2019

Statistical Modeling of Influenza-Like-Illness in Montana using Spatial and Temporal Methods

Benjamin A. Stark
University of Montana, Missoula

Statistical Modeling of Influenza-Like-Illness in Montana using Spatial and Temporal Methods

By

Benjamin August Stark

B.A., Mathematics, Statistics option. University of Montana. Missoula, Montana.

Professional Paper

Presented for fulfillment of the degree of Master of Arts in Mathematics, Statistics option.

University of Montana
Missoula, MT.

May 2019

Reviewed By:

Dr. Jonathan Graham
Mathematical Sciences

Dr. Erin Landguth
Division of Biological Sciences

Dr. David Patterson
Mathematical Sciences

Abstract

Studying air pollution and public health has been a historically important question in science. It has long been hypothesized that severe air pollution conditions lead to negative implications in basic human health. Primarily, areas thats are prone to severe degrees of human pollution are the focus of such studies. Such research relating to less populated areas are scarce, and this scarcity raises the question of how such pollution dynamics (human-made and natural) influence human health in more rural areas.

The aim of this study is to explore this hole in research; in particular we explore possible links between air pollution and Influenza-like-illness in Montana. We begin with a discussion of our starting hypotheses, the data we have accumulated to test these hypotheses, and some exploratory analysis of these data. The body of this research is based on modeling of the natural factors that influence influenza dynamics in general and how these factors apply in the state of Montana. Here, we will explore different modeling approaches and how to apply them to the given data. To conclude this research, a summary is provided and the implications this has for the state of Montana.

Acknowledgements

I would like to acknowledge Stacey Anderson, MPH, State Epidemiologist from the Communicable Disease Epidemiology department at Montana DPHHS for permission and use of the county-wide influenza data and acknowledge Zack Holden, PHD, USDA Forest Service for creating the PM 2.5 data. I also thank the INBRE (IDeA Network of Biomedical Research Excellence) group for their funding and confidence in our research.

I would further like to thank Dr. Erin Landguth for her time and patience in teaching me the fundamental aspects of performing basic research, and for our meetings, discussions, and general pleasant correspondence. Also, I wish to thank Dr. Jonathan Graham for his invaluable insight and guidance in modeling and exploratory analysis involved in this research.

Contents

Introduction

Particulate Matter

Studies that relate air pollution to influenza are not at all rare; there have been countless studies that have aimed to form links between measures of air pollution and influenza incidence. Such studies fall in the realm of Epidemiology, the study of incidence, cause, and control of disease. Often, these studies are trying to relate a specific kind of pollution (such as 'smog') to a specific aspect of human health (mortality, susceptibility, etc.). In this study, we are primarily interested in air pollution as measured by Particulate Matter 2.5. Particulate Matter 2.5 (denoted PM2.5) is defined as 'fine inhalable particles, with diameters that are generally 2.5 micrometers and smaller' [1]. In figure 1, the scale of PM 2.5 is illustrated.

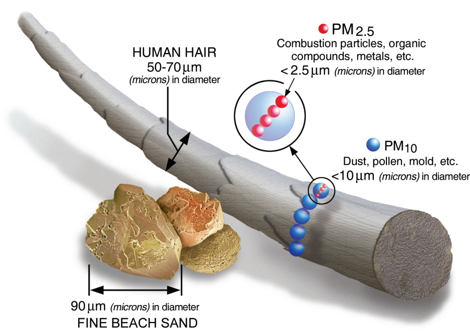
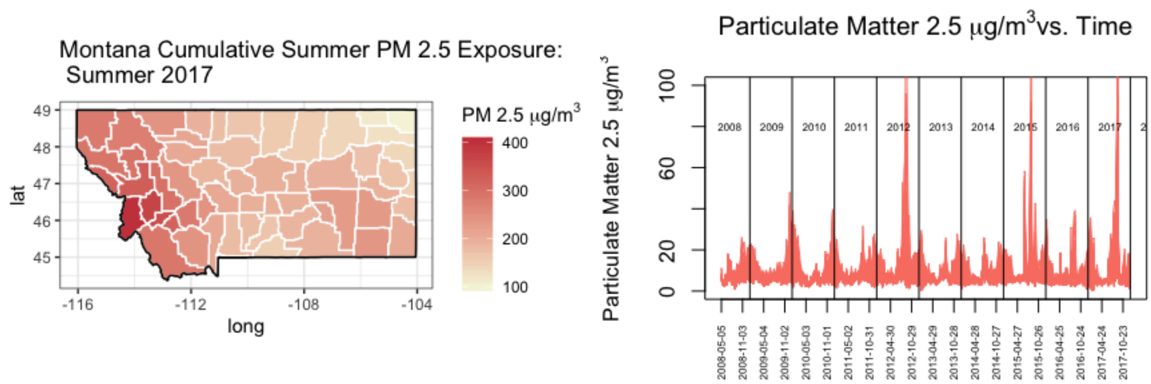


Figure 1: An illustration of the scale of PM2.5 [1]

A large body of studies was reviewed in 2016 in the journal *Environmental Health Perspectives* in which they conclude "Consistent evidence from a large number of studies indicates that wildfire smoke exposure is associated with respiratory morbidity with growing evidence supporting an association with all-cause mortality. More research is needed to clarify which causes of mortality may be associated with wildfire smoke, whether cardiovascular outcomes are associated with wildfire smoke, and if certain populations are more susceptible"[2].

PM2.5 is of particular interest in Epidemiological studies because it is among those pollutants which are projected to increase in terms of density in the future. This increase leads to natural curiosities about how regional health will be influenced or even altered.

In Montana, PM2.5 pollution from wildfires is causing massive problems. For example, in the summer of 2017 a devastating string of 21 wildfires ravaged Montana landscapes burning roughly 438,000 acres of land [4]. During the summer of 2017 in Montana, PM2.5 readings got as high as 109 $\mu g/m^3$ in some regions. For reference, The Environmental Protection Agencies threshold for 'unsafe' levels of PM2.5 exposure in a 24-hour window is $35\,\mu g/m^3$[5]. This level of exposure occurred at least 7 times in a 3 month period in the summer of 2017 according to our data. Studies have shown that there have been immediate health impacts to this level of exposure ($> 35\,\mu g/m^3$), and that there exists an association between respiratory admissions and such intense conditions [6].

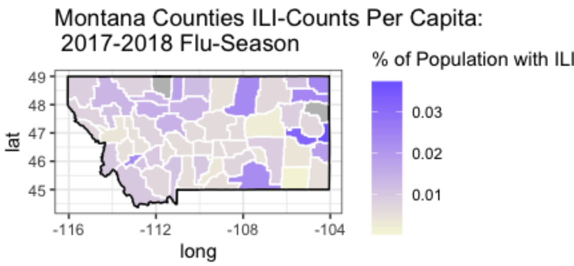


(a) (b)

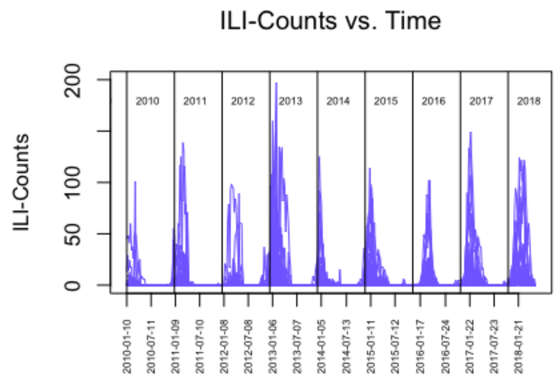
Figure 2: (a.) Cumulative PM2.5 Exposure for Summer 2017 (b.) Weekly Average PM2.5 vs. Time for all Montana Counties

Influenza-Like-Illness

In this study, our primary response of interest is counts of Influenza-Like-Illness (ILI). The World Health Organization (WHO) defines an ILI as having one of the following symptoms: a measured fever of $38\,C^\circ$ ($100.4\,F^\circ$), a cough, with onset within the last 10 days [7]. Generally, reports of ILI tend to be highest in what we call the 'flu-season'. This period of time often starts in October, peaks in February and March, and fades in late April (This trend can be seen in Figure 3b).



(a)



(b)

Figure 3: (a.)Total Influenza Counts Per Capita for 2017-2018 Flu-Season (b.) ILI-Counts vs. Time for all Montana Counties

Problem Statement and Hypotheses

The focus of this research is to explore possible associations between PM2.5 and ILI. We approach this exploration in a variety of ways including spatial statistical methods, temporal statistical methods, and general statistical modeling. Our encompassing hypothesis is that there exists a positive association between PM2.5 and ILI, as the levels of PM2.5 density rise to unsafe levels we expect to see a higher number of reported ILIs. We are primarily interested in exploring the effect that PM2.5 produced during summer seasons (that are prone to wildfires in the region) have on corresponding incidence of ILI in the following influenza season. However, we also consider the possibility that there is a more immediate impact of PM2.5 on ILI counts during the flu season. Some studies have suggested that short term effects (some on the scale of days, others on weeks) of PM2.5 density exist on multiple health related issues [8].

Influenza, PM2.5 Data and Associated Covariates

Data Format

In this section, the collected data and the formatting of these data are discussed. Beginning with our response, we have collected total weekly influenza-like-illness reports from county hospitals in nearly all of Montana [9]. We say 'nearly all' because we are missing ILI data from two counties (Toole and Richland County) and we also group 6 counties (Musselshell, Petroleum, Judith Basin, Wheatland, Golden Valley, and Fergus County) into the 'Central Montana Health District' (CMHD). ILI counts are only consistently collected from the beginning of September to the end of May the following year. In the summer months, ILI counts are not recorded, as they are so minimal in this time period. Initially, we had collected ILI data from as far back as the 2009 flu-season, but ultimately had to dispose of these data due to suspicions of misreporting. We believe that this misreporting arose from not recording on a weekly basis, but instead accumulating reports over multiple weeks and reporting all such ILI cases on one particular week. It is unlikely that this happened past the 2009 flu-season. Thus, the credible date range for our collected ILI data is from 2010-01-01 to 2018-06-01.

Daily PM2.5 emissions from wildfires were provided by the Missoula Fire Lab Wildfire Emission Inventory [10]. These data were paired with data from the upper atmosphere characterization of transport wind direction to wildland fire smoke transports from the North American Regional Reanalysis [11]. These variables were used to model PM2.5 concentrations measured at air quality monitoring stations throughout the state on a weekly basis and develop a PM 2.5 geographic layer over time. We are primarily interested in exploring two possible functions of PM2.5 in relation to ILI: a long-term effect experienced from summer exposure, and a short-term effect due to winter inversion experienced during influenza season. We considered multiple functions of the PM 2.5 data to express these different kinds of exposure in Table 1. Examples of these variables for Gallatin County are shown in Figure 4.

		Variable Name	Description
PM 2.5 Variables Tested	**Short-term Effects**	n-week lag	Lag PM 2.5 density up to n-weeks before current week of ILI.
		n-week moving window sum	Sum total PM 2.5 exposure up to n-weeks before current week of ILI
		Cumulative PM2.5 exposure	Sum total PM 2.5 exposure over entirety of flu-season.
	Long-term Effects	Total PM2.5 exposure	Sum total PM2.5 exposure from summer months preceding Influenza Season

Table 1: A table displaying the different PM 2.5 variables being tested in modeling for association with ILI

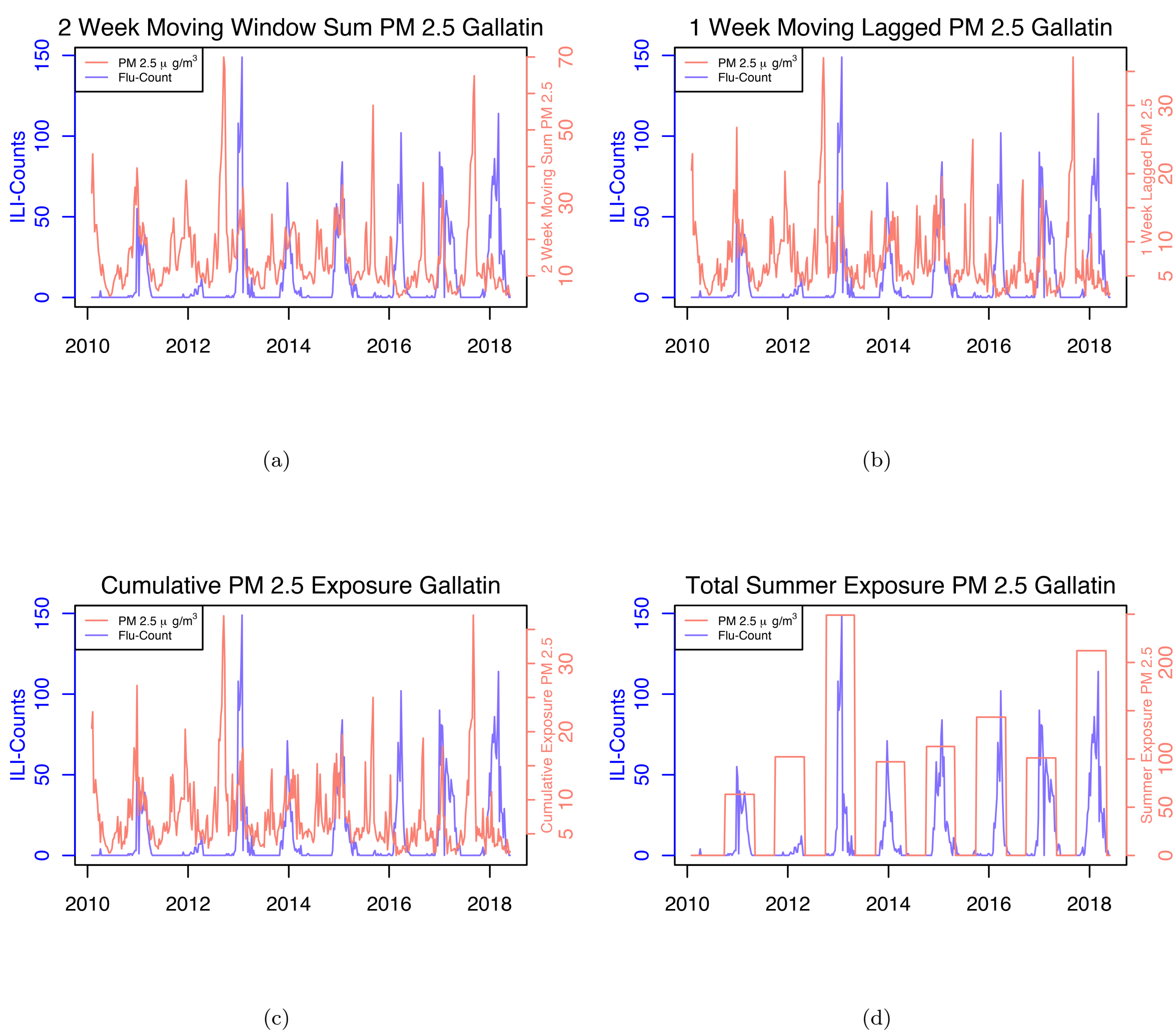


Figure 4: (a.) A time-series of the 2 week PM2.5 Moving Sum (b.) the 1 week PM 2.5 Lag (c.) the total influenza-season PM 2.5 exposure, (d.) the cumulative summer PM 2.5 exposure for Gallatin County.

