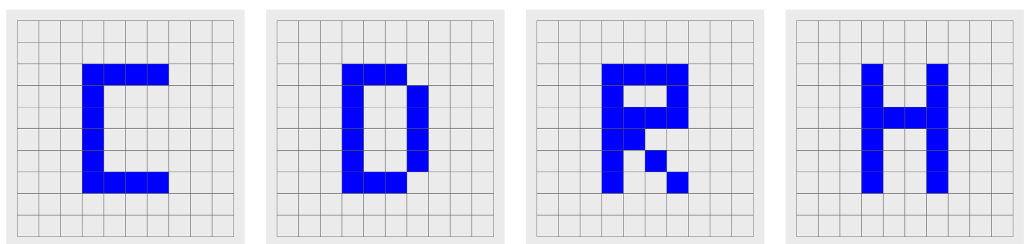
Bayesian Spatial Cluster Signal Learning with Application to Adverse Event (AE)

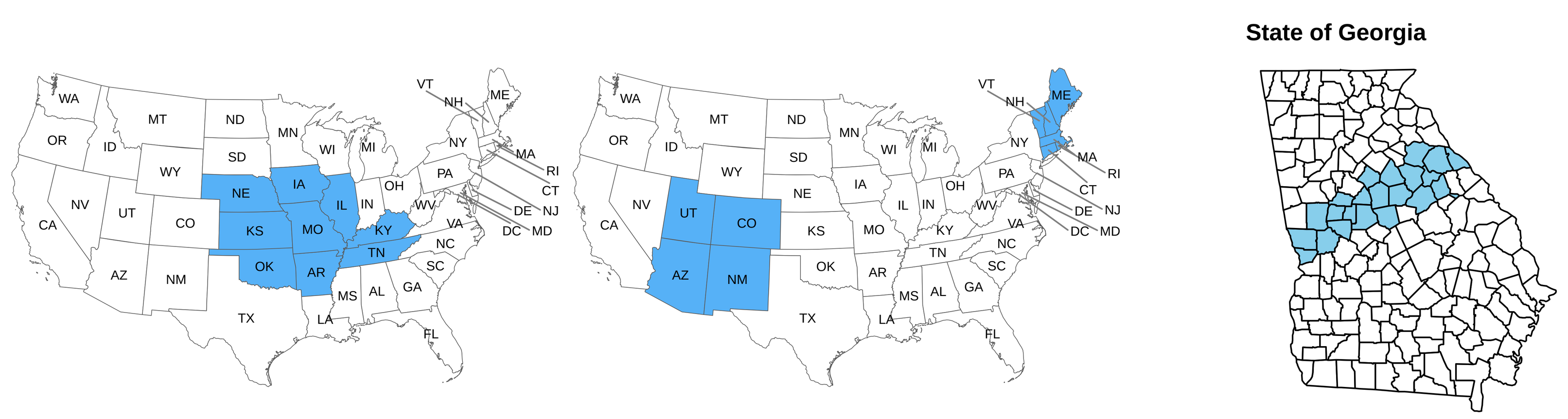
Hou-Cheng Yang [1] Guanyu Hu [2]

[1]U.S. Food and Drug Administration [2]University of Missouri

Motivation

The medical device industry plays an crucial role in the healthcare of patients worldwide. For example, it can do prevention, diagnosis, or even treatment for the patients. The safety, effectiveness, and security of medical devices in the United States are monitored by the U.S. Food and Drug Administration (FDA). Many manufacturers, and regulators are interesting in monitoring the safety of those medical devices. They tried to find whether there exist a geographic pattern of those adverse events (AEs). By exploring those finding pattern, it would be helpful for manufacturers and regulators to take corresponding actions.





Background, Challenges, Goals, and Contributions

- Background, and Challenges.
  - Hu et. al (2021) used a frequentist spatial scan statistics to find the potential spatial cluster for medical device safety.
  - Although spatial scan statistics is a useful tool but it suffers some obstacles.
    - Overfitting issue.
    - Contiguous moving window assumption.
    - Computational cost.
- Goals, and Contributions.
  - Our model provide an alternative way to identify the spatial signal region in an efficient manner.
  - Provide an ability to capture both locally spatially contiguous clusters and globally discontiguous clusters.

Notations

Assuming only one device and a fixed adverse event. Existing sub-region $s_1, \ldots, s_n$ in an entire geographical area Θ. We have the corresponding pair data $(y(s_i), t(s_i))$,

- $y(s_i)$ denotes the count of adverse event occurrences.
- $t(s_i)$ denotes the device exposure information.

