Final Report: Vibrio *spp* in Chesapeake Bay
Mariana I. Fernández Correa

AI-Disclosure: GPT 5.1 and 5.2 were used to help translate the methods into an R script, but all statistical methods are my own, based on what we learned in class. GPT was used throughout to debug and refine R scripts for figures and to clean up the math formatting in this report.

Abstract

Vibrio spp. is a flesh-eating bacterium that is contracted from either eating raw seafood, mainly oysters, or from swimming in water with open wounds. In the US, an area of concern for Vibrio infections is the Chesapeake Bay, where warm water and lower salinity create favorable conditions for higher concentrations of Vibrio. In this report, I use data from (Urquhart et al., 2014) to add a spatial component via SAR and CAR, and as a covariate in Bernoulli GLMs. I also explore different methods of adding a time component to create a few models and compare their performance in fit and prediction tests. Ultimately, a binomial GLM with a time lag and a spatial lag performed best in terms of fit and prediction. This spatial and time-lagged GLM used distances only over water and a 2-month influence over a 50km radius.

Introduction:

The Chesapeake Bay, located between Virginia, Delaware, and Maryland, serves as an economic hub, a recreational destination, and an ecological asset to the mid-Atlantic region. However, instances of *Vibrio spp,* a flesh-eating family of bacteria, in Chesapeake Bay have nearly doubled in the past decade (Maryland Department of Health, 2013). The paper I based this analysis on (Urquhart et al., 2014) develops a statistical model relating seasonality, temperature, and salinity to Vibrio *vulnificus*. There is strong evidence that seasonality plays an important part in detecting a Vibrio outbreak. Vibrio tends to thrive in water above 15 °C and between 5 and 25 practical salinity units (Urquhart et al., 2014).

Over the past 30 years, water temperatures in the Chesapeake Bay have increased by approximately 0.3 °C, extending the warm season that promotes Vibrio growth. In contrast to the open ocean, which typically has a salinity of about 35 Practical Salinity Units (PSU), the Chesapeake Bay maintains lower salinity levels due to freshwater input from tributaries such as the Susquehanna River (Urquhart et al., 2014). Salinity within the bay is highly variable in space and time due to depth and proximity to freshwater tributaries (Chesapeake Bay Program, n.d).

About the Data

The dataset was collected from July 2011 until September 2011, and again in 2012, from March until June 2012, and recorded instances of two Vibrio species, V. *vulnificus* (VV), and V. *parahaemolyticus* (VP). Both species of Vibrio are harmful to humans, but VV is much more hazardous to humans than VP. VV and VP infections usually present with similar symptoms, including cramps, nausea, fever, diarrhea, and skin infections. Usually, VV skin infections can be more severe (Leighton, R.E., 2023). The “Optimal Salinity” variable is a calculated parameter representing the absolute difference between the measured salinity and the optimal salinity for VP growth (11.5 PSU). Table 1 contains all variable descriptions.

Table 1: Variable Description - Adapted from variable description assignment

Variable	Units / description	Min	Max	Mean	Std
Lat	Degrees	37.9101	39.5479	38.8	0.4
Long	Degrees	-76.502	-76.008	-76.3	0.1
Cruise	Integer	1	6	3.6	1.6
Month	Integer	3	9	5.7	1.8
VV	Binary	0	1		
VP	Binary	0	1		
Temperature	C	8	30.95	20.3	6.6
Salinity	Practical Salinity Units (PSU)	0.1	13.9	8.0	3.8
Water Depth	Meters sources from R package *marmap*	0.22	27.544	12.6	6.75
Optimal Salinity	SalnOpt = 11.5 - Salinity	0	11.4	3.8	3.5

Summary of previous paper methods

The paper I gathered the data from used three models: 2 Generalized Linear Models (GLMs) for each of the months and one Generalized Additive Model (GAM). Dr. Urquhart's models can be summarized in Table 2.

