Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ July 31, 2020

Thomas Arthur Burke Date

Food Flows to Assess Community Exposure to *Escherichia Coli* O157:H7 in Romaine Lettuce

By

Thomas Arthur Burke
MPH

Department of Epidemiology

_____

Timothy L. Lash, DSc, MPH
Committee Chair

Food Flows to Assess Community Exposure to *Escherichia Coli* O157:H7 in Romaine Lettuce

By

Thomas Arthur Burke

Bachelor of Science, Microbiology
Kansas State University
2013

Thesis Committee Chair: Timothy L. Lash, DSc, MPH

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
MPH
in Epidemiology
2020

Abstract

Food Flows to Assess Community Exposure to *Escherichia Coli* O157:H7 in Romaine Lettuce

By Thomas Burke

Length:

In 2018, there were two E coli O157:H7 nationwide outbreaks tied to romaine lettuce consumption with product originating from Yuma, Arizona and the Salinas, CA growing regions.[1, 2] Principles in "First in, First out" (FIFO) and other industry practices to maximize sellable product may create conditions in which some consumers have more risk of consuming romaine subject to temperature abuse and/or subject to other food safety risks. Proximity to food distribution centers may affect quality and potential safety given transit time to the ultimate retail destination. This ecological study utilizes a Poisson regression to evaluate the hypothesis that proximity to primary distribution node (DN) (i.e. the immediate first destination of product originating from Yuma County, Arizona) influences case counts at the county level. The model utilizes the Food Flows dataset created by Lin et al to find primary distribution nodes and calculated proximity distances through Python. Average county temperatures in March 2018 were also included to account for its influence on cold chain integrity. Other covariates evaluated were age proportions for persons under 15 and over 60, which are ages of special vulnerability to illness from *E coli* O157:H7. The final model includes average March temperature, the exposure of interest, and an effect modifier of temperature on DN proximity. We find a relation between DN proximity and case counts in the Yuma outbreak, with a risk ratio of 5.86 (95% CI 3.08, 11.2) for an increase of 250 kilometers, while holding temperature constant. Overall, increased distance from a DN portends increased risk for cases in the Yuma outbreak. Higher temperatures were also associated with an increase in risk. Because of the inability to include potential confounders, ecological bias is a primary concern to the interpretation of the results. Further exploration using Bayesian methods for mapping supply chain risk may better account for inter county variations in underlying covariates. Public health entities should consider attaching supply chain characteristics as a component of data collection in epidemiologic analyses to improve evaluation of future food safety outbreaks.

Food Flows to Assess Community Exposure to *Escherichia Coli* O157:H7 in Romaine Lettuce

By

Thomas Arthur Burke

Bachelor of Science, Microbiology
Kansas State University
2013

Thesis Committee Chair: Timothy L. Lash, DSc, MPH

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2020

Table of Contents

Table of Figures

Table of Tables

Introduction and Concept

In recent years, the United States has seen significant foodborne outbreaks associated with fresh leafy greens, especially romaine lettuce. In 2018, there were two *Escherichia coli* O157:H7 nationwide outbreaks tied to romaine lettuce consumption with product originating from Yuma, Arizona and the Salinas, CA growing regions.[1, 2] Investigations of both of these outbreaks identified the pathogen strain and vehicle relatively quickly, but inefficiencies and industry practices necessitated officials to issue broad recalls of all romaine lettuce originating from these growing regions.[3] Through environmental assessments and review from the traceback investigations, it is hypothesized that contamination of leafy greens happens through irrigation water, being in close proximity with Concentrated Animal Feeding Operations (CAFOs).[1, 2] Due to a variety of factors including zoning laws, it is unlikely the systemic issues that cause these types of contamination will be resolved in the near future.[4] Fresh produce, in addition to these systemic factors, inherently is a riskier food group due to the lack of microbial kill steps.

Because of the ongoing risk to the population, there would be benefit in supplementing supply chain information for risk profiles in fresh produce. If food safety authorities could use supply chain characteristics or factors of the "built environment" of the food distribution system to establish risk profiles based on geography, surveillance systems and food industry practices could be improved. Herein is a converging of knowledge between supply chain management, food science of cold storage, and epidemiology to better understand how supply chains affect consumers.

Hypothesis

Fresh romaine consumed by Americans is primarily grown in two regions of the United States: Yuma, AZ and Salinas, CA. During their respective growing seasons, up to 75% of Romaine lettuce available to consumers originates from these regions.[5] The hypothesis for this project is that risk of consuming contaminated romaine is not uniformly distributed based on supply chain flows. Principles in "First in, First out" (FIFO) and other industry practices to maximize sellable product may create conditions in which some consumers have more risk of consuming romaine subject to temperature abuse and/or subject to other food safety risks. Proximity to distribution centers may affect quality and potential safety given transit time to the ultimate retail destination. Ultimately the research question is "Do industry practices based on the geography of distribution in food supply chains influence the case counts in county level data in the Yuma valley associated *E. coli O157:H7* outbreak in Romaine lettuce?"

Background

E coli O157:H7 and Outbreaks Associated with Leafy Greens including Romaine Lettuce

An often-cited statistic from the Centers for Disease Control and Prevention (CDC) is that the United States annually experiences 48 million illnesses, 128,000 hospitalizations, and 3,000 deaths resulting from foodborne pathogens.[6] Most types of food carry the potential for foodborne disease, but fresh food and vegetables are a prominent commodity type that have been implicated in multi-state foodborne illness outbreaks. Pathogens associated with fresh fruits and vegetables include *Escherichia coli, Clostridium spp., Bacillus cereus, Listeria monocytogenes, Salmonella,*

and Vibrio.[7] Among outbreaks identified by CDC, 390 outbreaks from 2003 to 2012 were associated with *E. coli*, including 4,928 illnesses, 1,272 hospitalizations, and 33 deaths.[8] In outbreaks of *E coli*, 65% came from consumption of food versus environmental and other avenues of infection.[8]

E coli has only been recognized as an important food safety pathogen since the 1980s.[9] Its primary reservoir is cattle, with primary transmission through consumption of contaminated food.[9] Transmission vehicles include food and water, direct human-to-human, and animal contact (including their environment).[8] Sources of contamination at pre-harvest include contaminated irrigation and surface water, contaminated soils, and introduction from domestic and wild animals.[7] The epidemiologic picture has changed over recent decades due to the confluence of several food factors.[10] International food commodity trade flows have influenced the shape and configuration of food supply chains.[10] Centralized and rapid food distribution systems, as well as food consumption patterns, have considerably influenced the epidemiology of produce associated foodborne disease outbreaks.[10]

As noted above, foodborne outbreaks associated with leafy greens, such as romaine lettuce, are complex public health problems. The short perishability, supply chain conditions, raw consumption, and production environment of romaine lettuce mean that the commodity has some inherent risk that requires public health infrastructure vigilance. Exemplifying this, there has been seasonality and regularity of romaine lettuce outbreaks since 1988 (seen in Figure 1).[5] When analyzing outbreaks of *E coli*, leafy greens accounted for 7% of all outbreaks from 2003-2012.[8] In CDC's studies of *E coli* outbreaks, leafy greens had the highest median outbreak size at

16 cases, with 0.8% of illnesses resulting in death, the highest rate among food categories studied[8] All foodborne E coli outbreaks in this analysis were multi-state.[8]

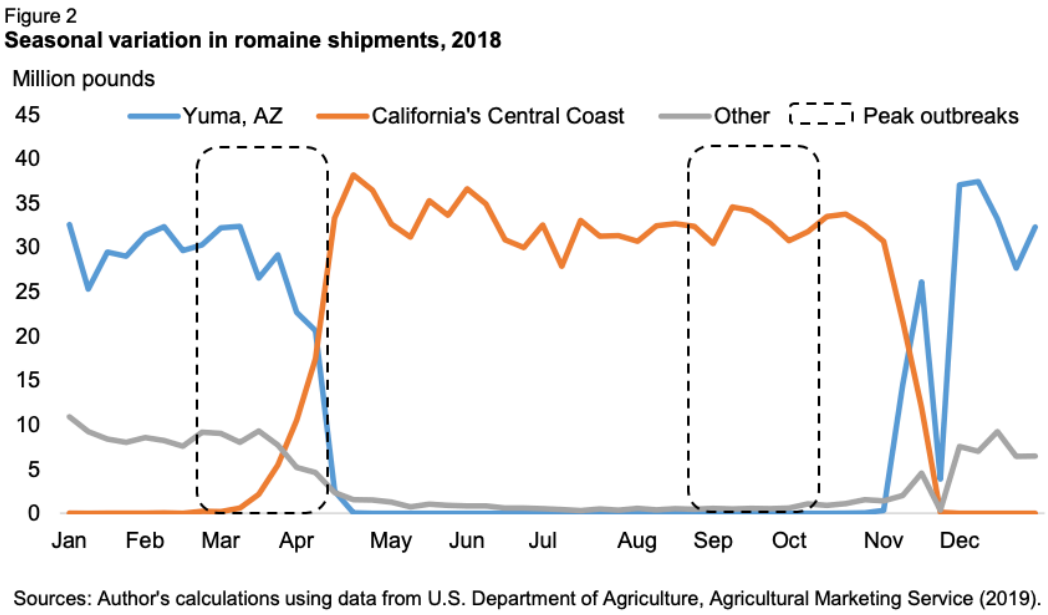


Figure 1: Seasonal Variation in Romaine Shipments, 2018 from USDA. [5].

From 1998-2018, outbreaks associated with romaine lettuce most frequently occurred in the months of March, April, September and October.[5] USDA also notes that Yuma and the California Central Valley account for the vast majority (>75%) of romaine production for the US.[5] These conditions create a scenario in which public health officials need to better understand how supply chain conditions may portend outbreak counts. Due to the entrenched challenges in seasonality, regularity, and rigidity of produce supply chains, being able to utilize knowledge on the built environment may improve surveillance and traceback investigations.