We only need $y(s_i)$ to finding the potential signal clusters.

$$y(s_i) \sim \text{Poisson}(\lambda(s_i)), \quad i = 1, \cdots, n.$$

We assume the n data vectors can be clustered into k potential signal clusters.

Nonparametric Clustering

- What if the number of clusters is not fixed?
- Nonparametric: can grow if data need it.
- Probabilistic distribution over number of clusters.
  - We used Mixture of Finite Mixture (MFM) over Dirichlet Process Mixture (DPM) for clustering.

Markov Random Field (MRF)

Inspired by Orbanz and Buhmann (2008), we apply the pairwise Markov random filed into our model to handle spatial dependency. Consider an undirected random graph $G = (V, E, W)$, where V is the vertex set while E is the set of graph edges, with weights W on the corresponding edges. The pairwise MRF model is defined as follows,

$$\Pi(\lambda_1, \ldots, \lambda_k) = \exp\left\{\sum_{i \in E} H_i(\lambda_i) + \sum_{(i,j) \in E, j \neq i} H_{ij}(\lambda_i \lambda_j) - A(W)\right\}$$
$$\propto P(\lambda_1, \ldots, \lambda_k)\, M(\lambda_1, \ldots, \lambda_k),$$

where P is a vertex-wise term and M is an interaction term.

Bayesian Hierarchical Model: MRF-MFM-Poisson

Adapting MRF, and used Gamma distribution as the base measure, the Bayesian hierarchical model can be expressed as follows,

Data Model: $y(s_i) \mid \lambda(s_i) \sim \text{Poisson}(\lambda(s_i)), i = 1, \ldots n$

MRF: $(\lambda(s_1), \cdots, \lambda(s_n)) \sim M(\lambda(s_1), \cdots, \lambda(s_n)) \prod_{i=1}^{n} G(\lambda(s_i))$

MFM: $G(\lambda(s_i)) = \sum_{j=1}^{k} \pi_j \lambda_j,$

$\lambda_1, \ldots, \lambda_k \sim \text{Gamma}(a, b),$

$P(z_i = j) = \pi_j,$

$\pi_1, \ldots, \pi_k \mid k \sim \text{Dirichlet}(\ \ , \ldots, \ \ ),$

$k \sim p(\cdot)$, where $p(\cdot)$ is a p.m.f on $\{1, 2, \ldots\}$.

Likelihood Ratio Test (LRT)

Recall, for each sub-region, we have the corresponding pair data.

- First step only use $y(s_i)$ to finding potential signal clusters.
- In this step, both $(y(s_i), t(s_i))$ will be involved.

Assuming that,

$$n_Z \sim \text{Poisson}(p_Z * t_Z),$$
$$n_\Theta - n_Z \sim \text{Poisson}(q_Z * (t_\Theta - t_Z)), \quad n_Z \perp\!\!\!\perp (n_\Theta - n_Z).$$

The hypothesis testing as follows,

$$H_0: p_Z = q_Z \equiv p_0, \quad \forall Z \subseteq \Theta$$
$$H_A: p_Z > q_Z, \quad \text{for at least one zone.}$$

Then, the region associated with maximum log-likelihood ratio is identified as the most likely cluster signal.

References

[1] Tingting Hu, Lan Huang, Jianjin Xu, and Ram Tiwari. Spatial-cluster signal detection in medical devices using likelihood ratio test method. *Therapeutic Innovation & Regulatory Science*, 55:56–64, 2021.