In addition to these environmental variables we also include a seasonal variable to control for seasonal fluctuations in Influenza. This seasonal component is called a 'Fourier Component' [18]; in the most general sense this is a sum of sine and cosine terms of varying periodicity that is meant to capture seasonal fluctuations in ILI counts. The basic idea is that every year, typically one flu-season 'peak' is experienced with some varying smaller influenza spikes. This component is intended to account for this natural cyclic seasonality in ILI prior to estimating the effects of other variables of interest. In our study, we find that three sine functions and three cosine functions at periodicity 52 weeks (one spike a year) 26 weeks (two spikes a year) and 13 weeks (4 spikes a year) captured the majority of the seasonal fluctuations in ILI-counts in our models. These six seasonal fluctuations are visualized in Figure 5.

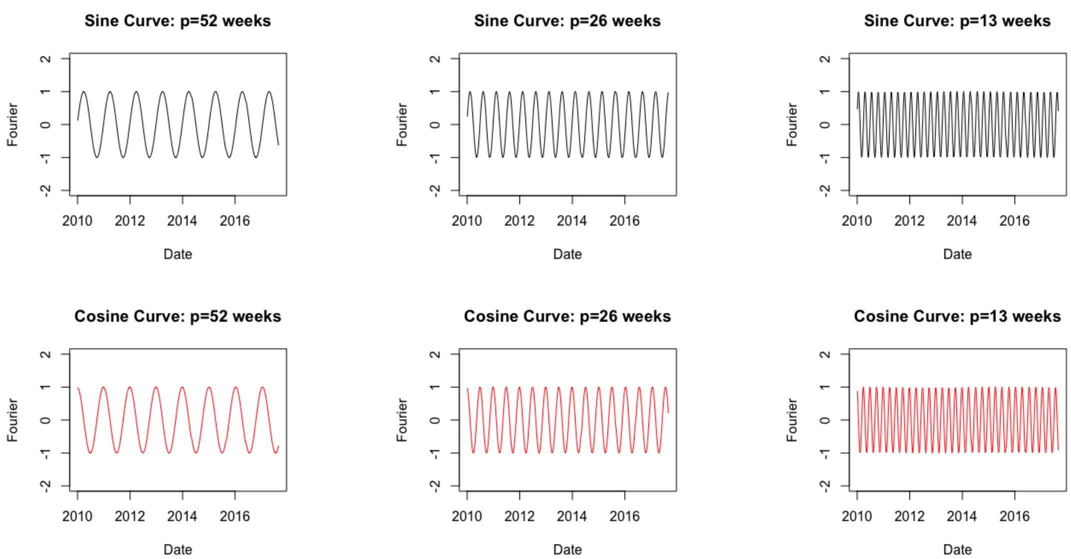


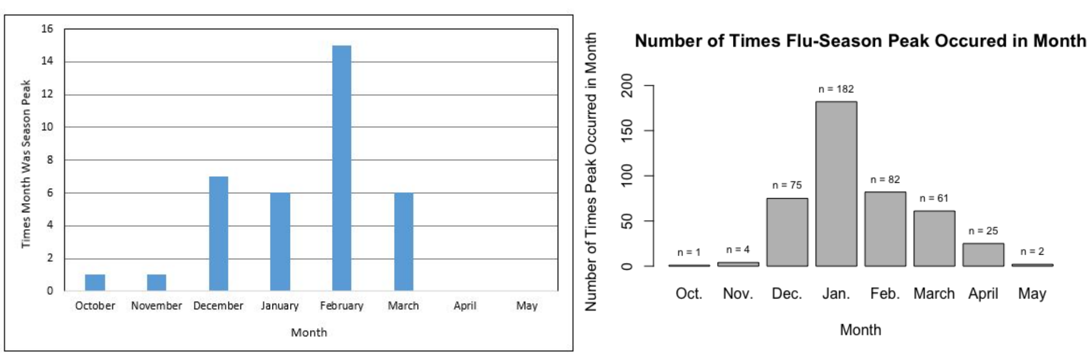
Figure 5: A visualization of different periodicities in the Fourier Components

Exploratory Analysis

Here we establish some basic properties of the observed Influenza and PM2.5 data for counties in Montana. Of particular interest is exploratory analysis in three statistical senses: The traditional sense where we observe ILI counts and PM2.5 via visualizations and look for possible issues that may arise in modeling, the spatial sense where we seek to explore spatial relationships or patterns present in the ILI data or the PM2.5 data, and the temporal sense where we will observe if there are persisting relationships through time that could be used in modeling.

Basic Exploratory Analysis

One of the primary issues with modeling ILI incidence is the sporadic and spontaneous nature of when the flu season starts to accelerate. We have sufficient evidence to say that flu seasons happen in 'peaks', usually as one or two per flu-season. This is evident in figure 3b, where we see one prominent peak per year and some secondary peaks of less magnitude. The Center for Disease Control conducted an interesting study where they observed in which month this 'peak' occurred for flu-seasons as far back as 1982 [12]. Their study produced the histogram in Figure 6a. In Figure 6b, the same plot is shown for our collected data separately for all Montana counties.This shows that the time in which flu-season peaks is highly variable; anytime between October and May is possible. With this kind of uncertainty comes some modeling issues due to the sparsity of ILI.



(a) (b)

Figure 6: (a.)Month in which the Flu-Season Peaked Since 1982 [12] (b.) Month in which Flu-Season Peaked in Montana Counties 2010-2018.

There are many weekly recordings in our observational time frame (October-May every year since

2010) in which no cases of ILI are recorded at all. There are multiple reasons for this abundance of null counts. One such reason is that Montana is the 43th most-populated state in the United States with only 1,062,000 residents and the 47th most-populated per land area (mi^2)[13]. To summarize, Montana has very few people and they tend to be concentrated in a small number of cities. In a recent paper by Dr. Benjamin D. Dalziel et. al. (2018) it was discovered that population density has an influential role in the characteristics of flu-season. To summarize their paper, they concluded that 'epidemics in smaller cities were focused on a shorter period of the influenza season, while the incidence was more diffuse in larger cities.' [14]. This discovery applies well to Montana cities; consider for example a comparison between a highly populated county and a less populated county. In Figure 7, ILI counts for Yellowstone county (home to Billings,MT, the most populated city in the state) and its' neighboring county Carbon county (21st most populated county in Montana) are shown. It is clear to see that the flu-season persists for a longer number of weeks in Yellowstone County than in Carbon. This is evident because the peaks for Yellowstone County are much wider in general than the corresponding peaks in Carbon County.

This property of ILI causes some issues in this study. We maintain a consistent time frame from October to May of the following year for each flu season, but a majority of those weekly values are very low or zero. Observing histograms from these counties in terms of ILI counts reflects this problem. In figures 8a, 8b, and 8c the ILI for all weeks in the study can be observed for the three most populated counties in Montana: Yellowstone, Missoula, and Gallatin County. It is clearly seen that a majority of weeks are in the very low count range (0-10 ILI), with only sparse occurrences of spikes. To illustrate this sparsity of ILI counts, consider the three most populated counties in Montana listed earlier. The percentage of weeks with 0 ILI reports within our study time frame for these three counties is: 33.9% for Yellowstone County, 40.4% for Missoula County, and 29.8% for Gallatin County. Contrastingly, in figures 8d, 8e, and 8f the corresponding ILI histograms for three smaller population counties (Broadwater, Sweet Grass, Wibaux) are shown. In all cases a distinct right skew is apparent; in larger populations (figures 8a, 8b, and 8c) we see longer tails due to the increased population. This emphasizes the point that a large number of ILI counts in our study range are near zero for all counties and especially small ones. This issue will be accounted for in the statistical models for ILI considered later in this paper.

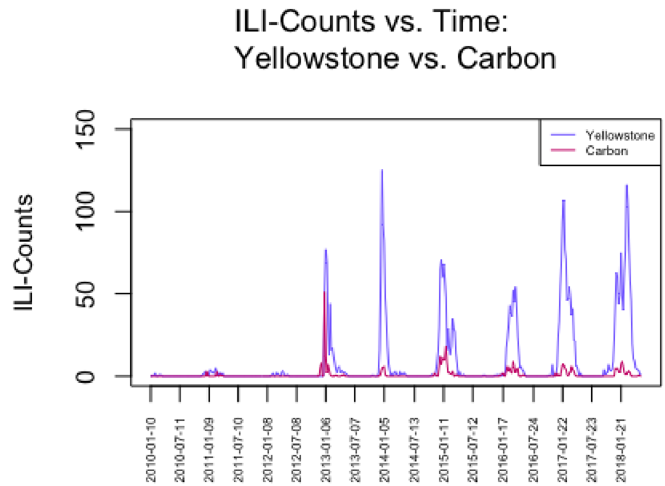
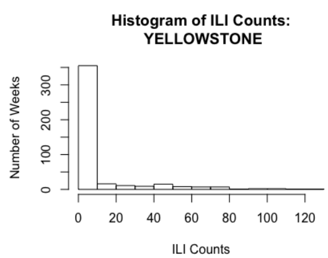
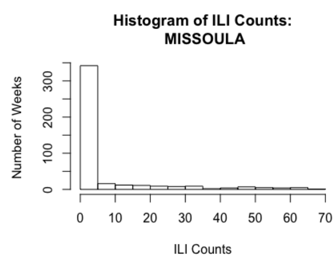


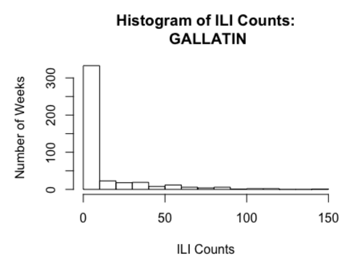
Figure 7: Comparison of Yellowstone and Carbon County ILI counts.



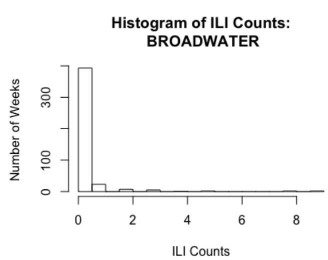
(a)



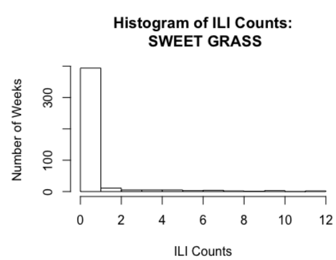
(b)



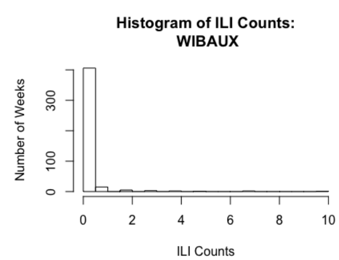
(c)



(d)



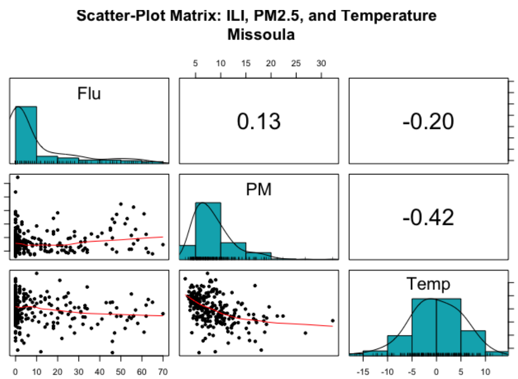
(e)



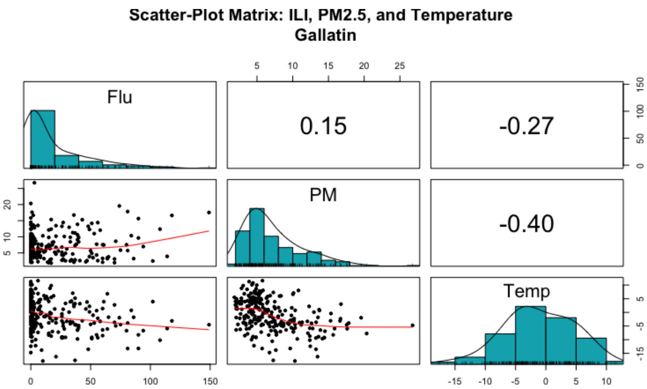
(f)

Figure 8: (a.)Histogram of ILI for Yellowstone (b.) Histogram of ILI for Missoula (c.) Histogram of ILI for Gallatin (d.) Histogram of ILI for Broadwater (e.) Histogram of ILI for Sweet Grass (f.) Histogram of ILI for Wibaux

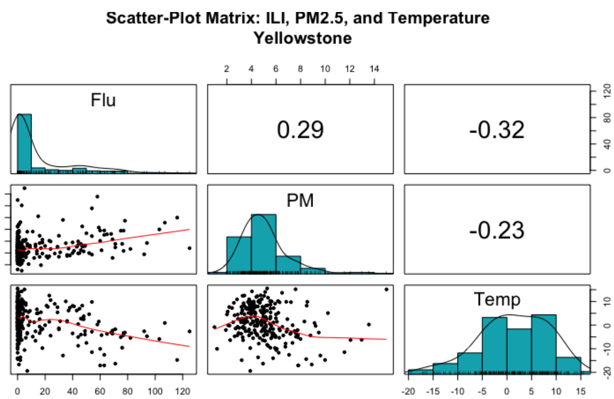
We are further interested in understanding the basic distributions of our covariate data, including PM2.5 (and functions thereof) and temperature. In figure 9, we see scatterplot matrices for these three primary quantities of interest for a small handful of counties to illustrate their stochastic behavior. This is intended to explore for possible issues in collinearity that may arise in modeling or any irregularities in the covariates of interest. For example, it is possible that PM 2.5 density and temperature are collinear; when the temperature is high, PM 2.5 density tends to also be high.