Table 2: Summary of models and training data from XX

Model	Formula	Description
NOAA_GLM	$\beta_0 + \beta_1 Temp + \beta_3 \lvert SalnOpt \rvert$,	Trained using a different dataset (n=235) with only *V. vulnifincus* (VV) and collected between 2007 and 2008.
JHU_GLM	$\beta_0 + \beta_1 Temp + \beta_3 \lvert SalnOpt \rvert$	Trained using the same dataset as the one here
JHU_GAM	$\beta_0 + s_1(Temp) + s_2(Saln)$	Trained using the same dataset used here

Similarities and differences between the original models and the report models

All models in Dr. Urquhart's paper used a logit link function, and the paper focused on the probability of detecting VV rather than VP at the bay sampling locations. The models described in my report rely on whether the model correctly fits the covariates to either detect or fail to detect *Vibrio spp.*, so instead of using a probability, I focus only on the corresponding binary response. From what I gathered, their models accounted for months using the temperature and salinity ranges for each month. In all my models, I focus on the presence or absence of either VV or VP at a specific location, rather than only VV as in the 2014 paper.

Methods

I used R Studio for analysis, and a list of packages used is below in Table 3.

Table 3: Packages used for the Rscript.

Package Name	Description
Readxl	Reading Excel file
Dplyr	Manipulating the formed data frame
Ggplot2	Plotting figures
Maps	Boundary of polygons for ggplot2 figures
spdep	Used for creating the neighborhood structures
Spatialreg	Spatial and Conditional Autoregressive models
car	VIF
marmap	Access to NOAA Bathymetry and division of land and water
raster	Helps working with the grids in the "over water only" distances
gdistance	Used to help compute the distance overwater
sp	Used for projection settings and specifications
Rsample	Used for hold-out testing

Variable sorting

I decided to remove Salinity and keep only Optimal salinity in all my models because of high (Variable Inflation Factor) VIFs in early model iterations. Also, since Optimal Salinity was derived solely by subtracting the Optimal Salinity of VP, there was sufficient information to remove one of the salinity variables confidently. Example table of VIF values below VIFF Values for early iterations of the model.

Table 4: Early iterations of the VIF show that dropping one of the salinities is recommended.

Temp	SalnOpt	Month	Saln
1.178648e+01	7.724541e+09	1.179246e+01	7.724541e+09

Time Lag.

I approached two different methods for designing a relation to time. The only variable that provided time was the months category, and, given how the data is collected, it's possible to create relationships with previous months. The first iteration of including space used a sine and cosine decomposition of the months (Stolwijk, 1999). The idea being that months (labeled 1-12) as a regular covariate would simplify the cyclical nature of months to a linear number scale. So, December would be taken as the furthest month away from January, when in reality these months are much closer together. Breaking the months down into Sine and Cosine sections could capture the seasonality, so January and December are ideally close together. Determining these transformations is described below:

$$\text{MonthSin} = \sin\left(\frac{2\pi\,\text{Month}}{12}\right)$$
$$\text{MonthCos} = \cos\left(\frac{2\pi\,\text{Month}}{12}\right)$$

These two modifications would be added as two covariates, replacing the month covariate. However, after further consideration of the dataset and its collection, this approach could be improved. The dataset was collected in two periods: months 7-9 in 2011 (July to September) and months 3-6 in 2012 (March to June). Beyond the month, the better the association is looking at the Vibrio prevalence in the past. In short, the month itself should not have as strong a correlation as the prevalence or lack thereof of Vibrio in the month(s) prior. This is very similar to the approach taken in the original paper, where months play a role when associated with temperatures and salinity in the bay, not by themselves (not included as a covariate).

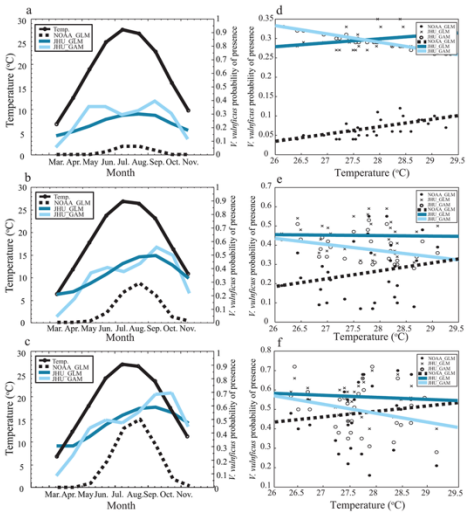

Figure 1: Representation of how months and time were captured in Dr. Urquhart's 2014 paper.

Literature research indicates that Vibrio infections are most common during warm months (CDC, 2024). This corresponds to the data-collection months. It also mentions that Vibrio presence usually lasts a few weeks, which is a time resolution finer than my dataset can resolve. However, I still want to experiment with time lags as I have never done so before. Given that the residence time of the Chesapeake Bay is about 6 months, it's still worth examining whether a spatial lag component is relevant to this dataset.