Despite advances in understanding *E coli* O157:H7, the epidemiology of the pathogen is still unclear.[9] The pathogen is ubiquitous in agricultural environments and may be spread through a

variety of species, wild and domestic. Cattle tend to shed *E coli* O157:H7 into the environment, with shedding peaking in the spring-summer months, but that knowledge does not necessarily extend to being able to better respond to outbreaks among food.[9] It is uncertain whether other species are also involved in its spread, and its survival on various foodstuffs is still being studied.[9] It has also been found to persist in the environment for some time, up to 6 months in sedimentary soils.[9]

Public Health Consequences

Challenging to the food safety considerations of *E coli* is the extremely low infective dose, meaning that any proliferation of the pathogen through a compromised cold chain may mean illnesses.[9] Severe illnesses resulting from *E coli* include hemorrhagic colitis (HC) and hemolytic uremic syndrome (HUS).[9] *E coli* variants are categorized by their virulence factors: enterotoxigenic, enteropathogenic, enteroaggregative, cytolethal distending toxin producing, and lastly enterohemorrhagic (EHEC), the subject to this study.[9] Raw foods with EHEC had higher rates of hospitalization.[8] Signs and symptoms include severe diarrhea (often bloody), severe stomach cramps, and vomiting.[8] Additionally, hemolytic uremic syndrome is a contributing factor to severe illness and death.[8] O157:H7 is an archetypal EHEC strain, producing verotoxin (Shiga-like toxins), which are responsible for the HC and HUS manifestations.[9] Because fresh produce is constantly exposed to the environment, it has a complex microflora picture, with microbial populations exceeding 10^8 CFU/G at times.[11] Coliform counts are known to be up to 10^4 CFU/g in surveyed products.[11] These can make routine screening and testing of food samples difficult, thereby making this study more relevant.[11]

Description of Supply Chains

When modeling perishable fresh vegetables, considerations that go into supply chain strategy include postharvest behavior of crops, effects of weather, transportation time, postharvest decay, labor, and delivery costs.[12] The enactment of the Food Safety Modernization Act and subsequent rules, such as the FSMA Final Rule on Sanitary Transportation of Human and Animal Food, show that a preventative model to retain cold storage of food has been a priority of the US government.[13] It is notable that leafy green vegetables have potential contamination points all along preharvest, harvest, processing, packaging, distribution, and preparation/consumption by the consumer.[7] [14] In handling fresh produce, temperature abuse can occur, which impacts its quality and microbial safety.[14] Postharvest, improper food handling by harvesters and others along the supply chain can contaminate produce.[7] In supply chain management, perishability is defined as the number of units of product which outdate or perish.[15] Decay is generally modelled in units lost rather than value lost.[15] Longer retail storage of lettuce is associated with *E coli* O157:H7 growth.[14] Long-term temperature abuse is easier to model than short term (i.e. defrost cycles of refrigeration units).[14]

The above information pertains to the overarching goal, which is to preserve the value of perishable crops as much as possible.[12] For produce brokers, there is a tradeoff between time to reach the consumer and cost.[12] To balance this tradeoff, cold storage, biocides, and other food processing techniques are used to preserve value while also maintaining food safety.[12] Along with market uncertainties and yield and maturation of perishable crops, the shelf life and logistics cold chain are limiting factors.[12] Food producers, distributors, and retailers use monitoring to manage fresh produce as it is moving through its cold chain.[14] However, it is

challenging to keep this cold chain consistent throughout the distribution chain especially when there are multiple handoffs of product along the value chain.[14] Other factors that may influence temperature abuse include the physical location in a truck and pallet on E coli growth.[14]

US agricultural and food trends have influenced the food safety picture especially regarding fresh produce.[16] With an emphasis on export markets, US agriculture moved to dependence on urban centers and coastal export facilities for transportation of goods, many of which are distant from the production region itself.[16] For fresh produce, the US primarily sees the highway system as the transportation means for distribution, with 94% of fresh vegetables being transported by truck.[16] Major shipping areas, as noted by the US Department of Transportation, are located along the coastal rim of the US: California, Florida, Texas, and the East Coast.[16] However, there is seasonal variation to these shipment patterns.[16] Generally, studies have found that cold storage is maintained, though this can be compromised in warmer climates.[14] Retail storage is where the most significant temperature abuse occurs and where 1-3 days of storage can occur.[14] Fresh cut romaine can have visual appearance of freshness while still being contaminated.[15]

Population at Risk

Host factors that may impact acquiring infection by STEC include age, immunity, health status, use of antibiotics and antimotility agents, stress, and genetic factors.[10] Food consumption patterns may also influence risk of foodborne illness resulting from *E coli* contaminated romaine lettuce.[10] Risk groups, and therefore confounding variables, include children (0-14) and the elderly (>60).[9]

Model Justification

Taking all of this background into account unveils a situation for which a model that analyzes supply chain characteristics can inform our understanding of risk on a county level in multi-state foodborne outbreaks. When building the model, factors that should influence the case counts include age, sex, temperature at the county of interest where retail storage occurs, and the exposure of interest, proximity to primary distribution node.

Methodology

This project models the supply chain characteristics that may contribute to an influence of county geographic location on county case counts in the 2018 *E coli* O157:H7 outbreak associated with romaine lettuce from Yuma county, Arizona. To account for the spatial relationship of the supply chain, the project includes the framework developed by Lin et al.[17] The Food Flow Framework was created to understand direct food flows on a county level basis in the United States, heavily leveraging the Freight Analysis Framework.[17] While the Food Flow Framework was explored as the primary spatially defined exposure, in this case, intensity of food freight at a county level, the ultimate research question is whether geographic location from primary distribution nodes (DNs) influences case counts in a multi-state outbreak.[18] Because the study is assessing this exposure at an aggregate level, it is an ecological study.

In forming the methodology, first referenced was Statistical Methods in Environmental Epidemiology by Thomas.[18] In the spatial methods section, Thomas references a methodology for accounting for spatial variation of natural radiation and its relationship to childhood leukemia

using a hierarchical Bayesian model with Poisson regressions accounting for local variability and a log-linear mixed model to account for the spatial structure between districts.[19]

With this as a beginning basis to build a methodology, the study aims were evaluated against the criteria for a justified ecological study design. In Chapter 25 of Modern Epidemiology, 3rd edition, there are 5 rationale for conducting ecological analyses with low cost and convenience, measurement limitations of individual-level studies, interest in ecologic effects, and simplicity of analysis and presentation being pertinent to this study.[20] The analysis takes advantage of secondary data sources to evaluate an ecologic effect (i.e. supply chain patterns on foodborne illness risk). Choices and practices among industry supply chain actors contribute to the overall system. This type of measure is not one that can be measured on an individual basis. Indeed, the analysis is particularly focused on the ecological effect from primary distribution; the supply chain distribution node from production and processing to regional distribution centers; and those patterns on the overall exposure profile. An ecological study is well justified for this study's hypothesis, given the above reasons.

One aim of this ecological study is to ascertain the contextual effect of supply chain topology on foodborne illness incidence in multi-state outbreaks traced to Romaine lettuce. The hope is that this model may be used for other multi-state outbreaks or aggregates of multi-state outbreaks to see if geographic location influenced the incidence of cases. As is known, ecological studies have application, but are liable to misinterpretation and/or bias. The chosen covariates are focused on reducing potential from unequal confounding distributions on case count frequency and geographic distribution.

More complicated models were explored before arriving at using a Poisson regression. In assessing spatial effects of an aggregate exposure, a hierarchical Bayesian model was used in several studies and methods papers.[19, 21, 22] Bayesian analysis uses prior knowledge (prior distribution) in combination with the sample distribution (posterior) to make an inference about the data. This is well suited in spatial analyses where varying levels of individual risk are uncertain but can be guessed.[23] The given model utilizes only a Poisson regression from a frequentist perspective with a view to explore a more complicated analysis in the future.

At its core, this study relies on a Poisson distribution, assuming the event of becoming ill from contaminated Romaine during this period to be a rare event. While foodborne illness is common, the vast majority come from cross-contamination and temperature abuse at the point of preparation or consumption, rather than food ingredients and products themselves.

Data Source Description

Population

As an ecological study, the population studied is the US population in the lower 48 states (excluding Alaska and Hawaii) in the 84 days of the Romaine lettuce outbreak. Because we are studying distribution of food from Yuma, Arizona, especially in the context of trucking traffic and supply chain practices thereof, the food flows to the non-contiguous states are not relevant for this study. Product shipped to Hawaii and Alaska utilize a different supply chain configuration relying on primarily air freight. County demographic information, including sex and age, were obtained through the US Census Bureau's American Community Survey.[24] These

data are estimates on a county level basis demarcated by age and sex.[24] By using this dataset, population data were obtained for every US county, identified through their FIPS code.[24, 25]

Food Flows Description and Primary Distribution Node Proximity

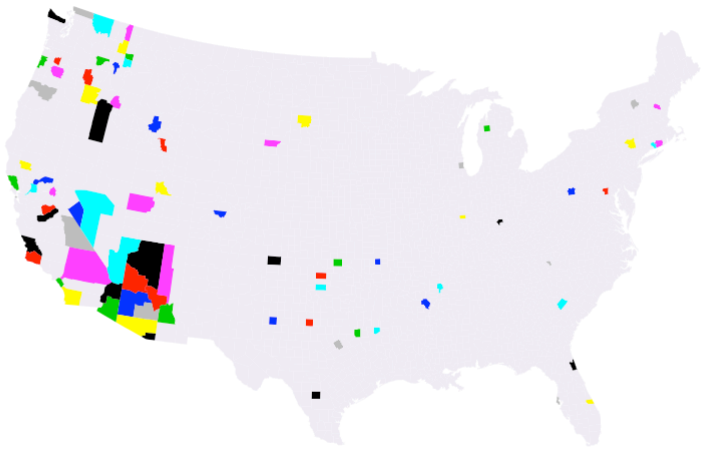
The food flows data from Lin et al. encompasses a comprehensive effort to describe food commodities on a county to county transfer basis. The intention of the database is to understand the structure of US food distribution. They have 7 SCTG codes which they modeled:

Table 1: Standard Classification of Transported Goods Used for Food Flows Data

SCTG	Model
1	Animals and Fish (live)
2	Cereal grains (including seed)
3	Agricultural Products (excludes animal feed, cereal grains, and forage products)
4	Animal feed, eggs, honey, and other products of animal origin
5	Meat, poultry, fish, seafood, and their preparations
6	Milled grain products and preparations, and bakery products
7	Other prepared foodstuffs, fats and oils

While these categories seem vague, for our purposes, they help articulate the picture of food flows coming from Yuma County. Restricting the food flow data to originating from Yuma county resulted in 81 counties (excluding Hawaiian counties), the map (Figure 2) below shows the presumptive primary distribution counties. This distribution is well aligned with our hypothesis. Primary and secondary processing would commonly occur in nearby counties, illustrated below. Additionally, there is to be expected primary distribution nodes leading north towards Canada. The other counties seen are near population centers across the US, though surprisingly few in the northern Midwest.

Figure 2: Primary Distribution Nodes from Yuma County Arizona



The data reveal the food flow from the Yuma region show diverse agricultural output, but with 7 the highest, which includes Romaine lettuce among products (Table 2). Standard Classification of Transported Goods Codes are at the right.

Table 2: Standard Classification of Transported Goods Volumes in Kilograms for Yuma Primary Distribution Nodes

SCTG	*Volume (in Kilograms)*
1	82822
2	204405717
3	189201926
4	208906867
5	0
6	57457286
7	495893478

While the flow volume data itself is interesting, this model is studying the effect of the food miles and cold chain integrity contributing to the risk of *E coli* infection in the Yuma outbreak. It is assumed that trade nodes are delivering a roughly equivalent amount of romaine lettuce per capita, though there may be some regional variation. To derive the proximity to nearest distribution node, Python was used to generate a temporary dictionary with key value pairs of DN FIPS code to distance for each given FIPS code in the dataset. After calculating each 81 distances, the algorithm selected the lowest one and its corresponding distance as the exposure.