(a)



(b)



(c)

Figure 9: Scatterplot matrices for (a.) Missoula (b.) Gallatin (c.) Yellowstone

In general scatterplot matrices such as those in figure 9 tend to be similar among counties. There do not appear to be any significant collinearity issues among the covariates that will be used in the modeling of ILI counts for each individual county. This is supported by the low correlations among the two covariates, PM 2.5 (and functions thereof) and temperature, PM 2.5 density curves are somewhat right skewed and centered at about $10 - 15\,\mu g/m^3$ and temperature tends to be fairly symmetric centered at about $0 - 5\,^{\circ}C$.

Spatial and Temporal Analysis

We wish to explore the existence of spatial and temporal relationships in ILI count data and PM2.5 data in Montana. This can be accomplished through the use of spatial and temporal correlation statistics but also basic intuition. In this section we discuss the mathematical forms of these descriptive statistics and interpret the meaning of these statistics in context with our analysis. The general hypothesis we started with is that ILI counts by county are spatially correlated, i.e. if county A has a high level of ILI counts at some time, it is plausible that neighboring county B would as well. If this is the case, this spatial information could be modeled and incorporated into spatial models which use neighborhood information to produce predictions in space (see [25] for details).

Spatial Analysis: Moran's I Spatial Statistic

In Figure 3 at the beginning of this study, we displayed a visualization of the spatial distribution of ILI counts at a particular date; from this visualization it is somewhat unclear if there is a spatial relationship we can utilize. In order to assess the presence of spatial autocorrelation between ILI counts and PM2.5 density at the county level in Montana we make use of Moran's I statistic. This value gives a correlation metric on the spatial autocorrelation of a single quantitative variable. More details about the Moran's I statistic can be found in 'Statistics for Spatial Data' [25] by Cressie (1993). In the most basic sense, Moran's I measures how one response in space is similar to other responses in space surrounding it. The statistic has the following mathematical definition:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2}$$

where

$$y_i = \text{ILI counts of county i divided by population of county i}$$

$$\bar{y} = \sum_{i=1}^{n} y_i \text{ the arithmetic mean of flu rates of all counties}$$

$$w_{i,j} = \begin{cases} 1 & \text{of county i borders county j}, i \neq j \\ 0 & \text{otherwise} \end{cases}$$

Note that I is contained in the interval $[-1, 1]$ just as the traditional Pearson's correlation coefficient. The closer to 1 the statistic I is, the stronger positive spatial correlation there tends to be (i.e. regions closer to each other tend to be positively associated), the closer to -1 the statistic I is the the more negatively associated neighboring regions are. It is important to note that there are many weighting metrics that can be used to express the relationships between neighboring counties, but here a binary weighting system was chosen over differing neighborhood weighting systems such as distance to centroid of county or proportion of shared county border. It is my belief that these more complicated weighting schemes could be more valuable at finer resolutions, such as zip-code level, but for county level analysis do not have enough influence.

We also scale the ILI's at time t by the population of the county in order to have a similar scale for the response of interest statewide. Later, when modeling is discussed for a single county we generally do not perform this transformation because we are interested more in developing count models rather than rate models. At the individual county level, this scaling by population is not necessary as the types of models we are using a resistant to constant scale changes. The reasons for this distinction are discussed more in the modeling sections of this paper. We estimate Moran's I statistics at every week for which we have ILI data available for, Figure 10 shows the estimates for each week over the 8 years of data.

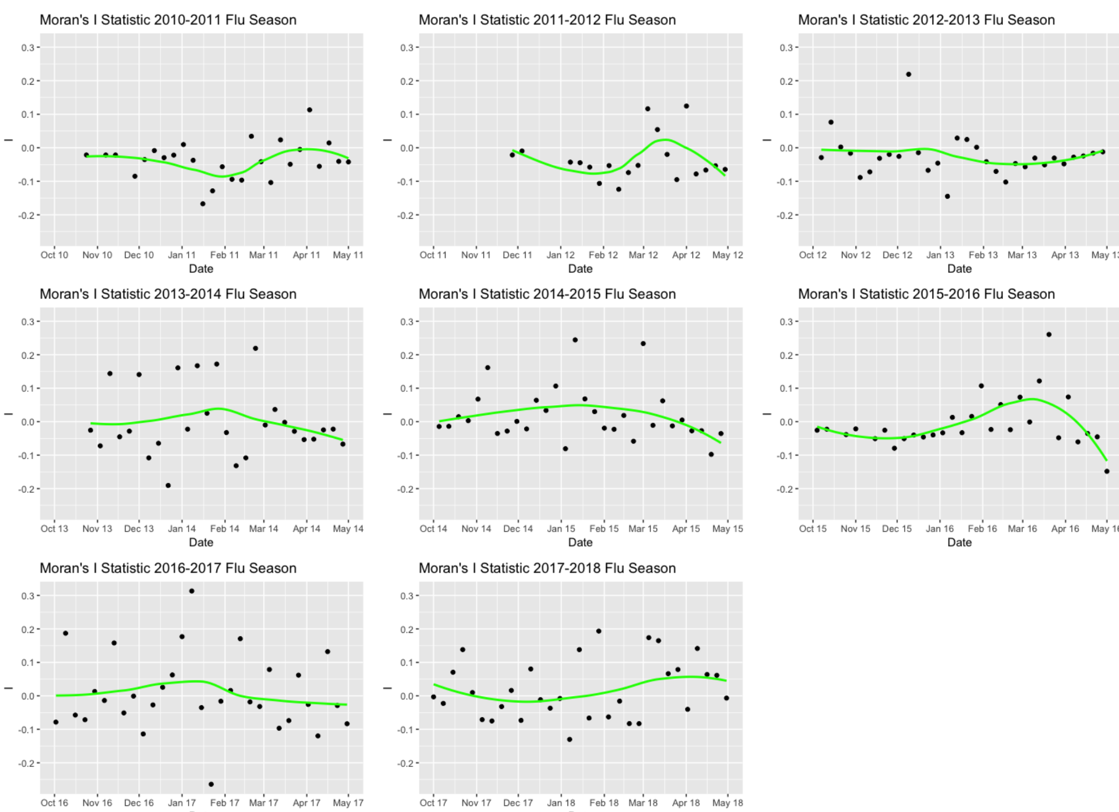


Figure 10: The estimated Moran's I statistics for every week for which the calculation was acceptable stratified by influenza-season.

The consensus is that there is very little if no spatial correlation present in these ILI counts at the county level. I believe the reason for this to be multifold. First and foremost ILI, spatial relationships at the county level are not a fine enough resolution; preferably zip-code level would be used. I believe this to be due to the ways in which Influenza is traditionally transmitted, as covariates such as socio-demographic variables, distance to public schools, and public transport are the primary drivers of Influenza transmission [26]. For modeling purposes, persistent spatial autocorrelation in the response of interest require the use of spatial modeling methods, but this analysis suggests that any spatial modeling would be less useful than other methods. However, this spatial relationship is certainly worth exploring more at the zip-code level or even the census track level as indicated by other studies [27].

Moran's I statistic was also applied to the PM2.5 variable on a week-by-week basis for all counties in Montana. Figure 2 suggests that it is highly plausible that there is spatial autocorrelation in PM2.5 densities across Montana. In Figure 11, the estimates of Moran's I for PM2.5 density over time can be visualized.

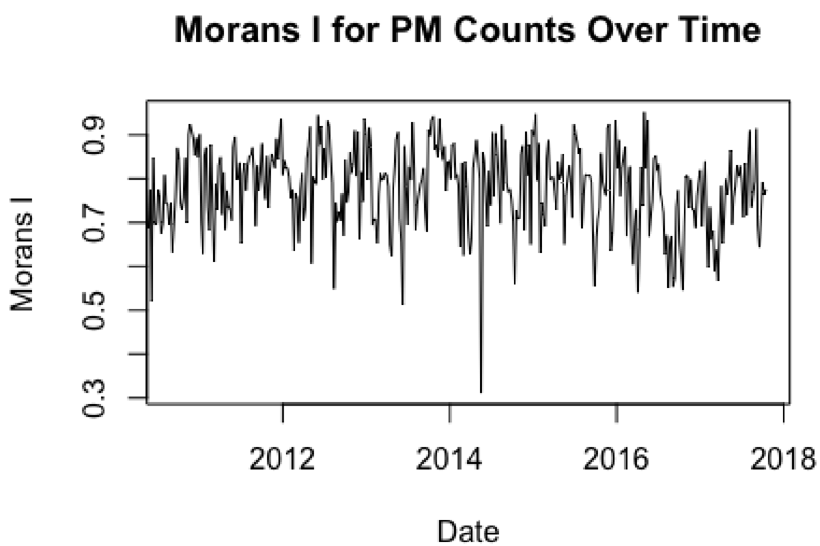


Figure 11: Estimated Moran's I statistics for PM2.5 density in all of Montana on a weekly basis.

Clearly, the spatial auto-correlation in PM2.5 densities is substantially higher than the corresponding ILI values. In general, the estimate of Moran's I tends to hover around a spatial correlation of 0.8. Note that the large spike at '2015-03-15' is due to missing data. This analysis suggests that as counties experience higher levels of PM2.5 density, generally so do neighboring counties. In contrast, our response is not strongly spatially related at the county level.

Temporal Analysis: Auto- and Cross- Correlations

In our study, it is of great interest to understand the temporal dynamics at play relating to both ILI counts and PM2.5 density at the county level. To do this, we compute the autocorrelation function (ACF) to assess temporal autocorrelation in the observations. Letting $y_1, y_2, ...y_n$ represent the time series of ILI counts for a Montana county, the auto correlation function at a lag of l is defined as follows [28]:

$$r_l = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(y_{i-l} - \bar{y})}{\sqrt{\sum_i^n (y_i - \bar{y})^2}\sqrt{\sum_i^n (y_{i-l} - \bar{y})^2}}$$

This is essentially Pearson's correlation coefficient with lagged components being used in place of some other variable of interest. Figure 12 shows the estimated auto-correlations for the ILI counts and PM 2.5 densities for every county in Montana. Referencing figure 12, there tend to be high amounts of auto-correlation for both ILI and PM2.5 counts for Montana counties. On average, Montana counties showed estimated ILI auto-correlations of .597, .465, and .353 for lags of 1 week, 2 weeks, and 3 weeks respectively. A similarly strong result is obtained for PM2.5 auto-correlations exhibited by estimates of .542, .391, and .292 for 1 week, 2 weeks, and 3 week lags. Clearly in some counties these temporal relationships are stronger than others. For example, in Yellowstone County, the estimated one-week auto correlation is .90, a very strong positive association. Typically, lags of up to two weeks in ILI counts seem to have significant temporal association, using the significance cut-off below described in [28]:

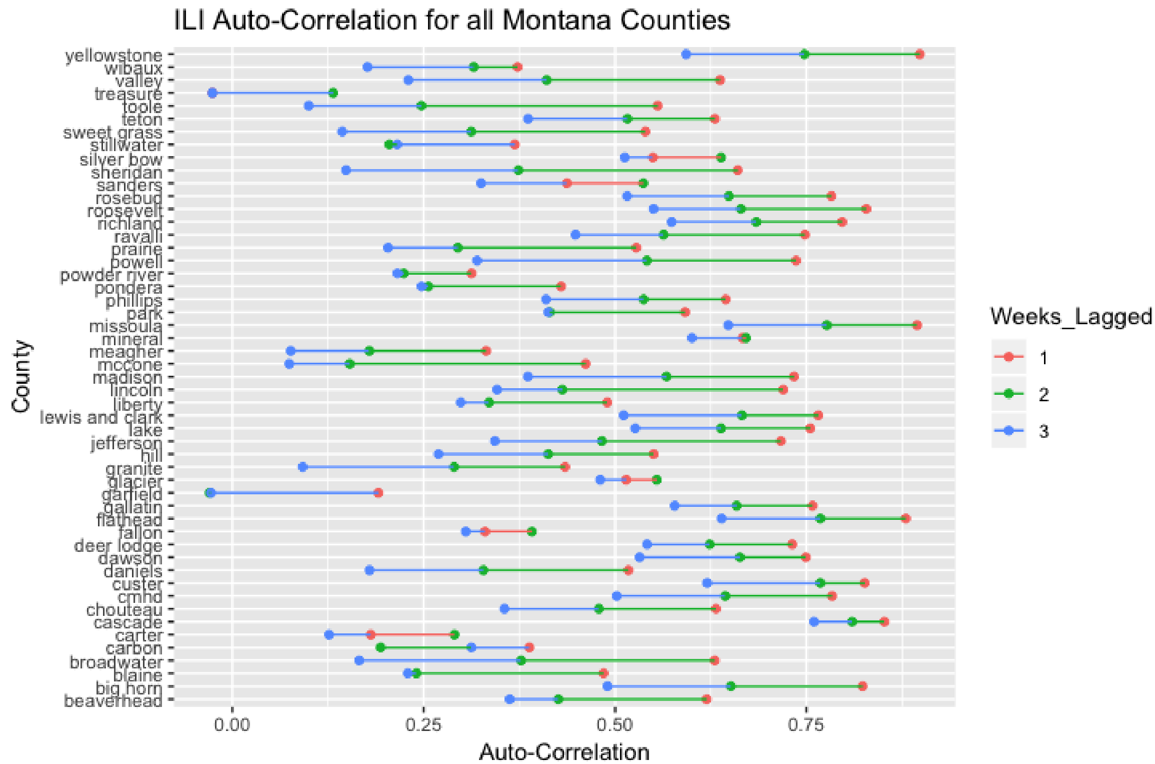
$$\frac{1.96}{\sqrt{\underbrace{T}_{\text{Number of Obs.}} - \underbrace{l}_{\text{lag}}}} = \frac{1.96}{\sqrt{238-1}} = .127$$

Note in Figure 12 plots (a) and (b) (both with sample size 238 after removing non-flu season data and accounting for lagged dependencies) that a majority of the estimated auto-correlations for 1 week lags exceed this significance threshold. We have reasonable evidence to say that there is a temporal association in both ILI counts and PM 2.5 densities for a majority of Montana counties (with the exception of some lightly populated areas such as Treasure County). This relationship is vitally important to modeling counts in ILI on a weekly basis; this suggests to us that an auto-regressive model of lag 1 (AR1) could be highly useful in controlling for natural auto-correlation in the response. We take this into account and discuss in length during the modeling section of this paper.