I have chosen to include a 50 km radius of influence and a mostly arbitrary time lag of the previous 2 months. Making the time lag structure as follows. *Note: I used GPT 5.1/2 to help implement this.*

$$t_i = \text{month index of observation } i$$
$$\Delta t_{ij} = t_i - t_j$$
$$N_i^{(t)} = \{j : 0 < \Delta t_{ij} \leq 2\}$$
$$w_{ij}^{(t)} = \frac{1}{\Delta t_{ij} + 1}$$

$$\text{TempLag}_{\text{i}} = \frac{\sum_{j \in N_i^{(t)}} w_{ij}^{(t)} y_j}{\sum_{j \in N_i^{(t)}} w_{ij}^{(t)}}$$

$$\text{TempLag}_{\text{i}}^{(\text{z})} = \frac{\text{TempLag}_{\text{i}} - \mu_i}{\sigma_i}$$

To incorporate temporal dependence, I created a temporal lag variable based on observations from the previous two months. Each observation includes a month as a covariate. I chose to calculate the 2-month time lag over a 50 km distance.

The two-month lag considers measurements within the same month, one month prior, and two months prior that were part of the neighborhood (within a 50 km water distance) for the data point. Measurement of time differences was defined as indicated. Temporal neighbors are the observations occurring within the prior two months of the reference observation.

Temporal weights were assigned using an inverse-lag structure, so that more recent observations had a greater influence than older ones. The temporal lag at each observation was calculated as a weighted average of neighboring responses, then standardized using z-score normalization. The raw monthly measurements are shown below.

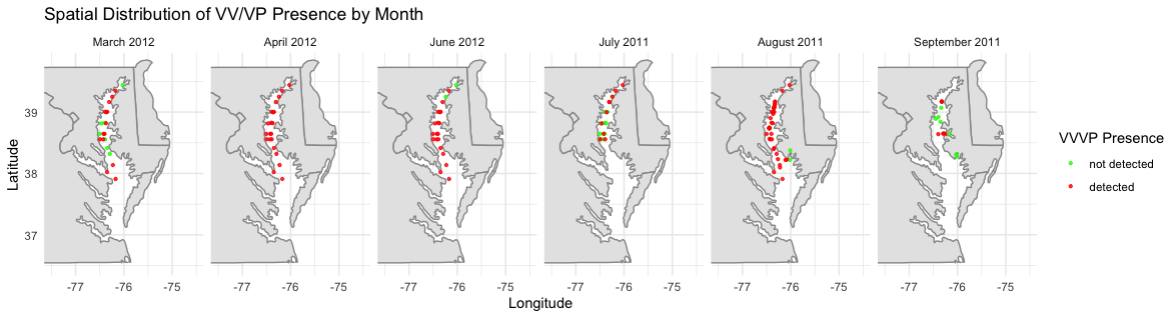

Figure 2: Monthly samples of Vibrio presence. Repeat sample locations are visible.

Defining "water-only" distances – I used GPT 5.1/2 to help brainstorm and implement a plan for this.

I first presented plots using Euclidean distances as their spatial distance structure. This was the same approach I used in the final exam:

$$w_{ij} = \left\{ \frac{1}{{d_{ij}}^2} \;\; if \; d_{ij} \leq r \; and \; i \neq j \; ; \;\; 0 \;\; otherwise \right\}$$

However, as mentioned in class, one of my goals was to define the distance structure, acknowledging that the dots that are close together but also have land between them would not be as close in VV or VP concentrations as one might think. So, for the Bernoulli GLMs

described below, I used water-only distances. I used NOAA's bathymetry R package *marmap*. Bathymetry values greater than 0 were removed (leaving only water). The resolution of this raster is one arcminute (roughly 1.45 km x 1.45 km, OpenDEMinfo, n.d.). Then the raster is given a "Value" of 1 for water and NA for land based on bathymetry "B(s)" at each grid "s".

$$V(s) = \begin{cases} 1, B(s) < 0 \text{ (water)} \\ NA, B(s) \geq 0 \text{ (land)} \end{cases}$$

$$g_{pq} = \frac{1}{\frac{C_p + C_q}{2}}$$

$$g_{pq} = \begin{cases} 1, if\ V(p) = 1\ and\ V(q) = 1 \\ NA, if\ V(p) = NA\ or\ V(q) = NA \end{cases}$$

$$\ell_{pq} = distance\ between\ centers$$

$$cost_{pq} = \frac{\ell_{pq}}{g_{pq}}$$

$$d_{ij}^{\text{water}} = \min_{\gamma \in \Gamma_{ij}} \sum_{(p,q) \in \gamma} \text{cost}_{pq}$$

