Foodborne disease outbreak detection

Healthcare case study

Objective

- Attribution of Food borne illnesses to Food Commodities
- Measuring probability and magnitude of disease outbreaks at a specific region and time based on historical outbreak data
  - Government has a target for disease control – what is the probability that the targets can be met?

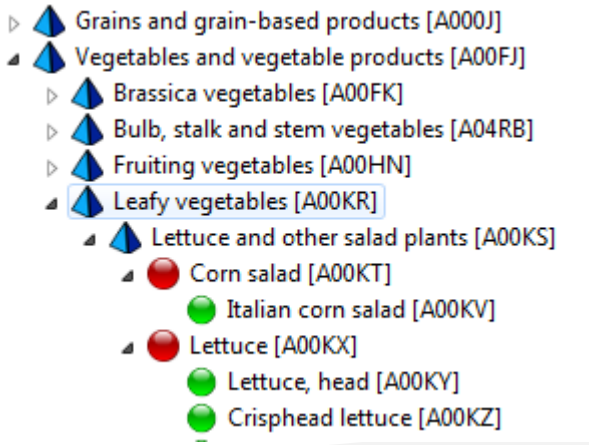
Pathogen / Syndrome	Year																2010 National health objective§	2020 National health objective¶
	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	
Surveillance population (millions)†††	14.27	16.13	20.71	25.86	30.64	34.85	37.86	41.75	44.34	44.77	45.32	45.84	46.33	46.76	47.14	47.51	47.51	
Campylobacter	23.59	24.55	19.42	14.82	15.36	13.63	13.38	12.63	12.82	12.71	12.73	12.81	12.64	12.96	13.52	14.28	14.30	12.3	8.50
Listeria**	0.43	0.43	0.53	0.40	0.33	0.26	0.25	0.31	0.26	0.29	0.28	0.26	0.26	0.32	0.27	0.28	0.25	0.24	0.20
Salmonella	14.46	13.55	13.61	16.07	14.08	15.04	16.24	14.46	14.65	14.53	14.76	14.89	16.09	15.02	17.55	16.45	16.42	6.8	11.40

Data Sources

- *CDC Foodborne disease data*

Year	Month	State	Genus_Species	Status	Location Of Consumption	Total Ill	Total Hospitalizations	Total Death	FoodVehicle
1998	June	Washington			Private home	2			chicken, unspecified
1998	August	Washington	Vibrio cholerae	Confirmed	Restaurant	2	0	0	oysters, unspecified
1998	September	Vermont	Salmonella enterica	Confirmed	Other	4	0	0	

- *European Food classification Standard*

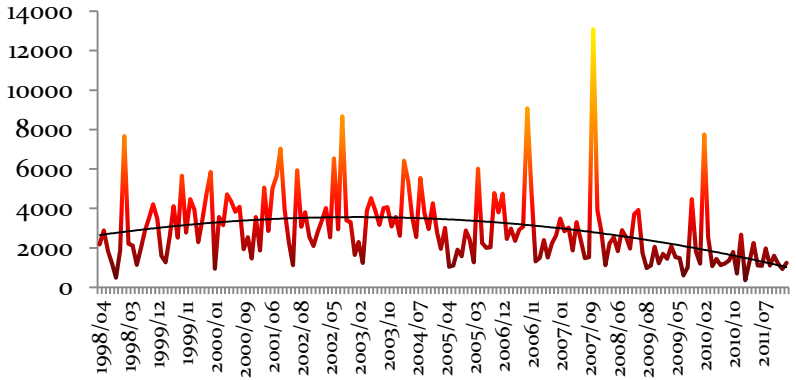


- *Pathogen to Etiology Mapping*

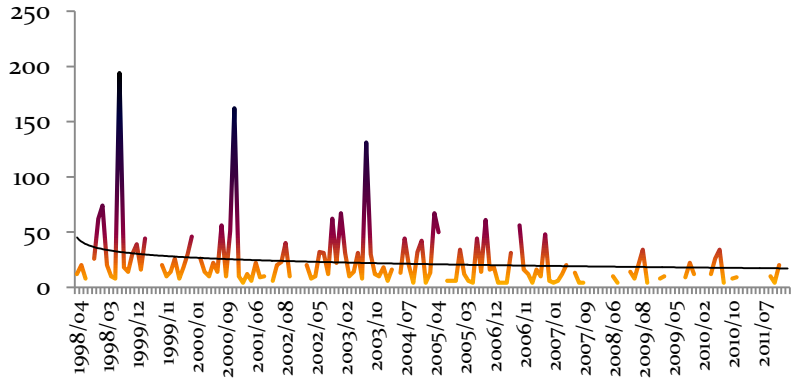
Etiology Type	**Pathogen - Genus Species (e.g.)**
Bacterial	Bacillus cereus
Chemical	Scombroid toxin
Viral	Staphylococcus aureus
Parasitic	Giardia intestinalis

