Campylobacter Geospatial Cluster Analysis, Colorado, 2001 – 2010

Ben White, MPH

Fall EIP Meeting 2013

Colorado Department
of Public Health
and Environment

Campylobacter 101

Symptoms (1-10 Days After Exposure):

- Diarrhea
- Ab. Pain
- Fatigue
- Nausea
- Vomiting
- Fever

Common Sources:

- Undercooked poultry, beef and pork
- Unpasteurized dairy products
- Un-chlorinated water
- Raw produce
- Food or drink contaminated with feces
- Animal contact

Diagnostic Testing:

- Culture from clinical specimen (confirmed)
- Antigen-based or PCR test (suspect)
- 7 day reportable condition in CO

Campylobacter Research Objectives

- Analyze sporadic case data from the Colorado Electronic Disease Reporting System (CEDRS) to understand disease trends and patterns.

- Identify clusters or ‘hot spots’ of sporadic cases in the state of Colorado across space and time using geospatial methods.

- Attempt to model the incidence rate differences with independent variables from existing statewide datasets at a meaningful spatial resolution.

HYPOTHESES

- That rates differ significantly by geography across Colorado

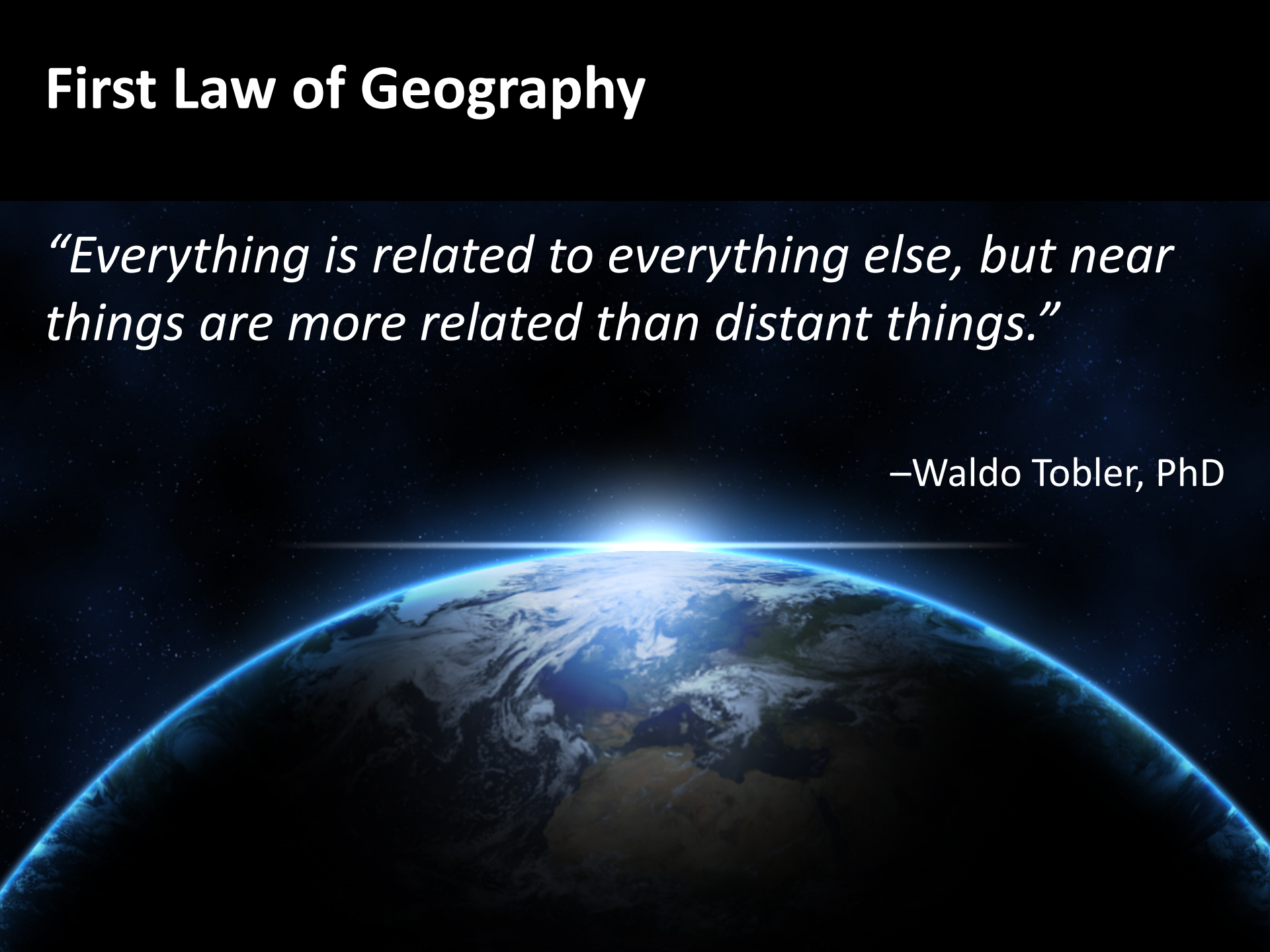
- Campylobacter rates are higher in rural areas due to differences in urban/rural living (geography, number of restaurants, water systems, SES, employment type)

- Intensity of agriculture (particularly livestock) and associated animal contact is a driving force behind higher Campylobacter rates.

First Law of Geography

"Everything is related to everything else, but near things are more related than distant things."