Temperature Data

Because the clear influence of environmental temperatures on cold chain integrity, the average county temperature in March of 2018 was used as a covariate. These values were obtained from NOAA[26].

Model Analysis

Proposed Model

Given the exploratory nature of this study, a Poisson regression with counties as subunits and aggregate covariates was chosen to evaluate the influence between distance from primary distribution center and US county case counts in the 2018 Yuma outbreak. The basis for this model is below:

$$Pois(\lambda) = \frac{e^{\lambda^k}}{k!}$$

For counts y_i in area i (in this case US counties), there is an independently identically Poisson distribution of cases, with the expectation in area i as e_i. Multiplied by the θ_I as the area risk, we get:

$$y_i, iid \sim Pois(e_i\theta_i), \;\; i = 1, \ldots, n$$

In the literature review, potential confounding variables, associated with the exposure, which is a measure of supply chain configurations and thereby geographic distribution, and the outcome, county counts of *E coli* O157:H7. These covariates are described in the data dictionary, but are briefly described below in Table 3.

Table 3: Variable Descriptions for Model

Variable name in dataset	Variable Description	Reason for inclusion
Distances	Distance to Food Flow Center	Exposure of interest
Mar_y	March Average Temperature in 2018	Environmental temperature at point of retail was highlighted as potential reason for increased *E coli* growth in produce supply chains.
PropU15	County Proportion of population under age 15 in 2018	Severe illness among persons under 15 are more probable and thereby more likely to contribute to case counts
PropOver60	County Proportion of population over 60 in 2018	Severe illness among persons over 60 are more probable and thereby more likely to contribute to case counts
PropFemale	County Proportion of population that is female sex in 2018	Dietary habits and risk among women who are pregnant may contribute to greater risk of illness form *E coli*
Offset		
TotalPop	County population in 2018	Used for offset in Poisson Regression

The overall full model with covariates is below:

$$\ln(\lambda_i) = \ln\left(\frac{E(Y_i)}{\ell_i}\right)$$

$$= \beta_0 + \beta_1 Distances + \beta_2 Mar\_y + \beta_3 PropU15 + \beta_4 PropOver60$$

$$+ \beta_5 PropFemale + \gamma_1 PropOver60 * PropFemale + \gamma_2 PropU15$$

$$* PropFemale + \delta_1 Mar\_y * Distances$$

To model this in SAS 9.4, the "offset" is carried over to model the loglinear association of the expected value. Because the model is explicitly evaluating the effect on this outbreak, the duration of the outbreak was used to estimate the person time exposed to contaminated lettuce from this specific public health event. The offset in this model is Person-Years calculated below:

$$\ell = TotalPop * \left(\frac{84}{365}\right)$$

With the offset, the model being derived from the dataset is thus:

$$\ln(E(Y_i)) = \beta_0 + \beta_1 Distances + \beta_2 Mar\_y + \beta_3 PropU15 + \beta_4 PropOver60 + \beta_5 PropFemale + \gamma_1 PropOver60 * PropFemale + \gamma_2 PropU15 * PropFemale + \delta_1 Mar\_y * Distances + \ln(\ell)$$

The null and alternative hypotheses for this model are:

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

The model is seeking to determine whether the distance to primary distribution nodes of the food flows from Yuma County Arizona explains variation in county-level case counts within the 2018 *E coli* 0157:H7 outbreak associated with romaine lettuce originating from Yuma County while controlling for local county average temperature in March, vulnerable population proportions (under 15 and over 60), and proportion of female sex. These group level proportions to adjust for confounders have been used in other spatial regression analyses.[27, 28]

Analysis of Exposure

The Lin et al. food flows dataset was chosen to evaluate the research question about the influence of supply chain topology on the occurrence of cases in multistate foodborne disease outbreaks, especially in fresh produce. The supply chains of fresh produce, especially like the products implicated in this outbreak, follow a fairly uniform pattern of Critical Tracking Events. Produce is harvested at farms, then packed and shipped to copackers, processors, and distributors. At these junctures, fresh produce is comingled and copacked with a multitude of

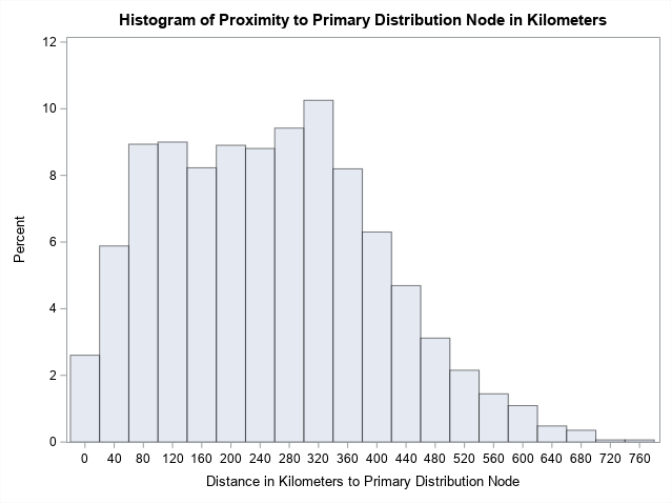
potential sources (though all from Yuma County). These are the points where products enter their form in which consumers will eventually consume the product. After this, cold chain storage, shelf life, practices among distributors, and environmental factors may influence the ability of *E coli* to proliferate and make a product more likely to create a case of foodborne illness.

This variable was calculated through the use of midpoint data of the latitude and longitude of county midpoints. The algorithm, described in the Python code in Appendix 2, finds the nearest DN and calculates the distance between county midpoints. Below the variable is described in Table 4 and Figure 3.

Table 4: Basic Statistical Measures for Distribution Node Proximity in Kilometers

Basic Statistical Measures			
Location		**Variability**	
Mean	253.8	**Std Deviation**	146.2
Median	247	**Variance**	21376
Mode	0	**Range**	761
		Interquartile Range	219

Figure 3: Histogram of Distribution Node Proximity in Kilometers



The most common value is 0 at 81, which accounts for the 81 DNs identified through the food flows dataset. The median and mean are centered around 250 kilometers. The data seem relatively normally distributed, though with a slight right skew. There appear to be no outright outliers. The three highest values are in Montana.

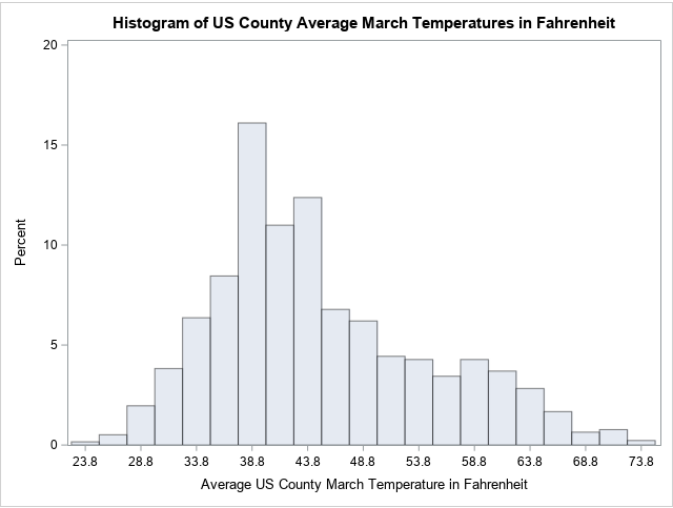
Analysis of Covariates

Temperature in March

Table 5: Statistical Measures for Average County Temperatures in Degrees Fahrenheit for March 2018

Basic Statistical Measures			
Location		**Variability**	
Mean	44.8	**Std Deviation**	9.66
Median	42.7	**Variance**	93.3
Mode	43.4	**Range**	51.2
		Interquartile Range	12.4

Figure 4: Histogram for Average County Temperatures in Degrees Fahrenheit for March 2018



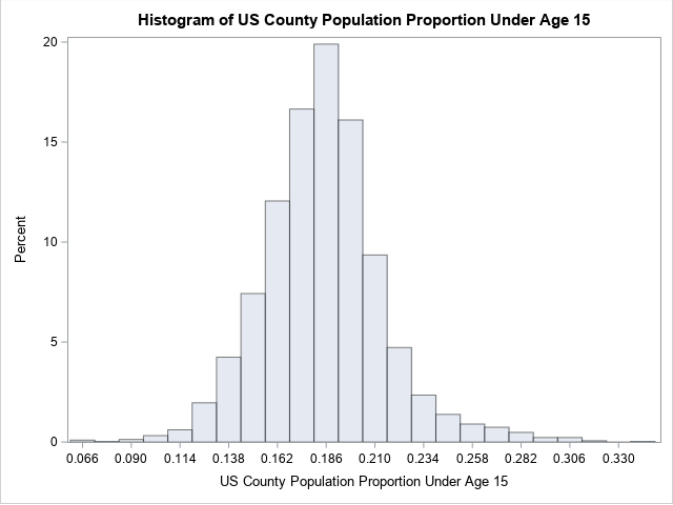
March temperature is skewed right, which makes sense given a small fraction of southern US lower 48 counties having relatively warm temperatures in March. There are no extreme outliers and the mean and median are close to each other, suggesting uniformity in the distribution. Temperature was kept in Fahrenheit so that all temperatures would be positive for ease of parameter interpretation.

Proportion Under 15

Table 6: Statistical Measures for US County Proportion of Population Under Age 15 in 2018

Basic Statistical Measures			
Location		**Variability**	
Mean	0.1844	**Std Deviation**	0.02991
Median	0.1838	**Variance**	0.0008948
Mode	0.1713	**Range**	0.2815
		Interquartile Range	0.03332

Figure 5: Histogram for US County Proportion of Population Under Age 15 in 2018



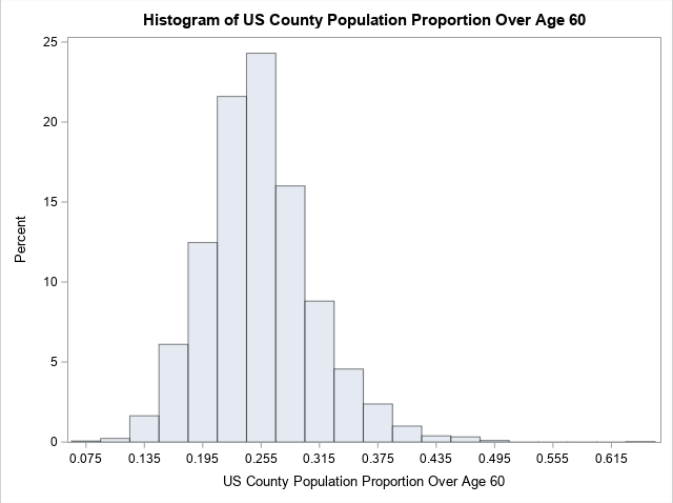
Proportion of the population under the age of 15 is normally distributed with a mean centered at around 20%.

