OAK RIDGE National Laboratory

Project 1: Unmeasured Spatial Risk in Disease Modeling

A Topological Data Analysis Approach

H.D. Fernando

Advanced Computing in Health Sciences,
Oak Ridge National Laboratory, Oak Ridge, USA
University of Tennessee, Knoxville, USA
03/10/2026

ORNL is managed by UT-Battelle LLC for the US Department of Energy

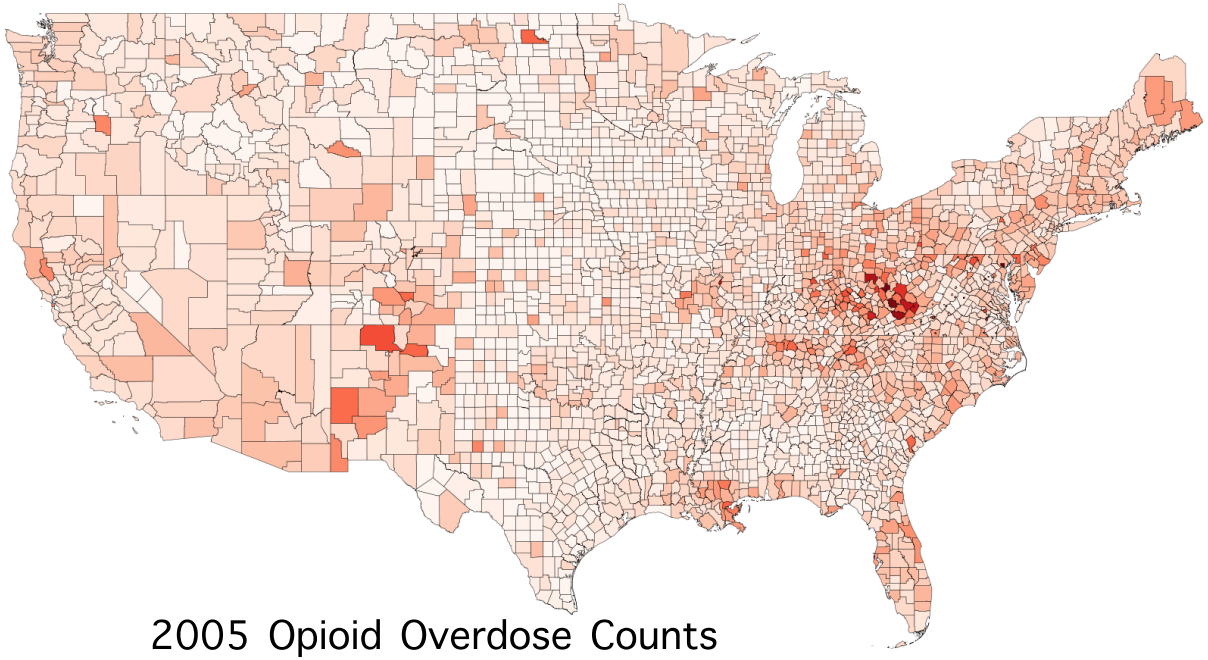
U.S. DEPARTMENT OF ENERGY

THE UNIVERSITY OF TENNESSEE KNOXVILLE

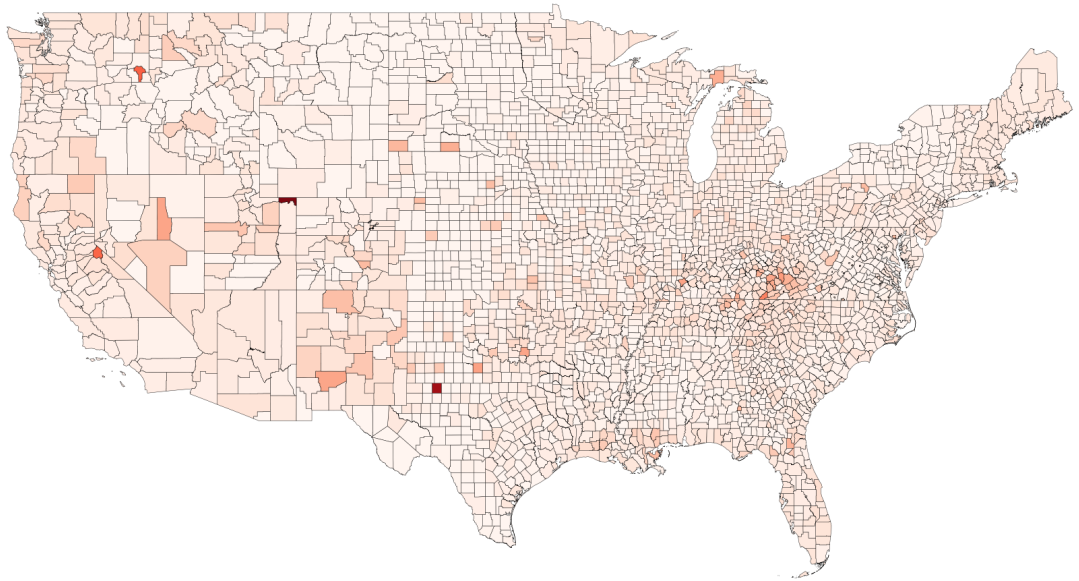
The Opioid Epidemic

- A complex public health crisis influenced by multiple factors.
  - Social connections.
  - Availability of opioids.
  - Socio-economic factors.
  - Geographical factors.
- How do researchers measure the risk of opioid overdos
  - Standardized Mortality Ratio (SMR).
    - Variance is high in small populations.
  - Spatial Bayesian Modeling.

2020 Opioid Overdose Counts



2005 Opioid Overdose Counts



Disease Mapping – Spatial Bayesian Modeling

Deaths can be modeled as

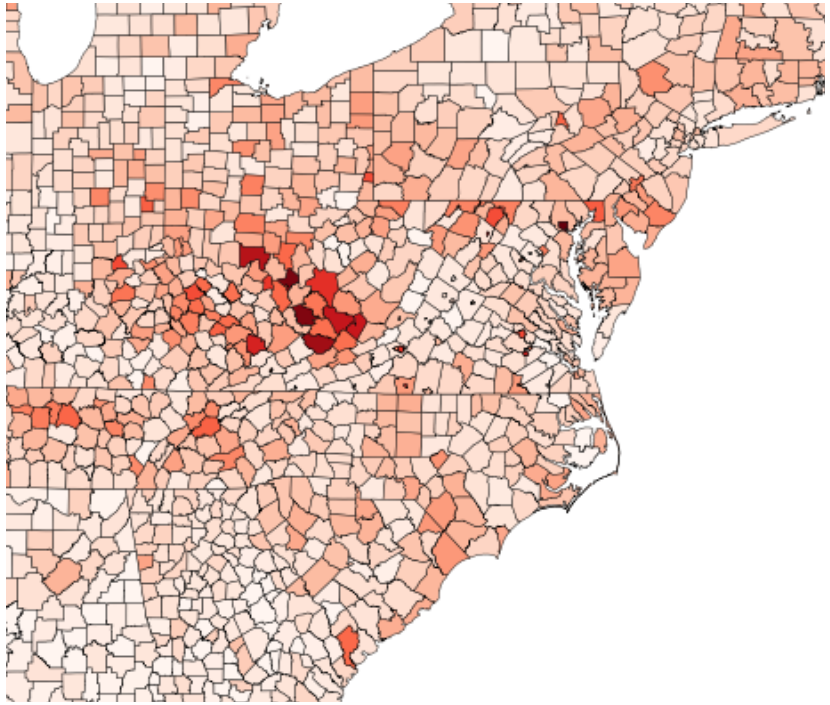
$$y_i|\theta_i \sim \text{Poisson}(E_i\theta_i)$$