–Waldo Tobler, PhD

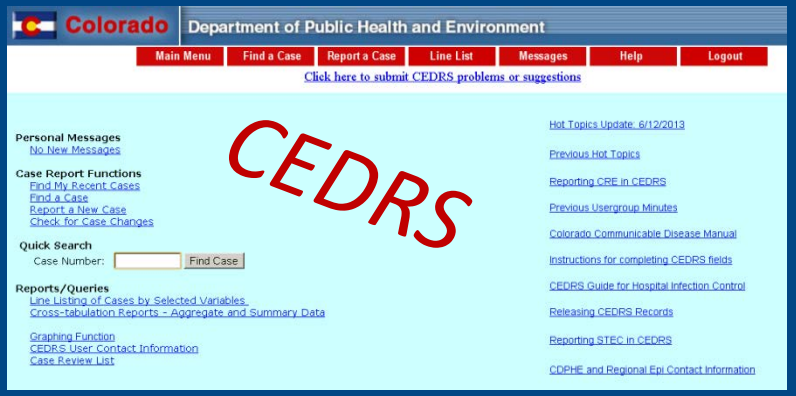


Literature Review of Spatial Statistical Methods

Article	Disease	Cases	Cluster Detection	Resolution	Spatial analysis	Variables	Software
Green Krause Wylie 2006	Campy	Laboratory Confirmed C	Spatial Scan Statistic	census sub-divisions and local neighborhood (city)	Lorenz Curve, Gini Coefficient, Poisson Regression, Smoothing, Spearman's Rank and Pearson's R	age group, gender, urban/rural, SES, and ag occupation but not animals densities i	SigEpi, S-Plus, ArcGIS, SatScan
Jepsen et al 2009	Campy	Lab Positive	Modified Local Moran's I	One county, and then country	Poisson, Cube Neighborhood weights matrix. Smoothing	unknown	ArcGIS, SatScan, NetLogo
Kistemann et al 2004	STEC/EHEC	all human cases	Moran's I	county level	Poisson, Chi Square, Joint Count Statistics, Pearson's R, Stepwise Multiple Linear Regression	farm density, cattle density, % farms with cattle	ArcGIS, SPSS
Michell et al 1999	STEC/EHEC	Laboratory Confirmed	Moran's I and G Statistic	county level and centroids	OLS and MLE regression with Inverse Distance Matrix	total ag land, cattle density, livestock density	ArcGIS, SpaceStat
Morgan 2002	Campy	reported cases	kerneling	collection districts and centroids	smoothing of rates, log linear modeling of case data	age, gender,	ArcGIS, CrimeStat, SatScan
Nygard et al 2004	Campy	All domestic reported cases	NONE	municipalities and county	Pearson's R, Poisson,	Percipitation, temperature, water supply, ag	ArcGIS, Stata
Odoi et al 2004	Giardia	reported cases	Spatial Scan Statistic	census sub-divisions or county	Poisson, Bayesian smoothed, Spearman's Rank	animal density, age, gender, etc.	ArcGIS, SatScan
Valcour et al 2002	STEC/EHEC	Sporadic reported cases		census sub-division		ag land, soil, drainage, age, gender	ArcGIS,
Jonsson 2010	Campy	lab confirmed cases and positive broiler flocks	Spatial Scan Statistic	municipalities and county	Relative Risk	population density of broilers, humans,	ArcGIS, SatScan

Colorado Campylobacter Case Finding Process

CEDRS tables joining and extraction



7750 Cases Between 2001 and 2010

-Lab confirmed campy cases (culture)
- CollectDt 01/01/2001 to 12/31/2010
-Not part of an outbreak
-Colorado resident at time of diagnosis

-Address verification
-Case status verification
-Outbreak verification

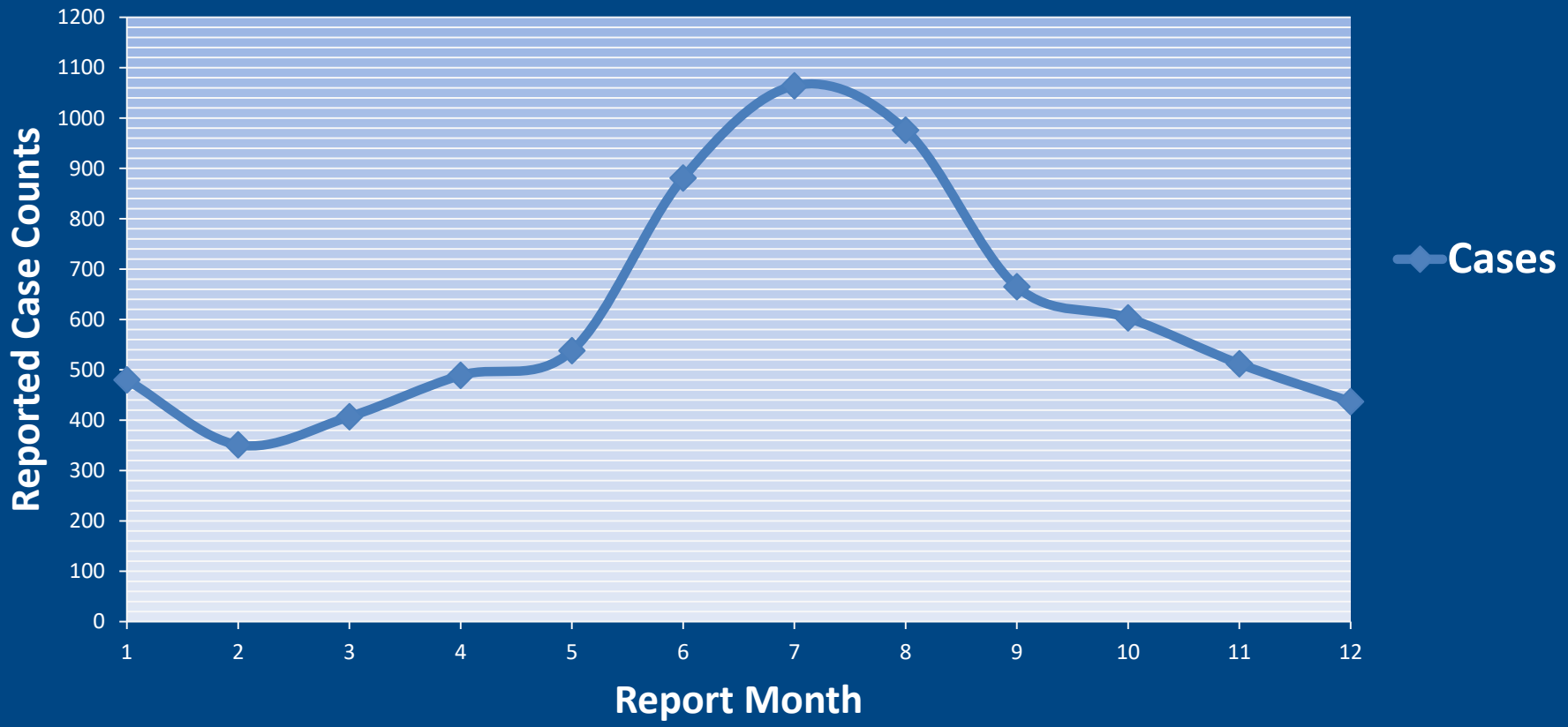
-7403 Cases
-Calculate yearly crude incidence rates by county