Etiology progression

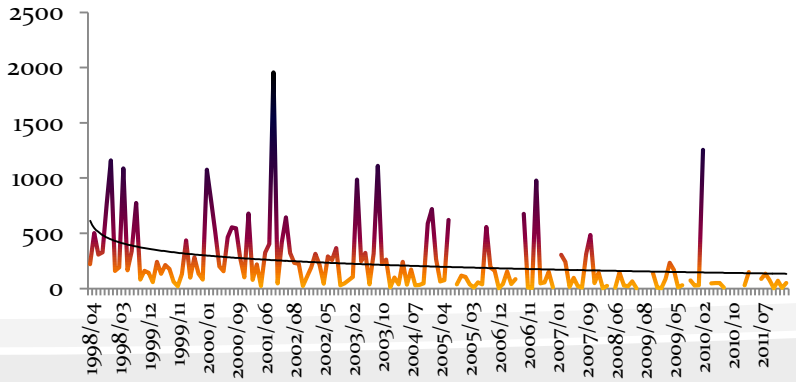
Bacterial



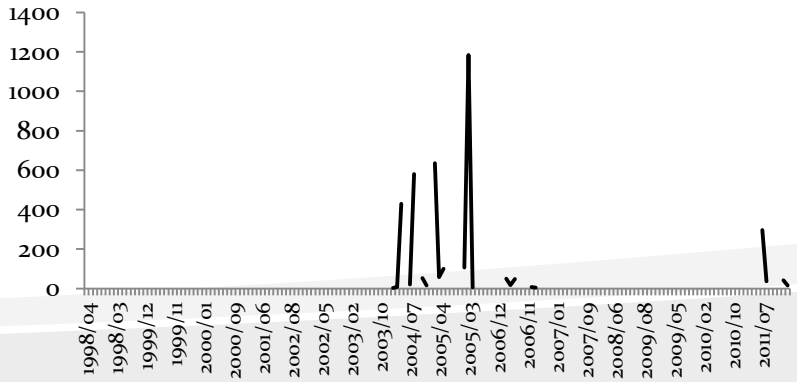
Chemical



Viral



Parasitic



Solution Approach

- Map the outbreaks to their food vehicle
  - Find the Root food group based on European Standard
- Cluster the states into similar outbreak regions
  - K-Means clustering
- Hierarchical clustering time periods within State clusters
- Within each state-time cluster group – compute the outbreak probability
  - Bayesian networks

Attribution of Food Groups to Disease outbreaks

	Etiology Type				
Food Group	**Bacterial**	**Chemical**	**Viral**	**Parasitic**	**All Agents**
vegetables and vegetable products	**44794 (19.6%)**	**173 (12.7%)**	**4153 (26%)**	**924 (38.1%)**	**59587 (18.7%)**
aromatic herbs or flowers, fresh	19508 (43.5%)	162 (93.9%)	2052 (49.4%)	410 (44.3%)	27640 (46.3%)
fungi	20581 (45.9%)	9 (5.1%)	1752 (42.1%)	125 (13.5%)	25016 (41.9%)
garden vegetables	2013 (4.4%)	0 (0%)	283 (6.8%)	28 (3.1%)	3007 (5%)
legumes, vegetable fresh	1593 (3.5%)	1 (0.8%)	27 (0.6%)	3 (0.3%)	1696 (2.8%)
marine algae	578 (1.2%)	0 (0%)	7 (0.1%)	25 (2.7%)	1177 (1.9%)
sprouted beans and seeds	176 (0.3%)	0 (0%)	5 (0.1%)	331 (35.8%)	550 (0.9%)
vegetable products	22 (0%)	0 (0%)	0 (0%)	0 (0%)	66 (0.1%)
grains and grain-based products	**21422 (9.4%)**	**14 (1%)**	**1644 (10.3%)**	**233 (9.6%)**	**27490 (8.6%)**
bread and similar products	6593 (30.7%)	6 (45.8%)	394 (23.9%)	32 (13.7%)	7896 (28.7%)
cereal bars	4609 (21.5%)	0 (0%)	196 (11.9%)	130 (55.7%)	5711 (20.7%)
cereals and similar	3990 (18.6%)	0 (0%)	379 (23%)	0 (0%)	5667 (20.6%)
fine bakery wares	3291 (15.3%)	7 (49.4%)	393 (23.9%)	0 (0%)	4273 (15.5%)
other cereal-based sUnknowncks	1916 (8.9%)	0 (0%)	89 (5.4%)	31 (13.3%)	2462 (8.9%)
pasta and similar products	855 (3.9%)	0 (4.7%)	184 (11.1%)	40 (17.1%)	1265 (4.6%)
raw doughs and pre-mixes	162 (0.7%)	0 (0%)	6 (0.4%)	0 (0%)	198 (0.7%)
meat and meat products	**17490 (7.6%)**	**18 (1.3%)**	**2350 (14.7%)**	**22 (0.9%)**	**24659 (7.7%)**
animal fresh meat	15382 (87.9%)	18 (100%)	1622 (69%)	22 (100%)	21134 (85.7%)
animal organs (edible offals non-muscle)	1301 (7.4%)	0 (0%)	659 (28%)	0 (0%)	2251 (9.1%)
animal other slaughtering products	661 (3.7%)	0 (0%)	67 (2.8%)	0 (0%)	1116 (4.5%)
meat products	142 (0.8%)	0 (0%)	1 (0%)	0 (0%)	154 (0.6%)
composite dishes	**8260 (3.6%)**	**58 (4.3%)**	**305 (1.9%)**	**44 (1.8%)**	**10445 (3.2%)**
dishes, incl. ready to eat meals (excluding soups and salads)	7881 (95.4%)	57 (98.8%)	272 (89%)	44 (100%)	9952 (95.2%)
soups and salads	379 (4.5%)	0 (1.1%)	33 (10.9%)	0 (0%)	493 (4.7%)
fish, seafood, amphibians, reptiles and invertebrates	**3416 (1.5%)**	**627 (46.3%)**	**853 (5.3%)**	**9 (0.3%)**	**5807 (1.8%)**
amphibians, reptiles, sUnknownils, insects	1069 (31.3%)	331 (52.8%)	59 (6.9%)	0 (0%)	1659 (28.5%)
crustaceans and products thereof	790 (23.1%)	2 (0.4%)	585 (68.6%)	1 (15.7%)	1605 (27.6%)
fish	545 (15.9%)	288 (45.9%)	12 (1.4%)	8 (84.2%)	996 (17.1%)
molluscs	706 (20.6%)	2 (0.3%)	75 (8.8%)	0 (0%)	910 (15.6%)
processed fish products	296 (8.6%)	2 (0.4%)	120 (14.1%)	0 (0%)	628 (10.8%)