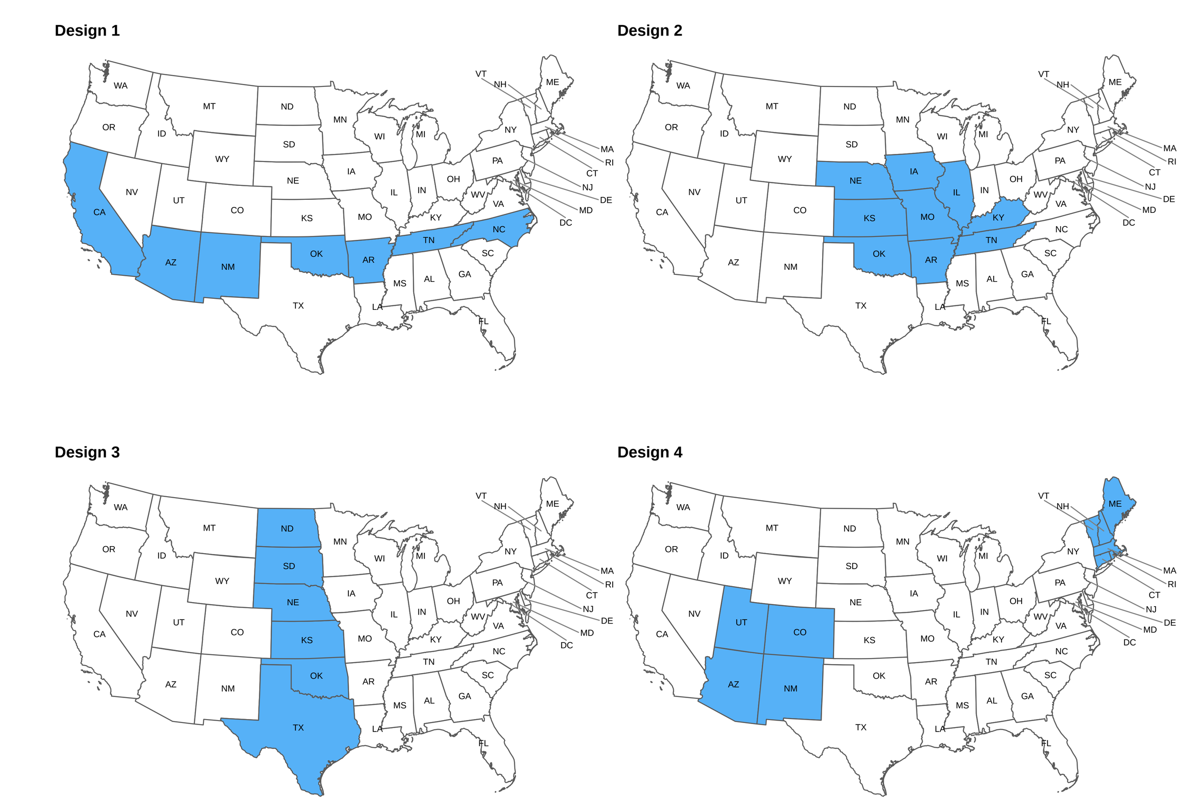
[2] Jeffrey W Miller and Matthew T Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356, 2018.

[3] Peter Orbanz and Joachim M Buhmann. Nonparametric bayesian image segmentation. *International Journal of Computer Vision*, 77:25–45, 2008.

[4] Peng Zhao, Hou-Cheng Yang, Dipak K Dey, and Guanyu Hu. Bayesian spatial homogeneity pursuit regression for count value data. *arXiv preprint arXiv:2002.06678*, 2020.

Real Data Application

Hypothetical data comprised total number of Left Ventricular Assist Device (LVAD) used in the state and the related number of stroke occurrences associated with LVAD use. Three combination cases of adverse event rates. The AE rate for the area outside the cluster is 0.1, and three different AE rates for the area inside the cluster is 0.297, 0.497, and 0,697.



		Design 1	Design 2	Design 3	Design 4
Case 1	Sen	0.927	0.975	0.908	0.944
	Spe	0.944	0.972	0.929	0.954
	PPV	0.802	0.906	0.798	0.884
	NPV	0.988	0.994	0.988	0.986
	Cover Rate	16.5%	34.1%	21.4%	22.8%
	Not Detected	0.4%	-	3.7%	1.4%
Case 2	Sen	0.990	0.997	0.996	0.995
	Spe	0.997	0.999	0.997	0.997
	PPV	0.988	0.994	0.988	0.988
	NPV	0.998	0.999	0.999	0.999
	Cover Rate	84.9%	92.5%	90.1%	83.4%
Case 3	Sen	0.998	0.999	0.999	0.999
	Spe	0.999	0.999	0.999	0.999
	PPV	0.999	0.999	0.999	0.999
	NPV	0.999	0.999	0.999	0.999
	Cover Rate	98.1%	99.3%	99.5%	98.8%

Discussions and Future Works

- Discussions
  - Our algorithm included two steps:
    1. First step, we used the proposed MRF-MFM-Poisson model to find potential signal clusters.
    2. Then we used likelihood ratio test (LRT) as second step on these potential clusters to identified the most likely spatial cluster signal.
  - Our model provide an alternative way to identify the spatial signal region in an efficient manner.
  - Our method provide an ability to capture both locally spatially contiguous clusters and globally discontiguous clusters.
  - A lot of investigations and strategies can be follow up depend on each specific scenario.
- Future Works
  - Extending to county-level, and city-level.
  - Considering sparsity issue.
  - Extending to spatio-temporal setting.