DESCRIPTIVE STATISTICS

Colorado Sporadic Campylobacter Cases: Incidence Rates per 100,000 2001-2010

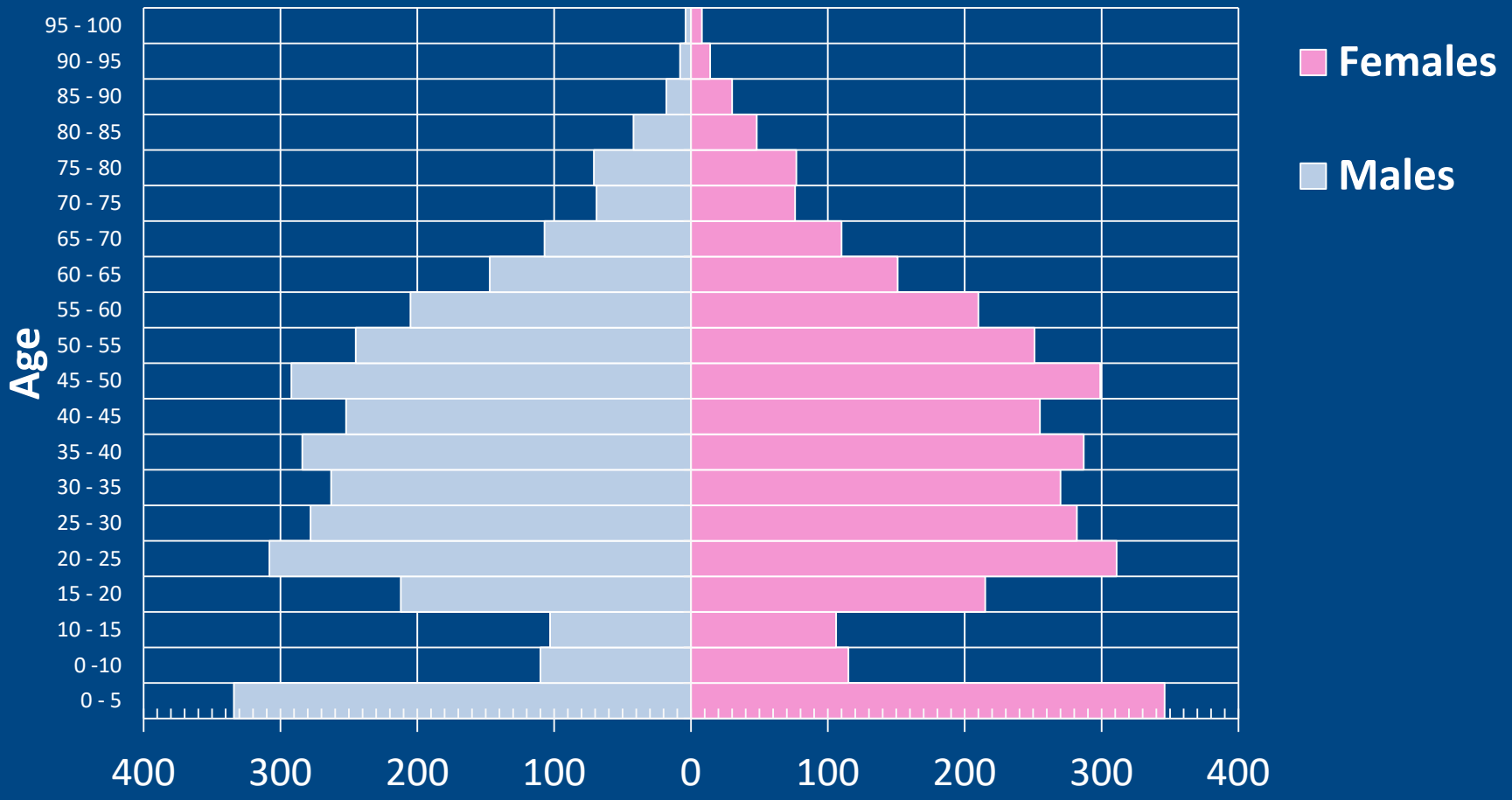
740 annual case average

Colorado Sporadic Campylobacter Cases by Month, 2001-2010



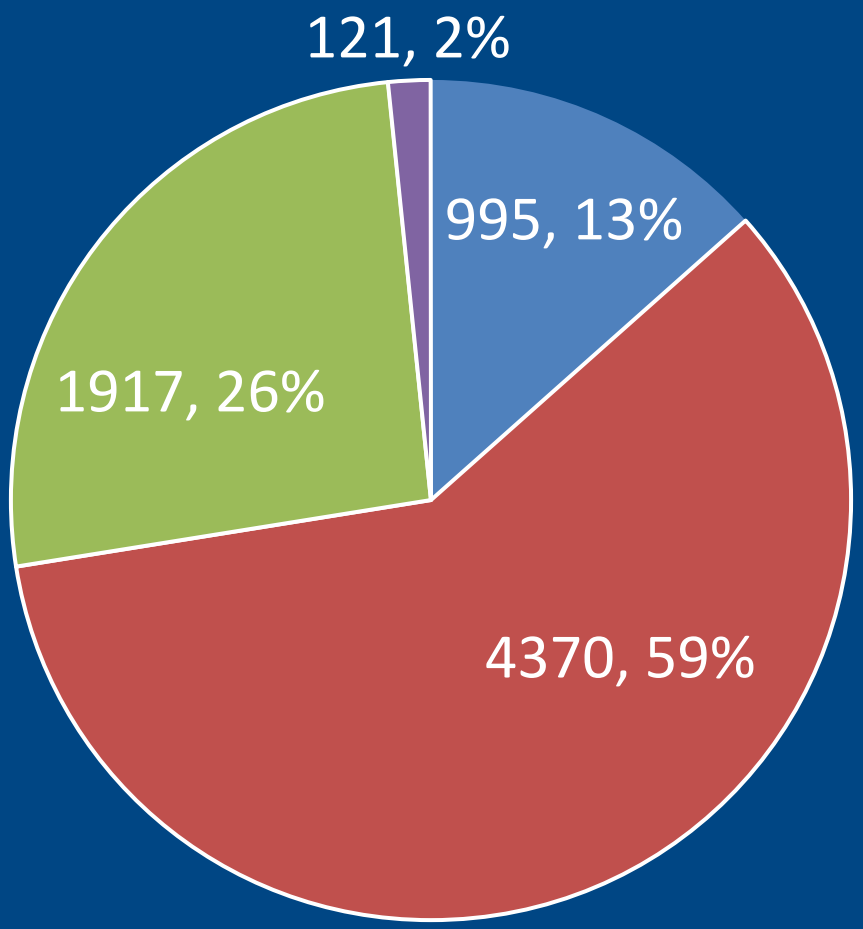
June , July and August are peak months for Campylobacter infection in Colorado, which mirrors national trends

Colorado Sporadic Campylobacter Cases, Age/Sex Distribution, 2001-2010



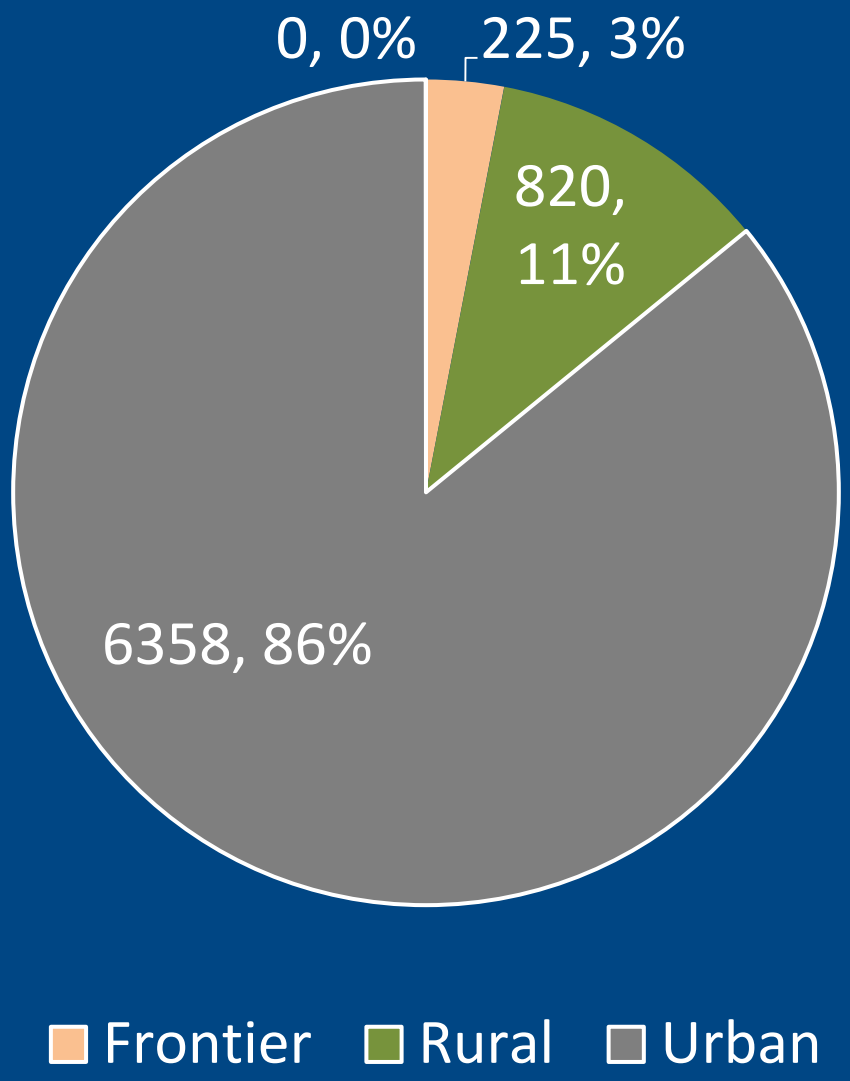
53.8% Male **45.6% Female** **0.6% Unknown**