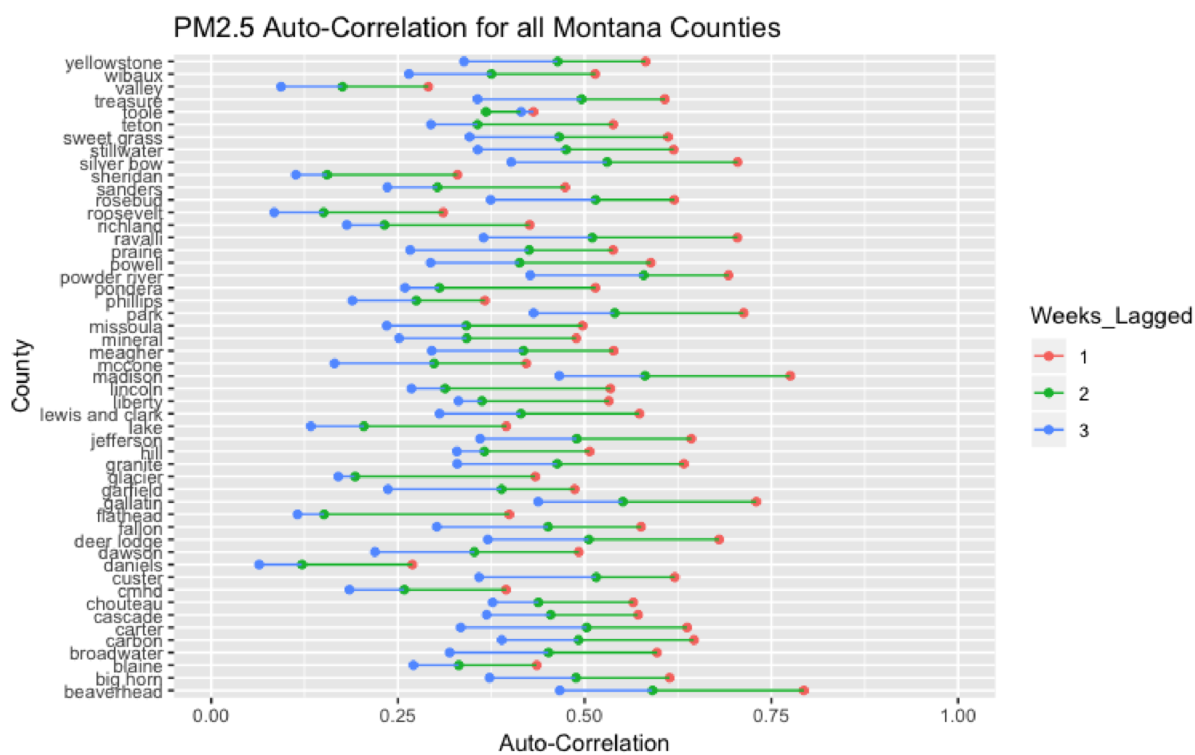
We are also highly interested in the relationship in both the short term and long term between the ILI counts at a current week in Montana counties, and both the immediate and distant past in PM2.5 density behavior. To assess this relationship is the core of our research; here we examine some possible short term associations. To explore this relationship, we make use of cross-correlation. This time-series statistic assess the strength of a relationship between two quantitative variables in time by correlating the current value of one of the variables with the other variable lagged some amount in time. Letting $x_1, x_2, ..., x_T$ represent the sequence of weekly PM2.5 density estimates for one Montana county, that county's estimated cross correlation function at a lag of l is given by:

$$r_l = \frac{\sum_{i=l+1}^{T}(x_i - \bar{x})(y_{i-l} - \bar{y})}{\sqrt{\sum_i^T (x_i - \bar{x})^2}\sqrt{\sum_i^T (y_i - \bar{y})^2}}$$

Human health and behavioral studies have highlighted the dangers of using cross-correlation over long periods of time to infer causal relationships [29,30]. From a practical standpoint, it is logical that cross-correlations at shorter lagged periods are less susceptible to confounding than cross-correlations at longer lagged periods. One reason for this intuition is that in general estimated cross-correlations for long lags tend to be based on fewer data points than for shorter lags [30]. Staying true to this philosophy, in Figure 12 we observe cross correlations between the current ILI weekly count and the PM2.5 density from 1 week, 2 weeks, and 3 weeks previous to the current week. We save longer term analysis for the modeling portion of the paper so possible confounding variables can be appropriately controlled for.



(a)



(b)

Figure 12: (a)Estimated ILI auto-correlations in Montana counties. Red dots represent the auto-correlation at a 1 week lag, green at a 2 week lag, and blue at a 3 week lag. (b) Estimated PM2.5 auto-correlations in Montana counties. Red dots represent the auto-correlation at a 1 week-lag, green at a 2 week lag, and blue at a 3 week lag.

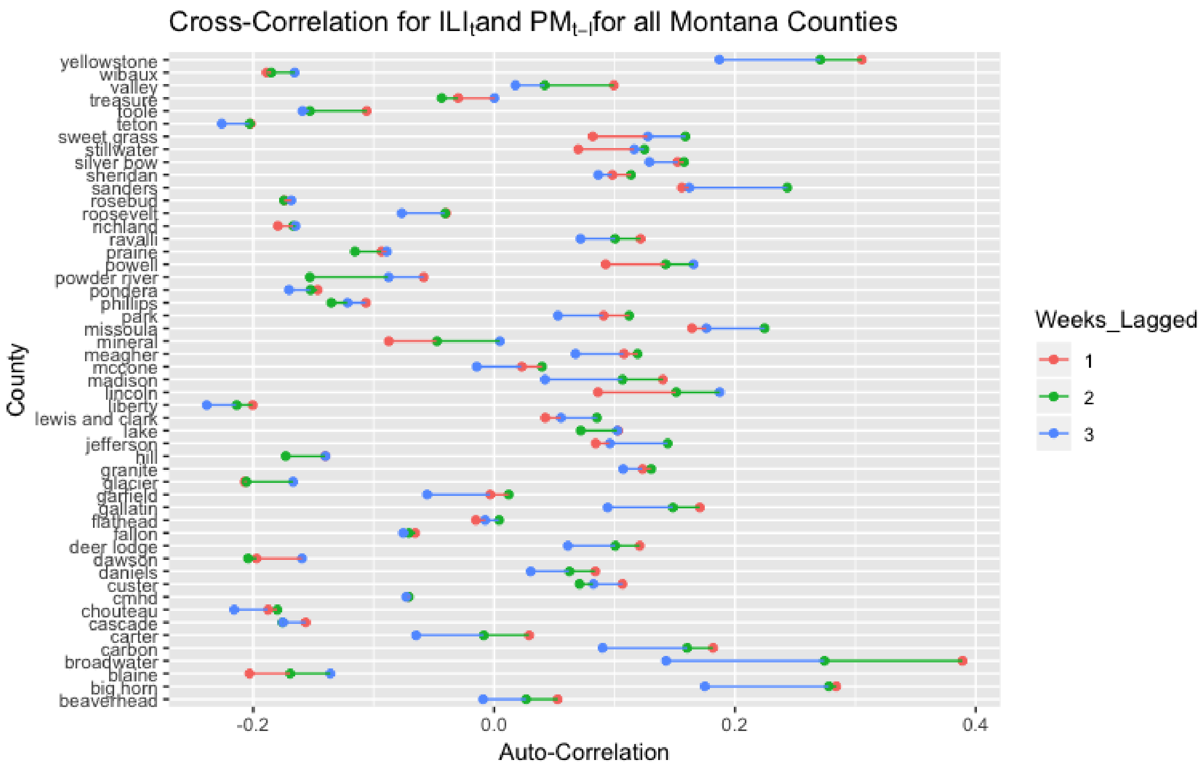


Figure 13: Cross-Correlations between current ILI weekly count and PM2.5 density from 1,2, and 3 weeks prior

In figure 13, it is immediately apparent that short-term cross correlations are highly erratic. In general, a majority of counties tend to have a positive estimate of cross correlation between ILI at time t and PM2.5 density at time $t-l$, but often these estimates fall right around the .10 mark, just below the $\alpha = .05$ significance threshold [28] for our study. This suggests that short-term associations between ILI and PM 2.5 are inconsistent, and vary from county to county. In the modeling section of this paper, we will explore these short term associations (as well as long term) in more detail.

Model Selection

In this section several topics are discussed including model selection, basic theory about types of models, and components of these models. These models are then interpreted and analyzed for performance and basic inference can be done. Following the work of Feng et. al. [15], Hooten et al. [16], and Imai et al [17] we propose a model of ILI that incorporates the over-dispersion of ILI counts discussed in earlier sections. We utilize the temporal correlation in ILI counts on a weekly basis, weekly average temperature, multiple functions of PM2.5 and seasonality components to estimate and predict ILI counts. We begin by utilizing a **generalized linear model** in order to impose different distributional assumptions on the response.

Generalized Linear Models: Definition and Basic Theory

A generalized linear model has the following form:

$$g(\boldsymbol{\mu}) = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\mu}$ is the average of the response of interest and $g(.)$ is some link function imposed on the average response to restrict the response to a suitable domain [18]. Common link functions are the 'logit' function (which restricts responses to the range $[0, 1]$) and the 'log' function (which restricts responses to the positive range $[0, \infty)$). $\mathbf{X}$ is a matrix of covariates of interest, $\boldsymbol{\beta}$ are the coefficients of the covariates, and $\boldsymbol{\epsilon}$ is a vector of errors for each observation. In typical linear modeling, we impose a normality assumption that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. This assumption is not necessarily made in generalized linear modeling; instead we can impose distributional assumptions that more accurately reflect the characteristics of our response. To clarify, a normal assumption on the errors, $\boldsymbol{\epsilon}$, would imply a normal assumption on the responses (ILI counts) for a glm for ILI counts. We know by previous exploratory analysis (see earlier sections) that this is clearly not the case. Instead we need to use distributions that are more well suited to our responses.

An appropriate choice for link function $g(.)$ would be the log function. This would ensure that $\boldsymbol{\mu}$, the input to $g(.)$, is strictly positive. This is a beneficial property for this research; clearly the average ILI count cannot be negative. We further assume that the covariates $\mathbf{X}$ are fixed and known quantities and the coefficients $\boldsymbol{\beta}$ are fixed but unknown quantities we need to estimate. Inferences in this study rely on the coefficients $\boldsymbol{\beta}$, as they will ultimately supply any evidence in relating PM2.5 density to ILI counts in Montana. Typically, the link function $g(.)$ is a non-linear function of the average response $\boldsymbol{\mu}$. This non-linearity forces us to estimate $\boldsymbol{\beta}$ differently than we would in the case of ordinary least squares where we would simply solve the normal equations $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$. The parameter estimates can be found using iteratively re-weighted least squares, a form of the classic Newton's Iterative Algorithm [19]. The basic expression of this algorithm is given below:

$$\boldsymbol{\beta}^{\boldsymbol{t+1}} = \boldsymbol{\beta}^{\boldsymbol{t}} + \mathbf{J}^{-1}(\boldsymbol{\beta}^{\boldsymbol{t}})\mu(\boldsymbol{\beta}^{\boldsymbol{t}})$$

Where $\mathbf{J}^{-1}$ is the observed information matrix (the negative of the Hessian Matrix) of the parameter estimates and $\mu(\boldsymbol{\beta}^{\boldsymbol{t}})$ is the score function of $\boldsymbol{\beta}^{\boldsymbol{t}}$ which indicates the gradient of the log-likelihood with respect to $\boldsymbol{\beta}^{\boldsymbol{t}}$ and t is the index of iteration [19]. This is a numeric-approximation algorithm that iterates until a convergence criterion is met and gives us a local-maxima of the likelihood function of $\boldsymbol{\beta}$, thus giving us our parameter estimate $\hat{\boldsymbol{\beta}}$. This is the basic algorithm that the R-statistical software function 'glm.fit' [20] uses, which is the function we will be utilizing most in this research.

Generalized Linear Models: Distributional Assumptions

Now we establish the distributional assumptions that characterize our generalized linear model. Recall that $\boldsymbol{\mu}$ is the average of the response of interest (in our case ILI counts). We now aim to impose a distribution on the response vector $\mathbf{Y}$ which has expected value $\boldsymbol{\mu}$ so that $E[y_i] = \mu_i$. In generalized linear models, a distribution from the exponential family of distributions is chosen. Common distributions from the exponential family are the normal, poisson, gamma, and binomial distributions. The exponential family of distributions is described as any distribution which can be written in the form:

$$f(y_i, \boldsymbol{\theta}) = exp[a(y_i)d(\boldsymbol{\theta}) + b(\boldsymbol{\theta}) + c(y_i)]$$

where y_i is the i^{th} observed response, $\boldsymbol{\theta}$ is a vector of parameters of the distribution (for example μ and σ in a normal distribution), and $a(.), b(.), c(.)$ and $d(.)$ are separate functions of the parameters and i^{th} observation. Distributions of this form have useful theoretical properties, in particular their maximum likelihood estimators are typically fairly easy to derive [18]. What we seek is the most appropriate choice of distribution to use in our generalized linear model. We will impose the assumption that $ILI_i \sim$ Exponential Family($\boldsymbol{\theta}$) and choose the distribution which conforms best with our responses. This will be discussed more when the research-specific model is formed.