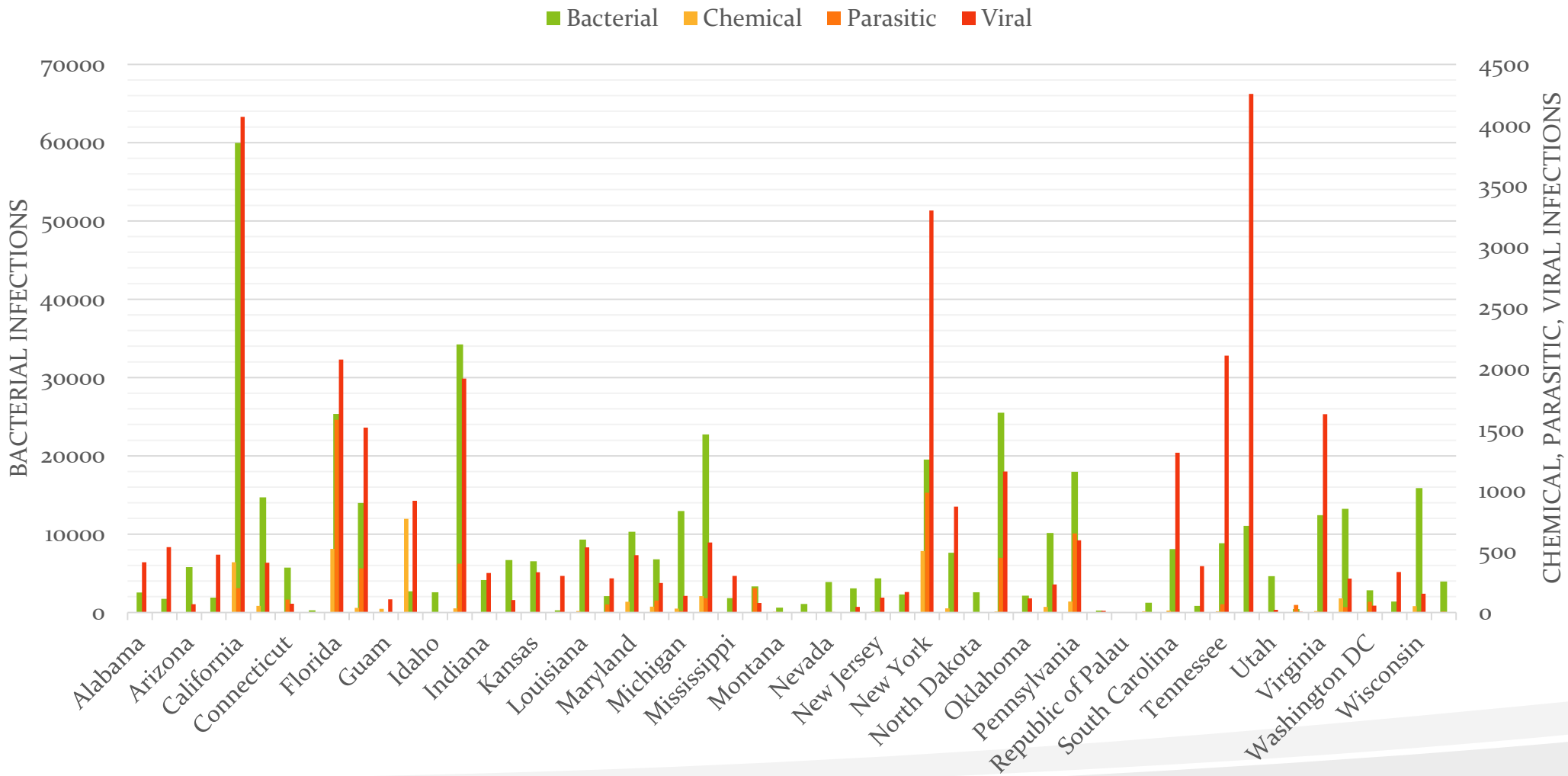
Attribution of Food Groups to Disease outbreaks (Contd..)

	Etiology Type				
Food Group	**Bacterial**	**Chemical**	**Viral**	**Parasitic**	**All Agents**
seasoning, sauces and condiments	**2610 (1.1%)**	**0 (0%)**	**498 (3.1%)**	**7 (0.2%)**	**3536 (1.1%)**
chutneys and pickles	1466 (56.1%)	0 (0%)	312 (62.7%)	0 (0%)	1971 (55.7%)
gravy ingredients	937 (35.9%)	0 (0%)	180 (36.2%)	7 (100%)	1311 (37%)
salt	110 (4.2%)	0 (100%)	5 (1%)	0 (0%)	134 (3.7%)
seasoning mixes	93 (3.5%)	0 (0%)	0 (0%)	0 (0%)	105 (2.9%)
table-top condiments	0 (0%)	0 (0%)	0 (0%)	0 (0%)	11 (0.3%)
milk and dairy products	**2639 (1.1%)**	**0 (0%)**	**148 (0.9%)**	**12 (0.5%)**	**3217 (1%)**
cheese	2300 (87.1%)	0 (0%)	144 (97%)	9 (76%)	2783 (86.5%)
dairy dessert and similar	331 (12.5%)	0 (100%)	4 (2.9%)	3 (24%)	407 (12.6%)
starchy roots or tubers and products thereof, sugar plants	**2211 (0.9%)**	**12 (0.9%)**	**158 (0.9%)**	**2 (0.1%)**	**3132 (0.9%)**
starchy root and tuber products	2099 (94.9%)	12 (100%)	158 (100%)	2 (100%)	2968 (94.7%)
starchy roots and tubers	111 (5%)	0 (0%)	0 (0%)	0 (0%)	163 (5.2%)
sugar plants	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
products for non-standard diets, food imitates and food supplements or fortifying agents	**2440 (1%)**	**0 (0%)**	**115 (0.7%)**	**41 (1.6%)**	**2921 (0.9%)**
food for particular diets	2388 (97.8%)	0 (0%)	115 (100%)	41 (100%)	2846 (97.4%)
meat and dairy imitates	51 (2.1%)	0 (0%)	0 (0%)	0 (0%)	75 (2.5%)
fruit and fruit products	**1324 (0.5%)**	**0 (0%)**	**109 (0.6%)**	**201 (8.3%)**	**1967 (0.6%)**
berries and small fruit	337 (25.5%)	0 (0%)	2 (1.8%)	201 (100%)	650 (33%)
citrus fruit	270 (20.4%)	0 (0%)	68 (63.1%)	0 (0%)	463 (23.5%)
dried fruit	313 (23.6%)	0 (0%)	34 (31.8%)	0 (0%)	369 (18.7%)
miscellaneous tropical and sub-tropical fruits	283 (21.4%)	0 (0%)	0 (0%)	0 (0%)	329 (16.7%)
pome fruit	80 (6%)	0 (0%)	3 (3.2%)	0 (0%)	118 (6%)
processed fruit products	36 (2.7%)	0 (0%)	0 (0%)	0 (0%)	36 (1.8%)
stone fruit	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
coffee, cocoa, tea and infusions	**1531 (0.6%)**	**0 (0%)**	**112 (0.7%)**	**0 (0%)**	**1888 (0.5%)**
coffee, cocoa, tea and herbal drinks	1143 (74.6%)	0 (0%)	96 (86%)	0 (0%)	1451 (76.8%)
coffee, cocoa, tea and herbal ingredients	213 (13.9%)	0 (0%)	10 (9.5%)	0 (0%)	245 (13%)