Sporadic Campylobacter Cases by Ethnicity Colorado, 2001-2010



2001-2010 Colorado Pops:
Hispanic – 19.3%
Non Hispanic – 80.7%

Sporadic Campylobacter Cases by CRHC Class Colorado, 2001-2010



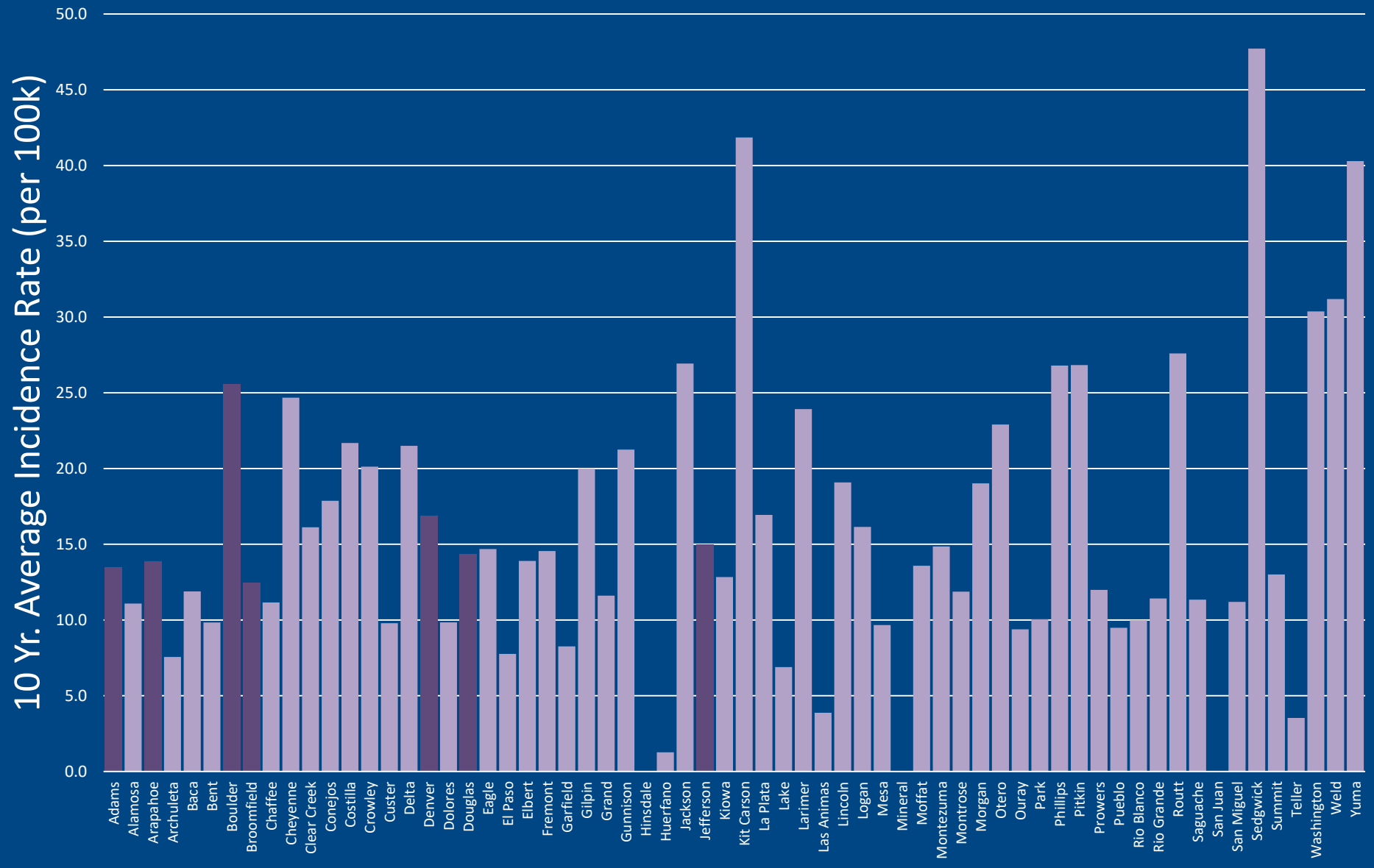
2001-2010 Colorado Pops:
Urban - 86.0 %
Ruran - 11.3%
Frontier - 2.7%

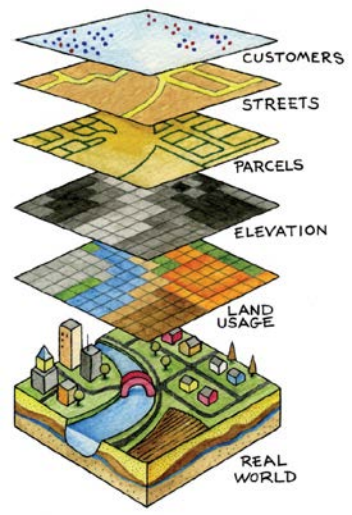
COLORADO RURAL HEALTH CENTER
The State Office of Rural Health

Qualitative Analysis of Case Notes from CEDRS

- Words that occur most frequently in notes are larger in word cloud.
  - Not phrase-based, words are out of context
    - Not statistically significant