Generalized Linear Models: Model for Predicting Influenza-Like-Illness Counts

We now define the form of the model in which we model weekly ILI counts by county in Montana. The generalized linear model has the following form:

$$log(\mu_{t,k}) = \underbrace{\beta_0 + \sum_{i=1}^{6} \beta_i F_i(t) + \beta_7 ILI_{t-1,k} + \beta_8 \text{Temperature}_{t,k}}_{\text{Control for Confounders}}$$
$$+ \underbrace{\beta_9 \text{Summer Total Exposure PM 2.5}_{t,k}}_{\text{Long-term Effects}} + \underbrace{\beta_{10} \text{Flu-Season Exposure PM 2.5}_{t,k}}_{\text{Short-term Effects}} \quad (1)$$

$$t = \text{ week number of study from 2010-01-03 to 2018-05-27, } t = 1, 2, ..., 245$$

$$k = \text{ index of Montana County, } k = 1, 2, ..., 50$$

$$\mu_{t,k} = \text{ Expected ILI count at time } t \text{ in county } k, \text{ assuming} ILI_{t,k} \sim \text{ Exponential Family}(\boldsymbol{\theta})$$

F_i = Sine(i=1,2,3) or cosine(i=4,5,6) function of t with period 52 weeks (i=1,4), 26 (i=2,5) weeks, and 13 (i=3,6) weeks

Total Exposure PM $2.5_{t,k}$ = Cumulative PM 2.5 density from previous summer at week t in county k.

Flu-Season Exposure PM $2.5_{t,k}$ = a function of short-term PM 2.5 exposure at week t in county k.

Notice that the first eight terms in the model are all variables typically associated with influenza dynamics. Studies have shown the benefits of including seasonal Fourier components to control for natural seasonality in ILI counts [17]; likewise temperature is added as another common controlling factor [21]. Earlier it was discussed that ILI counts have highly strong auto-correlations for lags up to 2 weeks prior to the current week in question, thus an auto-regressive component ($ILI_{t-1,k}$) is another important controlling variable; this is intended to 'model-out' (account for) the natural auto-correlation present in ILI counts by county. This model is run for all Montana counties in which we have reliable ILI data (49 total counting all counties in CMHD as one county).

It is important to understand the interpretation of these coefficients before any kind of inference is performed. Observing the model form in equation (1) above, solving for $\mu_{t,k}$ would give us:

$$\mu_{t,k} = exp(\beta_0) \cdot exp(\beta_1 F_1) \cdot \ldots \cdot exp(\beta_9 \text{Total Exposure PM2.5}) \cdot exp(\beta_{10} \text{Flu-Season Exposure PM 2.5}) \quad (2)$$

Letting $g^{-1}_{t,k}$ denote the right hand side of equation (2.) evaluated at some time t and some county k, consider a vector of the covariates at time t in county k:

$$\mathbf{x_{t,k}} = [1, F_1(t,k), F_2(t,k), ..., \text{Total Exposure PM2.5}_{t,k}, \text{Flu Season Exposure PM 2.5}_{t,k}].$$

Then the value of $\mu_{t,k}$ can be expressed as:

$$\begin{aligned}
\mu_{t,k} &= exp(\beta_0) \\
&\quad \cdot exp(\beta_1 F_1(t,k)) \ldots exp(\beta_9 \text{Total Exposure PM2.5}_{t,k}) exp(\beta_{10} \text{Flu-Season Exposure PM 2.5}_{t,k}) \\
&= exp(\boldsymbol{\beta} \mathbf{X}_{t,k}) \\
&= g^{-1}(t,k)
\end{aligned} \quad (3)$$

Now consider increasing the covariate 'Total Exposure PM $2.5'_{t,k}$(TEPM$2.5_{t,k}$) by 1 unit to produce a new estimated ILI count $\mu'_{t,k}$. Then equation 3 would remain exactly the same with the only change being $\mu'_{t,k} = \ldots exp(\beta_9(\text{Total Exposure PM2.5+1})) \ldots = \ldots exp(\beta_9(\text{Total Exposure PM2.5})exp(\beta_9(1)) \ldots$. Notice that this is equivalent to saying $\mu'_{t,k} = g^{-1}_{t,k} \cdot exp(\beta_9) = \mu_{t,k} \cdot exp(\beta_9)$. This gives us a natural interpretation of the coefficients. This means that the average ILI count at time t in county k is expected to change by a magnitude of $exp(\beta_9)$ for every one unit increase in 'Total Exposure PM 2.5' [22]. Note that this interpretation applies for all covariates listed in equation (1).

With this model in place, we now examine how to choose the right exponential family to model the ILI counts. Considering that our outcomes, ILI counts, are discrete-count variables we considered three distributions: the Poisson distribution, the Quasi-Poisson distribution, and the Negative-Binomial distribution. These distributions are traditionally used in counting occurrences of some outcome in a period of time, with the only difference between the Poisson and the latter two being that the latter two distributions account for over-dispersion of the response. "Over-dispersed" in this sense simply means that the variance is functionally larger than the mean; in the pure-Poisson model the mean and variance are equal. The only real difference between the quasi-Poisson model and the Negative-Binomial model is that the variance of a quasi-Poisson model is a linear function of the mean ($Var(Y) = k\mu$) while the variance of a negative binomial model is a quadratic function of the mean ($Var(Y) = \mu + k\mu^2$)[23]. Choosing among these distributions is not a trivial problem, standard model selection criteria such as AIC or BIC are not advised to compare quasi-Poisson and negative binomial models. Instead an evaluation of the mean-variance relationship of the responses is needed. In Figure 14, a visualization of ILI county mean vs. county ILI variance is displayed. It is clearly apparent that the negative binomial assumption appears to be the most accurate with respect to the variance mean relationship of the ILI counts. Certainly, a pure-Poisson model here is completely invalid, whereas a quasi-Poisson might be appropriate.

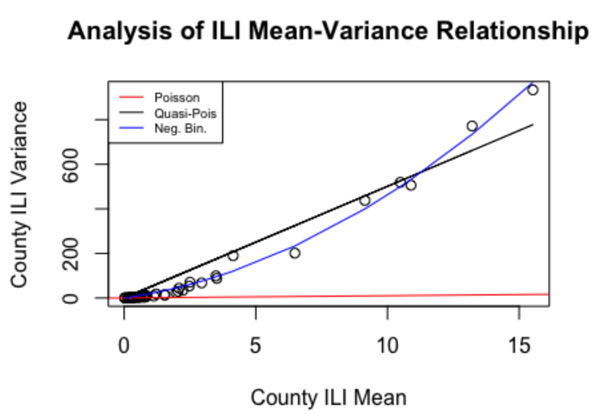


Figure 14: The Variance-Mean relationship of ILI counts for all counties.

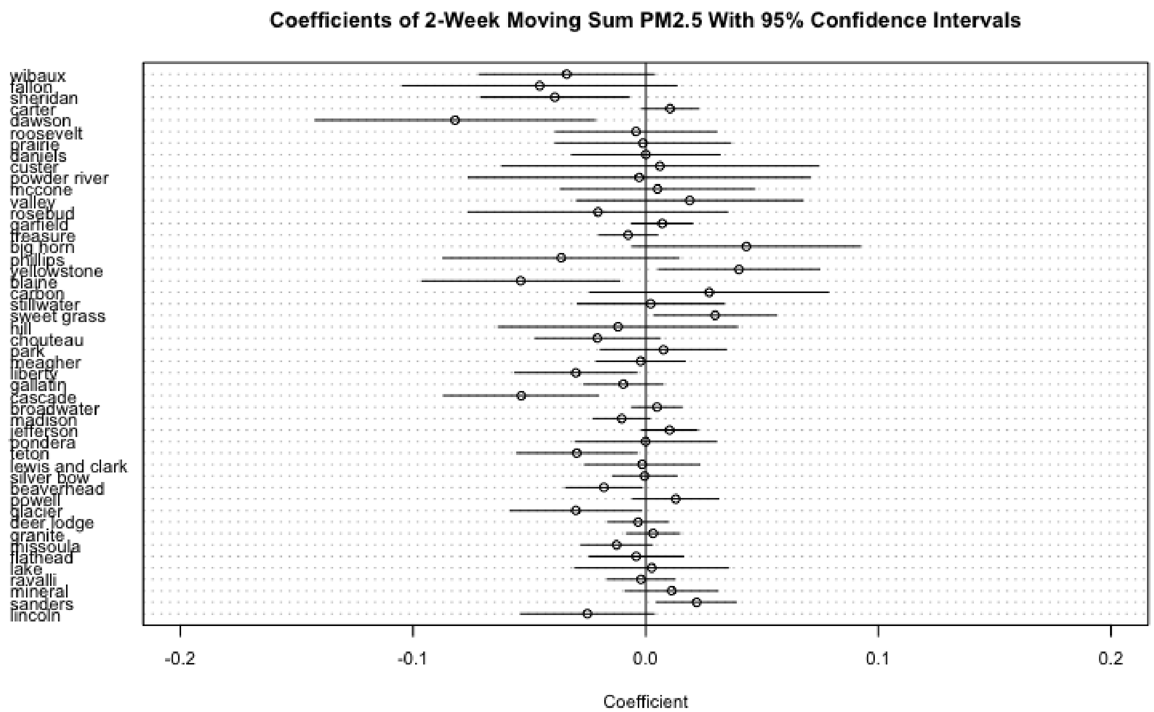
Inference

The model specified in equation (1) was applied to every county in Montana in which ILI counts were available. In this section, basic statistical inference will be performed on the coefficients of these models. We will interpret the coefficient estimates from equation (1) and discuss some resulting evidence that comes from these estimates. Further, we will discuss what functions of PM 2.5 seem to be related to ILI counts for Montana counties.

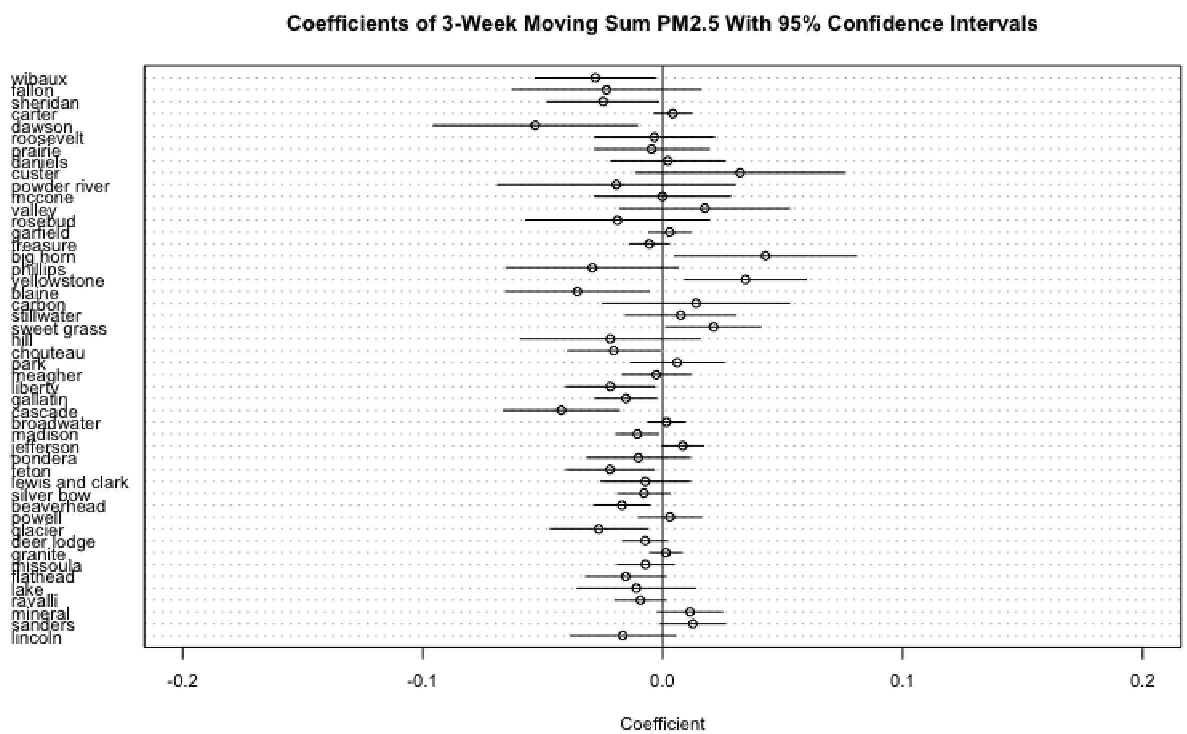
Coefficient Inference

Our generalized linear model is of the form from equation (1). We will first consider the coefficients produced by our covariates of interest: 'Total Exposure PM 2.5' ($\text{TEPM2.5}_{t,k}$) and 'Flu-Season Exposure PM 2.5' ($\text{FSEPM2.5}_{t,k}$). For the short-term exposure component ($\text{FSEPM2.5}_{t,k}$), recall from Table 1 that this could be one of many functions tested for best fit to the model data. We fit the model in equation (1) using each of the proposed short-term PM2.5 functions separately, evaluate the deviance explained in the model, and plot 95% confidence intervals for each of the short-term parameters for this variable among the counties.