Then I define a conductance (g_{pq}) across all eight cells neighboring the starting cell. However, corner cells are farther away (diagonal distance is $\sqrt{2}$) than cells that share an edge (distance 1), unlike corner cells. So, a cost for water grid cells ($g_{pq} = 1$) is assigned based on length. The minimum distance is defined by evaluating each path "γ" within the set of all possible paths, "Γ_{ij}". Effectively leaving the shortest distance overwater. Then I defined the weights (Bivand, 2013).

$$w_{ij} = \frac{1}{d_{ij}^{water}}$$

$$\widetilde{w_{ij}} = \frac{w_{ij} - \mu_i}{\sigma_i}$$

$$\text{SpatLag}_i = \sum_{j \in N_i} \widetilde{w_{ij}}\ y_j$$

I assigned spatial weights using an inverse-distance function and then normalized these weights using z-score standardization. The spatial lag at location i was computed as a weighted combination of responses from neighboring locations $j \in N_i$, where N_i denotes the set of neighbors of location i, $y_j$is the response observed at neighbor j, and $\widetilde{w}_{ij}$is the standardized spatial weight linking neighbor j to the location i. Because some sampling locations were repeated, the distance matrix was filtered to retain a single unique distance structure for each pair of locations $10^{-6} < d_{ij}^{\text{water}} < 70$ km. The small offset made the code not break. All this work results in paths shown in Figure 3.

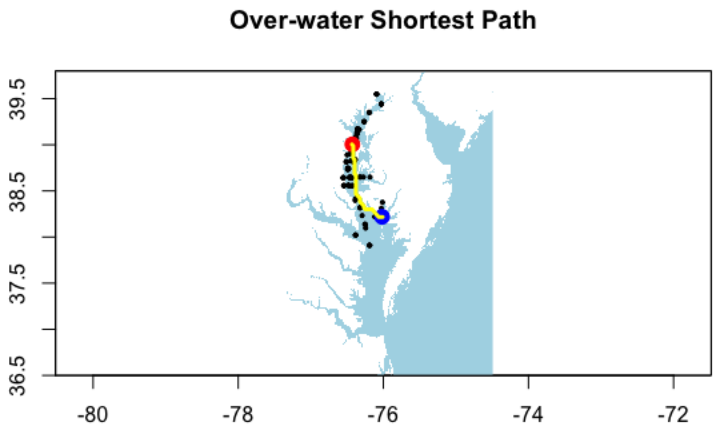

Figure 3: Cost-based distance approach path for sample points.

Moving away from SAR and CAR models

As presented in class, I tried a Spatial Autoregressive (SAR) & Conditional Autoregressive (CAR) model. However, I made a mistake of assuming that the model would output the same response variable type as the input (Binary). The output of both models was normally distributed. I tried to fix my model output using Markov Chain Monte Carlo (MCMC) using the "spatialprobit" library. However, after many iterations, the model did not converge. Although unsure, I believe this was due to a small dataset after duplicate locations were removed (n=59).

Instead, although not very robust, I chose to fit the normalized SAR and CAR outputs using min-max normalization (0-1), classifying values above 0.5 as 1's and below 0.5 as 0's. At this point, I wanted to practice and understand how a SAR or CAR model could work. I didn't quite try that in the exam, but I was already seeing there wasn't a clear benefit, so I didn't improve these models any further. The way the distance matrix is used is also outdated, only using Euclidean distance. The math is as follows:

CAR model:

$$y_i = x_i^\top \beta + u_i,\ u_i = \lambda \sum_j w_{ij} u_j + \varepsilon_i$$

$$y \sim \mathcal{N}(\mu, \sigma)$$

$$z = \frac{y - y_{min}}{y_{max} - y_{min}}$$

$$\tilde{y} = \{1 \text{ if } z > 0.5,\ 0 \text{ if } z \leq 0.5$$

$$w_{ij} = \begin{cases} \frac{1}{d_{ij}} \text{ if } 0 < d_{ij} \leq 400km,\ 0 \text{ otherwise} \end{cases}$$