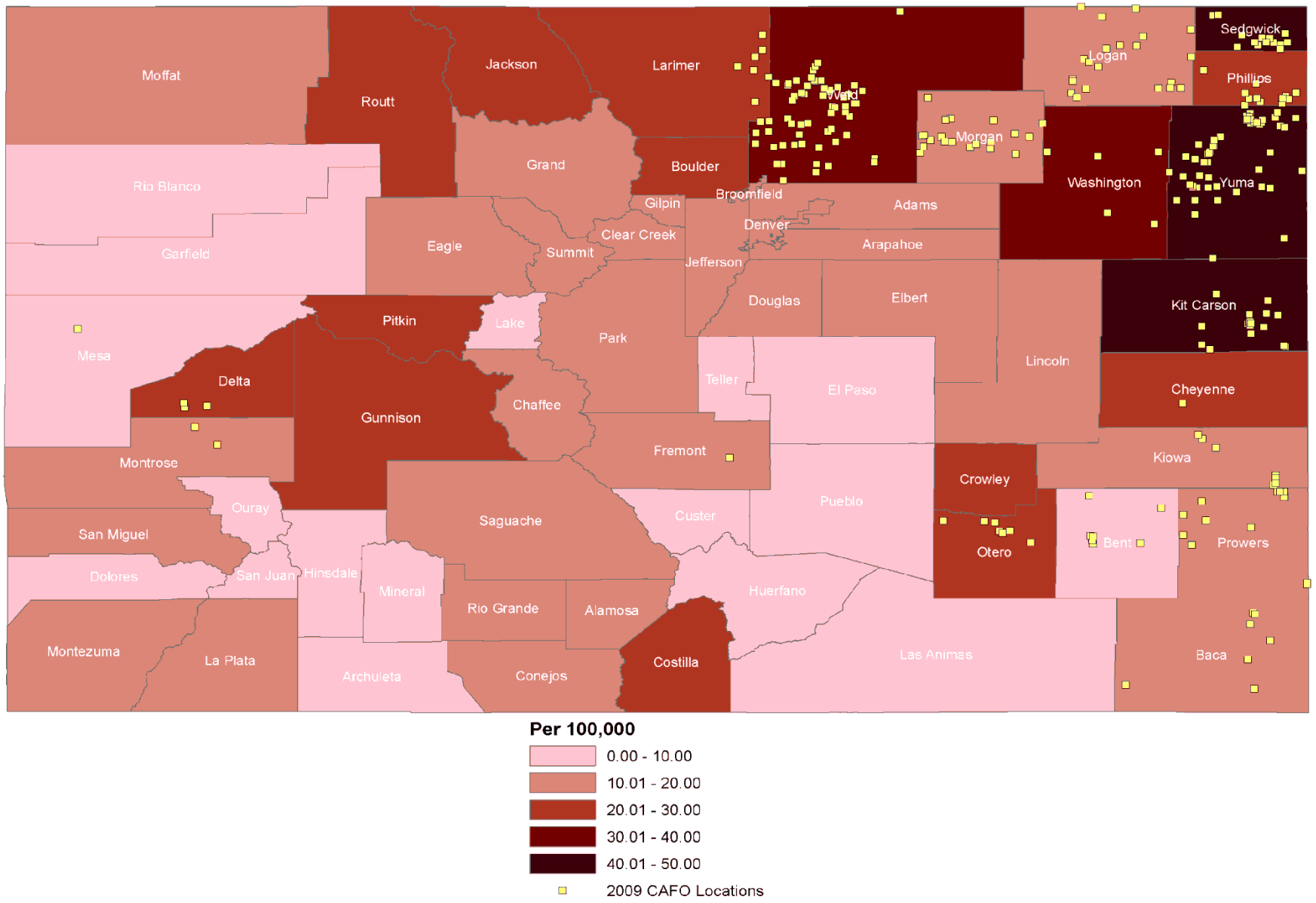
Campylobacter Incidence Rates per 100,000, by County, 2001-2010 Average





GIS Cluster Analysis

Colorado: Sporadic Campylobacter Cases, Incidence Rates Per 100,000, 2001-2010



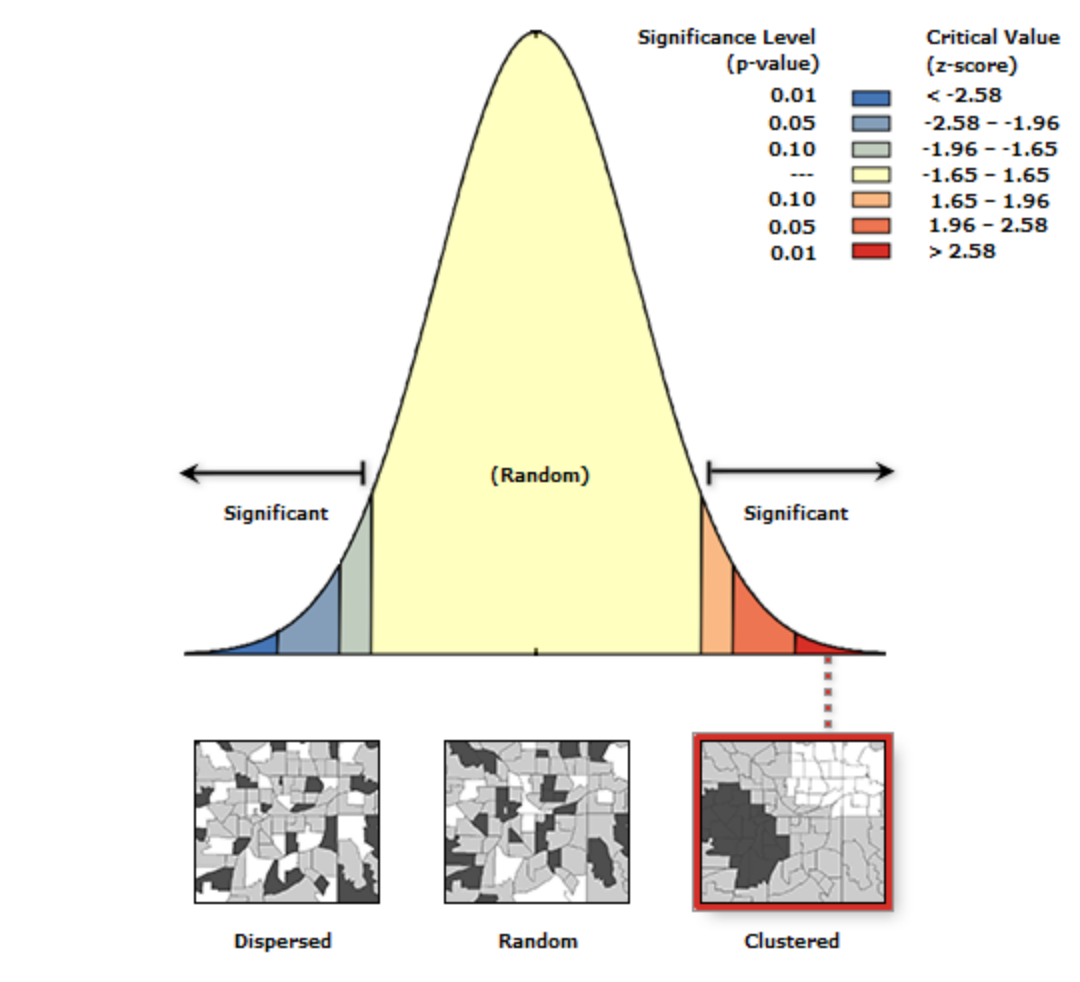
Global Moran's I

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

- Test for spatial autocorrelation.
- Compares whole dataset against one location, looks at variable and its variance
- Varies between -1 and 1, with 0 being completely random i.e. no association. Negative (dissimilar) and Positive (similar) spatial autocorrelation exists.
- Spatial weights matrix needed. This study uses 9 Nearest Neighbors (max. number counties bordering a CO county, Lincoln County)