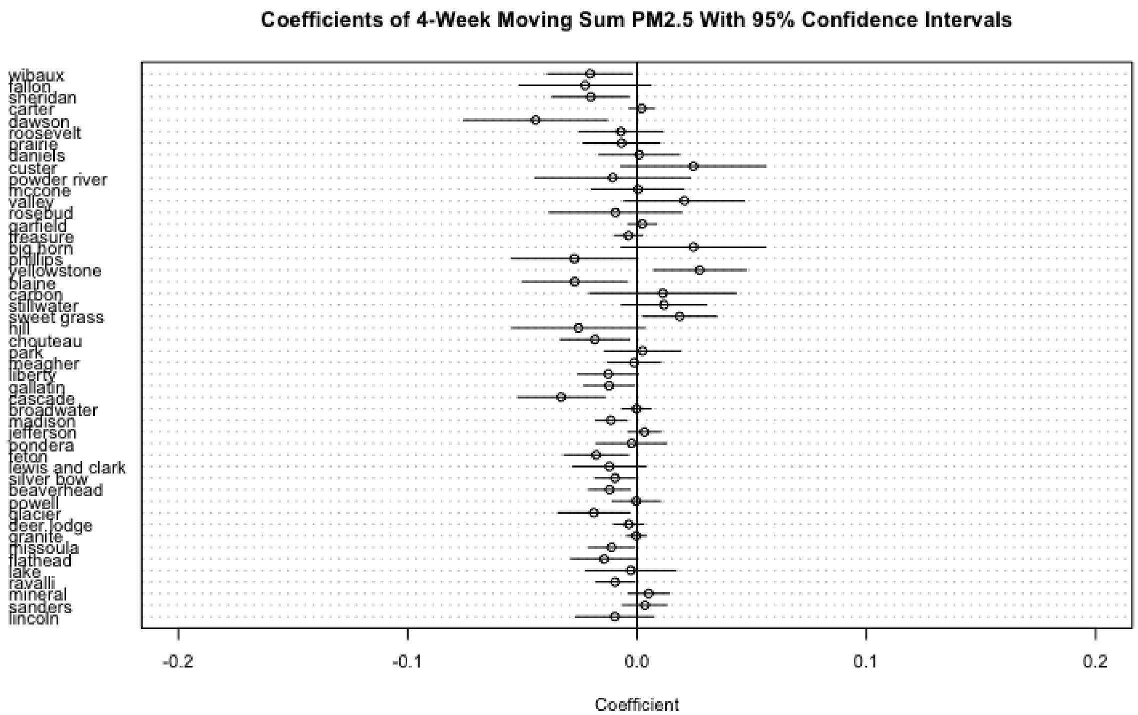
Figure 15 suggests a couple facts about the short-term PM2.5 effects. First, these models suggest that after accounting for the other variables in equation (1) there is very little that the short-term PM2.5 variables adds for these counties. Regardless of short-term function chosen for the model, the corresponding coefficients tend to have the similar estimates in terms of sign. For example take the first three counties in the confidence intervals in figure 15 (Wibaux, Fallon, Sheridan); regardless of short-term variable, all three coefficient estimates are negative. Further, we see that there is little evidence to suggest that these short-term variables are consistently influential across counties in predicting ILI counts. A majority of the 95% confidence intervals overlap 0, suggesting that these short-term effects tend to have very little effect if any on ILI counts. For all short-term PM functions, about half of the coefficient estimates are positive and half are negative. All functions are comparable in terms of residual deviance explained in the model from the short-term PM2.5 variable. Regardless of choice of short-term PM 2.5 function, they all tended to explain similar amounts of deviance in response. For simplicity, we keep the two-week moving window sum PM2.5 variable in the model as a predictor variable. Though the two-week moving window sum of PM2.5 was inconsistently influential across all Montana counties, there are still some counties which exhibit signs of a short-term effect (Carter County for example).


Coefficients of 2-Week Moving Sum PM2.5 With 95% Confidence Intervals
wibaux
fallon
sheridan
carter
dawson
roosevelt
prairie
daniels
custer
powder river
mccone
valley
rosebud
garfield
treasure
big horn
phillips
yellowstone
blaine
carbon
stillwater
sweet grass
hill
chouteau
park
meagher
liberty
gallatin
cascade
broadwater
madison
jefferson
pondera
teton
lewis and clark
silver bow
beaverhead
powell
glacier
deer lodge
granite
missoula
flathead
lake
ravalli
mineral
sanders
lincoln
-0.2
-0.1
0.0
0.1
0.2
Coefficient

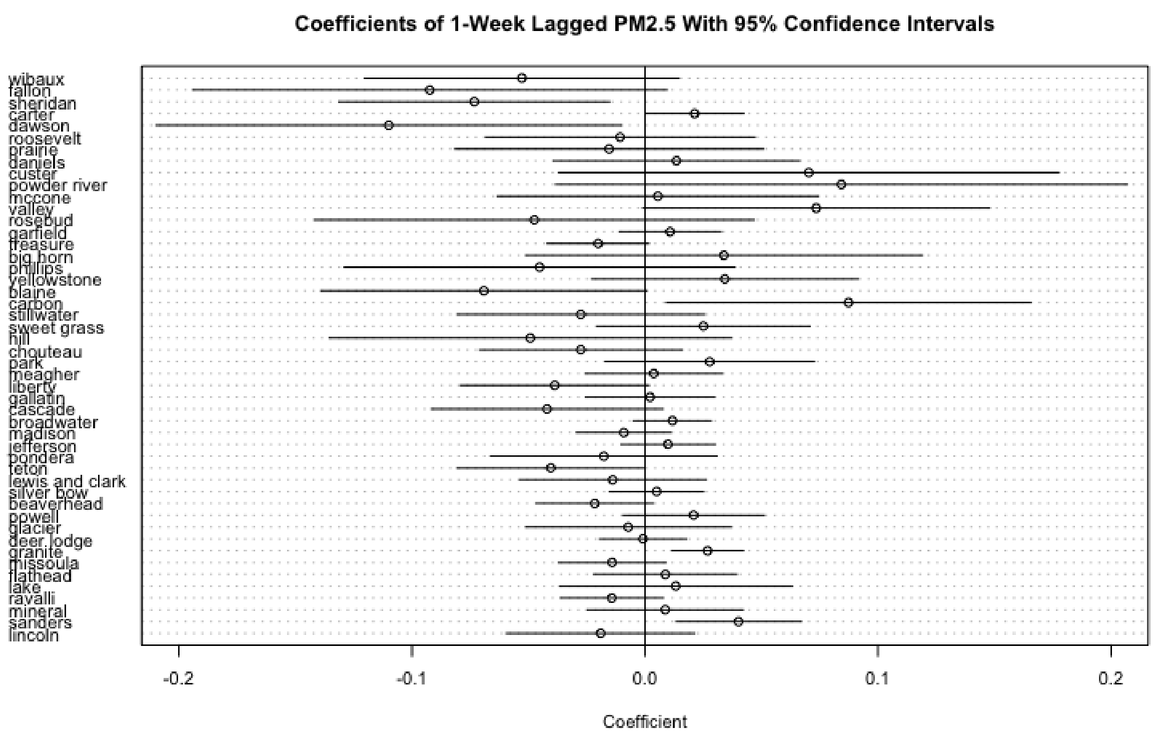
(a.)


Coefficients of 3-Week Moving Sum PM2.5 With 95% Confidence Intervals
wibaux
fallon
sheridan
carter
dawson
roosevelt
prairie
daniels
custer
powder river
mccone
valley
rosebud
garfield
treasure
big horn
phillips
yellowstone
blaine
carbon
stillwater
sweet grass
hill
chouteau
park
meagher
liberty
gallatin
cascade
broadwater
madison
jefferson
pondera
teton
lewis and clark
silver bow
beaverhead
powell
glacier
deer lodge
granite
missoula
flathead
lake
ravalli
mineral
sanders
lincoln
-0.2
-0.1
0.0
0.1
0.2
Coefficient

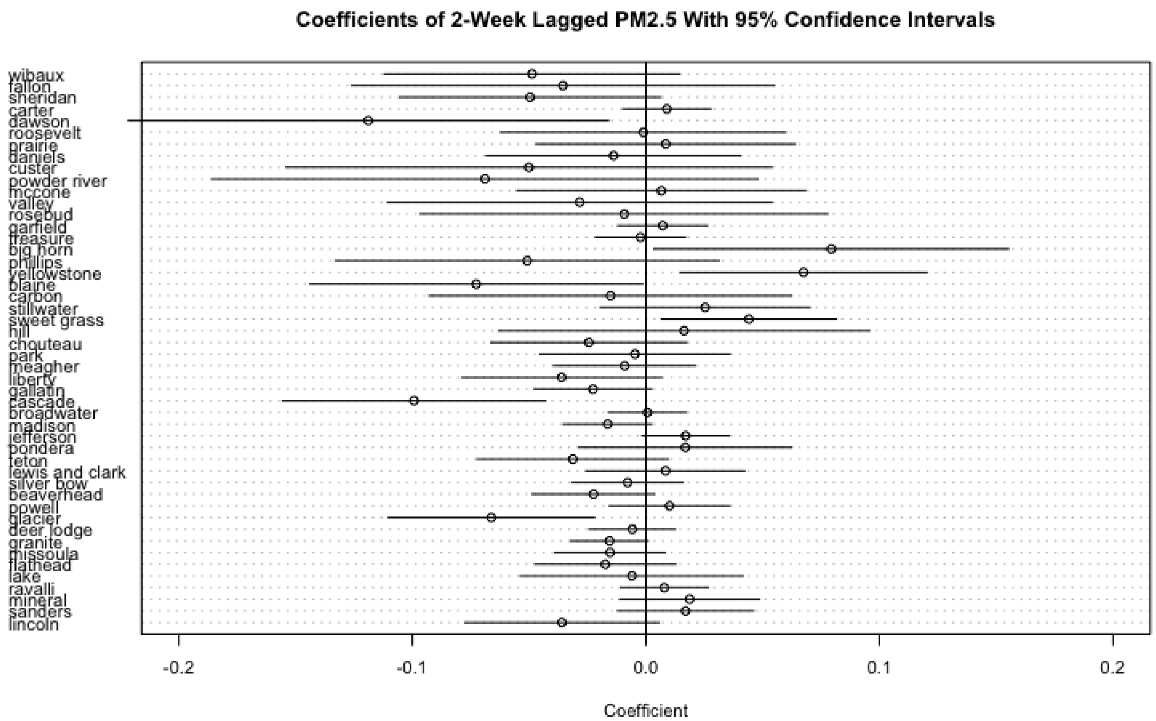
(b.)


Coefficients of 4-Week Moving Sum PM2.5 With 95% Confidence Intervals
wibaux
fallon
sheridan
carter
dawson
roosevelt
prairie
daniels
custer
powder river
mccone
valley
rosebud
garfield
treasure
big horn
phillips
yellowstone
blaine
carbon
stillwater
sweet grass
hill
chouteau
park
meagher
liberty
gallatin
cascade
broadwater
madison
jefferson
pondera
teton
lewis and clark
silver bow
beaverhead
powell
glacier
deer lodge
granite
missoula
flathead
lake
ravalli
mineral
sanders
lincoln
-0.2
-0.1
0.0
0.1
0.2
Coefficient

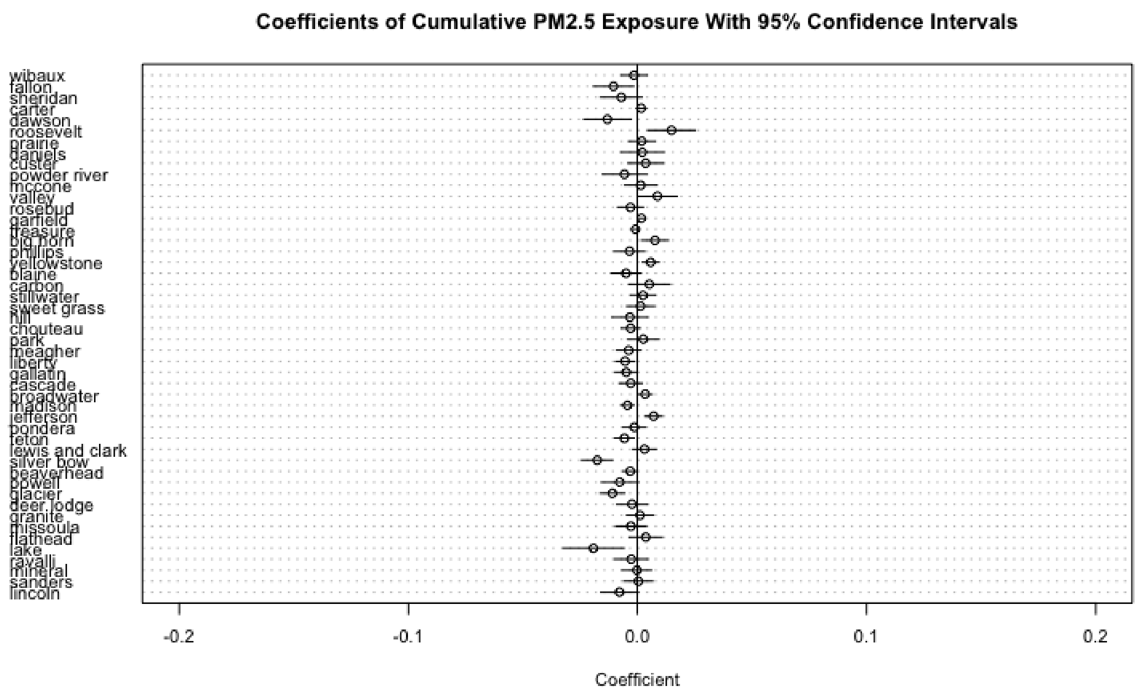
(c.)


Coefficients of 1-Week Lagged PM2.5 With 95% Confidence Intervals
wibaux
fallon
sheridan
carter
dawson
roosevelt
prairie
daniels
custer
powder river
mccone
valley
rosebud
garfield
treasure
big horn
phillips
yellowstone
blaine
carbon
stillwater
sweet grass
hill
chouteau
park
meagher
liberty
gallatin
cascade
broadwater
madison
jefferson
pondera
teton
lewis and clark
silver bow
beaverhead
powell
glacier
deer lodge
granite
missoula
flathead
lake
ravalli
mineral
sanders
lincoln
-0.2
-0.1
0.0
0.1
0.2
Coefficient

(d.)



(e.)



(f.)

Figure 15: (a) 2-Week PM 2.5 Moving Sum Coefficient Confidence Intervals (b) 3-Week PM 2.5 Moving Sum Coefficient Confidence Intervals (c) 4-Week PM 2.5 Moving Sum Coefficient Confidence Intervals (d) 1-Week PM 2.5 Lag Coefficient Confidence Intervals (e) 2-Week PM 2.5 Lag Coefficient Confidence Intervals (f) Cumulative PM2.5 Exposure Coefficient Confidence Intervals

Far more interesting than the short-term PM2.5 effects in this study are the apparent long-term PM2.5 effects. This is measured by the TEPM2.$5_{t,k}$ variable. After accounting for all other variables in equation (1), figure 16 displays 95% confidence intervals for the coefficients of the long-term PM 2.5 effects on Montana counties. It is clearly apparent that these estimates are predominantly positive for almost all Montana counties. All together, 47 of the 48 Montana counties modeled produced a positive coefficient estimate for this long-term effect (only Roosevelt county gave a negative estimate). Further, 35 of the counties had positive coefficients at the $\alpha = .10$ significance level. With a majority of coefficient estimates being positive, this provides our baseline evidence that summer PM2.5 density levels primary driven by wildfire smoke have a positive relationship with corresponding winter ILI incidence in Montana counties.