Logarithms of relative risk is defi$\eta_i = log(\theta_i)$

Where:

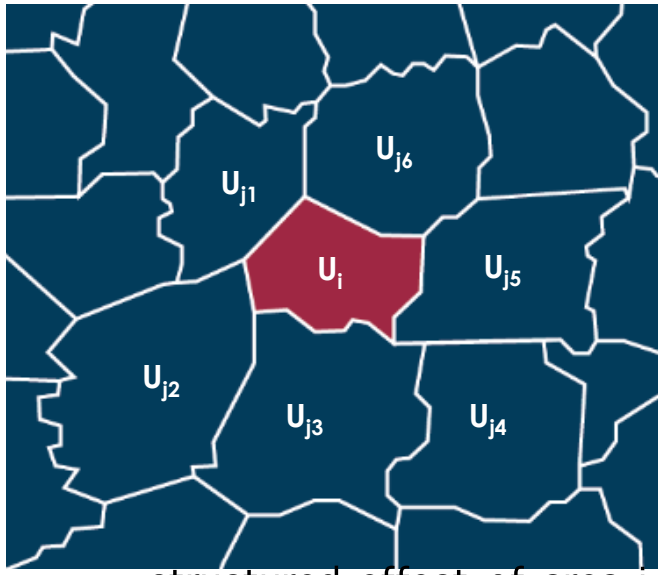
η_i $m + \beta \sum_{r=1}^{R} f(X_{ir}) + b_i$

- relative risk of area i.
- observed mortality count in area i.
- θ_i expected number of cases in area i.
- y_i overall intercept.
- E_i regression coefficient(s).
- m covariate for area i / measurable factor(s).
- β spatial random effect of area i (captures unmeasurable influence).
- X_{ir}
- b_i



Spatial Distribution of Observed Death Counts.

Spatial Random Effect (b)



- Structured Effect(u)

$$u_i | u_{i\sim j} \sim N\left(\frac{\sum_{i\sim j} u_i}{d_i}, \frac{\sigma_i^2}{d_i}\right)$$

  - Areas that are close to each other may have similar values.
  - Use conditional autoregressive (CAR) models.
  - Smooths the data according to a certain neighborhood structure.
  - Joint specification of u_i is a multivariate normal distribution[1].

- Unstructured Effect(v)

$$u_i \sim N(0, Q^{-1})$$

  - Random variation.
  - Independent across each area $v_i \sim N(0, \tau^{-1} I)$
  - Variation that cannot be explained by the rest of the model (including any covariates).

- u_i structured effect of area i.
- d_i number of neighbors of area i.
- σ_i^2 unknown variance.
- v_i unstructured effect if area i.
- τ precision parameter.
- Q Precision matrix.

OAK RIDGE National Laboratory

[1] - Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society: Series B (Methodological), 36(2), 192-225.

Spatial Random Effect (b_i) In Bayesian Disease Models

- The Besag-York-Mollie (BYM) Model.

$$b_i = u_i + v_i$$

- The Besag-York-Mollie 2 (BYM2) Model (2016).

$$b_i = \frac{1}{\sqrt{\tau_b}}(\sqrt{1-\phi}v_i + \sqrt{\phi}u_{*i})$$

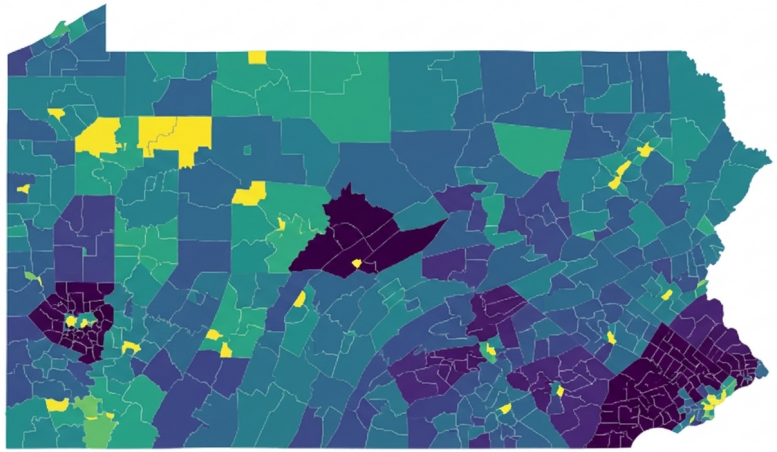
Where

- is the structured effect.
- is the unstructured effect.
- u_i
- controls the precision.
- v_i
- is the mixing parameter.
- τ_b
- ϕ

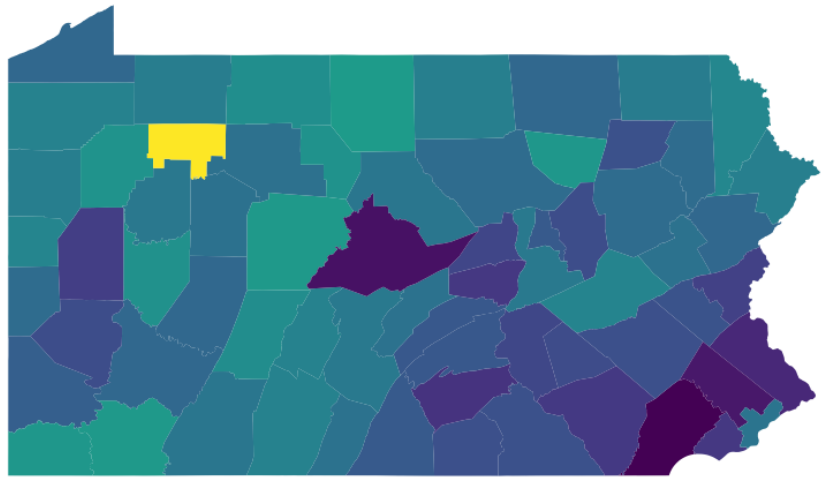
Practical Problems in Estimating Relative Risks in Large Areas

- Most disease outcome data available at the county level.
  - Example: Outcome data (mortality) available at the county level, whereas covariate data (socioeconomics) available at the census tract level.
- Disease risks at a regional level (with large spatial units) are likely to show both areas of smooth evolution and places where sudden changes happen.
- Standard Bayesian spatial models introduce global spatial smoothing across the entire study region.

Pennsylvania

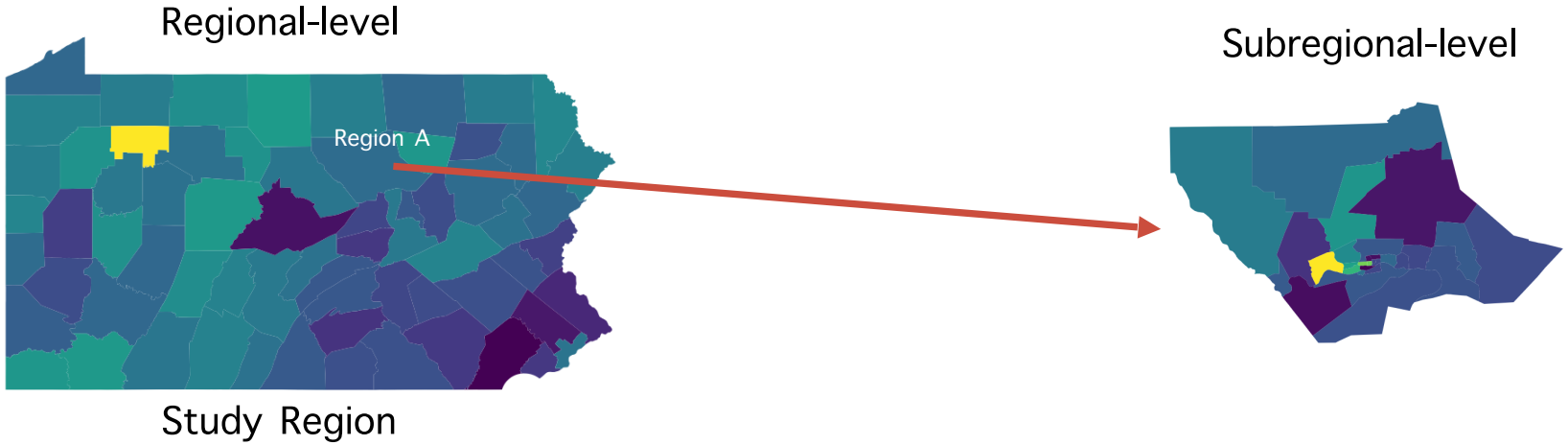


Subregional-level – Census tracts



Regional-level - Counties

Our solution?