Global Moran's I

Moran's Index:	**0.370353**
z-score:	6.575428
p-value:	0.0000001

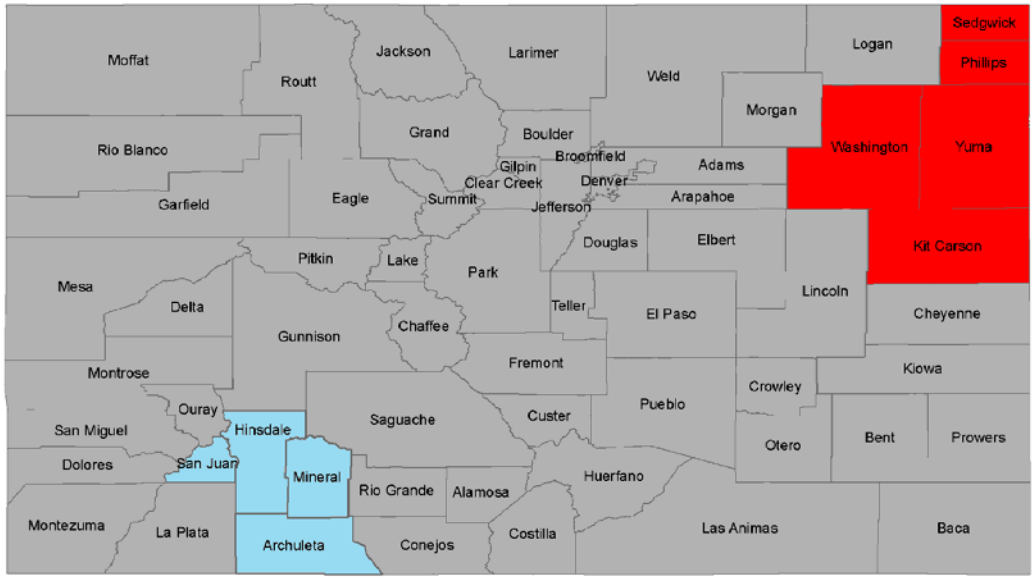


Given the z-score of 6.58, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

Local Moran's I (Anselin's LISA)

- Calculates statistic for each single location (polygon) and compares it to index
- Index is the spatial weights matrix, calculated based on previously defined k nearest neighbors (NN=9 in our case)
- Since each index has an associated statistic, we can compare statistical significance of relationship between polygon and its neighbors
- Better at discerning spatial patterns than Global Moran's

Colorado Sporadic Campylobacter Cases 2001-2010, 10 Year Average Unsmoothed Incidence Rates: Local Moran's I Results



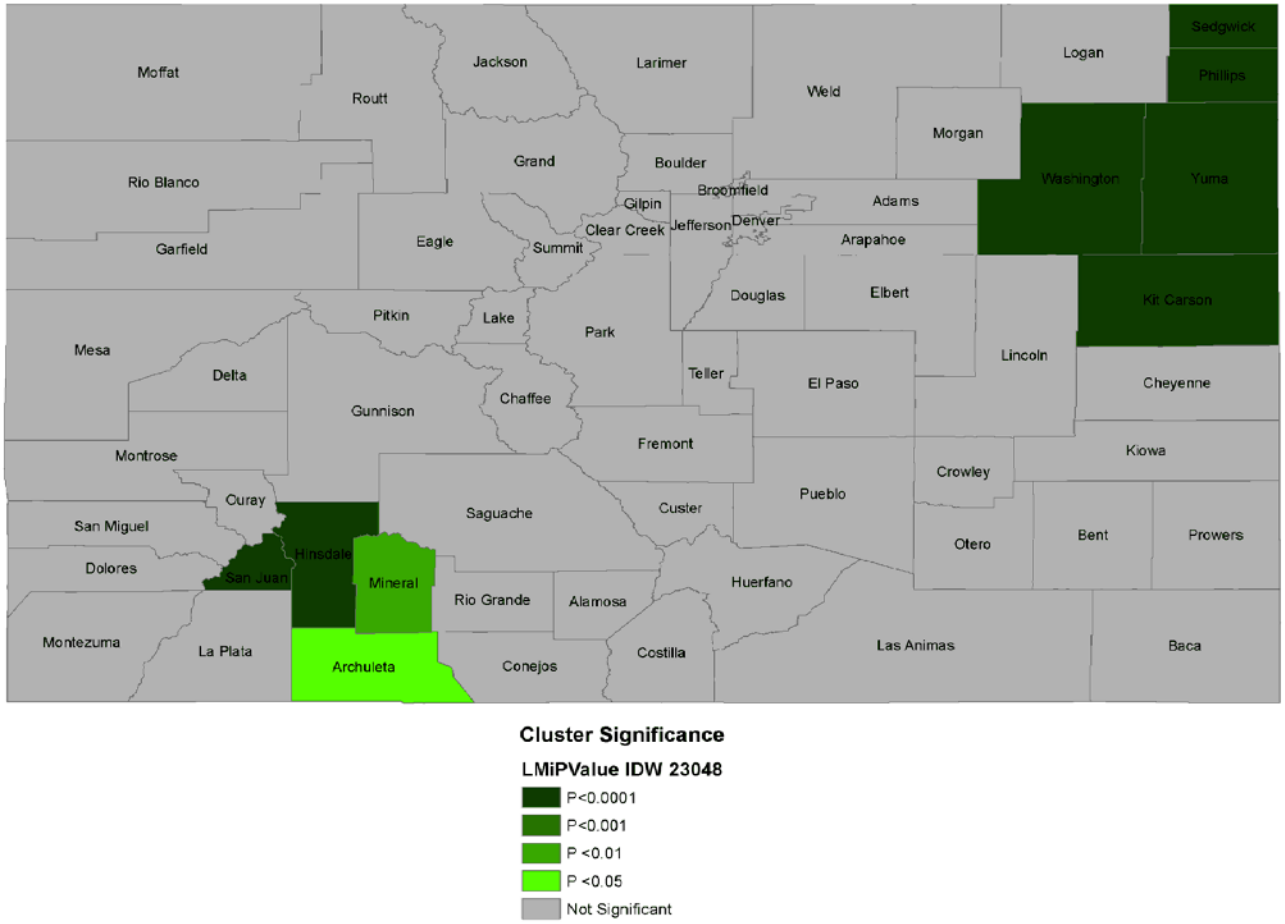
P<.05

- Not Significant
- High/High
- High/Low
- Low/High
- Low/Low

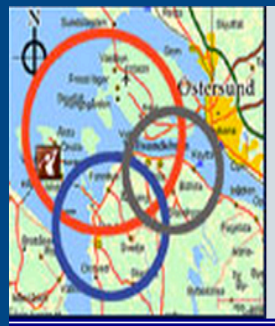
The Local Moran's I results can be interpreted in this way High/High are counties with high campy inc. rates surrounded by counties that also show high incidence rates. Low/Low are counties with low campy inc. rates surrounded by counties that also have low inc. rates.