Proportion Over 60

Table 7: Statistical Measures for US County Proportion of Population Over Age 60 in 2018

Basic Statistical Measures			
Location		**Variability**	
Mean	0.2532	**Std Deviation**	0.05606
Median	0.2493	**Variance**	0.00314
Mode	0.1846	**Range**	0.5857
		Interquartile Range	0.06577

Figure 6: Histogram for US County Proportion of Population Over 60 in 2018



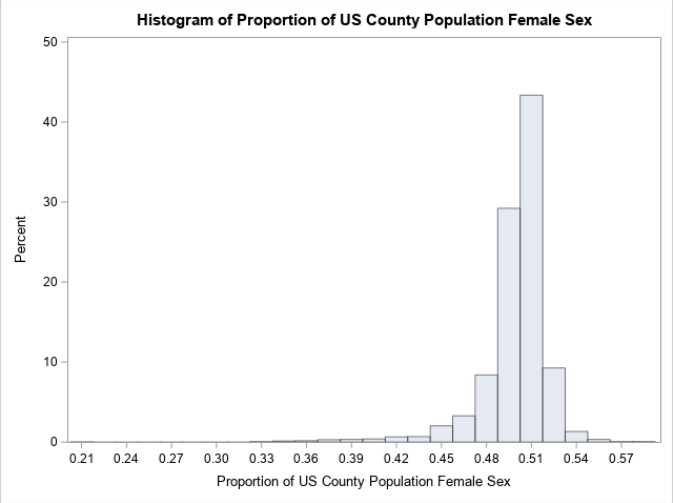
Proportion over 60 have some outliers, mostly due to retirement communities in Florida, but the overall distribution of the proportion over 60 follows a normal distribution. None of the outliers appear to be due to error.

Proportion Female Sex

Table 8: Statistical Measures for US County Proportion of Population Female Sex in 2018

Basic Statistical Measures			
Location		**Variability**	
Mean	0.4995	**Std Deviation**	0.02353
Median	0.5040	**Variance**	0.0005538
Mode	0.5000	**Range**	0.3761
		Interquartile Range	0.01685

Figure 7: Histogram for US County Proportion of Population Female Sex in 2018



The proportion of female sex have fairly expected values with some extreme values skewed towards high male populations, which is known, especially in rural US counties. These outliers also do not seem to be in error.

Analysis of Case Count Data

Table 9: Frequency Data by US County

Yuma Outbreak Case Counts				
count	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	2999	96.40	2999	96.40
1	71	2.28	3070	98.68
2	14	0.45	3084	99.13
3	13	0.42	3097	99.55
4	7	0.23	3104	99.77
6	2	0.06	3106	99.84
7	2	0.06	3108	99.90
9	2	0.06	3110	99.97
10	1	0.03	3111	100.00

Table 10: Statistical Measures for US County Case Counts

Basic Statistical Measures			
Location		**Variability**	
Mean	0.0707	**Std Deviation**	0.498
Median	0	**Variance**	0.248
Mode	0	**Range**	10.0
		Interquartile Range	0

Data that are publicly available on foodborne disease outbreaks are typically aggregated to the state level. The data obtained from CDC gave more granular detail on the geographic location of cases included in the outbreak workup. When fitting this model, all possible exposed counties were included besides Alaska and Hawaii. 202 cases were included in the dataset, with 8 cases from Alaska excluded. The CDC case count data also included environmental samples, which were also excluded. Figures 8 and 9 show the epidemiologic profile of this outbreak, with the somewhat prolonged epidemic curve and the diverse geographic picture in the case count map.

Figure 8: Epidemic Curve for 2018 E Coli O157:H7 Outbreak Associated with Romaine Lettuce from Yuma Arizona[2]

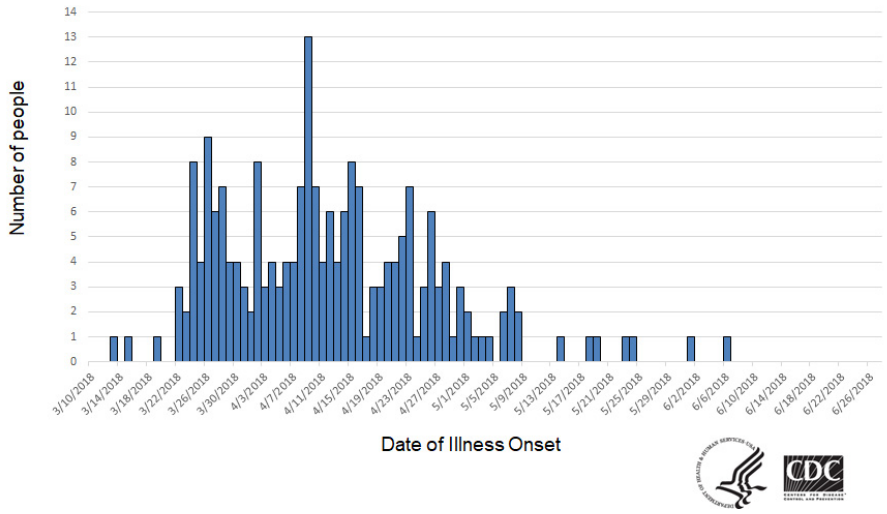
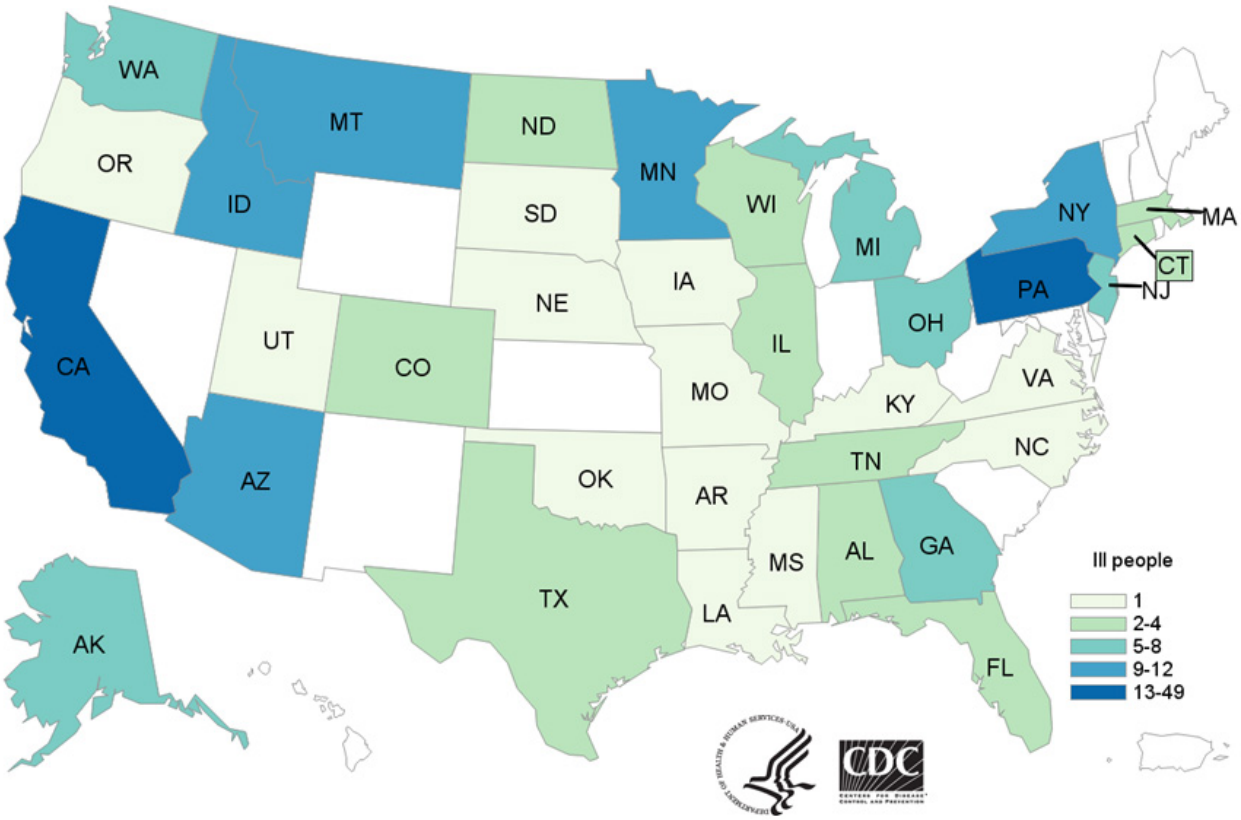


Figure 9: Case Count Map for 2018 E Coli O157:H7 Outbreak Associated with Romaine Lettuce from Yuma Arizona[2]



Model Building

First collinearity was assessed through the use of the "Collin" macro analyzing the covariate matrix generated from proc gen mod. In running the collinearity macro, it was found that all population proportions had collinearity issues with the model. Various attempts to reduce collinearity were conducted including combining "at risk" population proportions (under 15 and over 60) into one variable and multiple log transformations. Each resulted in condition indices (CNIs) over 30. To preserve parameter estimate stability, these variables were excluded from the model.

After accounting for collinearity, the remaining variables in the model include ND distance and March average temperature. Interaction of temperature and DN proximity was examined due to the potential influence of higher temperatures on compromising cold chain integrity. The effect modifier was found to be significant using a Wald test. The SAS code can be found in Appendix 1. The final model after assessing for interaction is below.

$$\ln(E(Y_i)) = \beta_0 + \beta_1 Distances + \beta_2 Mar\_y + \delta_1 Mar\_y * Distances + \ln(\ell)$$

The resulting output for the model is below:

Table 11: Parameter Estimates for Final Model

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-14.6598	0.2516	-15.1529	-14.1668	3395.89	<.0001
distances	1	0.0071	0.0013	0.0045	0.0096	28.93	<.0001
mar_y	1	0.0427	0.0051	0.0327	0.0527	70.11	<.0001
distances*mar_y	1	-0.0002	0.0000	-0.0002	-0.0001	31.82	<.0001
Scale	0	0.5284	0.0000	0.5284	0.5284		

Table 12: Goodness of Fit Statistics for Final Model

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	3107	867.4183	0.2792
Scaled Deviance	3107	3107.0000	1.0000
Pearson Chi-Square	3107	2874.8931	0.9253
Scaled Pearson X2	3107	10297.5616	3.3143
Log Likelihood		-1580.7582	
Full Log Likelihood		-567.4793	
AIC (smaller is better)		1142.9587	
AICC (smaller is better)		1142.9716	
BIC (smaller is better)		1167.1295	

The deviance over degrees of freedom metric shows an underdispersed dataset that is opposite the issue usually seen in Poisson regressions. To account for the underdispersion, the Deviance/DF statistic was fixed to 1.

