we need to make guesses based on weather from other years. There were two levels of indirection in the imputation. One is that we imputed things like average temperature based on measurements from the same day in other years, but the second is that these guesses are for the weather station's location, but the drying beds are in a different location with potentially different weather.

The procedure was this. Find the range of dates over which we need weather data. For each day in the range, get the data (e.g., temperature or solar radiation) for that day from all available other years. (There were three such years available.) Average the data for a given day. So for example, in all three years get the temperature values for May 3 and average them. Record that as the guess for the temperature for May 3. Do the same thing for relative humidity, solar radiation, pressure, wind speed, and rain.

Fecal Coliform And Salmonella With Weather

First we just modeled the total amount of fecal coliform or salmonella as decaying exponentially with time/days. In this section we discuss seven different models, each where the predictor variable is some weather variable instead of days. To try out the weather data this way we were forced to create some odd predictors, like the integral of relative humidity over time, something we ended up calling cumulative relative humidity. We created similar cumulative predictors for the other weather variables as well. So for instance, cumulative air temperature is roughly the integral over time of the average air temperatures, and cumulative wind speed is the integral over time of the wind speeds. In the end, although cumulative relative humidity ended up being the single best predictor, it was not significantly better than days. The results are shown in Table 3.

	Coefficient	P-value	R-squared	AIC
Days	-0.13	0.00	0.82	64.04
Tair	-0.01	0.00	0.76	73.52
RH	-0.00	0.00	0.82	63.44
RS	-0.00	0.00	0.81	66.26
Press	-0.00	0.00	0.82	64.05
WindSpd	-0.03	0.00	0.81	65.34
Rain	-0.80	0.01	0.24	108.95

Table 3: Weather Predictors for Fecal Coliform

The results for all the single predictor exponential decay models for salmonella using the imputed and attested weather data are shown in Table 4. Once again, the best fitting model has relative humidity as its predictor, though as before, it is not necessarily better than days.

	Coefficient	P-value	R-squared	AIC
Days	-0.06	0.00	0.61	18.00
Tair	-0.00	0.00	0.53	22.28
RH	-0.00	0.00	0.64	16.21
RS	-0.00	0.00	0.60	18.49
Press	-0.00	0.00	0.61	18.01
WindSpd	-0.02	0.00	0.63	17.15
Rain	-0.27	0.42	0.03	38.05

Table 4: Weather Predictors for Salmonella

Aging Ponds

We conducted similar analyses for the aging ponds using cumulative number of days, and accumulated weather data as predictors. We had data for cumulative solar radiation and wind speed, (attested, not imputed), so we used them in models with one predictor, and also used days. Below are tables summarizing the model fit to the data. For fecal coliform, the R^2 are all around 0.49. For Salmonella, the pattern held, except now the R^2 are between .38 and .40.

	Coefficient	P-value	R-squared	AIC
Days	-0.05	0.00	0.48	47.04
RS	-0.00	0.00	0.49	46.55
WindSpd	-0.01	0.00	0.49	46.78

Table 5: Predictors for Fecal Coliform in Aging Ponds

	Coefficient	P-value	R-squared	AIC
Days	-0.03	0.07	0.39	11.32
RS	-0.00	0.07	0.40	11.08
WindSpd	-0.01	0.08	0.38	11.49

Table 6: Predictors for Salmonella in Aging Ponds

Based on the these results we plotted the log pathogen counts against days, with lines in the plots representing 95% and 99% prediction intervals (Figure 11, Figure 12). As before, the 99% prediction interval is the wider one. Because data were not collected beyond 49 days, we cannot estimate how long it takes for the pathogens to reach safe levels (without extrapolation).

Finally, note that the fecal coliform observations from day 49 are quite low with respect to the regression line. Indeed, they nearly exceed the lower bound of the 99% prediction interval, and might be considered outlier observations. Based on these data, it is not clear whether the mean level of fecal coliforms reached less than 1000. It is possible that these two observatons reflect measurement variability, and that the actual pathogen load in the aging ponds remained greater than 1000 at day 49. However, we cannot know. This figure illustrates a risk of relying on a single noisy measurement for decision-making when a longer-term history is available.

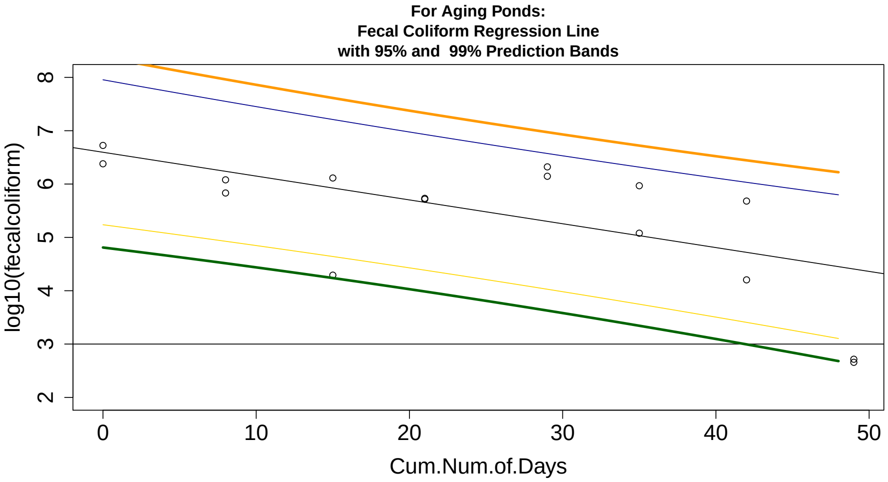

Figure 11: For Aging Ponds: Fecal Coliform Regression Line with 95% and 99% Prediction Bands

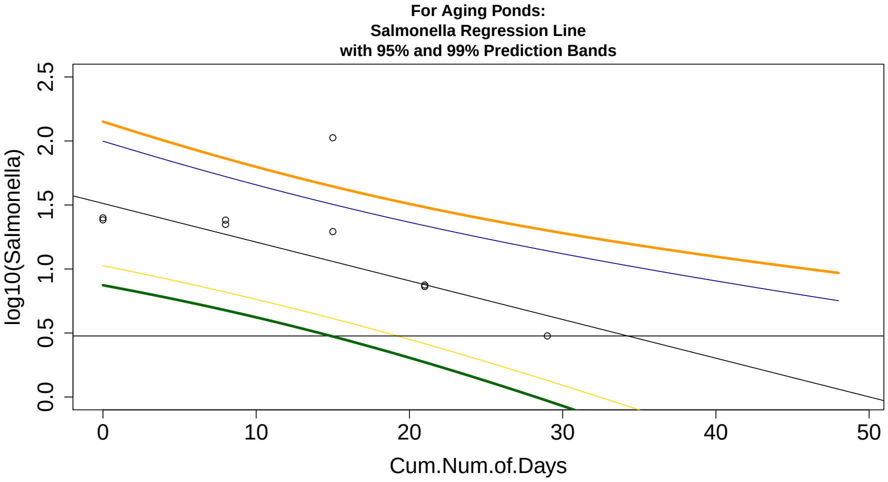

Figure 12: For Aging Ponds: Salmonella Regression Line with 95% and 99% Prediction Bands

Conclusion

Based on these data we believe that number of days provides the best predictive model of pathogen load. Because of the homogeneity of the observed weather data, we don't want to extrapolate this model to other parts of the year where weather conditions are different.