Cluster of high incidence rates in the agrarian NE corner of the State. Cluster of low incidence rate in the SW corner of the state. Counties assigned “H” and “L” in dataset cluster variables.

Colorado Sporadic Campylobacter Cases 2001-2010, 10 Year Average Unsmoothed Incidence Rates: Cluster Significance



Statistically significant cluster relationships as determined by LISA (Sedgwick, Phillips, Washington, Yuma, and Kit Carson counties) (Mineral, Archuleta, Hinsdale & San Juan C=counties)



SaTScan™

Software for the spatial, temporal, and space-time scan statistics

- Home
- Download [SaTScan v9.1.1 *March 9 2011*]
- Technical Documentation
- Bibliography
- Data Sets
- Contact Us

Purpose

SaTScan™ is a free software that analyzes spatial, temporal and space-time data using the spatial, temporal, or space-time scan statistics. It is designed for any of the following interrelated purposes:

- Perform geographical surveillance of disease, to detect spatial or space-time disease clusters, and to see if they are statistically significant.
- Test whether a disease is randomly distributed over space, over time or over space and time.
- Evaluate the statistical significance of disease cluster alarms.
- Perform repeated time-periodic disease surveillance for early detection of disease outbreaks.

The software may also be used for similar problems in other fields such as archaeology, astronomy, botany, criminology, ecology, economics, engineering, forestry, genetics, geography, geology, history, neurology or zoology.

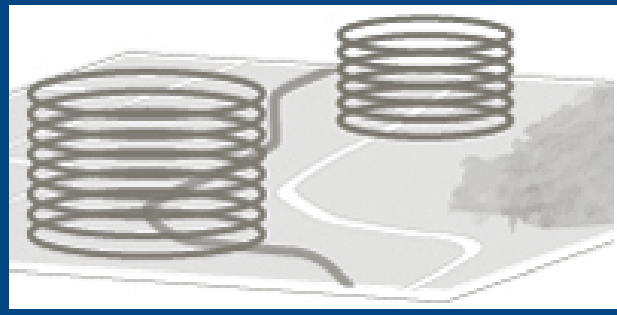
Data Types and Methods

SaTScan uses either a Poisson-based model, where the number of events in a geographical area is Poisson-distributed, according to a known underlying population at risk; a Bernoulli model, with 0/1 event data such as cases and controls; a space-time permutation model, using only case data; an ordinal model, for ordered categorical data; an exponential model for survival time data with or without censored variables; or a normal model for other types of continuous data. The data may be either aggregated at the census tract, zip code, county or other geographical level, or there may be unique coordinates for each observation. SaTScan adjusts for the underlying spatial inhomogeneity of a background population. It can also

SaTScan™

Software for the spatial, temporal, and space-time scan statistics

- Yearly campylobacter case counts by county (2001-2010)
- Yearly total population by county (2001-2010)
- US Census Population- based centroids
- Discrete Poisson Space/Time Model, 999 Monte Carlo iterations
- 6 different cylinder limitations explored:
  -1% of population and no distance limit
  -1% of population and 100km
  -5% of population and 100km,
  -10% of population and 100km
  -50% of the population and 50km
  - 50% of the population and 100km

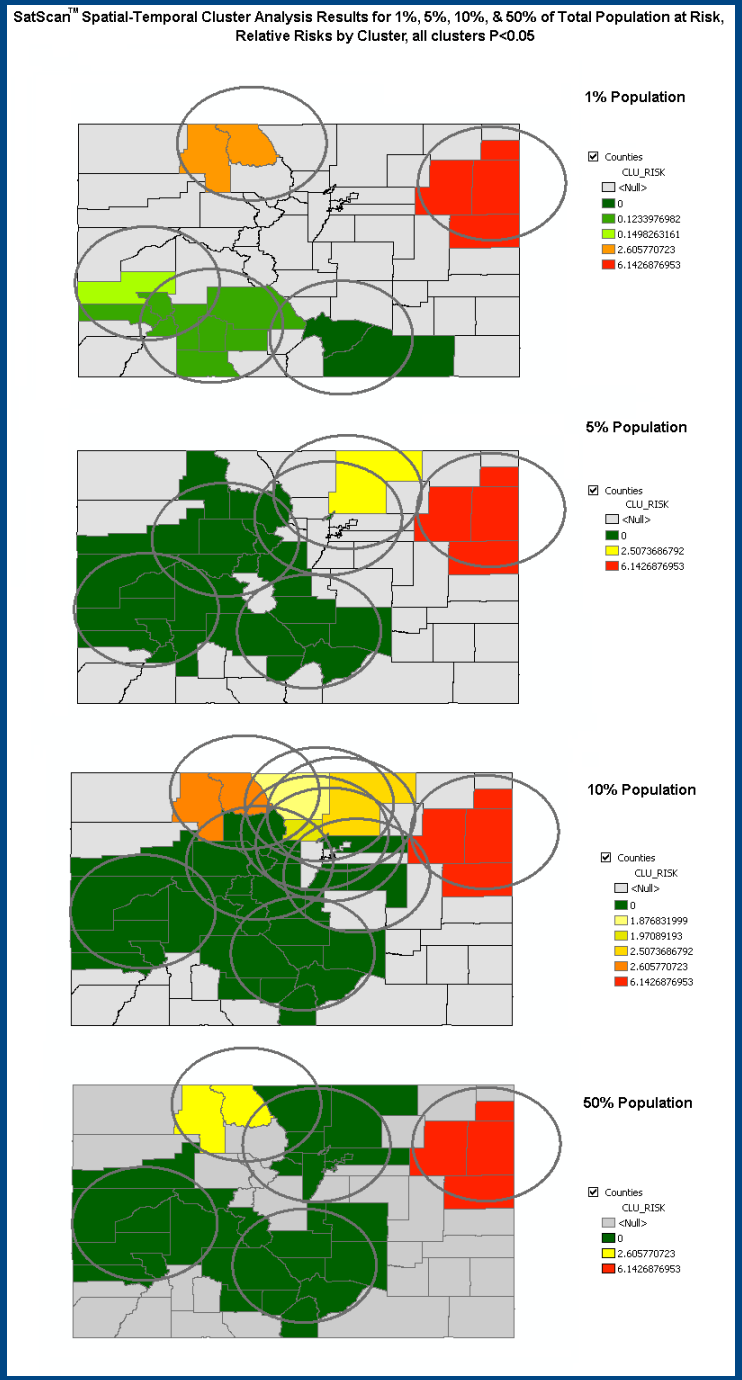


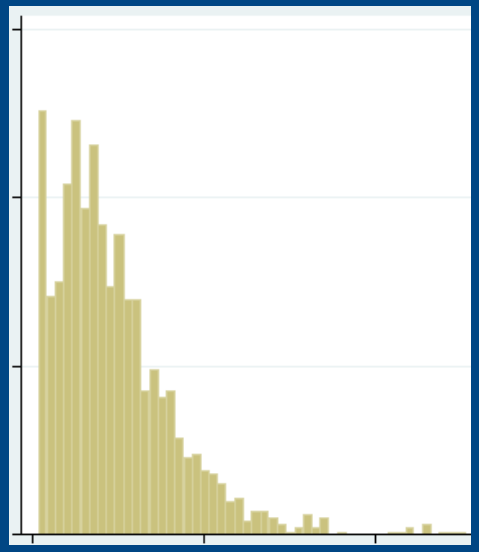
(adapted from Sugumaran, Larson & DeGroote 2009)