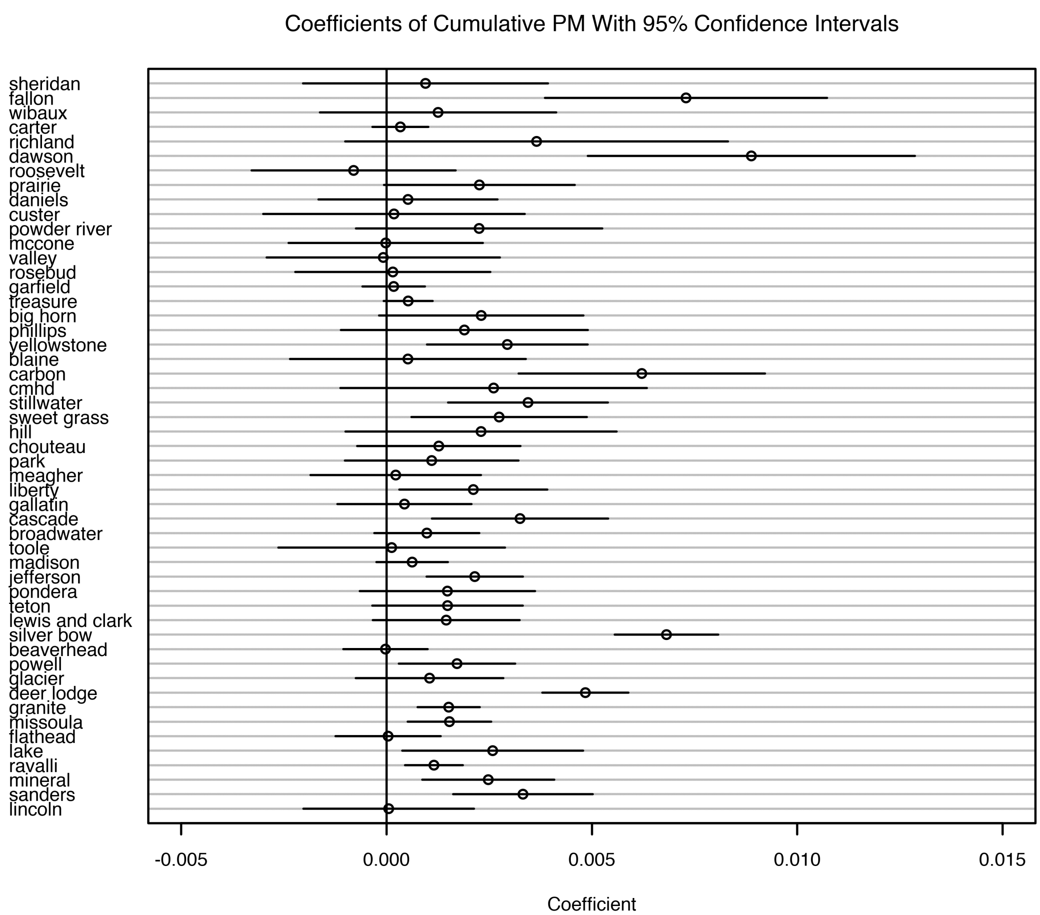
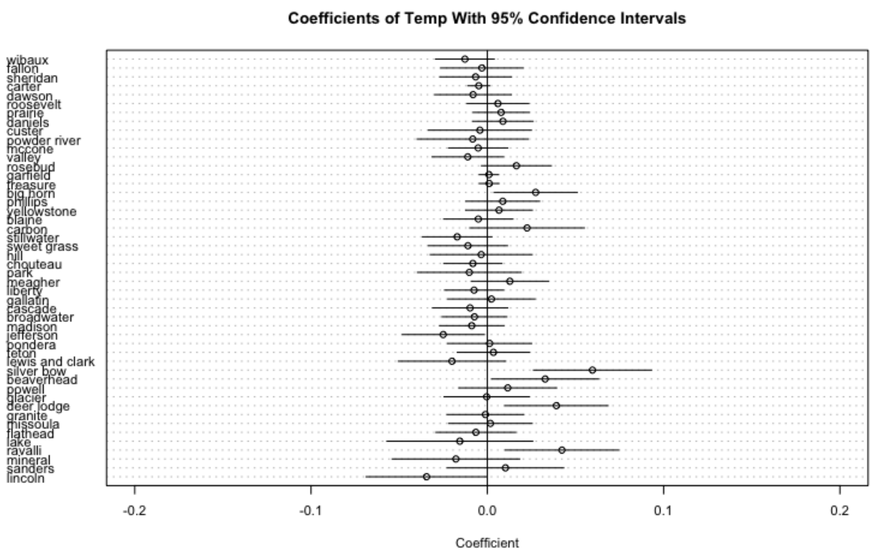
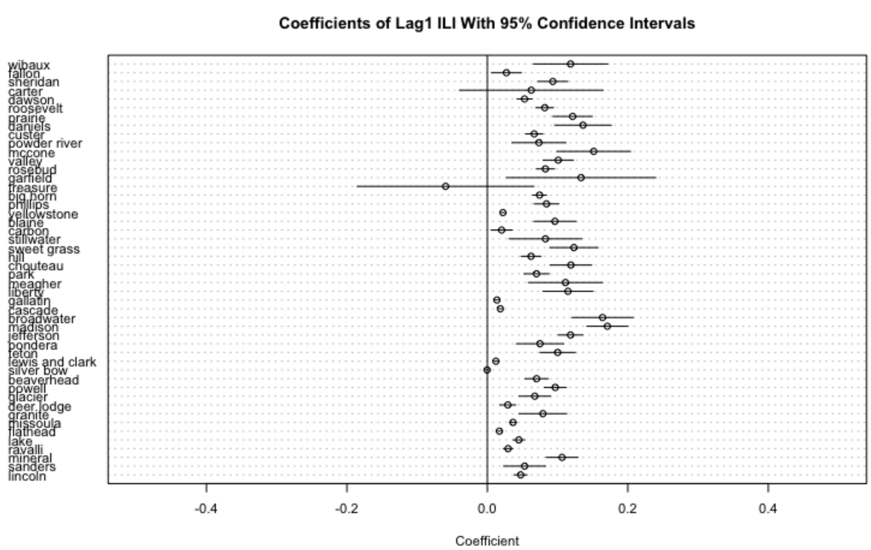


Figure 16: The estimated coefficients of the TEMP2.5 variable for every Montana county with 95% confidence intervals

Some other coefficients of interest including those corresponding to the time series lagged ILI counts and temperature are visualized in Figure 17. We can see that the relationships noted in the temporal analysis section earlier and from past research in the field persist in models for Montana. First, the majority of Montana counties show a negative relationship between temperature and ILI counts during influenza season as expected, although most parameter did not not differ from 0. Further, nearly all of the counties (with the exception of Treasure county) exhibit a strong positive autocorrelation with the previous week's ILI counts. This makes contextual sense in the setting of ILI modeling: knowing how many reports of ILI were made in the previous week gives us information about what will happen in the current week.



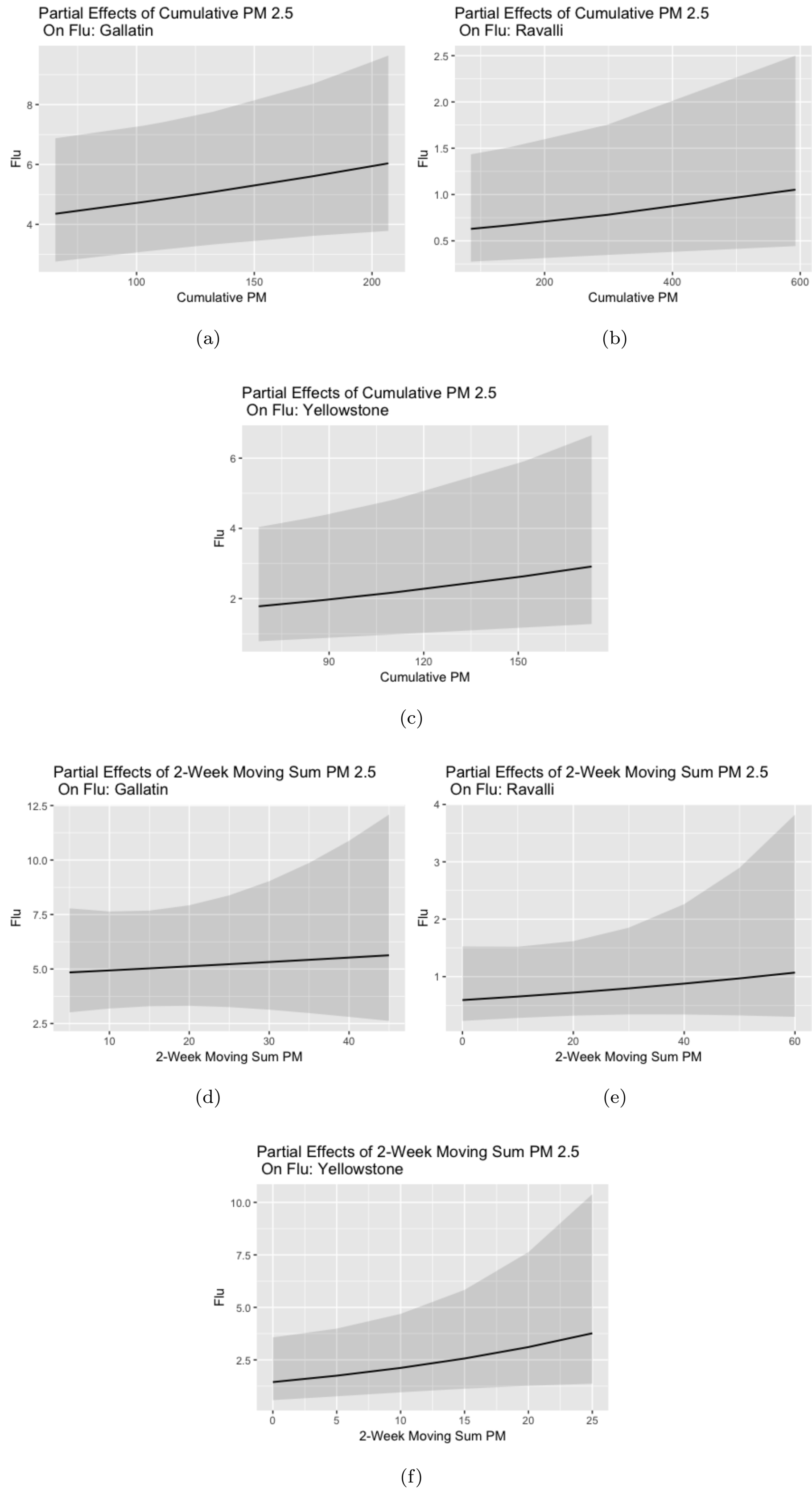
(a)



(b)

Figure 17: (a.) The estimated coefficients of the weekly temperature variable for every Montana county with 95% confidence intervals (b.) The estimated coefficients of the 1-week lagged ILI count variable for every Montana county with 95% confidence intervals

With these models in place, we can observe partial effects plots in Figure 18 resulting from the model fits. These plots allow us to view how the predictions of the ILI counts change for changing values of the covariates of interest holding all other covariates constant. We do this for a couple of the larger population counties in Montana to illustrate the effects of these variables. The positive nature of these associations is clearly observable; however the estimated standard errors sometimes make this positive association negligible.



(a) (b) (c) (d) (e) (f)

Figure 18: (a,b,c) Partial effects plots of Cumulative Summer PM2.5 Exposure on ILI counts for Gallatin, Ravalli, and Yellowstone County. (d,e,f) Accompanying partial effects plots for the 2-week moving sum PM 2.5 variable.

Model Performance

In this section we analyze several different model diagnostic checks to assess model performance. Among these are basic residual analysis where we will see if the residuals coincide with our assumptions, and some basic model fit statistics.

Residual Analysis

We can visually inspect the goodness-of-fit of these models via analysis of the resulting residual plots. In Figure 19, residual plots for three heavily populated Montana counties are displayed with accompanying residual density. The residual plots exhibit a large string of consistent residuals that occur in the bottom left of the plots. These patterns occurred in other studies [17] and indicate poor handling of zero count ILI cases. The issue of zero counts is difficult to address, but there are models (i.e. zero-inflated poisson models) that could possibly account for this issue. We see that these models tend to show larger over-estimations than underestimations. On the natural log scale, these over estimations can get as high as 30-40 cases too large. However, a majority of the residuals are within close proximity of the zero line indicating better model fit. The accompanying residual density plots reflect the nature of the abundant zero-counts by being centered slightly left of zero (indicating that in general we over-predicted). It is also clear that the residual density plots are somewhat right skewed.