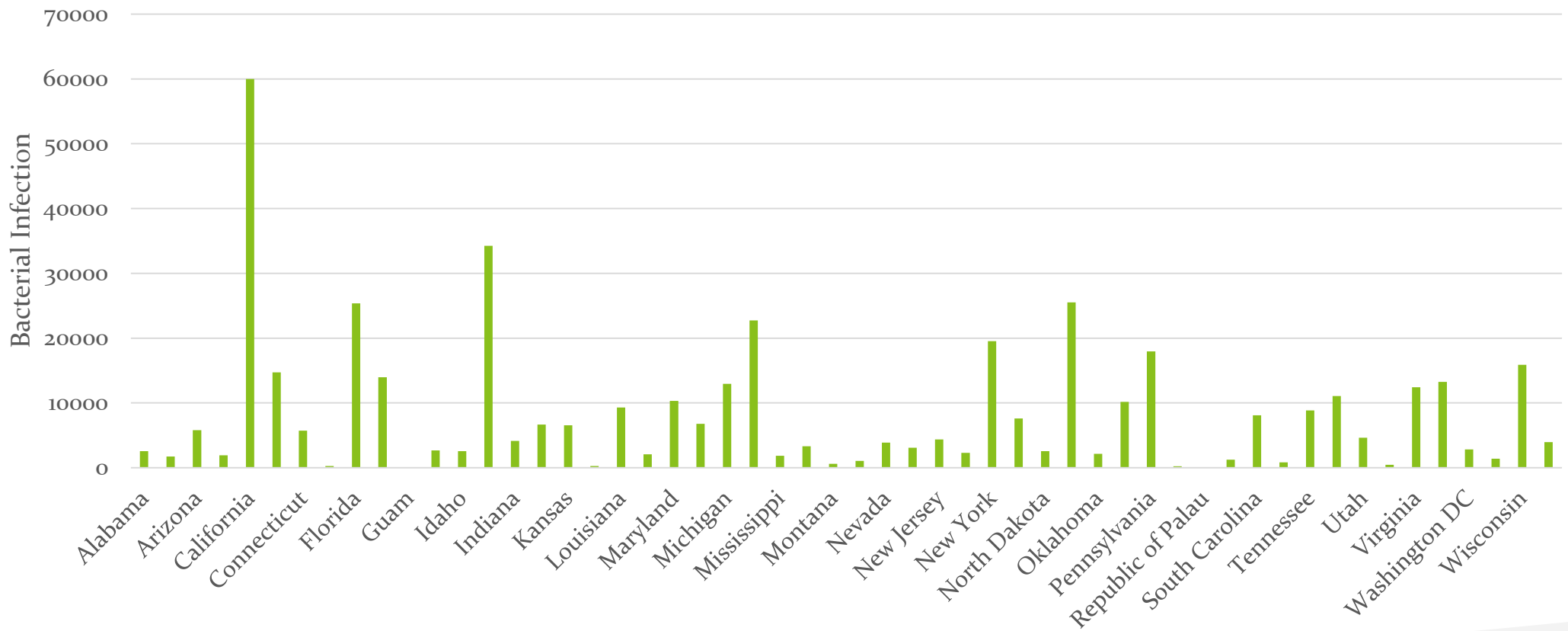
Attribution of Food Groups to Disease outbreaks (Contd..)