SaTScan™

Software for the spatial, temporal, and space-time scan statistics

- Phillips, Yuma, Washington, and Kit Carson identified as cluster (Sedgwick left out)
- Found with multiple scan window sizes
- Relative Risk (RR) of this cluster is 6.14, representing how much more common disease is in this location and time period compared to the baseline
- SW corner of lower (RR) picked up in 1% Pop window only





Poisson Regression Analysis

Poisson Regression Analysis Steps

Acquiring Datasets

Combine variables into one dataset

Pearson's R Correlation Statistic, keep only variables with $R > \pm.30$, $P<0.05$

Test for Colinearity between variables, remove variables that seem to be too similar (Example: Total Cattle in county and Cattle/Km^2)

SAS GENMOD Stepwise Poisson Regression, paying attention to Chi Square and Deviance, and Type I & III Errors

Final Model Output and Interpretation

Modeling Disease, Host, Environment

Years: 1997, 2002, 2007
-Livestock totals
-Total farms
-Farm size

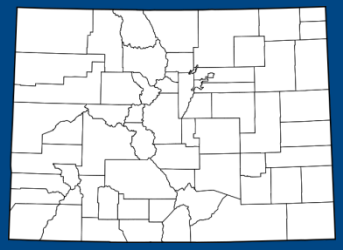
Years: 2001-2010
-Domestic wells
-Municipal wells
-Commercial wells

Years: 2001 -2010
-Employment counts (% of total state) by worker industry

FINAL DATASET

Years: 2010
-Urban, Rural, Frontier County classifications based on available health services

Years: 2006-2010
-Race/Ethnicity
-Age
-Gender



Years: 2001-2010
ArcGIS Shapefiles
-County boundaries
-County area
-Yearly Incidence Rates

Years: 2001-2010
-Unemployment rates

Pearson’s R – Continuous Variables

Variable	Pearson's R	P Value	Type
Avg # of Hogs	0.4627	0.0009	Strong Direct Relationship
Avg. # of Farmland (Acres)	0.43133	0.0005	Strong Direct Relationship
Avg. # Cattle	0.40183	0.0012	Strong Direct Relationship
COWS/KM2	0.43731	0.0004	Strong Direct Relationship
HOGS/KM2	0.44679	0.0015	Strong Direct Relationship
Avg. Broiler Chickens	0.40745	0.0074	Strong Direct Relationship
Median Farm Size (Acres)	0.31905	0.0108	Moderate Direct Relationship
FARMS/KM2	0.28804	0.021	Moderate Direct Relationship
Avg. # of Farms	0.28653	0.0217	Weak Direct Relationship
Avg # Farmer Employees	0.28487	0.0225	Weak Direct Relationship
10 Year Unemp. Rate (Mean)	-0.3694	0.0027	Moderate Inverse Relationship
10 Year Unemp. Rate(Median)	-0.34235	0.0056	Moderate Inverse Relationship

ANOVA – Categorical Variables

ANOVA: Single Factor

SUMMARY

Groups	*Count*	*Sum*
Urban	17	257.15
Rural	24	374
Frontier	23	379.31

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	19.84	2	9.92	0.1058	0.8997	3.15
Within Groups	5716.73	61	93.72			
Total	5736.57	63				

F is 0.10, F Critical is 3.14, .10 < 3.14 so we fail to reject H_0 that they don't differ

Model Output: Significant Variables

Parameter	URFClass	Wald Chi Sq.	P-Value	IRR
Intercept		**665.83**	**<0.0001**	
Farms/Km2	-	103.02	<0.0001	7.95463
Farm Employees (Avg)	-	12.91	0.0003	1.00012
Food Service Emp (Avg)	-	93.15	<0.0001	1.00003
% Pop.Under 5yo (2010Census)	-	84.73	<0.0001	0.84714
Median 10yr Unemp. Rate (%)	-	59.8	<0.0001	0.78437
Broiler Chickens (Avg)	-	6.69	0.0097	1.00018
URFClass-Urban	1	21.29	<0.0001	0.58295
URFClass- Rural	2	10.92	0.0009	0.74026
URFClass-Frontier	3			1.00000
Wells (# of)	-	46.74	<0.0001	0.99965
% Pop. Hispanic (2010 Census)	-	32.18	<0.0001	1.01385

Model Output: IRR Interpretations

- Increasing the number of farms/km^2 by one (1) farm increases the estimated incidence rate by a factor of 7.9.

- Increasing the number of broiler chickens in a county by 1000 increases the estimated incidence rate by a factor of 1.2.

- The incidence rate for living in an urban county is 0.59 times the incidence rate of living in a frontier county."

Model Output: Case Predictability

County	Cases	Predicted	residual
Cluster Counties			
Yuma	40	36.4	3.6
Washington	15	12.2	2.8
Kit Carson	34	18.7	15.3
Front Range Counties			
Adams	540	505.1	34.9
Jefferson	789	790.5	-1.4
Douglas	354	346.4	7.6
Weld	706	712.1	-6.2
Arapahoe	741	773.2	-32.2

Model Output: Goodness of Fit

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	31	152.7181	4.9264
Scaled Deviance	31	152.7181	4.9264
Pearson Chi-Square	31	139.6977	4.5064
Scaled Pearson X2	31	139.6977	4.5064
Log Likelihood		30476.1267	
Full Log Likelihood		-191.3600	
AIC (smaller is better)		404.7200	
AICC (smaller is better)		413.5200	
BIC (smaller is better)		423.8343	

Model Issues

- Too Small of N (42/64 counties)?
- Urban/Rural/Frontier, ANOVA = NO but Poisson model = YES?
- Inverse relationship Median Unemployment rate and Incidence?
- Missing important variable(s)?
- Incomplete data on existing variables?
- Other Methods?
- Modeling Case Counts instead of Rates?

Conclusions:

- Discernible clustering of campylobacter in the NE Region of the State across space and time (i.e. persistent)

- Employment and Ethnicity play a factor (proxy for SES)

(e.g. +Unemployment, - Reported Disease)

- The interaction between agent, host, and environment is complex -> It's just a model.

Study Implications and Actions

- Public health education in rural counties regarding illness and animal/farm contact
- Build relationships between rural local health agencies, their hospitals, and local agriculture businesses
- More in depth interview questionnaires in ‘cluster counties’ to identify more specific exposures
- Need for accurate and timely laboratory methods to confirm diagnosis so that similar studies can be conducted

THANK YOU!

Alicia Cronquist , CDPHE

Devon Williford, CDPHE

Russ Rickard, CDPHE

Local County Health Departments

Labs, Clinics, and Hospitals

QUESTIONS

Ben White, MPH

Phone: 303.691.4920

Email: Benjamin.White@state.co.us