β = regression coefficients
u = spatial error component
W = spatial weights matrix

$\lambda =$ spatial autocorrelation parameter
$\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$

SAR Model

$$y_i = \rho \sum_j w_{ij} y_j + x_i^\top \beta + \varepsilon_i$$

$$y \sim \mathcal{N}(\mu, \sigma)$$

$$z = \frac{y - y_{min}}{y_{max} - y_{min}}$$

$$\tilde{y} = \{1 \text{ if } z > 0.5,\ 0 \text{ if } z \leq 0.5$$

$$w_{ij} = \begin{cases} \frac{1}{d_{ij}} \text{ if } 0 < d_{ij} \leq 400km, \ 0 \text{ otherwise} \end{cases}$$

$\beta =$ regression coefficients
$u =$ spatial error component
$W =$ spatial weights matrix
$\rho =$ spatial autoregressive parameter
$\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$

For these two models, my covariates were Optimal Salinity, Temperature, and Month, transformed by sin and cosine to make them cyclical. Because of the complications and complexity of forcing binomial outputs on SAR and CAR models, I have decided to explore binomial models instead.

Bernoulli GLM models:

I decided to continue the project using Bernoulli Generalized Linear models. Once defined, I used these for holdout tests and AIC using a logit link. The general form of the Generalized linear model is shown below.

$$\text{VVVPpresent}_i \sim \text{Bernoulli}(\pi_i) \quad , \epsilon_i \sim N(0, \sigma^2)$$

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_n + \ldots . \beta_{n-1} + \epsilon_i$$

I chose four iterations of the Bernoulli GLM in addition to the rudimentary SAR and CAR models, all differenced by their covariates. A comparison table is shown below.

Table 5: Covariates (all normalized) used in each of the models.

Model	Temperature	Optimal Salinity	Depth	Spatial Lag (Water-only)	Spatial Lag (Linear distance)	Time Lag (sine and cosine decomposition)	Time Lag (using a distance of influence)
Bernoulli GLM (BGLM)	x	x	x				
Spatial Bernoulli GLM (SBGLM)	x	x	x	x			
Time lagged Bernoulli (TBGLM)	x	x	x				x

Spatial and Time lagged Bernoulli (STBGLM)	x	x	x	x			x
SAR	x	x	x		x	x	
CAR	x	x	x		x	x	

holdout test, T-Test, and AIC

To create a holdout test, I followed the steps of the SabMann file from class. I split the data 80/20 between the training and prediction sets and repeated the split 50 times using only the Bernoulli models. To classify the predictions, as I did with the SAR and CAR models, I used a 0.5 threshold. Predictions under 0.5 were considered 0s, and above were considered 1s.

For the T-test, I still followed the SabMann file from class and compared the Mean Absolute Error (MAE) between the two using the command. t.test(mae(x),mae(y)) from the "stats" package. The Akaike Information Criterion (AIC) was also calculated for all models, using the same commands as throughout the course.

Results:

Best fit model

The GLMs, as expected, outperformed the SAR and CAR models. A visual description is shown in Figure XX. As there are repeat sampling locations, the bar chart in Figure XX can much more clearly show that the best fit is the model with the most covariates, the spatially and time-lagged model. Figures 4 and 5 show the difference between