Analysis Conclusion

Due to lack of additional attributes to the underlying case count data, proportions of at-risk groups were included in the model building to account for their confounding with the exposure and the outcome. Because the exposure, which is proximity to primary distribution nodes, and the outcome, which are case counts of *E coli* O157:H7, could both be associated with age and sex demographics, they were chosen to be examined as part of the model. However, in assessing for collinearity, all of these proportions could not be included. This collinearity is not entirely surprising, but it brings some caveats to the interpretation of the model, especially when trying to

account for ecological bias. When including the exposure and the confounders that were not collinear, we arrive at a model that does show that distance from primary distribution node contributes to the frequency of case counts seen in the 2018 *E coli* O157:H7 outbreak when controlling for environmental temperature. Below are a range of expected values with various temperatures and distances.

Table 13: Expected Values with Hypothetical Temperature and Distance Data with Varying Person-Time

Temperature (Fahrenheit)	Distance (km)	Expected Value (1 PY)	Expected Value (1 million PY)
30	0	1.55E-06	1.55
30	250	4.09E-05	2.04
30	500	0.00108	2.68
35	0	1.92E-06	1.92
35	250	6.50E-05	1.96
35	500	0.00221	2.01
40	0	2.37E-06	2.37
40	250	0.000103	1.89
40	500	0.00451	1.51

Figure 10: Graph of Expected Value of 1 for 100,000 PY

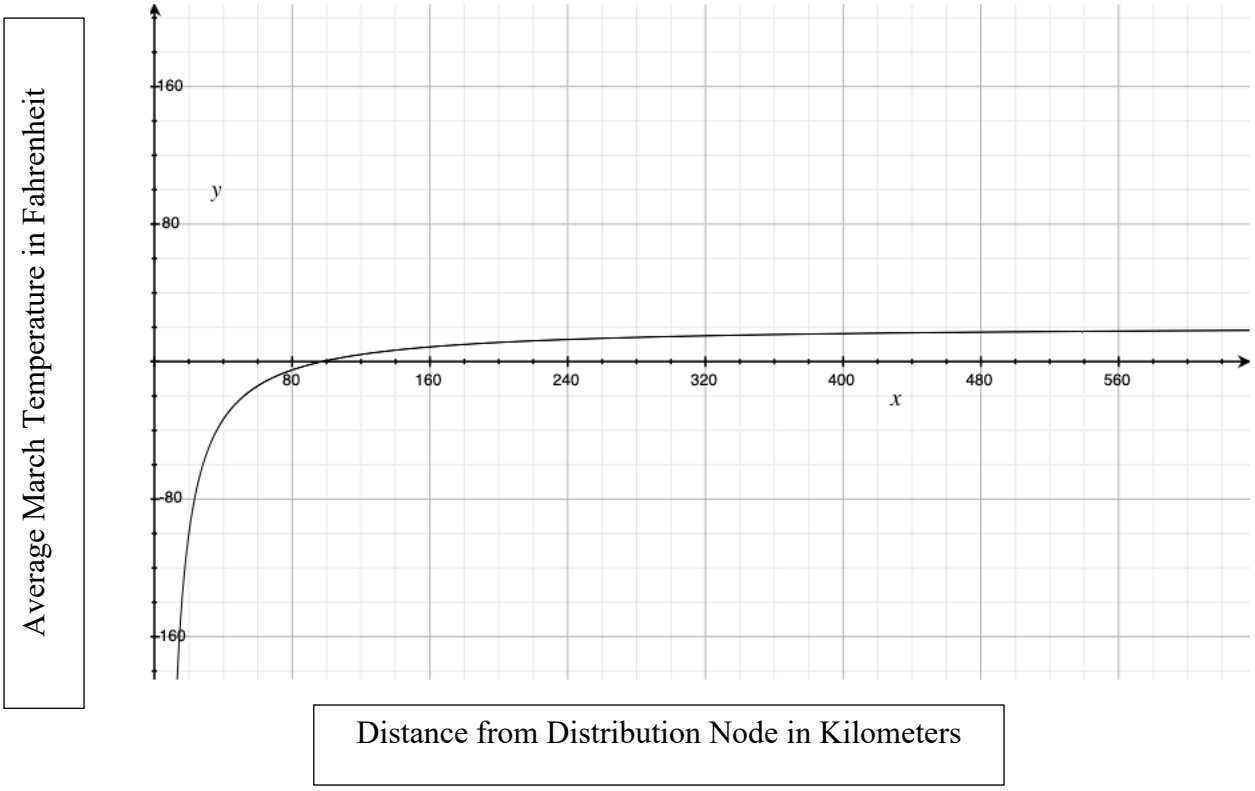


Table 14: Risk Ratios for Hypothetical Temperature and Distance Data

Contrast Estimate Results									
Label	**Mean Estimate**	**Mean Confidence Limits**		**L'Beta Estimate**	**Standard Error**	**L'Beta Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
RR 0 km	1.00	1.00	1.00	0.00	0.00	0.00	0.00	.	.
RR 250 km	5.86	3.08	11.2	1.77	0.329	1.12	2.41	28.93	<.0001
RR 500 km	34.3	9.47	125.	3.54	0.657	2.25	4.82	28.93	<.0001
RR 0 km 5 F	1.24	1.18	1.30	0.214	0.0255	0.164	0.26	70.11	<.0001
RR 250 km 5 F	5.92	3.21	10.9	1.78	0.312	1.17	2.39	32.50	<.0001
RR 500 km 5 F	28.3	8.64	92.6	3.34	0.605	2.16	4.53	30.51	<.0001
RR 0 km 10 F	1.53	1.39	1.69	0.427	0.051	0.327	0.528	70.11	<.0001
RR 250 km 10 F	5.98	3.34	10.69	1.79	0.297	1.21	2.369	36.37	<.0001
RR 500 km 10 F	23.3	7.87	69.0	3.15	0.554	2.06	4.23	32.32	<.0001
RR 0 km 20 F	2.35	1.92	2.87	0.855	0.102	0.65	1.05	70.11	<.0001
RR 250 km 20 F	6.10	3.58	10.37	1.81	0.271	1.28	2.34	44.51	<.0001
RR 500 km 20 F	15.8	6.46	38.6	2.76	0.456	1.87	3.65	36.60	<.0001

Discussion

From the start of this project, ecological bias was the primary concern. Given that the data were mostly count data and the units of analysis are counties rather than individuals, being able to account for county to county variations in confounders is important. As Wakefield describes, covariates in aggregate counties could easily account for the association between the exposure and the outcome.[22] In this analysis, we have not been able to include population demographics as a constituent to the model, which may influence the susceptibility of a county to have cases in a multi-state outbreak. Additional covariates could also conceivably influence this model, such as number of physicians, laboratory testing capacity, socioeconomic status, or number of retail stores. These attributes were not found in the literature as potential confounders, but the list illustrates how ecological bias may be introduced into this model.

In epidemiology, studies try to model the counterfactual; attempting to account for most to all plausible factors to elucidate the causality between the exposure and the outcome. Here, the model has one intended measure of interest, proximity to DN, with one confounder, average county temperature in March, and an effect modification term between these two variables. In terms of the variables that may truly impact the susceptibility of supply chain configurations in case counts, these two factors do have the potential to give a model that gives us some useful information, but the model only gives enough for us to provide a path forward for further investigation.

As described above, the interpretation of this model is complicated by the inability to model some probable confounders and interaction terms. Looking at the final model, when controlling for average US county temperature in March, a 250-kilometer difference in primary DN proximity results in a risk ratio of 5.86 (95% CI 3.08, 11.2) of having a case of *E coli* 0157:H7 in Romaine lettuce coming from Yuma County, Arizona. Both DN proximity and temperature are positively associated with increased case counts, but there is negative interaction between these two. As ambient temperature increases, the expected value of cases decreases for similar distances. This model does corroborate findings in the literature review where integrity of the cold chain increases as both distances and environmental temperatures increase.

For generalizability, this model was based on a specific outbreak for a reason. The contamination of product was at the point of production but was not necessarily localized to a single or few farms. The prevailing hypothesis for the origins of the Yuma Romaine Outbreak is contaminated irrigation water resulting in product containing *E coli*. In the traceback investigation, the product was only deduced to be produced in Yuma County Arizona; there was no single implicated product type, processing facility or other convergence point. This outbreak gave a more generalizable picture of the influence on supply chain factors than other outbreaks for these reasons. Products of a variety of types were being shipped to every corner of the United States. Ultimately, the preclusion of underlying covariates in the aggregated jurisdictions does not allow us to really make a claim on the magnitude or direct effect of these factors but can give us a path forward to investigate further.

Even without elucidating the true causality, the model does suggest there may be a relationship between supply chain configuration and temperature that could have implications for bettering food safety systems. First, there has been a reevaluation in the food system with the advent of Covid-19 on the reliance of highly integrated, consolidated supply chains towards more localized food processing. Second, cold chain integrity still seems to be a challenge for fresh produce coming from Yuma County.

This study has several caveats including data collection and model complexity. The aggregate level statistics including case counts, temperatures, and supply chain information are crude measures. The model is also perhaps too simple for the ultimate research question.

The initial model helped delineate whether a relationship exists between proximity to supply chain primary distribution node and a county experiencing cases in a multi-state outbreak among Romaine lettuce originating from Yuma County, Arizona. As explored in the methods section, using Bayesian methods over a frequentist approach may better model this data, providing a better method of modeling aggregate level confounders. This project also could inform the approach that epidemiologists take when considering data collection. By systematically incorporating supply chain data, surveillance systems and investigation techniques could be more effectively modeled.

Acknowledgements

The author would like to thank several who helped this project come to fruition. First, the author would like to acknowledge and thank the PulseNet participating laboratories of the Centers for Disease Control and Prevention, whose data was used for the creation of this publication. Second, the author thanks with gratitude Timothy L. Lash DSc, MPH for his patient guidance, advisement, and encouragement over the course of the author's MPH studies and thesis project. The author also thanks Grace Leu-Rasmusson, Brett Amidon, Alex Wang, Noe Valdes Vega, Curtis Weathersby, Noni Bourne, Jena Black, and the Sisters of St. Benedict of Ferdinand, Indiana for all their support.

References

1. FDA. Environmental Assessment of Factors Potentially Contributing to the Contamination of Romaine Lettuce Implicated in a Multi-State Outbreak of E. coli O157:H7. In: FDA, editor.: FDA; 2018.

2. FDA. Investigation Summary: Factors Potentially Contributing to the Contamination of Romaine Lettuce Implicated in the Fall 2018 Multi-State Outbreak of E. coli O157:H7. In: HHS, editor.2019.

3. Gottlieb SO, Stephen. FDA Update on Traceback Related to the E. coli O157:H7 Outbreak Linked to Romaine Lettuce. In: HHS, editor.: FDA; 2018.

4. Taylor MR. What Will Drive Future Food Safety Progress? Food Safety Summit; 2019 May 2019; Rosemont, IL.

5. Astill G. Special Article: Seasonality in Romaine Outbreaks and Regional Shipments. In: USDA, editor. USDA.gov: USDA; 2019.

6. CDC. Burden of Foodborne Illness: Overview. Atlanta, Georgia2018 [updated November 5, 2018; cited 2020]; Available from: https://www.cdc.gov/foodborneburden/estimates-overview.html.