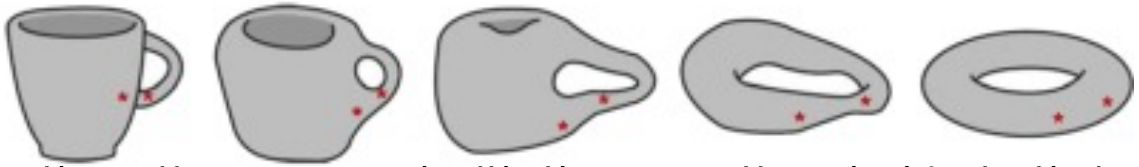


Let's consider a regional-level disease risk model,

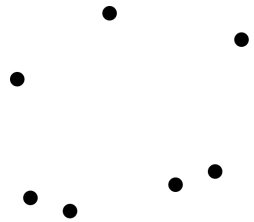
- Each individual region has multiple subregions.
- If we take a covariate in the model, each region has subregional-level distribution of the covariate data.
- We use persistence homology (a method in topological data analysis), to extract topological insights (summaries) from subregional-level covariate data.
- Then we integrate regional-level model with topological insights to reduce the uncertainty in their regional-level risk estimates, thereby helping the model to identify sharp risk shifts among regions.

OAK RIDGE National Laboratory

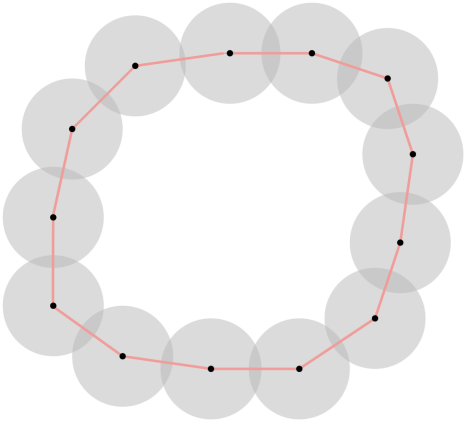
Topological Data Analysis (TDA)



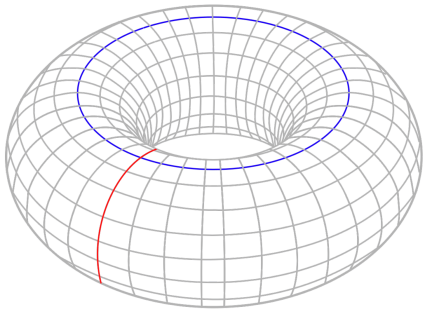
- Topology is a branch of mathematics concerned with the properties of objects that remain unchanged under continuous deformation, such as stretching, bending, and twisting.
- TDA identifies topological features within the data.
- Homology counts the different types of topological features (holes) in the data, helping to understand the shape.