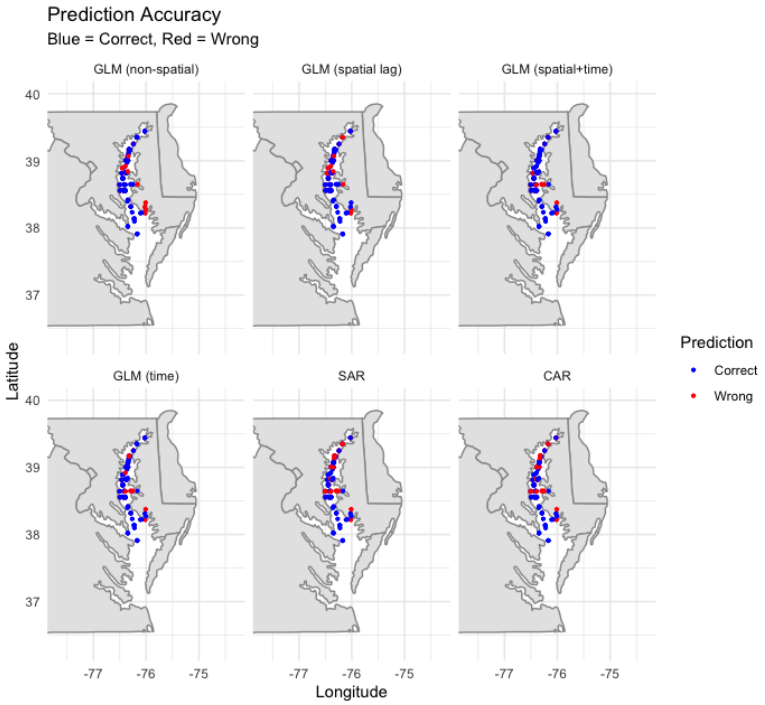

Figure 4: Map visual of model performance, because of repeat locations counting the dots is not an accurate representation of model performance.

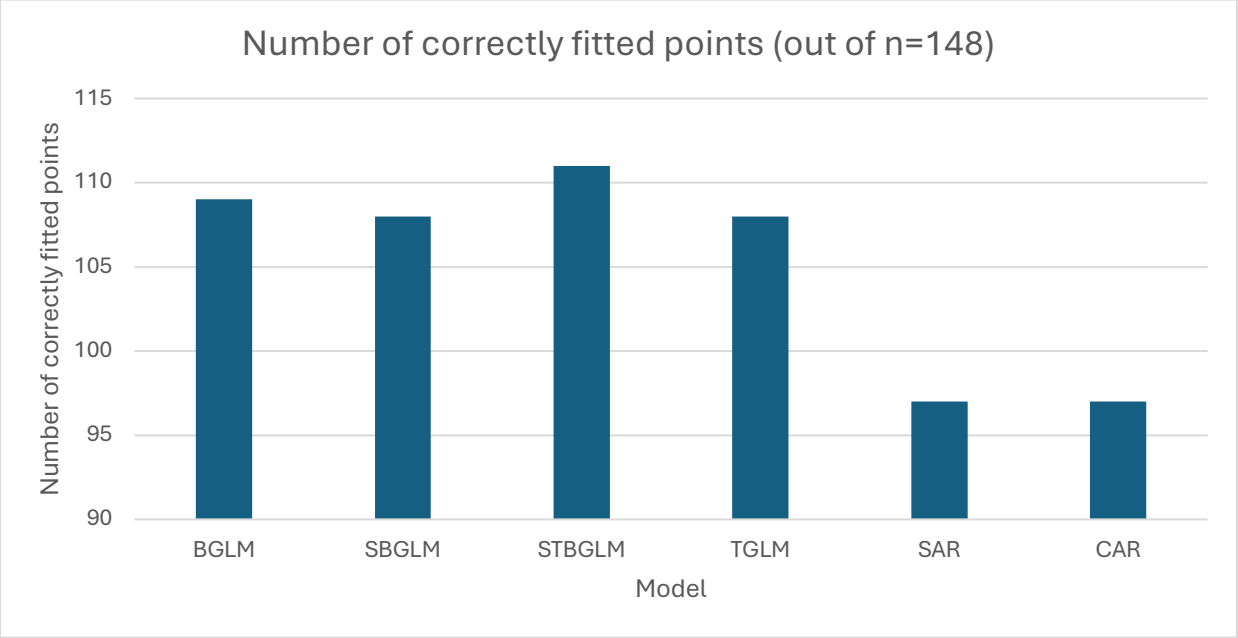

Figure 5: Model comparison of correctly fitted points.

Prediction results: Holdout Tests and T-Test

None of the models tested is particularly good at predicting outcomes. Differences between models are statistically significant; however, there isn't a large distinction between models. The best model for predicting VV or VP detection is once again the Spatially lagged and time-lagged model, but using only time is a close second.

Table 6: Mean Absolute Error and T-test with Bonferroni P value correction

Model 1	Model 2	Mean MAE (M1)	Mean MAE (M2)	t	p value	Best
BGLM	STBGLM	0.365	**0.300**	16.30	**1.19e-20**	STBGLM
BGLM	TBGLM	0.365	0.318	14.56	**1.19e-18**	TBGLM
SBGLM	STBGLM	0.348	**0.300**	13.45	**2.72e-17**	STBGLM
STBGLM	TBGLM	**0.300**	0.318	−8.55	**1.66e-10**	STBGLM
BGLM	SBGLM	0.365	0.348	7.64	**4.10e-09**	SBGLM
SBGLM	TBGLM	0.348	0.318	7.55	**5.70e-09**	TBGLM

AIC – Model fit vs. Model Complexity

My model does not have many covariates that are any more difficult to find than the detection of *Vibrio*, so I have chosen only to use AIC for reference; however, it's still good to note that the most complex model had the lowest AIC, meaning it had the best balance between complexity and model fit.