7. Beuchat LR, Ryu JH. Produce handling and processing practices. Emerg Infect Dis. 1997; 3:459-65.

8. Heiman KE, Mody RK, Johnson SD, Griffin PM, Gould LH. Escherichia coli O157 Outbreaks in the United States, 2003-2012. Emerg Infect Dis. 2015; 21:1293-301.

9. Rogers MC, Peterson ND. E. coli infections causes, treatment and prevention. Hauppauge, N.Y.: Hauppauge, N.Y. : Nova Science Publishers Inc.; 2011.

10. Rivas M, Chinen I, Miliwebsky E, Masana M. Risk Factors for Shiga Toxin-Producing Escherichia coli-Associated Human Diseases. Microbiology spectrum. 2014; 2.

11. Feng P. Shiga Toxin-Producing Escherichia coli (STEC) in Fresh Produce—A Food Safety Dilemma. Microbiology spectrum. 2014; 2.

12. Ahumada O, Villalobos JR. Operational model for planning the harvest and distribution of perishable agricultural products. International Journal of Production Economics. 2011; 133:677-87.

13. Sanitary Transportation of Human and Animal Food; Final Rule, 66 (2016).

14. Zeng W, Vorst K, Brown W, Marks BP, Jeong S, Perez-Rodriguez F, et al. Growth of Escherichia coli O157:H7 and Listeria monocytogenes in packaged fresh-cut romaine mix at fluctuating temperatures during commercial transport, retail storage, and display. Journal of food protection. 2014; 77:197-206.

15. Luo Y, He Q, McEvoy JL. Effect of Storage Temperature and Duration on the Behavior of Escherichia coli O157:H7 on Packaged Fresh-Cut Salad Containing Romaine and Iceberg Lettuce. Journal of Food Science. 2010; 75:M390-M7.

16. AMS U. The Importance of Freight Transportation to Agriculture. In: AMS U, editor. Study of Rural Transportation Issues: USDA; 2008.

17. Lin XW, Ruess PJ, Marston L, Konar M. Food flows between counties in the United States. Environ Res Lett. 2019; 14:17.

18. Thomas DC. Statistical Methods in Environmental Epidemiology. New York: Oxford Press; 2009.

19. Richardson S, Monfort C, Green M, Draper G, Muirhead C. Spatial variation of natural radiation and childhood leukaemia incidence in Great Britain. Stat Med. 1995; 14:2487-501.

20. Rothman KJ. Modern epidemiology. 3rd edition, thoroughly revised and updated.. ed. Greenland S, Lash TL, editors. Philadelphia: Philadelphia : Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.

21. Wakefield J. Disease mapping and spatial regression with count data. Biostatistics. 2006; 8:158-83.

22. Wakefield J. Ecological inference for 2x2 tables. J R Stat Soc Ser A-Stat Soc. 2004; 167:385-426.

23. DiMaggio C. Spatial Epidemiology Notes: Applications and Vignettes in R 2014. Available from: http://www.columbia.edu/~cjd11/charles_dimaggio/DIRE/resources/spatialEpiBook.pdf.

24. U.S. Census Bureau PD. Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin: April 1, 2010 to July 1, 2019. In: Bureau UC, editor. Online2019.

25. Bureau C. Gazetteer Files. In: Bureau C, editor.2018.

26. NOAA. nClimDiv COUNTY TEMPERATURE-PRECIPITATION. In: NOAA, editor.2018.

27. Loomis D, Richardson DB, Elliott L. Poisson regression analysis of ungrouped data. Occupational and Environmental Medicine. 2005; 62:325-9.

28. Pham HV, Doan HTM, Phan TTT, Minh NNT. Ecological factors associated with dengue fever in a Central Highlands province, Vietnam. BMC Infect Dis. 2011; 11:172-.

Appendix 1: SAS Code

```
libname yuma "H:\My Documents\Thesis\";

PROC IMPORT OUT= WORK.YUMA_final
            DATAFILE= "H:\My Documents\Thesis\Yuma_final_added.xlsx"
            DBMS=EXCEL REPLACE;
     RANGE="Sheet1$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;


data yuma;
      set work.yuma_final;
      ln_n = log(TotalPop*(84/365));
      PropAtRisk = PropU15 + PropOver60;
      ln_risk = log(PropAtRisk);
      ln_female = log(PropFemale);
run;


proc contents data = yuma;
run;

*Analysis of distances;

proc univariate data = yuma;
      var distances;
      histogram / normal;
run;

proc freq data = yuma;
      tables distances;
run;

*Analysis of mar_y (temperature data);


proc univariate data = yuma;
      var mar_y;
      histogram;
run;


*Analysis of propU15;


proc univariate data = yuma;
      var PropU15;
      histogram;
run;
```

```
*Analysis of propOver60;


proc univariate data = yuma;
      var PropOver60;
      histogram;
run;

proc freq data = yuma;
      tables PropOver60;
run;

*Analysis of PropFemale;


proc univariate data = yuma;
      var PropFemale;
      histogram;
run;

*Analysis of count;

proc univariate data = yuma;
      var count;
      histogram / midpoints = 0 to 10 by 1;
run;

proc freq data = yuma;
      tables count;
run;

*run full model;

proc genmod data = yuma;
      model count = distances mar_y PropU15 PropOver60 PropFemale
PropU15*PropFemale PropOver60*PropFemale distances*mar_y / dist=poisson link
= log offset = ln_n dscale;
run;

*run reverse bionomial;

proc genmod data = yuma;
      model count = distances mar_y PropU15 PropOver60 PropFemale
PropU15*PropFemale PropOver60*PropFemale distances*mar_y / dist=nb link = log
offset = ln_n;
run;

*Assess collinearity;


proc genmod data = yuma;
      model count = distances mar_y PropU15 PropOver60 PropFemale
PropU15*PropFemale PropOver60*PropFemale distances*mar_y / dist=poisson link
= log offset = ln_n dscale covb;
      ods output genmod.parminfo=parms;
      ods output genmod.covb=covdsn;
run;
```

```
%COLLIN(COVDSN=COVDSN, PROCDR=GENMOD, PARMINFO=Parms, OUTPUT=collintable);

*female sex is collinear;

proc genmod data = yuma;
      model count = distances mar_y PropU15 PropOver60 distances*mar_y /
dist=poisson link = log offset = ln_n dscale covb;
      ods output genmod.parminfo=parms;
      ods output genmod.covb=covdsn;
run;

%COLLIN(COVDSN=COVDSN, PROCDR=GENMOD, PARMINFO=Parms, OUTPUT=collintable);

*proportion of U15 and Over60 correlated;
*New variable which is population at greater risk;

proc genmod data = yuma;
      model count = distances mar_y PropAtRisk distances*mar_y / dist=poisson
link = log offset = ln_n dscale covb;
      ods output genmod.parminfo=parms;
      ods output genmod.covb=covdsn;
run;

%COLLIN(COVDSN=COVDSN, PROCDR=GENMOD, PARMINFO=Parms, OUTPUT=collintable);

*log transformed variables;

proc genmod data = yuma;
      model count = distances mar_y ln_risk ln_female distances*mar_y /
dist=poisson link = log offset = ln_n dscale covb;
      ods output genmod.parminfo=parms;
      ods output genmod.covb=covdsn;
run;

%COLLIN(COVDSN=COVDSN, PROCDR=GENMOD, PARMINFO=Parms, OUTPUT=collintable);



*log transformed at risk population;

proc genmod data = yuma;
      model count = distances mar_y ln_risk distances*mar_y / dist=poisson
link = log offset = ln_n dscale covb;
      ods output genmod.parminfo=parms;
      ods output genmod.covb=covdsn;
run;

%COLLIN(COVDSN=COVDSN, PROCDR=GENMOD, PARMINFO=Parms, OUTPUT=collintable);

*without any population;

proc genmod data = yuma;
      model count = distances mar_y distances*mar_y / dist=poisson link = log
offset = ln_n dscale covb;
      ods output genmod.parminfo=parms;
      ods output genmod.covb=covdsn;
```

```
run;

%COLLIN(COVDSN=COVDSN, PROCDR=GENMOD, PARMINFO=Parms, OUTPUT=collintable);


*Assess for interaction;

proc genmod data = yuma;
      model count = distances mar_y distances*mar_y / dist=poisson link = log
offset = ln_n dscale covb;
run;

*estimates;
proc genmod data = yuma plots;
      model count = distances mar_y distances*mar_y / dist=poisson link = log
offset = ln_n dscale covb;
      estimate "RR 0 km" distances 0;
      estimate "RR 250 km" distances 250;
      estimate "RR 500 km" distances 500;
      estimate "RR 0 km 5 F" distances 0 mar_y 5 distances*mar_y 0;
      estimate "RR 250 km 5 F" distances 250 mar_y 5 distances*mar_y 1250;
      estimate "RR 500 km 5 F" distances 500 mar_y 5 distances*mar_y 2500;
      estimate "RR 0 km 10 F" distances 0 mar_y 10 distances*mar_y 0;
      estimate "RR 250 km 10 F" distances 250 mar_y 10 distances*mar_y 2500;
      estimate "RR 500 km 10 F" distances 500 mar_y 10 distances*mar_y 5000;
      estimate "RR 0 km 20 F" distances 0 mar_y 20 distances*mar_y 0;
      estimate "RR 250 km 20 F" distances 250 mar_y 20 distances*mar_y 5000;
      estimate "RR 500 km 20 F" distances 500 mar_y 20 distances*mar_y 10000;
run;

proc univariate data = yuma;
      var climate_distance;
      histogram;
run;
```