(a) Connected components



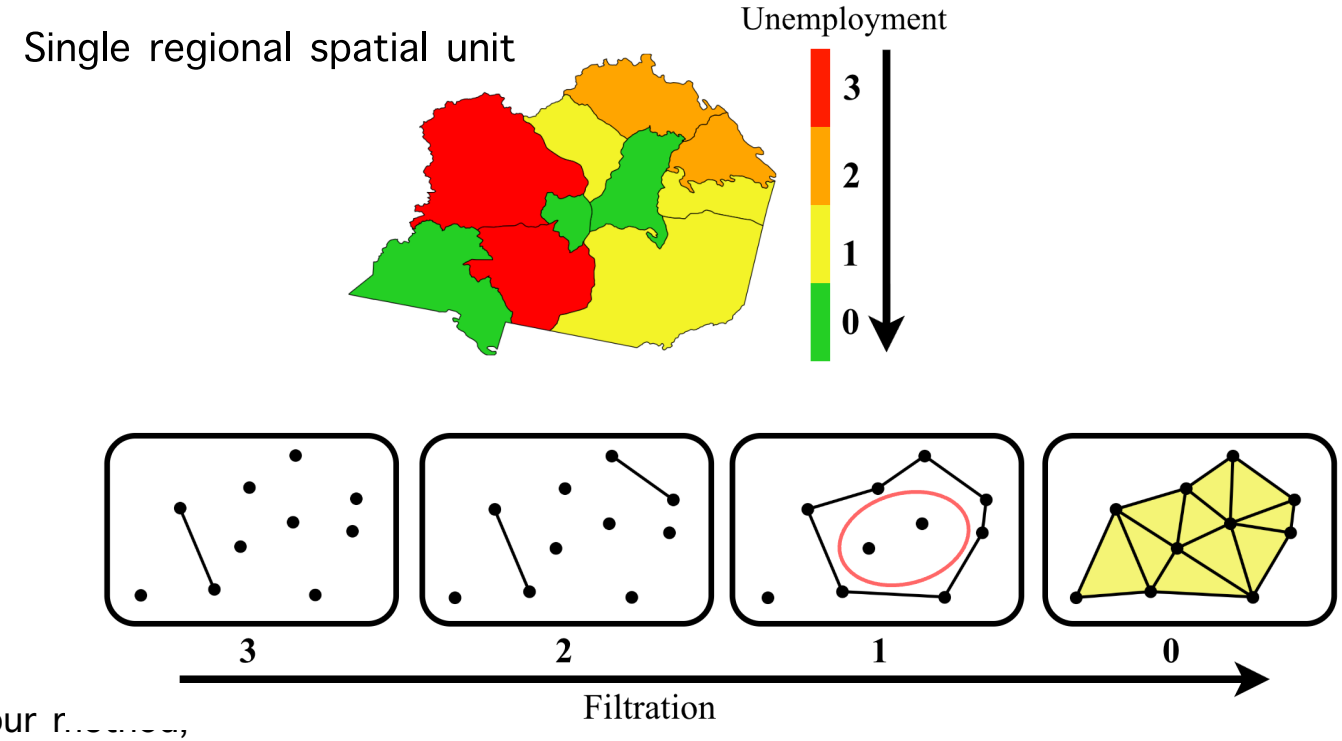
(b) Loops

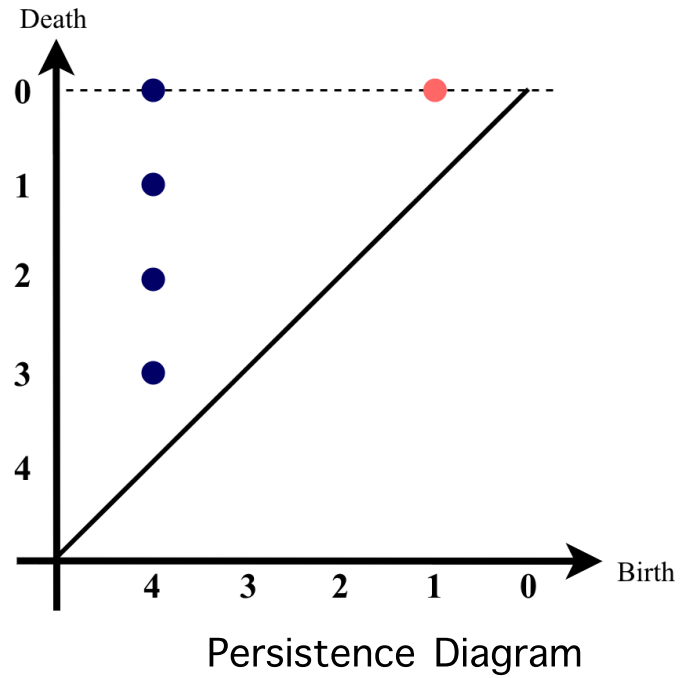


(c) Voids

OAK RIDGE National Laboratory

Persistent Homology of Spatial Data





- In our r...,
  - We consider distance in variable space.
  - We create edges by geographical adjacency and variable similarity.
  - We employ super-level set filtration (specifically looking at "peaks").
- A persistence diagram is a collection of the pairs of birth and death times of topological features obtained from persistent homology.

OAK RIDGE National Laboratory

Topological Summaries

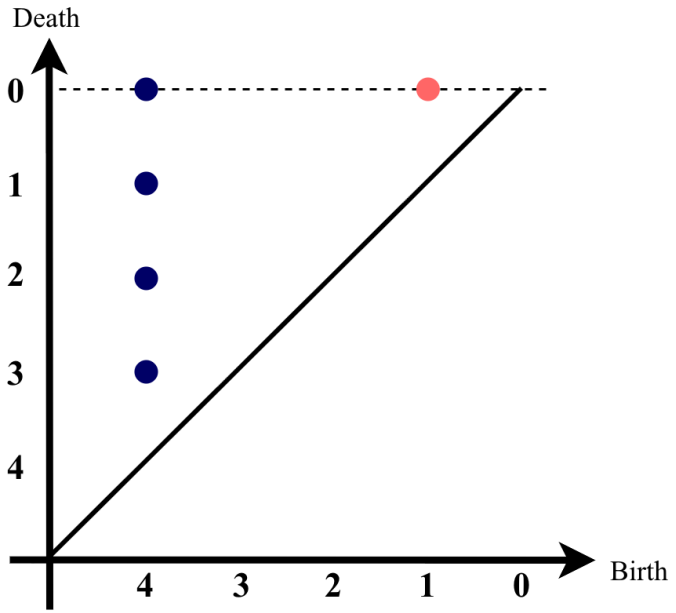
Summarizes the persistence of the topological features.

1. Average Lifespan (AL) = $\frac{1}{T}\sum_{t=1}^{T}(z_t - w_t)$
2. Average Midlife (AML) = $\frac{1}{T}\sum_{t=1}^{T}(z_t + w_t)/2$

Here z_t is the birth of a topological feature t, and w_t is the death of a topological feature t. Total number of features denoted as T.

We are only considering connected components in our study.

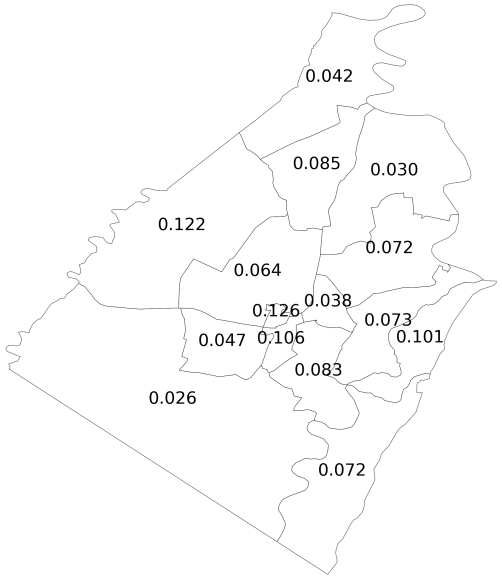
- This choice is motivated by the fact that persistence pairs of the connected components in persistence diagram serve as identifiers for peaks, or local maxima, in spatial data [2, 3] .



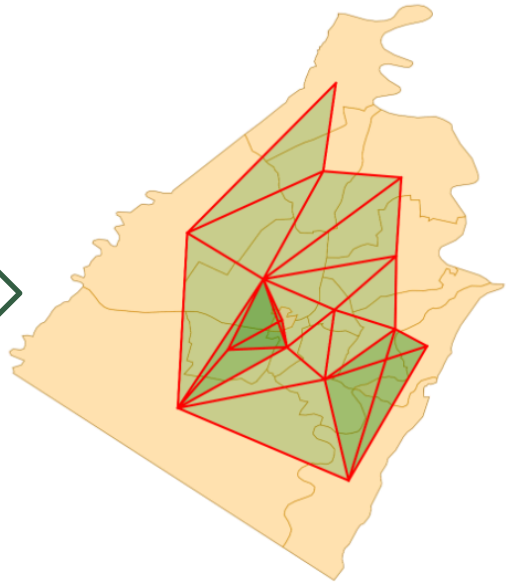
Persistence Diagram

[2] - S. Huber, Persistent homology in data science, in International Data Science Conference, Springer, 2020, pp. 8
[3] - K. Limbeck and B. Rieck, Detecting spatial dependence in transcriptomics data using vectorised persistence diagrams, arXiv preprint arXiv:2409.03575, (2024).

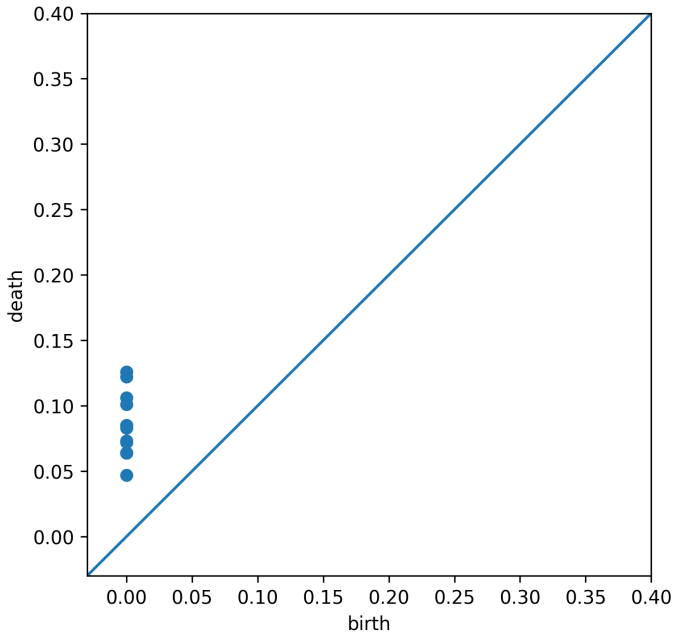
In Summary



Region A; Covariate: unemployment



Filtered simplicial complex



Persistence diagram

Topological summary for unemployment In Region A

OAK RIDGE National Laboratory

The BYM-Topological (BYM-T) Model

Risk defined as

Spatial random effect defined as $$\eta_i = \xi + x_i^T \beta + b_i^{new}$$

Where: $$b_i^{new} = \frac{1}{\sqrt{\tau_b}}(\sqrt{\alpha_1} v_i + \sqrt{\alpha_2} u_i^* + \sqrt{\alpha_3} h_i)$$