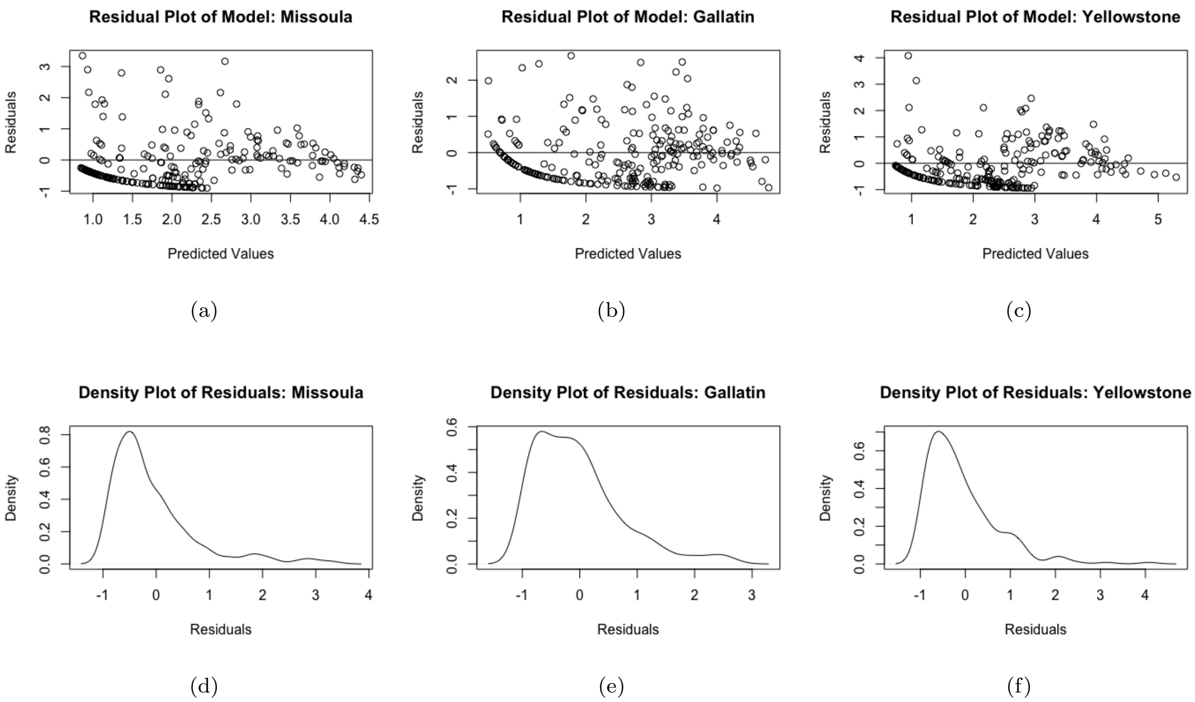


Figure 19: Residual plots from model fits in (a)Missoula County (b)Gallatin County (c)Yellowstone Coutny. Residual histograms from model fits in (d)Missoula County (e)Gallatin County (f)Yellowstone County

Assessing the goodness of fit quantitatively is still somewhat of an open question. There is no uniformly 'best' way to assess the goodness of fit. For this study, we will make use of a psuedo-R^2 statistic based on residual and null deviance defined by A. Colin Cameron [31] given in the equation below. This measure assesses the degree of improvement (in terms of decrease) in the model deviance given by including the covariates as opposed to only modeling using an intercept. Critics of this pseudo-R^2 suggest that it is easily inflated by over-parameterization; however with only 10 parameters and upwards of 450 observations per model, we may assume this issue does not arise here. Figure 20 shows the pseudo-R^2 for every county-level model fit.

$$R^2_{dev} = 1 - \frac{Residual\ Deviance}{Null\ Deviance}$$

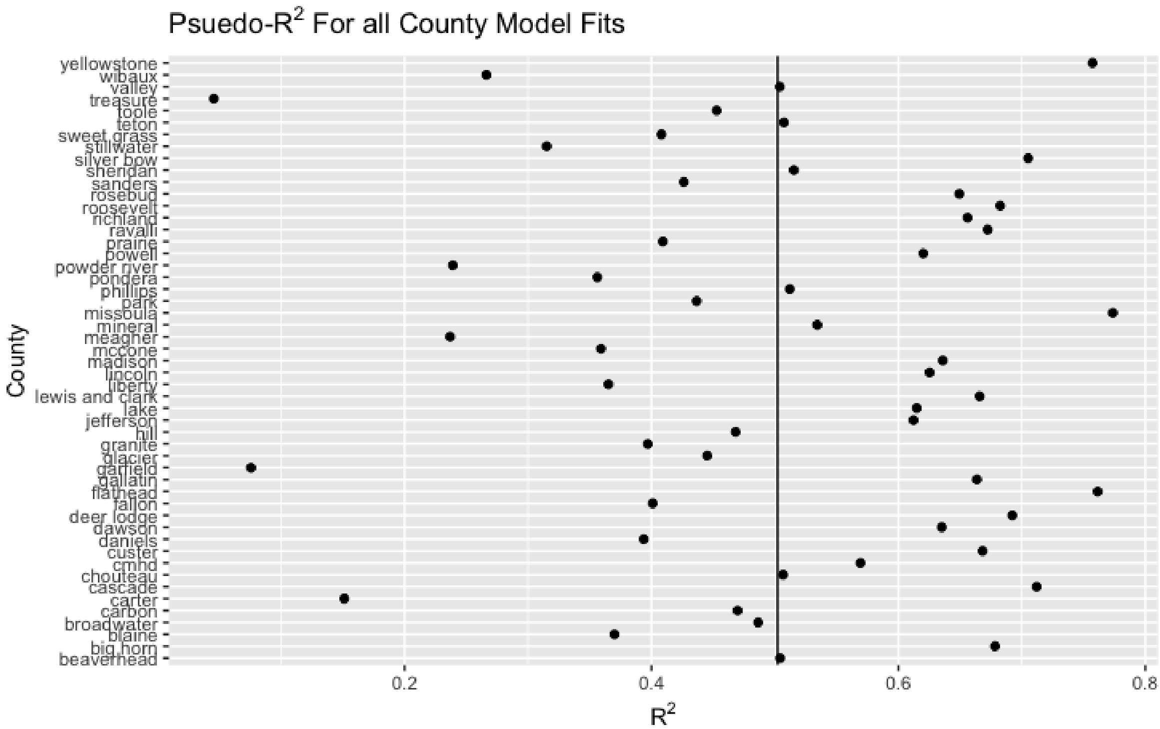


Figure 20: A plot of Pseudo-R^2 statistics. The vertical black line represents the average pseudo-R^2 for all county models at .501.

In general, we can expect for this model to account for roughly half of the variation in ILI counts for a given county in Montana. Granted, this is not a fantastic fit, but it is sufficient to show that the model form and covariates have some significant associations with the response. It is particularly interesting to observe that counties with very low pseudo-R^2 statistics tend to be very sparsely populated (Treasure county $R^2_{dev} = .045$, Garfield county $R^2_{dev} = .075$, Carter county $R^2_{dev} = .151$) and counties with very high populations tend to perform a bit better in terms of deviance explained (Yellowstone $R^2_{dev} = .757$, Missoula $R^2_{dev} = .773$, Flathead $R^2_{dev} = .761$). This reflects the difficulty in modeling small population sizes, where there are few reported ILI cases on a weekly basis, rarely exceeding 10 reports. This analysis suggests that this model has the ability to account for the variation in ILI counts for certain counties, but lacks the flexibilities in Montana counties with smaller populations and thus smaller reported ILI incidence.

Conclusions

In this section we conclude the analysis of ILI counts in Montana at the county level and their relationship to environmental covariates including PM2.5 density. We discuss the weaknesses of our study and possible future work that can remedy these weaknesses. Further we discuss the impact models of this type can have on the study of human health.

Weaknesses and Future Works

This study illustrated some of the difficulties of working with human health data in lightly populated and rural areas. Our model had difficulty accounting for the excess of weeks in which no ILI cases were reported, which possibly distorted the relationship between the ILI response and the covariates of interest. Remedies to this problem come in the form of zero-inflated models that account for zero-counts which come in the form of 'structural' zeros and 'count' zeros [32,33]. Other remedies to this problem included more auto-regressive time series components, which were not included here to guard against over-parameterization.

Another weakness is the segmented nature of the modeling. We constructed our models in such a way that we assume that county level covariates do not 'talk' to each other, i.e. they are independently behaved. Our exploratory analysis suggested that neighboring counties tend to not be spatially associated in our response of interest; however the are somewhat spatially related in some of the covariates (PM2.5 density, Temperature, etc.). A model with the flexibility to allow parameters from one county model fit to learn from other county model fits could be beneficial in improving our predictions and our understanding of county level interactions with respect to our ILI count response and covariates of interest. Recommended approaches to incorporating this information include Bayesian 'Gaussian-Process' modeling in which the intercept and slope of multi-level terms can be related through a gaussian distribution [34]. While on the topic, Bayesian methods would be fascinating to examine with respect to this problem so we could discuss the 'probability' of a positive PM2.5 coefficient effect rather than take the more conventional significance level method. Such Bayesian models have been lightly tested in our setting and in other research [16,17] but were not included in this study for the interest of length and time.

Summary

Our study examines the relationship between Influenza-Like-Illness in Montana and it's relationship to county-level covariates of interest such as PM2.5 density, temperature, and seasonality. We found that there is significant evidence of a positive effect between the amount of exposure to PM2.5 pollution during the wild-fire seasons and corresponding incidence in Influenza-Like-Illness counts for 49 of 50 counties available for modeling in Montana. We found little to no evidence of consistent short-term associations between PM2.5 and ILI counts during the typical influenza seasons. Further, our study demonstrated that there appears to be little spatial association in ILI counts at the county level in Montana, but we suggest rerunning spatial analysis on finer resolution scales (such as zip-code or census-tract level). Our proposed generalized linear model performs well in predicting ILI counts in higher population counties (with pseudo-R^2's as high as .79) but struggles with lower population counties. It is our intent that this research adds value to the knowledge of human health, as it pertains to Influenza-Like-Illness, in the following areas: The relationship of environmental covariates (specifically PM2.5 and temperature) with ILI, the spatio-temporal relationship of ILI in Montana, and model construction and evaluation techniques for similar studies.

Bibliography

[1] *Environmental Protection Agency*, https://www.epa.gov/pm-pollution/particulate-matter-pm-basics.

[2] Reid, E. Colleen et. al . 'Critical Review of Health Impacts of Wildfire Smoke Exposure'. *Environmental Health Perspectives.* 2016 Sep;124(9):1334-43.

[3] McClure, Crystal D. and Daniel A. Jaffe. 'US particulate matter air quality improves except in wildfire-prone areas'. *PNAS* July 31, 2018 115 (31) 7901-7906.

[4] Fortin, Jacey. 'Montana Battles Wildfires Amid a Severe Drought'. *The New York Times.* Sept. 7, 2017.

[5] *Environmental Protection Agency.* https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matter-pm, Accessed 19 March, 2019.

[6] Liu, Jia Coco et. al. 'Wildfire-specific Fine Particulate Matter and Risk of Hospital Admissions in Urban and Rural Counties'. *Epidemiology* 2017 Jan; 28(1): 77-85.

[7] *World Health Organization*, https://www.who.int/influenza/surveillance_monitoring/ili_sari_surveillance_case_definition/en/.

[8] Feng, Cindy et. al. 'Impact of ambient fine particulate matter (PM2.5) exposure on the risk of influenza-like-illness: a time-series analysis in Beijing, China'. *Environmental Health.* 2016; 15: 17.

[9] Anderson, Stacey. *Montana DPHHS*

[10] Urbanski SP, Reeves MC, Corley RE, Hao WM, Silverstein RP. Missoula Fire Lab Emission Inventory (MFLEI) for CONUS. Fort Collins, CO: Forest Service Research Data Archive. 2017. https://doi.org/10.2737/RDS-2017-0039

[11] Mesinger F, et al. North American regional reanalysis. Bulletin of the American Meteorological Society. 2006;87:343-360

[12] *Center For Disease Control*, 'The Flu Season' .https://www.cdc.gov/flu/about/season/flu-season.htm. July 12, 2018.

[13] "Annual Estimates of the Population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2015. 2015 Population Estimates. United States Census Bureau, Population Division. December 2015. Archived from the original (CSV) on December 23, 2015. Retrieved March 5, 2016.

[14] Dalziel, Benjamin D. et. al. 'Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities'. *Science* Vol. 362, Issue 6410, pp. 75-79. 05 Oct 2018.

[15] Feng, Cindy et. al. 'Impact of ambient fine particulate matter (PM2.5) exposure on the risk of influenza-like-illness: a time-series analysis in Beijing, China' *Environmental Health* 2016; 15: 17.

[16] Hooten, Mevin B. et. al. " Assessing North American influenza dynamics with a statistical SIRS model". *Spatial and Spatio-temporal Epidemiology 1.* (2010) 177?185.

[17] Imai, Chisato et. al. 'Time series regression model for infectious disease and weather' *Environmental Research* 142 (2015) 319?327.

[18] Rencher, Alvin C. and G. Bruce Schaalje. 'Linear Models in Statistics Second Edition'. *Wiley-Interscience.* 2008 pg. 514.

[19] Nelder, John; Wedderburn, Robert (1972). "Generalized Linear Models". Journal of the Royal Statistical Society. Series A (General). Blackwell Publishing. 135 (3): 370?384. doi:10.2307/2344614. JSTOR 2344614.

[20] Lewis. B. 'glm.fit'. *R Statistical Software.* https://www.rdocumentation.org/packages/scidb/versions/1.2-0/topics/glm.fit.

[21] Lowen, Anice C. and John Steel. 'Roles of Humidity and Temperature in Shaping Influenza Seasonality'. *Journal of Virology* 2014, 10.1128/JVI.03544-13.

[22] Popovic, Gordana. 'Interpreting Coefficients in GLMs'. *Environmental Computing.* http://environmentalcomputing coefficients-in-glms/.

[23] Ver Hoef, Jay M and Peter L. Boveng. 'QUASI-POISSON VS. NEGATIVE BINOMIAL REGRESSION: HOW SHOULD WE MODEL OVERDISPERSED COUNT DATA?' . *Ecology,* 88(11), 2007, pp. 2766?2772.

[24] Brookshire, Bethany. 'Statisticians want to abandon science?s standard measure of ?significance?'. *Science News* April 17, 2019.

[25] Cressie, Noel A. 'Statistics for Spatial Data'. *Wiley Series in Probability and Statistics.* 10 September 1993.

[26] Russell, Elizabeth S. et al. "Reactive School Closure During Increased Influenza-Like Illness (ILI) Activity in Western Kentucky, 2013: A Field Evaluation of Effect on ILI Incidence and Economic and Social Consequences for Families ". *Open Forum Infectious Diseases.* Volume 3, Issue 3. 25 May 2016.

[27] Yousey-Hindes, Kimberly M. "Neighborhood Socioeconomic Status and Influenza Hospitalizations Among Children: New Haven County, Connecticut, 2003?2010". Am J Public Health. 2011 September; 101(9): 1785?1789.

[28] Box, George E.P. and Gwilym M. Jenkins. “Time-Series-Analysis Forecasting and control”. *Wiley.* 1970.

[29] Dean, Roger T. “Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models” *Behavior Research Methods.* June 2016, Volume 48, Issue 2, pp 783?802.

[30] Penny, William D. 'Multiple Time Series Chapter 7'. *University College London.* https://www.fil.ion.ucl.ac.uk/ wpenny/course/array.pdf.

[31] Cameron, A Collin. “R-Squared Measures for Count Data Regression Models With Applications to Health Care Utilization”. *Journal of Business and Economic Statistics.* April 1995.

[32] Lambert, Diane (1992). ”Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing”. Technometrics. 34 (1): 1?14. doi:10.2307/1269547. JSTOR 1269547.

[33] McElreath, Richard. “Statistical Rethinking: A Bayesian Course with Examples in R and Stan”. Chapter 13. *CRC Press* Published December 22, 2015.