Appendix 2: Python Data Cleaning Code

```
#!/usr/bin/env python
# coding: utf-8
#Python code for Data Cleaning and Confounder Calculations
#Thomas Burke
#MPH Epidemiology Thesis
#July 21 2020


# Summary: This code was used to assign FIPS (Federal Information Processing
Standards) codes to the obtained Excel spreadsheet obtained
# from CDC. It also took the food flows (primarily from US DoT Framework),
Census and NOAA data which were identified as potential
# confounders. All these relied on FIPS identification on counties to merge
datasets.

# Because some complicated calculations regarding geography were to be used,
Python was chosen to clean the data. Data were loaded to
# dataframes in pandas.
import pandas as pd
import numpy as np


#uploading excel of CDC Case Count County Data of Romaine Lettuce Outbreak
into dataframe
df_yuma = pd.read_excel('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma.xlsx')
print(df_yuma)


#Load FIPS codes into dataframe. This dataset is merely the fips code,
county, and state.
df_fips = pd.read_excel('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/fips.xlsx')
print(df_fips)

#split polyname (county and state were formatted as strings) into dictionary
split_cells = {}
for i in df_fips['polyname']:
    split_cells[i] = i.split(",")
print(split_cells)

#map split_cells dictionary to FIPS dataframe by FIPS code
df_fips["split_cells"] = df_fips["polyname"].map(split_cells)
print(df_fips)

#change split cells to state code. Because the CDC data used state codes
rather than names
states = pd.read_csv('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/states.csv')
states['State'] = states['State'].str.lower()
print(states)
```

```
#join the states dataframe to the yuma one. Performed an "outer" join in case
of errors/misspellings
df_yuma_2 = pd.merge(df_yuma, states, left_on = "SourceState", right_on =
"Code", how = 'outer')
df_yuma_2["SourceCounty"] = df_yuma_2["SourceCounty"].str.lower()
print(df_yuma_2)

#match fips code in yuma dataset. This operation was not perfect, but
automated the FIPS assigning based on county and state name
#This code works by cycling through each entry of the CDC data, finding a
match by county, then by state, and then assigning
#the corresponding FIPS code. While statements were used to index the data
element for each iteration. The superceding while statement
#cycles this logic through each FIPS code.
j=3084
i=0 #fips
x=0 #yuma
y=314
df_yuma_2['fips'] = 0
while i <= j:
    while x <= y:
        if df_fips['split_cells'][i][1] == df_yuma_2['SourceCounty'][x]:
            print(x)
            print(i)
            if df_fips['split_cells'][i][0] == df_yuma_2['State'][x]:
                df_yuma_2['fips'][x] = df_fips['fips'][i]
                print(i)
                print(df_fips['fips'][i])
                print(x)
        x +=1
    x = 0
    i +=1

df_yuma_2

#write to excel to save file. For future projects, will make an immutable
dataset, but used excel saves to keep progress
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma2.xlsx') as writer:
     df_yuma_2.to_excel(writer)

#Example of reloading if need be.
df_yuma_2 = pd.read_excel('/Users/tburke/Documents/Emory Class Files/Thesis
and Research/E Coli Papers/untitled folder/Yuma2.xlsx')

#load data frame of Census population data, specifically from American
Community Survey. Gives Age, Gender data on every county
df_pop = pd.read_csv('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/cen_age_data.csv')

#Loading to new dataframe
#Cleaning FIPS data for misspellings, ambiguous data or missing values. The
FIPS code 55079 was assigned to errors or missing values.
#There was one legitimate 55079 (Milwaukee, Wisconsin), so cleaning was by
index.

#missing values
df_yuma_3 = df_yuma_2
```

```
df_yuma_3['fips'][16] = np.nan
df_yuma_3['fips'][19] = np.nan
df_yuma_3['fips'][94:136] = np.nan
df_yuma_3['fips'][139] = np.nan
df_yuma_3['fips'][203:213] = np.nan
df_yuma_3['fips'][218:226] = np.nan
df_yuma_3['fips'][230] = np.nan
df_yuma_3['fips'][240] = np.nan
df_yuma_3['fips'][242:251] = np.nan
df_yuma_3['fips'][252:254] = np.nan
df_yuma_3['fips'][258] = np.nan
df_yuma_3['fips'][265] = np.nan
df_yuma_3['fips'][288] = np.nan
df_yuma_3

#Cleaning FIPS data for misspellings and other errors
df_yuma_3['fips'][65] = 53053
df_yuma_3['fips'][73] = 36061
df_yuma_3['fips'][79] = 36061
df_yuma_3['fips'][81] = 12057
df_yuma_3['fips'][83] = 51059
df_yuma_3['fips'][152] = 6059
df_yuma_3['fips'][159] = 6095
df_yuma_3['fips'][160] = 6095
df_yuma_3['fips'][163] = 6095
df_yuma_3['fips'][165] = 6095
df_yuma_3['fips'][166] = 6095
df_yuma_3['fips'][171] = 6095
df_yuma_3['fips'][175] = 6095
df_yuma_3['fips'][194] = 6059
df_yuma_3['fips'][199] = 6055
df_yuma_3['fips'][229] = 6037
df_yuma_3['fips'][241] = 26159
df_yuma_3['fips'][267] = 27137
df_yuma_3['fips'][271] = 27007
df_yuma_3['fips'][295:298] = 1097
df_yuma_3['fips'][231:240] = 0


#delete rows from previous merge. From states which had no cases (artefact
from previous merge)
#print(df_yuma_3.index[301:])
df_yuma_3 = df_yuma_3.drop(df_yuma_3.index[301:315])
df_yuma_3

#write to excel file to examine
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yumax.xlsx') as writer:
     df_yuma_3.to_excel(writer)

#Census ACS data had extra digits in front. Stripped these first 5 to get
fips codes
df_pop['fips'] = df_pop['id'].apply(lambda x: x[-5:])
df_pop


#merge (all US counties) pop sex data with CDC data by fips. Converted them
to floats for merging.
```

```
df_yuma_3['fips'] = df_yuma_3['fips'].astype(float)
df_pop['fips'] = df_pop['fips'].astype(float)

df_yuma_4 = pd.merge(df_yuma_3, df_pop, left_on = "fips", right_on = "fips",
how = 'outer')
df_yuma_4


#write yuma4 to excel file to examine
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma4.xlsx') as writer:
     df_yuma_4.to_excel(writer)

#attach food flows data
#load data frame of population data
df_flows = pd.read_csv('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/yumadata.csv')

#merge combined data on yuma
df_yuma_5 = pd.merge(df_yuma_4, df_flows, left_on = "fips", right_on = "des",
how = 'outer')
df_yuma_5

#write yuma5 to excel file to examine
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma5.xlsx') as writer:
     df_yuma_5.to_excel(writer)

df_yuma_5.info

#drop unnecessary columns
#del df_yuma_5['Unnamed: 0_x']
#del df_yuma_5['Unnamed: 0.1']
del df_yuma_5['TypeDetails']
del df_yuma_5['Exposure']
del df_yuma_5['Traveled_To']
del df_yuma_5['Abbrev']
del df_yuma_5['Code']
del df_yuma_5['polyname']
df_yuma_5

#write new dataframe to excel file to examine
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma6.xlsx') as writer:
     df_yuma_5.to_excel(writer)

#add latlong data to a dataframe
df_latlong = pd.read_excel('/Users/tburke/Documents/Emory Class Files/Thesis
and Research/E Coli Papers/untitled folder/fipslatlong.xlsx')
df_latlong

#merge County demographic data plus cases to latlong data by fips code
df_yuma_7 = pd.merge(df_yuma_5, df_latlong, left_on = "fips", right_on =
"fips", how = 'outer')
#the latlong information were in strings, so they needed to be converted to
floats for calculations
df_yuma_7['lat'] = df_yuma_7['lat'].str.strip("°")
df_yuma_7['long'] = df_yuma_7['long'].str.strip("°")
```

```
df_yuma_7['lat'] = df_yuma_7['lat'].astype(float)
df_yuma_7['long'] = df_yuma_7['long'].str.slice(1)
df_yuma_7['long'] = df_yuma_7['long'].astype(float)
df_yuma_7['long'] = df_yuma_7['long']*-1
df_yuma_7

#assigning latlong data to food flows dataframe
df_flows_2 = pd.merge(df_flows, df_latlong, left_on = "des", right_on =
"fips", how = 'left')
df_flows_2['lat'] = df_flows_2['lat'].str.strip("°")
df_flows_2['long'] = df_flows_2['long'].str.strip("°")
df_flows_2['lat'] = df_flows_2['lat'].astype(float)
df_flows_2['long'] = df_flows_2['long'].str.slice(1)
df_flows_2['long'] = df_flows_2['long'].astype(float)
df_flows_2['long'] = df_flows_2['long']*-1
df_flows_2 = df_flows_2.drop(df_flows_2.index[81])
df_flows_2

#save dataframes to excel files to examine
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma7.xlsx') as writer:
     df_yuma_7.to_excel(writer)

with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/flows_2.xlsx') as writer:
     df_flows_2.to_excel(writer)

#To find the distance between primary distribution site and county, needed to
make a dictionary with food flows FIPS code and latlong

df_flows_dict = df_flows_2
df_flows_dict['latlong'] = list(zip(df_flows_dict.lat, df_flows_dict.long))
#df_flows_dict = df_flows_dict[['lat', 'long']].values.tolist()
flows_dict = df_flows_2.groupby('fips')['latlong'].apply(list).to_dict()
flows_dict


#find distance lists. Geopy finds distances between geographic points taking
into account the Earth's curvature.
import geopy.distance

#example
coords_1 = flows_dict[4005][0]
coords_2 = flows_dict[53005][0]

print(geopy.distance.distance(coords_1, coords_2).km)

#make all county plus case data with latlong tuple
df_yuma_8 = df_yuma_7
df_yuma_8['latlong'] = list(zip(df_yuma_8.lat, df_yuma_8.long))
df_yuma_8


#This calculation took the dictionary of all primary distribution nodes and
calculated the distance between them and the given county.
#It stored these FIPS:distance value pairs in a temporary dictionary and then
selected the smallest value. Distance to primary
```

```
#distribution was a variable of interest for this study. It used a boolean to
see if their were entries in the dictionary to account
#for missing values.

i = 0
j = 3411
df_yuma_8['distances'] = 0
df_yuma_8['nearest_dist'] = ''
while i<=j:
    temp_dict = {}
    for a in flows_dict:
        if np.isfinite(df_yuma_8['latlong'][i]).all():
            distance = geopy.distance.distance(flows_dict[a],
df_yuma_8['latlong'][i]).km
            temp_dict.update({ a: distance })
            print(geopy.distance.distance(flows_dict[a],
df_yuma_8['latlong'][i]))
    if bool(temp_dict):
        temp = min(temp_dict.values())
        res = [key for key in temp_dict if temp_dict[key] == temp]
        df_yuma_8['distances'][i] = temp
        df_yuma_8['nearest_dist'][i] = res
    i +=1

df_yuma_8


#write to excel to examine
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma9.xlsx') as writer:
     df_yuma_8.to_excel(writer)


#Average March temperature was also a variable of interest. Gathered average
county temperature data for March 2018.
df_climate = pd.read_csv('/Users/tburke/Documents/Emory Class Files/Thesis
and Research/E Coli Papers/untitled folder/climate.csv')

#reduced to just fips and march 2018 average temperature data
df_climate_18 = df_climate_19[['fips','mar']]
df_climate_18

#merging with overall dataset
df_yuma_10 = pd.merge(df_yuma_8, df_climate_18, left_on = "fips", right_on =
"fips", how = 'left')
df_yuma_10

#examining in excel
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma10.xlsx') as writer:
     df_yuma_10.to_excel(writer)

#NOAA does not have average county temperature for every county (about 30% of
entries do not have climate)
#A similar calculation was performed as above to calculate the closest county
where temp data is available

#first need to make a dataframe of latlong data for climate
```

```
df_climate_18['fips'] = df_climate_18['fips'].astype(float)
df_latlong['fips'] = df_latlong['fips'].astype(float)
df_climate_latlong = pd.merge(df_climate_18, df_latlong, left_on = "fips",
right_on = "fips", how = 'inner')

df_climate_latlong['lat'] = df_climate_latlong['lat'].str.strip("°")
df_climate_latlong['long'] = df_climate_latlong['long'].str.strip("°")
df_climate_latlong['lat'] = df_climate_latlong['lat'].astype(float)
df_climate_latlong['long'] = df_climate_latlong['long'].str.slice(1)
df_climate_latlong['long'] = df_climate_latlong['long'].astype(float)
df_climate_latlong['long'] = df_climate_latlong['long']*-1

df_climate_latlong['latlong'] = list(zip(df_climate_latlong.lat,
df_climate_latlong.long))
df_climate_latlong

#make a dictionary of fips, lat long associated with temperature data
climate_dict =
df_climate_latlong.groupby('fips')['latlong'].apply(list).to_dict()
climate_dict

df_yuma_11 = df_yuma_10
df_yuma_11

#Ran a similar algorithm as above, but for climate. Because the amount of
calculations is high, the algorithm checked if there
#was climate data already before finding the closest neighbor.
i = 0
j = 3413
df_yuma_11['climate_distance'] = 0
df_yuma_11['nearest_climate_fips'] = ''
while i<=j:
    temp_dict = {
    if np.isnan(df_yuma_11['mar'][i]):
        for a in climate_dict:
            if np.isfinite(df_yuma_11['latlong'][i]).all():
                distance = geopy.distance.distance(climate_dict[a],
df_yuma_11['latlong'][i]).km
                temp_dict.update({ a: distance })
                #print(geopy.distance.distance(climate_dict[a],
df_yuma_11['latlong'][i]))
                print(i)
        if bool(temp_dict):
            temp = min(temp_dict.values())
            res = [key for key in temp_dict if temp_dict[key] == temp]
            df_yuma_11['climate_distance'][i] = temp
            df_yuma_11['nearest_climate_fips'][i] = res
    else:
        df_yuma_11['nearest_climate_fips'][i] = df_yuma_11['fips'][i]
    print(i)
    i +=1

df_yuma_11

#output dataset to examine
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma11.xlsx') as writer:
     df_yuma_11.to_excel(writer)
```

```
#merge county dataset with climate data
#the calculation inserted a list with one data element rather than a float
df_yuma_12 = df_yuma_11
#fixing the datatype issue
i=0
j=3413

while i <=j:
    if type(df_yuma_12['nearest_climate_fips'][i]) == list:
        df_yuma_12['nearest_climate_fips'][i] =
df_yuma_12['nearest_climate_fips'][i][0]
    i+=1
#merging datasets
df_yuma_12 = pd.merge(df_yuma_12, df_climate_18, left_on =
"nearest_climate_fips", right_on = "fips", how = 'inner')
df_yuma_12


#do same datatype treatment for nearest_dist
i=0
j=3413

while i <=j:
    if type(df_yuma_12['nearest_dist'][i]) == list:
        df_yuma_12['nearest_dist'][i] = df_yuma_12['nearest_dist'][i][0]
    i+=1

#output file again
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma12.xlsx') as writer:
     df_yuma_12.to_excel(writer)


#Set missing data to 0
i=0
j=3413

while i <=j:
    if type(df_yuma_12['nearest_climate_fips'][i]) == str:
        df_yuma_12['nearest_climate_fips'][i] = 0
    i += 1

#examine file in excel
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma12x.xlsx') as writer:
     df_yuma_12.to_excel(writer)


#merge climate and nearest distance. Assigns temperature based on fips or
closest fips
df_yuma_12['nearest_climate_fips'] =
df_yuma_12['nearest_climate_fips'].astype(float)
df_yuma_13 = pd.merge(df_yuma_12, df_climate_18, left_on =
"nearest_climate_fips", right_on = "fips", how = 'left')
df_yuma_13
```

```
#examine file
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma13.xlsx') as writer:
     df_yuma_13.to_excel(writer)

#Data cleaning and reading the file for analysis

#drop unnecessary columns
df_yuma_14 = df_yuma_13[['fips_x', 'PFGE-XbaI-status', 'TotalPop', 'Female',
'under5','5to9','10to14','60over','medianAge','distances','nearest_dist','cli
mate_distance','nearest_climate_fips','mar_y']]

df_yuma_14
#The dataset is 3414 rows × 14 columns

#drop missing values of fips
df_yuma_15 = df_yuma_14[df_yuma_14['fips_x'].notna()]
df_yuma_15 = df_yuma_15[df_yuma_15['fips_x'] > 0]
#The dataset is 3324 rows × 14 columns. These are all non-county cases. This
means that 90 values are not included from 300
#positive IDs from CDC. These are largely environmental samples. More
information in the analysis.
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma15.xlsx') as writer:
     df_yuma_15.to_excel(writer)


#drop missing values for population data medianAge. This was mostly a merge
artefact.
df_yuma_16 = df_yuma_15[df_yuma_15['medianAge'].notna()]
df_yuma_16
#same size as Yuma 14 3324 x 14

#drop nan's for mar_y
df_yuma_17 = df_yuma_16[df_yuma_16['mar_y'].notna()]
df_yuma_17
#3245 rows × 14 columns. Some county population fips were of puerto rico. Not
relevant to this analysis.

#examine file in excel
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma17.xlsx') as writer:
     df_yuma_17.to_excel(writer)

#remove Alaska and Hawaii. Because this study was looking at examining the
trucking infrastructure, Hawaii and Alaska were excluded
#Though Alaska had significant cases (8), its distribution pattern is too
different for us to focus on the primary cold chain
#storage and other characteristics.
df_yuma_18 = df_yuma_17
df_yuma_18 = df_yuma_18.drop(df_yuma_18[(df_yuma_18.fips_x > 1999) &
(df_yuma_18.fips_x < 3000)].index)
df_yuma_18 = df_yuma_18.drop(df_yuma_18[(df_yuma_18.fips_x > 14999) &
(df_yuma_18.fips_x < 16000)].index)
#df_yuma_18 = df_yuma_17[df_yuma_17['fips_x'] < 2000 &
df_yuma_17['fips_x'] >2999]
df_yuma_18
```

```
#fully clean dataset examined
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma_clean.xlsx') as writer:
     df_yuma_18.to_excel(writer)

#Removing duplicate rows (counts from CDC) and adding counts as own column
df_yuma_19 = df_yuma_18
#df_yuma_19['count'] = 1
df_yuma_19

df_yuma_20 = df_yuma_19.drop_duplicates()
df_yuma_20

#use df_yuma_19 to count the duplicates. PFGE XbaI was a variable all counts
had but no other counties had. This was to prevent
#single counties being counted as cases but retained single county counts.
count_dict = df_yuma_19.groupby('fips_x')['PFGE-XbaI-
status'].apply(list).to_dict()
count_dict

#The dictionary counted how many 'PFGE XbaI status' entries. Then this
algorithm judged whether those dictionaries contained
#strings or not. If it did, it counted how many strings to arrive at the case
count.
count_dict_2 = {}
for i in count_dict:
    if type(count_dict[i][0]) == str:
        a = i
        b = len(count_dict[i])
        count_dict_2.update({ a:b })
    else:
        a = i
        b = 0
        count_dict_2.update({ a:b })
    i+=1

count_dict_2

#mapping case count dictionary to dataset and dropping duplicates
df_yuma_20 = df_yuma_19
df_yuma_20["count"] = df_yuma_20["fips_x"].map(count_dict_2)
df_yuma_20 = df_yuma_20.drop_duplicates()
df_yuma_20


#calculated dataset examination
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma_final.xlsx') as writer:
     df_yuma_20.to_excel(writer)


#make final variables
#proportion female
df_yuma_21 = df_yuma_20
df_yuma_21['PropFemale'] = df_yuma_21['Female'] / df_yuma_21['TotalPop']
#under 15
df_yuma_21['PropU15'] = (df_yuma_21['under5'] + df_yuma_21['5to9'] +
df_yuma_21['10to14'])/ df_yuma_21['TotalPop']
```

```
#over 60
df_yuma_21['PropOver60'] = df_yuma_21['60over'] / df_yuma_21['TotalPop']
#expected value for each cell
expected_PT = 202/((326500498-732438-1420491)*(84/365))
df_yuma_21['expectedValue'] = df_yuma_21['TotalPop'] * expected_PT
df_yuma_21

#final output for SAS
with pd.ExcelWriter('/Users/tburke/Documents/Emory Class Files/Thesis and
Research/E Coli Papers/untitled folder/Yuma_final_added.xlsx') as writer:
     df_yuma_21.to_excel(writer)
```