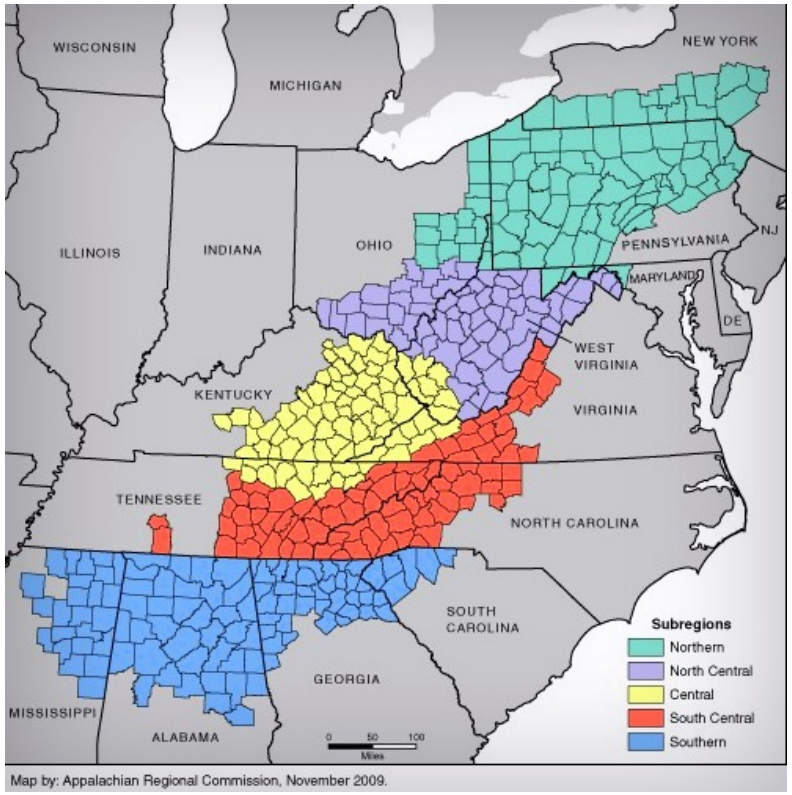
- A
- $h_i = \sum_{j=1}^{m} h_{ij}$
- is the ea i,
- $\alpha_1 + \alpha_2 + \alpha_3 = 1$.
- is the unstructured effect of area i,
- u_i is the topological effect of area i.

v_i

h_i

Single Covariate Study (Real-world data)

- Core Hypothesis: Integrating topological summaries of subregional-level covariate data into regional-level Bayesian disease modeling reduces uncertainty in regional-level risk estimates, thereby helping the model to identify sharp risk shifts among regions.
- Study Regions: West Virginia, Tennessee
- Data: Opioid overdose data (2018), and Social vulnerabilities index (SVI) data (2018).
- Experiment
  - Used Poverty as the single covariate to isolate the impact of topological summaries.
    - Fixed effect covariate uses county level poverty data.
    - Topological Input uses census tract-level poverty data to calculate topological summaries.
- Model Framework
  - MCMC simulations with PYMC.
  - Comparison
    - Baseline: BYM2 Model (Standard spatial model).
    - Proposed: BYM-T Model (Incorporates topological summaries).



Appalachian Region.

Single Covariate Study

Posterior parameter estimates and model comparison for single-covariate(poverty) studies in West Virginia and Tennessee, with predictive performance assessed using ELPD LOO between BYM-T and BYM2 relative risk maps.

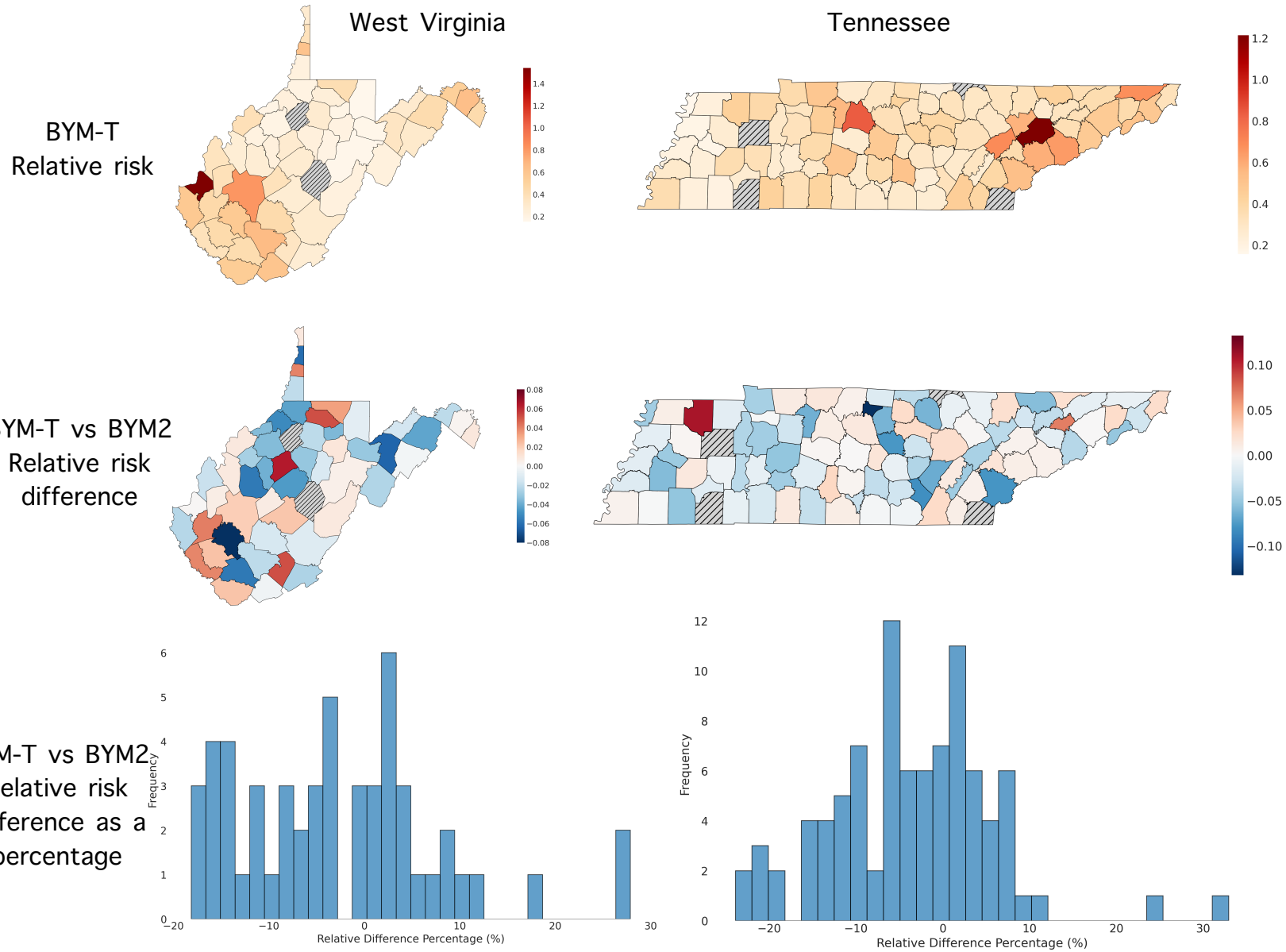
State	Model	$1/\sqrt{\tau_b}$	ϕ	$1-\phi$	$\alpha_1, \alpha_2, \alpha_3$	$\mathrm{ELPD}_{\mathrm{LOO}}$	$\mathrm{SE}(\mathrm{ELPD}_{\mathrm{LOO}})$
West Virginia	BYM-T	0.522	-	-	0.201, 0.516, 0.284	-143.230	6.904
	BYM2	0.544	0.502	0.498	-	-146.140	6.553
Tennessee	BYM-T	0.438	-	-	0.421, 0.340, 0.239	-246.736	9.509
	BYM2	0.452	0.275	0.725	-	-247.651	8.157

- Model parameters are directly comparable among both models.
  - and represents the same structured component.
  - ϕ and α_2 represents the same unstructured component.
- Redistribution of $1-\phi$ α_1 Spatial Weights
  - Significant Shift: Weights shift from "unstructured noise" to the "topological component."
  - Unstructured Component:
    - Decreased from 0.498 (BYM2) to 0.201 (BYM-T) in West Virginia.
    - Decreased from 0.725 (BYM2) to 0.421 (BYM-T) in Tennessee.
  - This decreased weight is shifted towards topological component .