Table 7: AIC Comparison across models

Model	AIC	ΔAIC from the best
STBGLM	115.34	0.00
TBGLM	117.03	1.69
SBGLM	131.81	16.47
BGLM	133.65	18.31

Moran's I against a non-spatial BGLM.
None of the models showed a statistically significant spatial structure explained by the spatial variable I used; however, models that included a spatial variable outperformed those that did not. Values shown in Figure 5. Moran's I value is shown in Table 8.

Table 8: Moran's I test for all Bernoulli Models.

Model	Moran's I	p-value
BGLM	0.003607	0.2714
SBGLM	0.003355	0.2755
TBGLM	0.003613	0.2712
STBGLM	0.007141	0.2166

Discussion and conclusion:

My goal with this dataset was to work with data that did not follow census tracts, water-based data, and, lastly, to try some of the analyses discussed in class that I did not try in homework or the final exam. I was successfully able to do all of this, including the extra analysis I wanted to run from my checklist in the final presentation. I worked with data that did not align with geopolitical boundaries, which was important because I would like to work with water bodies in the future. I also learned at least one more way of working with data where Euclidean distances do not tell the whole story, like the river networks paper we saw in class.

Working with SAR and CAR models was a new experience and an excellent opportunity to learn about the differences between them (which, in performance terms, seemed minimal), as I was pretty confused before performing the analysis myself. In the future, it would be helpful to see whether the MCMC could converge, or whether there was another way I could have created a binomial response from the SAR and CAR models.

My goal with this analysis was to detect the presence of Vibrio in the Chesapeake Bay under specific conditions. Initially, I was only planning to base it on model fit, but what's probably most useful is prediction, as this could help prevent public health vibrio outbreaks from swimming in the Chesapeake Bay or consuming raw oysters in the region.

According to the literature, Vibrio outbreaks last in a region for anywhere from a few days to a couple of weeks. However, the analysis still showed clear evidence that the spatial lag had a significant effect.

References:

Maryland Department of Health. (2013). *Cases of selected notifiable conditions reported in Maryland*. http://phpa.dhmh.maryland.gov/SitePages/disease-conditions-count-rates.aspx

Urquhart, E. A., Zaitchik, B. F., Waugh, D. W., Guikema, S. D., & Del Castillo, C. E. (2014). *Uncertainty in model predictions of Vibrio vulnificus response to climate variability and change: A Chesapeake Bay case study*. PLOS ONE, 9(5), e98256. https://doi.org/10.1371/journal.pone.0098256

Chesapeake Bay Program. (n.d.). *Physical characteristics of the Chesapeake Bay: Salinity and tributaries.* Retrieved https://www.chesapeakebay.net/discover/ecosystem/physical-characteristics

Leighton, R. E., Correa Vélez, K. E., Xiong, L., Creech, A. G., Amirichetty, K. P., Anderson, G. K., Cai, G., Norman, R. S., & Decho, A. W. (2023). *Vibrio parahaemolyticus* and *Vibrio vulnificus in vitro* colonization on plastics influenced by temperature and strain variability. *Frontiers in microbiology*, *13*, 1099502. https://doi.org/10.3389/fmicb.2022.1099502

Stolwijk, A. M., Straatman, H., & Zielhuis, G. A. (1999). Studying seasonality by using sine and cosine functions in regression analysis. *Journal of epidemiology and community health*, *53*(4), 235–238. https://doi.org/10.1136/jech.53.4.235

Centers for Disease Control and Prevention. (2024). *Vibrio (vibriosis)*. U.S. Department of Health & Human Services. https://www.cdc.gov/vibrio/about/index.html

OpenDem.info. (n.d.). *Arc-second to meters conversion*. Retrieved December 14, 2025, from https://www.opendem.info/arc2meters.html

Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2017). *Applied spatial data analysis with R* (J. Stat. Softw., 76(13)). Journal of Statistical Software. https://www.jstatsoft.org/article/view/v076i13