Appendix 3: Data Dictionary

Variable Name	Description	Variable Type	Values	Source	Original or Derived
fips_x	Federal Information Processing Standards (FIPS) code. A standardized way of categorizing state counties or county equivalents (e.g. Louisiana Parishes). Discontinued for most agencies, this identifier is still used by the Department of Transportation for their Freight Analytical Framework, the basis of the Food Flows dataset.	numeric	N/A	American Community Survey (Census); Food Flows; NOAA Climate Data	original
PFGE-Xbal-status	Pulse Field Gel Electrophoresis status. Not used in the Analysis; an artefact from data processing.	string	N/A	CDC	original
TotalPop	County total population in 2018	numeric	[102, 10098052]	American Community Survey (Census)	original
Female	County female population in 2018	numeric	[44, 5121264]	American Community Survey (Census)	original
under5	County population under 5 in 2018	numeric	[8, 624745]	American Community Survey (Census)	original
5to9	County population ages 5-9 in 2018	numeric	[5, 607905]	American Community Survey (Census)	original
10over14	County population ages 10-14 in 2018	numeric	[0, 626594]	American Community Survey (Census)	original
60over	County population over age 60 in 2018	numeric	[39, 1847041]	American Community Survey (Census)	original

medianAge	Median age of the county in 2018	numeric	[21.7, 67]	American Community Survey (Census)	original
distances	Distance in Kilometers to nearest Distribution Node	numeric	[0, 761]	Food Flows	derived
nearest_dist	FIPS code of the nearest Distribution Node	numeric	N/A	Food Flows	derived
climate_distance	Distance in Kilometers to closest climate data	numeric	[0, 402]	NOAA Climate Data	derived
nearest_climate	FIPS code to nearest climate data	numeric	N/A	NOAA Climate Data	derived
mar_y	Temperature in Fahrenheit of the county or the nearest county in March of 2018	numeric	[23.7, 74.9]	NOAA Climate Data	original
count	Number of cases in individual county in 2018 Yuma Romaine Lettuce Outbreak	numeric	[0, 10]	CDC Data	derived
PropFemale	Proportion of county population that are female in 2018	numeric	[0.210, 0.586]	American Community Survey (Census)	derived
PropU15	Proportion of county population that are under 15 years of age in 2018	numeric	[0.060, 0.342]	American Community Survey (Census)	derived
PropOver60	Proportion of county population over the age of 60 in 2018	numeric	[0.063, 0.649]	American Community Survey (Census)	derived
expectedValue	Expected value based on person time	numeric	[0.000, 27.3]	CDC Data	derived
ln_n	Offset used for Poisson Regression	numeric		American Community Survey (Census)	derived