α_3

Single Covariate Study: Impact on Relative Risk

- Based on the numerical difference, the BYM-T model adjusts risk estimates upward or downward compared to the baseline in both states.
- Percentage of relative risk difference max ~30% increase in some counties and ~20% decrease in some counties while most counties have slightly changed their relative risk.
- To identify the subregional heterogeneities driving these risk differences, we need to further examine counties with higher risk differences(%).



OAK RIDGE National Laboratory

Single Covariate Study

TABLE 2. Comparison of the top four increases (+) and decreases (-) in relative difference percentages for West Virginia and Tennessee.

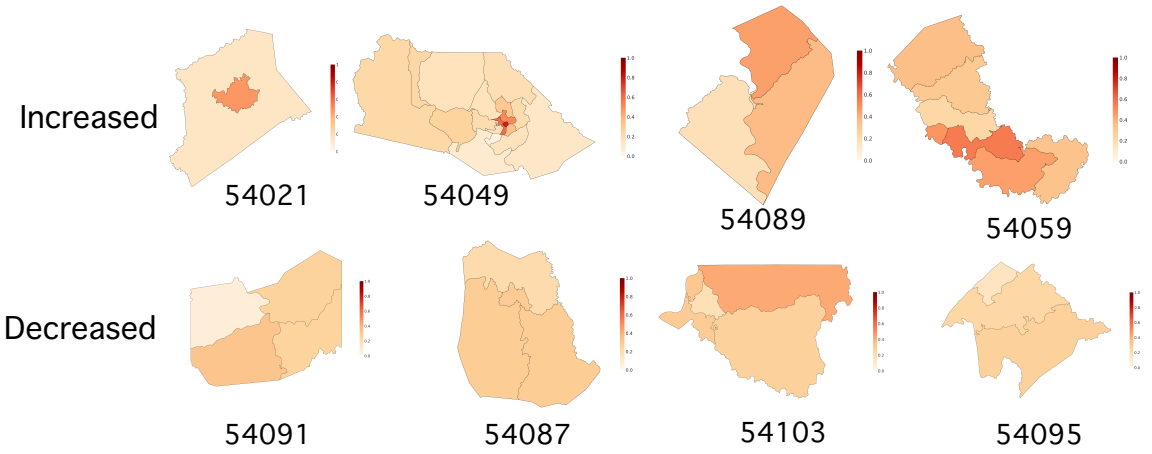
	West Virginia				Tennessee			
Trend	FIPS	Diff (%)	Lifespan	Poverty Range	FIPS	Diff (%)	Lifespan	Poverty Range
+	54021	27.875	0.902	0.137–0.473	47183	32.941	3.409	0.143–0.961
	54049	26.606	2.409	0.046–0.788	47063	27.708	1.448	0.088–0.654
	54089	17.238	0.133	0.174–0.446	47117	10.928	0.082	0.108–0.370
	54059	11.409	0.249	0.228–0.559	47087	9.136	0.073	0.190–0.418
–	54091	-16.252	-1.202	0.077–0.326	47069	-20.404	-0.970	0.179–0.299
	54087	-17.405	-0.236	0.205–0.296	47033	-21.801	-0.980	0.204–0.284
	54103	-17.453	-0.953	0.177–0.424	47095	-22.720	-0.932	0.309–0.370
	54095	-18.128	-0.722	0.151–0.264	47169	-23.870	-1.201	0.195–0.214

Note: Trend: (+) = Increase, (–) = Decrease. Diff = Relative Risk Difference.
Lifespan = Scaled Avg. Lifespan. Poverty Range denotes min-max census tract values.

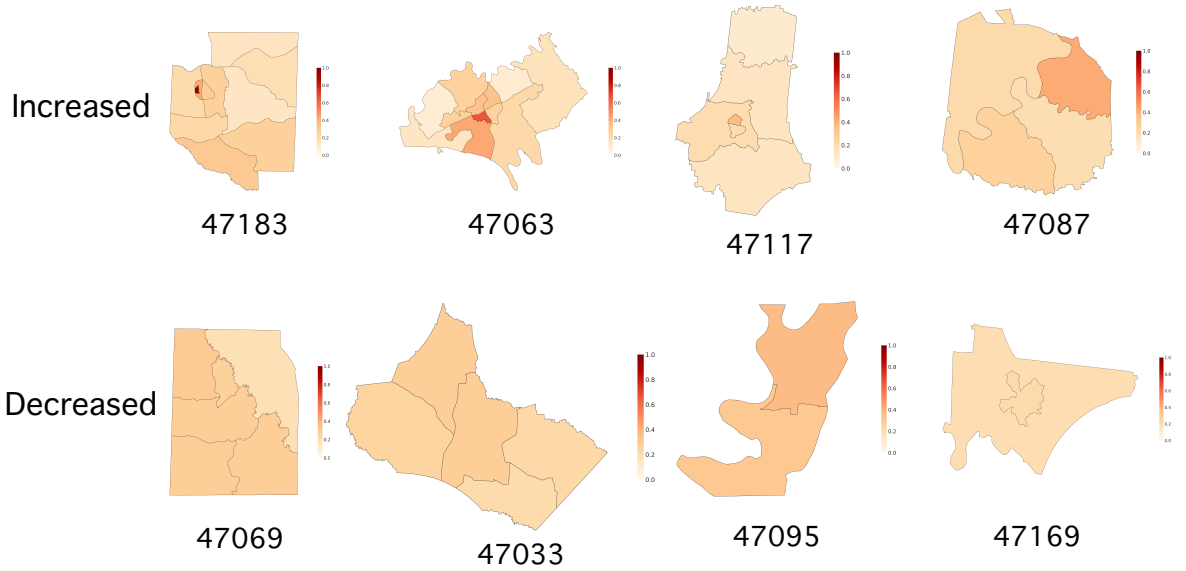
We examine top 4 with increased relative risk, and top 4 counties decreased relative risk in both West Virginia and Tennessee.

- Counties with increased relative risk.
  - High scaled topological summaries.
  - Census tract level poverty distribution with high poverty values or peaks.
- Counties with decreased relative risk.
  - Low scaled topological summaries.
  - Census tract level poverty distribution with low poverty values.