	Etiology Type				
Food Group	Bacterial	Chemical	Viral	Parasitic	All Agents
legumes, nuts, oilseeds and spices	**1170 (0.5%)**	**0 (0%)**	**4 (0%)**	**32 (1.3%)**	**1253 (0.3%)**
legumes, fresh seeds	735 (62.8%)	0 (0%)	1 (21%)	0 (0%)	747 (59.5%)
oilseeds and oilfruits	178 (15.2%)	0 (0%)	0 (0%)	0 (0%)	178 (14.2%)
processed legumes, nuts, oilseeds and spices	114 (9.7%)	0 (0%)	0 (0%)	0 (0%)	118 (9.4%)
spices	34 (2.9%)	0 (0%)	0 (0%)	32 (100%)	73 (5.8%)
tree nuts	49 (4.1%)	0 (0%)	0 (0%)	0 (0%)	72 (5.7%)
eggs and egg products	**839 (0.3%)**	**0 (0%)**	**1 (835.8%)**	**0 (0%)**	**882 (0.2%)**
food products for young population	**493 (0.2%)**	**0 (0%)**	**31 (0.1%)**	**0 (0%)**	**657 (0.2%)**
food for infants and young children	263 (53.3%)	0 (0%)	0 (0%)	0 (0%)	353 (53.8%)
fruit and vegetable juices and nectars	**526 (0.2%)**	**0 (0%)**	**18 (0.1%)**	**0 (0%)**	**603 (0.1%)**
concentrated or dehydrated fruit juices	344 (65.3%)	0 (0%)	0 (0%)	0 (0%)	372 (61.7%)
fruit juices	111 (21.1%)	0 (0%)	0 (0%)	0 (0%)	112 (18.5%)
mixed juices with added ingredients	71 (13.4%)	0 (0%)	0 (0%)	0 (0%)	73 (12.1%)
vegetable juices, ready to drink	0 (0%)	0 (0%)	18 (100%)	0 (0%)	45 (7.5%)
alcoholic beverages	**339 (0.1%)**	**9 (0.7%)**	**0 (0%)**	**123 (5%)**	**543 (0.1%)**
beer and beer-like beverage	169 (49.9%)	0 (0%)	0 (0%)	0 (0%)	183 (33.7%)
dessert wines	100 (29.7%)	9 (100%)	0 (0%)	0 (0%)	158 (29.1%)
mixed alcoholic drinks	19 (5.7%)	0 (0%)	0 (0%)	123 (100%)	145 (26.7%)
unsweetened spirits	49 (14.5%)	0 (0%)	0 (0%)	0 (0%)	49 (9.1%)
wine and wine-like drinks	0 (0%)	0 (0%)	0 (0%)	0 (0%)	6 (1.1%)
sugar, confectionery and water-based sweet desserts	**214 (0%)**	**0 (0%)**	**29 (0.1%)**	**0 (0%)**	**249 (0%)**
honey	195 (90.8%)	0 (0%)	29 (97.4%)	0 (0%)	227 (91%)
sugars	14 (6.5%)	0 (0%)	0 (0%)	0 (0%)	14 (5.6%)
sweet confectionery	5 (2.5%)	0 (0%)	0 (0%)	0 (0%)	7 (3%)
water-based ice creams	0 (0%)	0 (0%)	0 (2.5%)	0 (0%)	0 (0.3%)
water and water-based beverages	**149 (0%)**	**0 (0%)**	**0 (0%)**	**0 (0%)**	**159 (0%)**
diet soft drinks	143 (96.3%)	0 (0%)	0 (0%)	0 (0%)	153 (96.5%)
drinking water	3 (2%)	0 (0%)	0 (0%)	0 (0%)	3 (1.8%)
soft drinks	2 (1.6%)	0 (0%)	0 (0%)	0 (0%)	2 (1.5%)
animal and vegetable fats and oils	**50 (0%)**	**0 (0%)**	**0 (0%)**	**0 (0%)**	**50 (0%)**
animal fats and oils, processed	33 (67.1%)	0 (0%)	0 (0%)	0 (0%)	33 (67.1%)
spreadable fat emulsions and blended fats	16 (31.8%)	0 (0%)	0 (0%)	0 (0%)	16 (31.8%)
vegetable fats and oils, edible	0 (0.9%)	0 (0%)	0 (0%)	0 (0%)	0 (0.9%)
additives,flavours, baking and processing aids	**7 (325.8%)**	**0 (0%)**	**0 (0%)**	**0 (0%)**	**7 (233.9%)**
food additives	5 (78.4%)	0 (0%)	0 (0%)	0 (0%)	5 (78.4%)
miscellaneous composite agents for food processing	1 (21.5%)	0 (0%)	0 (0%)	0 (0%)	1 (21.5%)
Unknown	**115670 (50.8%)**	**439 (32.4%)**	**5415 (33.9%)**	**767 (31.7%)**	**167949 (52.9%)**
Unknown	115670 (50.8%)	439 (32.4%)	5415 (33.9%)	767 (31.7%)	167949 (52.9%)
Grand Total	**227603**	**1354**	**15951**	**2421**	**317011**