West Virginia counties top 4 (decrease/increase)



Tennessee counties top 4 (decrease/increase)

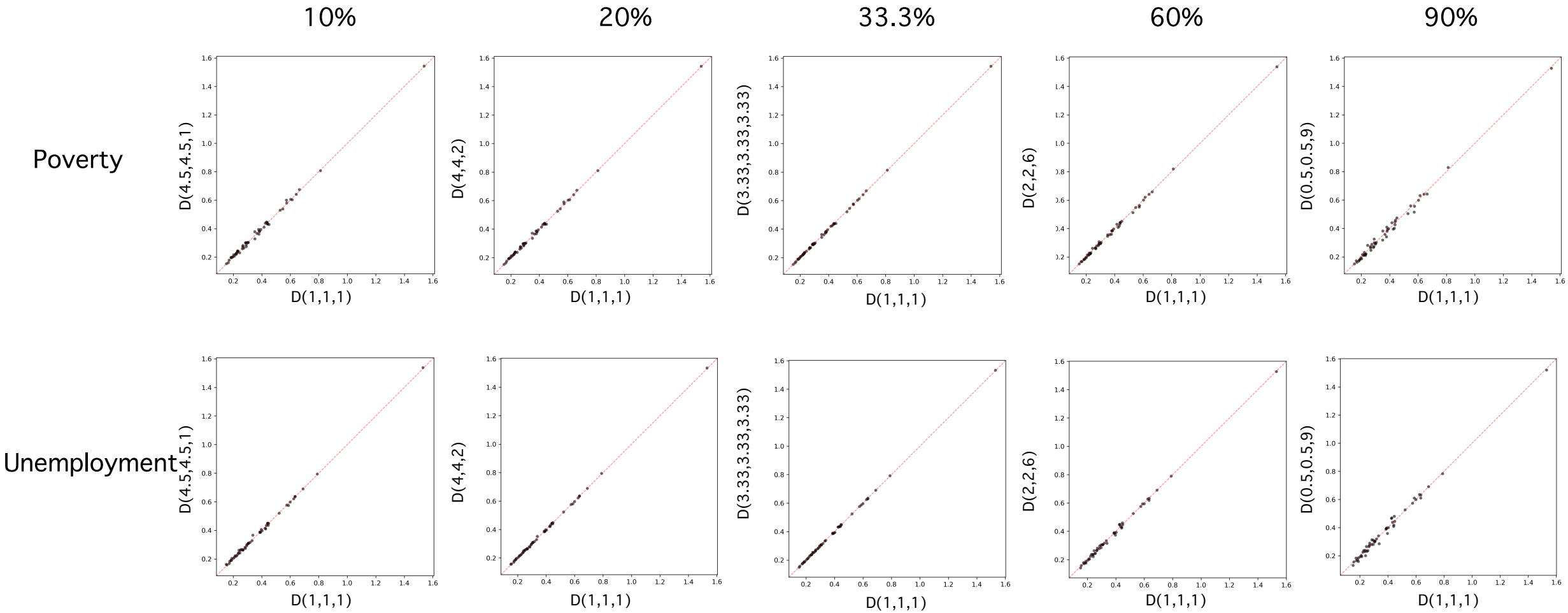


Sensitivity Analysis: Prior Selection

- Objective: Evaluate model stability and the influence of mixing parameters $\alpha_1, \alpha_2, \alpha_3$ prior selection.
- Models: Models including poverty covariate and models including unemployment covariate for West Virginia.
- Priors tested:
  - Non-informative: Dirichlet(1,1,1)
  - Informative (Topological weight): Topological component explains 10%, 20%, 33.3%, 60%, or 90% of the total random effect.
- Observations:
  - Overall risk and total standard deviation remained relatively unchanged across all specifications.
  - Structured component remained relatively constant.
  - Increasing topological weight resulted in a compensatory decrease in the unstructured component.
- Robustness: Posterior relative risk estimates showed minimal variation, indicating model robustness to prior choice.

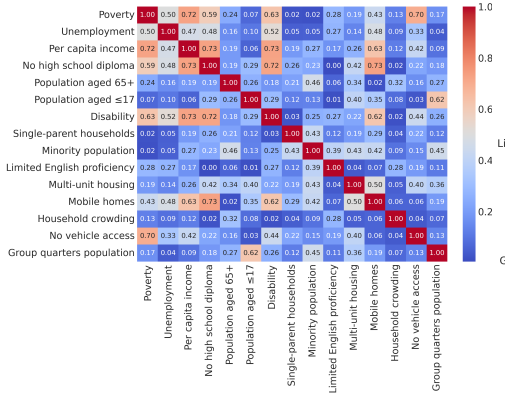
	Model Adjusted for Poverty				**Model Adjusted for Unemployment**			
Parameter	Mean	SD	3% quant.	97% quant.	Mean	SD	3% quant.	97% quant.
Prior for $(\alpha_1, \alpha_2, \alpha_3)$**:** $Dirichlet(1,1,1)$								
ξ	-1.051	0.200	-1.416	-0.668	-0.946	0.170	-1.268	-0.630
$1/\sqrt{\tau_b}$	0.522	0.071	0.395	0.657	0.539	0.075	0.403	0.681
α_1	0.201	0.144	0.00	0.457	0.276	0.165	0.000	0.651
α_2	0.516	0.153	0.241	0.804	0.521	0.166	0.206	0.822
α_3	0.284	0.101	0.094	0.471	0.203	0.091	0.034	0.365
Prior for $(\alpha_1, \alpha_2, \alpha_3)$**:** $Dirichlet(4.5,4.5,1)$								
ξ	-1.078	0.204	-1.468	-0.702	-0.957	0.169	-1.280	-0.648
$1/\sqrt{\tau_b}$	0.518	0.071	0.391	0.656	0.527	0.073	0.395	0.667
α_1	0.338	0.121	0.122	0.567	0.383	0.125	0.153	0.619
α_2	0.467	0.120	0.244	0.688	0.477	0.123	0.252	0.708
α_3	0.195	0.086	0.037	0.353	0.140	0.076	0.012	0.278
Prior for $(\alpha_1, \alpha_2, \alpha_3)$**:** $Dirichlet(4,4,2)$								
ξ	-1.060	0.204	-1.447	-0.682	-0.954	0.169	-1.266	-0.632
$1/\sqrt{\tau_b}$	0.517	0.070	0.390	0.649	0.529	0.072	0.395	0.661
α_1	0.307	0.118	0.098	0.530	0.351	0.121	0.138	0.583
α_2	0.449	0.118	0.227	0.665	0.462	0.120	0.240	0.687
α_3	0.245	0.088	0.077	0.403	0.187	0.079	0.048	0.335
Prior for $(\alpha_1, \alpha_2, \alpha_3)$**:** $Dirichlet(3.33,3.33,3.33)$								
ξ	-1.044	0.199	-1.440	-0.694	-0.943	0.166	-1.250	-0.630
$1/\sqrt{\tau_b}$	0.517	0.070	0.389	0.648	0.530	0.071	0.401	0.662
α_1	0.270	0.113	0.066	0.477	0.316	0.117	0.106	0.536
α_2	0.430	0.117	0.204	0.645	0.442	0.120	0.222	0.668
α_3	0.300	0.088	0.137	0.465	0.242	0.081	0.097	0.392
Prior for $(\alpha_1, \alpha_2, \alpha_3)$**:** $Dirichlet(2,2,6)$								
ξ	-1.018	0.199	-1.395	-0.648	-0.929	0.166	-1.239	-0.615
$1/\sqrt{\tau_b}$	0.513	0.067	0.397	0.644	0.536	0.071	0.410	0.672
α_1	0.202	0.106	0.023	0.396	0.255	0.118	0.042	0.466
α_2	0.403	0.114	0.190	0.615	0.413	0.123	0.178	0.637
α_3	0.395	0.086	0.237	0.555	0.332	0.081	0.184	0.484
Prior for $(\alpha_1, \alpha_2, \alpha_3)$**:** $Dirichlet(0.5,0.5,9)$								
ξ	-0.999	0.187	-1.352	-0.650	-0.915	0.158	-1.223	-0.626
$1/\sqrt{\tau_b}$	0.507	0.064	0.389	0.627	0.543	0.070	0.415	0.676
α_1	0.095	0.095	0.000	0.274	0.159	0.119	0.000	0.365
α_2	0.422	0.116	0.209	0.642	0.426	0.131	0.180	0.660
α_3	0.483	0.082	0.327	0.633	0.414	0.077	0.264	0.552

Sensitivity of the Relative Risk Estimates (Poverty and Unemployment)

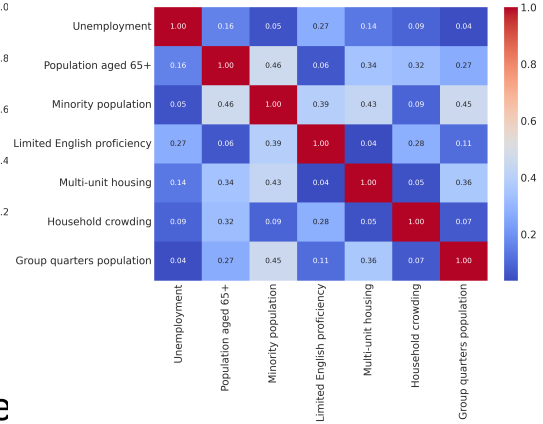


Multivariate Study – West Virginia

All variables



After variance inflation factor (VIF) screening

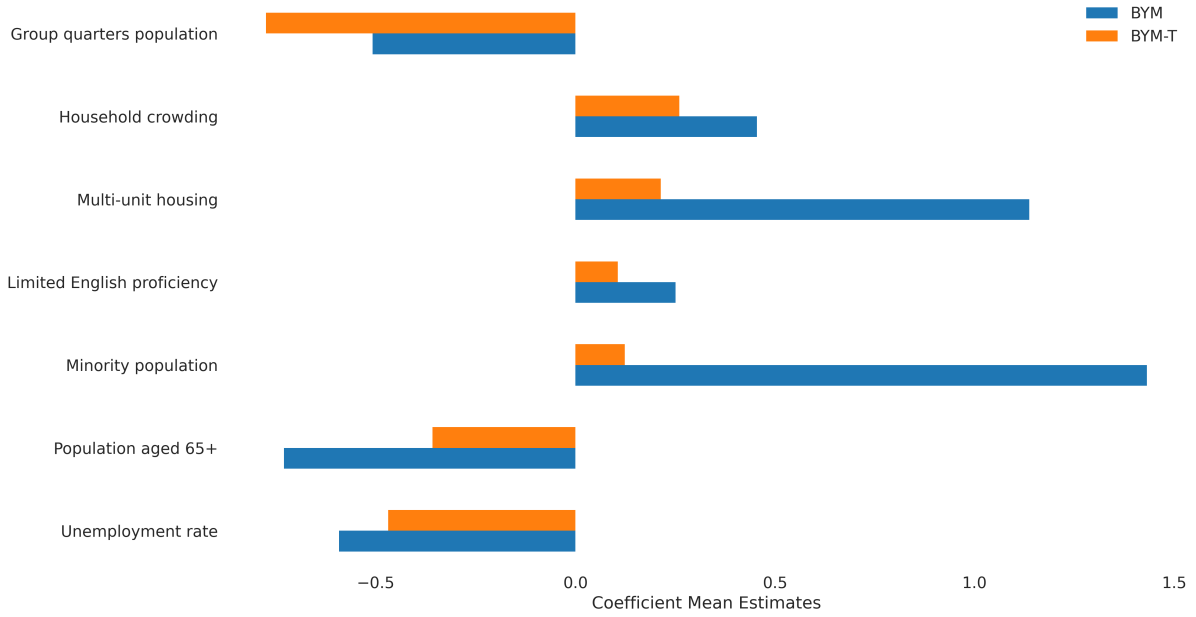


- As observed in the unstructured component weight decrease and, this decreased weight shifts towards topological component
  - Unstructured Weight decrease from 0.389 to 0.152.
  - Topological component accounts for 0.323.
- Inclusion of the topological summaries leads to a general shrinkage in the magnitude of fixed effects coefficients.

Posterior parameter estimates and model comparison for multiple-covariate studies in West Virginia, with predictive performance assessed using ELPD LOO between BYM-T and BYM2 relative risk maps.

Model	$\frac{1}{\sqrt{\tau}}$	ϕ	$1-\phi$	$\alpha_1, \alpha_2, \alpha_3$	LOO
BYM-T	0.482	-	-	0.152,0.525,0.323	-142.410
BYM2	0.467	0.611	0.389	-	-144.576

The fixed covariate coefficients between the BYM-T and BYM2 models.



OAK RIDGE National Laboratory

Conclusions and Next Directions

First Project:

- We successfully developed a framework that integrates topological insights into spatial disease modeling.
- Written a draft of the paper + working codebase.

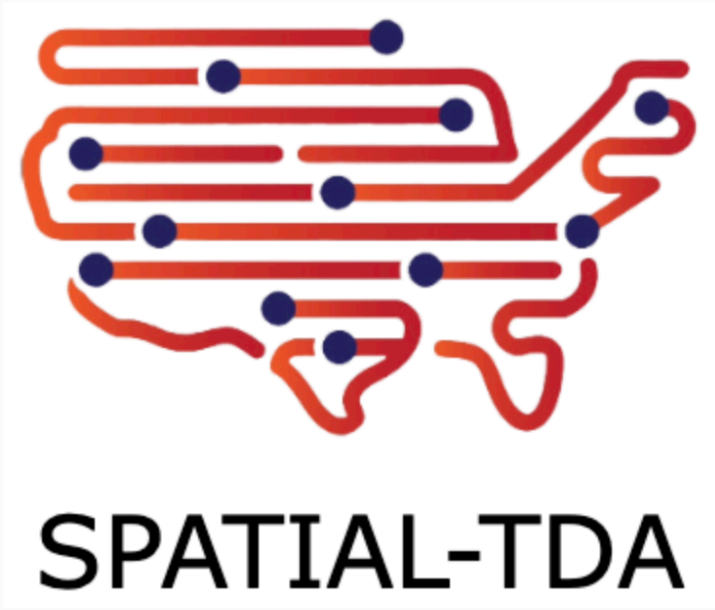
Second project: "The Use of Natural Language Processing to Investigate Opioid Epidemic: A Systematic Review".

  - Study yeas 2010-2025
  - First stage review 800 abstracts and selected 240 papers (last fall semester).

Next Directions

- Comprehensive exam needs to be done within 5 years:
  - In the fall semester including a proposal with preliminary results.
  - Final working thesis document as comprehensive document.

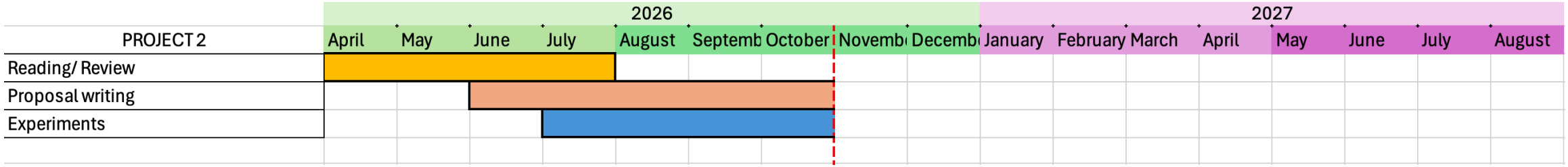
pypi v0.1.9 conda not found



A Python package for Topological Data Analysis(TDA) in spatial data.

- GitHub repo: https://github.com/hashansl/spatial-tda
- Documentation: https://hashansl.github.io/spatial-tda/
- PyPI: https://pypi.org/project/spatial-tda
- Free software: MIT license

Timeline


2026
2027
PROJECT 2
April
May
June
July
August
Septemb
October
Novembı
Decembı
January
February
March
April
May
June
July
August
Reading/ Review
Proposal writing
Experiments

Thank you!