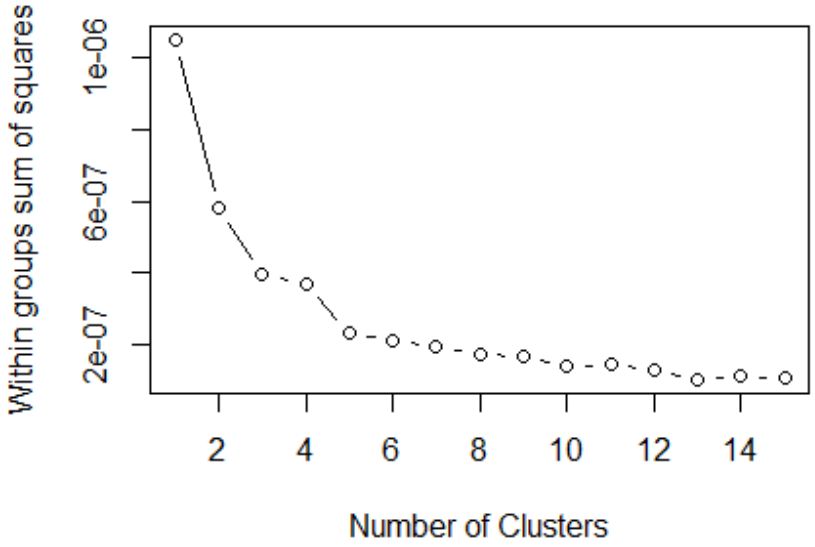
State wise distribution of Etiology



State wise distribution of Bacterial Etiology



Clustering

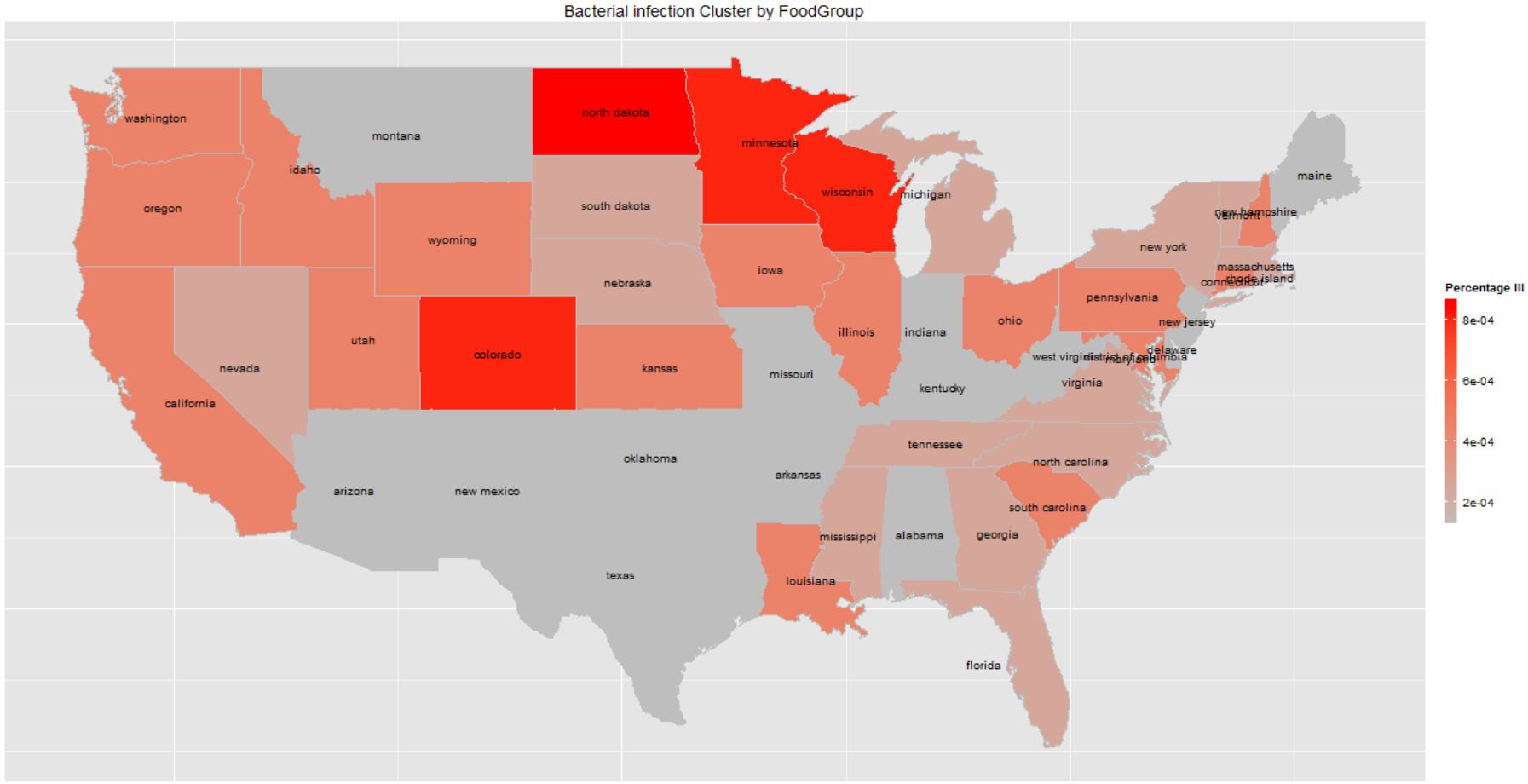


Cluster	Within Cluster Sum of Squares
1	0.00000002122
2	0.00000004837
3	0.00000009365
4	0.00000002656
5	0.00000004088

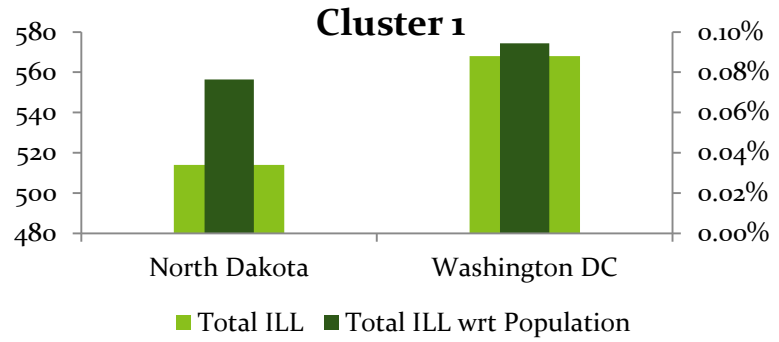
Cluster means:

Cluster	additives.flavours..baking.and.processing.aids_STD	alcoholic.beverages_STD	animal.and.vegetable.fats.and.oils_STD
1	0.00E+00	0.00E+00	0.00E+00
2	0.00E+00	2.67E-07	0.00E+00
3	1.88E-08	2.14E-06	2.50E-08
4	0.00E+00	1.31E-06	0.00E+00
5	0.00E+00	9.75E-07	2.05E-06

State Grouping based on Bacterial Infections due to Primary Food group

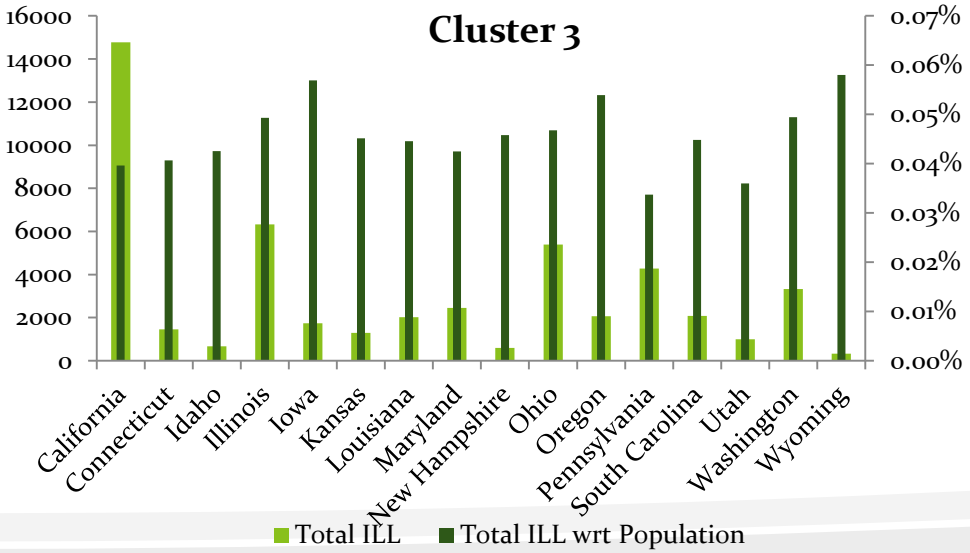


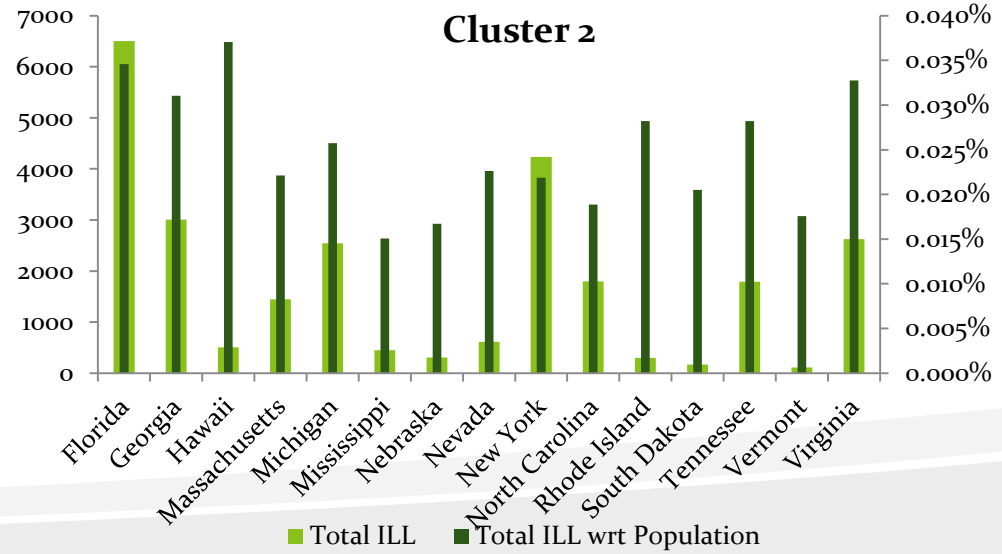
State Grouping based on Bacterial Infections due to Primary Food group (Contd..)



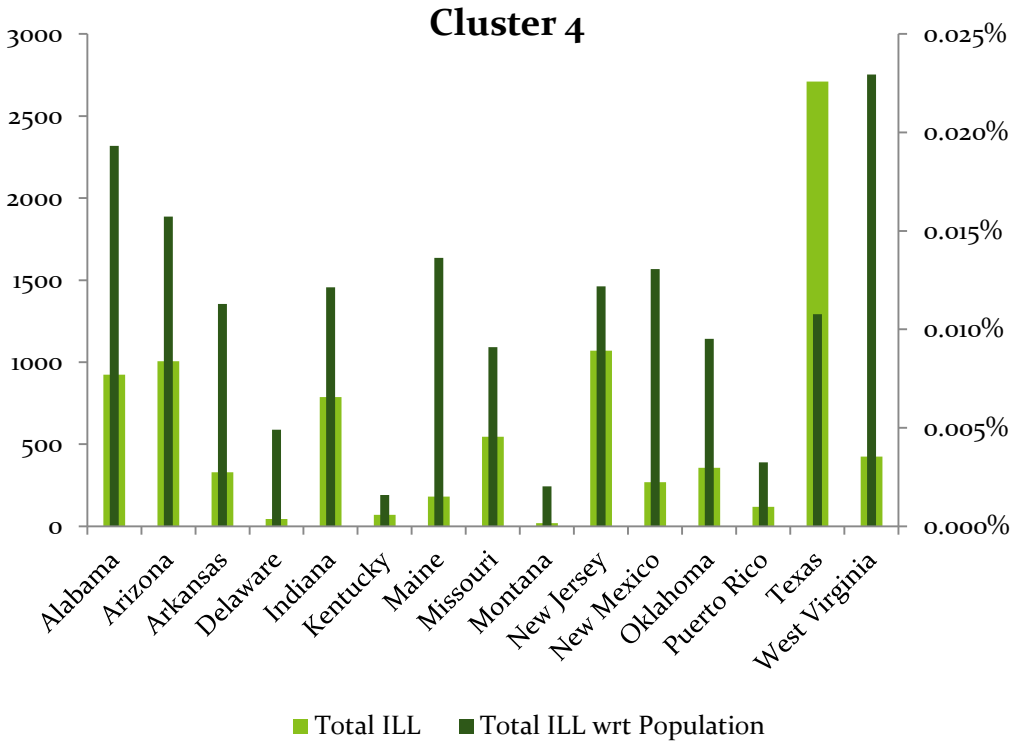
Total Ill Percentage

Cluster	Mean	Max	Min
1	0.085%	0.094%	0.076%
2	0.025%	0.037%	0.015%
3	0.046%	0.058%	0.034%



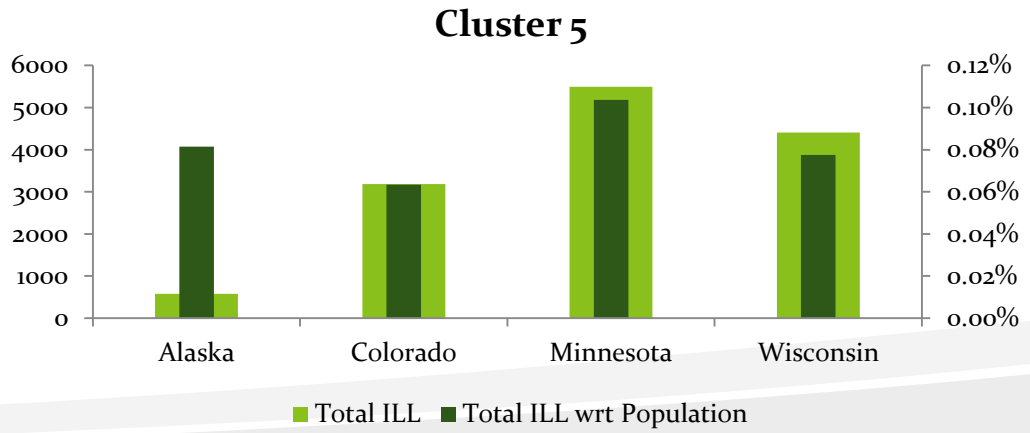


State Grouping based on Bacterial Infections due to Primary Food group (Contd..)

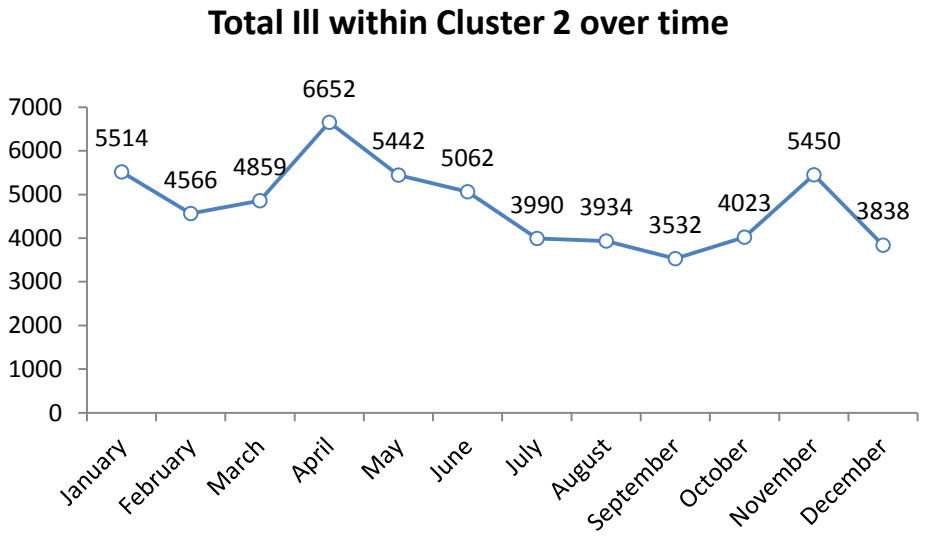


Total Ill Percentage

Cluster	Mean	Max	Min
4	0.011%	0.023%	0.002%
5	0.081%	0.104%	0.063%

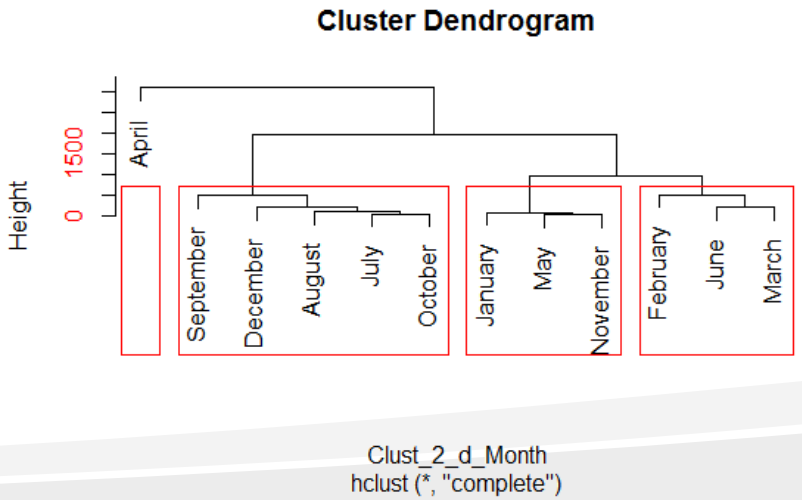


Outbreak Trend over time within each Cluster



Trends of Bacterial Infection spread within a cluster helps in identifying the likelihood of disease outbreaks for any given time frame for any Region in that Cluster.

Hierarchical Clustering on the disease outbreak volume helps in drilling down to the relevant subsets of disease outbreak.



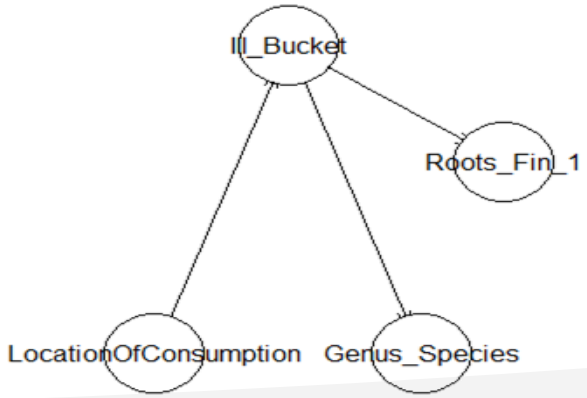
Bayesian Network to determine the probability of disease outbreak volume

Month	LocationOfConsumption	Genus_Species	Total_Ill	Food Group
July :477	Length:2013	Salmonella enterica :526	Min. : 0.000	Not present :717
August :442	Class :character	Norovirus Genogroup I :419	1st Qu.: 2.000	vegetables and vegetable products :447
October :386	Mode :character	Ciguatoxin :222	Median : 4.000	meat and meat products :234
September:365		Clostridium botulinum :171	Mean : 9.596	grains and grain-based products :226
December :343		Bacillus cereus :140	3rd Qu.: 10.000	fish, seafood:140
April : 0		E.coli, Enteropathogenic:118	Max. :232.000	composite dishes : 93
(Other) : 0		(Other) :417		(Other) :156

```
trainset1$Ill_Bucket
       n missing  unique
    2013       0       5

                1   2   3   4   5
Frequency 466 456 335 382 374
%          23  23  17  19  19
```

The network learned from the data using Hill search algorithm:



Cross Validation

* splitting 2013 datapoints in 5 subsets.

Cross Validation 1
> classification error for node Ill_Bucket is 0.3151365.
@ total loss is 0.3151365.

Cross Validation 2
> classification error for node Ill_Bucket is 0.2853598.
@ total loss is 0.2853598.

Cross Validation 3
> classification error for node Ill_Bucket is 0.3300248.
@ total loss is 0.3300248.

Cross Validation 4
> classification error for node Ill_Bucket is 0.2910448.
@ total loss is 0.2910448.

Cross Validation 5
> classification error for node Ill_Bucket is 0.2960199.
@ total loss is 0.2960199.

---

* summary of the observed values for the loss function:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.2854	0.2910	0.2960	0.3035	0.3151	0.3300

Prediction Model

Predicted Prob for Ill Bucket

Total_Ill	Ill_Bucket	1	2
23	2	0.24116	0.75884
27	2	0.147907	0.852093
27	2	0.126389	0.873611
143	2	0.656641	0.343359
3	1	1	0
3	1	1	0
3	1	1	0
3	1	1	0
2	1	0.314299	0.685701
2	1	0.325798	0.674202
2	1	0.317872	0.682128
9	2	0.134217	0.865783
60	2	0.086317	0.913683
20	2	0.090044	0.909956
17	2	0.556593	0.443407
17	2	0.450453	0.549547
30	2	0.3354	0.6646

ROC characteristic:
True vs. False : 0.8666667

Results

- The computed model strings enable one to predict the probabilities of different magnitudes of disease outbreak

- For example:
  - Predict the probability of 30 people getting ill due to consumption of leafy vegetables at picnic in Missisipi during September with Salmonella Entrica Infection

References

- Food Borne Disease Outbreak data: http://www.cdc.gov/outbreaknet/surveillance_data.html

- European Food safety Authority Food Classification System: http://www.efsa.europa.eu/en/datex/datexfoodclass.htm

- Predictive Analytics in Healthcare: https://www.researchgate.net/publication/236336250

Cenacle Research

Do More.