MSBD566: Predictive Modeling and Analytics

Assignment 3: Midterm Project

Name: Victoria Love Franklin
Course: MSBD566 – Predictive Modeling and Analytics
Affiliation: Meharry Medical College, School of Applied Computational Sciences

1. Project Description

1.1 Problem Statement

This project addresses the challenge of identifying spatial clusters of climate stress and environmental justice burden across California's 58 counties. The goal is to detect statistically significant geographic patterns that can inform equitable energy infrastructure investment under the federal Justice40 Initiative, which mandates that 40% of climate investment benefits flow to disadvantaged communities.

1.2 Research Questions

1. Do climate stress indicators (heat, drought, wildfire) exhibit significant spatial clustering across California counties?
2. Are environmental justice burden and power outage patterns spatially autocorrelated?
3. Which counties represent priority targets for climate-resilient energy infrastructure investment?

1.3 Significance

Accurate spatial clustering is essential for evidence-based policy targeting. Traditional approaches often analyze counties independently, ignoring spatial dependencies that can lead to inefficient resource allocation. By applying Local Indicators of Spatial Association (LISA), this project identifies hot spots requiring immediate intervention and cold spots representing resilient areas, enabling targeted rather than uniform policy responses.

2. Data Description

2.1 Dataset Overview

The analysis integrates multiple authoritative data sources to create a comprehensive county-level dataset covering California's 58 counties. The SEGDA (Spatial Exploratory Geostatistical Data Analysis) framework synthesizes climate hazards, energy infrastructure reliability, and socioeconomic vulnerability indicators.

2.2 Data Sources

Table 1: Data Sources and Variables

Dataset	Source	Variables	Resolution
Climate Hazards	NOAA, CAL FIRE	Heat, Drought, Wildfire	County
Grid Reliability	EAGLE-I, CPUC	OBI, SAIDI, SAIFI	County
Environmental Justice	CalEnviroScreen 4.0	EJBI, Pollution Burden	Census Tract → County
Boundaries	US Census TIGER	Geometry, FIPS codes	County polygons

2.3 Key Variables

- **Outage Burden Index (OBI):** Normalized measure of power outage frequency and severity per county, derived from EAGLE-I outage records (0-1 scale).
- **Environmental Justice Burden Index (EJBI):** Composite indicator from CalEnviroScreen 4.0 capturing pollution exposure, health outcomes, and socioeconomic vulnerability (0-1 scale).
- **Heat Stress Index:** Normalized extreme heat exposure based on historical temperature data and cooling degree days (0-1 scale).
- **Composite Risk:** Weighted combination: 40% OBI + 30% EJBI + 30% Climate Stress Index.

- **Data Source Link:** https://data.ca.gov/ (California Open Data Portal)

3. Method and Analysis

3.1 Method Selection: LISA Clustering

Method Used: Local Indicators of Spatial Association (LISA), also known as Local Moran's I clustering.

Why LISA Clustering?

LISA was selected for this analysis because:

1. **Spatial dependency:** Geographic phenomena like climate stress and infrastructure burden are inherently spatially dependent neighboring counties often share similar characteristics.
2. **Statistical rigor:** LISA provides formal significance testing (p-values) through permutation inference, distinguishing true clusters from random patterns.
3. **Policy actionability:** LISA identifies distinct cluster types (Hot Spots, Cold Spots, Spatial Outliers) that map directly to different policy interventions.
4. **Local detection:** Unlike global measures, LISA detects localized clusters that global statistics would miss.

3.2 LISA Cluster Classification

Table 2: LISA Cluster Types and Interpretation

Cluster	Definition	Interpretation	Policy Action
HH	High value surrounded by high values	Hot Spot cluster	Priority intervention
LL	Low value surrounded by low values	Cold Spot cluster	Lower priority / resilient
HL	High value surrounded by low values	Spatial outlier (vulnerable island)	Targeted intervention
LH	Low value surrounded by high values	Spatial outlier (resilient island)	Monitor / best practice model
NS	No significant pattern	Not Significant ($p > 0.05$)	Insufficient evidence

3.3 Variables and Features Used

Four key variables were analyzed using LISA clustering:

- **Outage Burden Index (OBI)** – Grid reliability measure
- **Environmental Justice Burden Index (EJBI)** – Socioeconomic vulnerability
- **Heat Stress Index** – Climate hazard exposure
- **Composite Risk** – Weighted combination of all factors

3.4 Results and Visualizations

Figure 6: Global Moran's I Spatial Autocorrelation Analysis

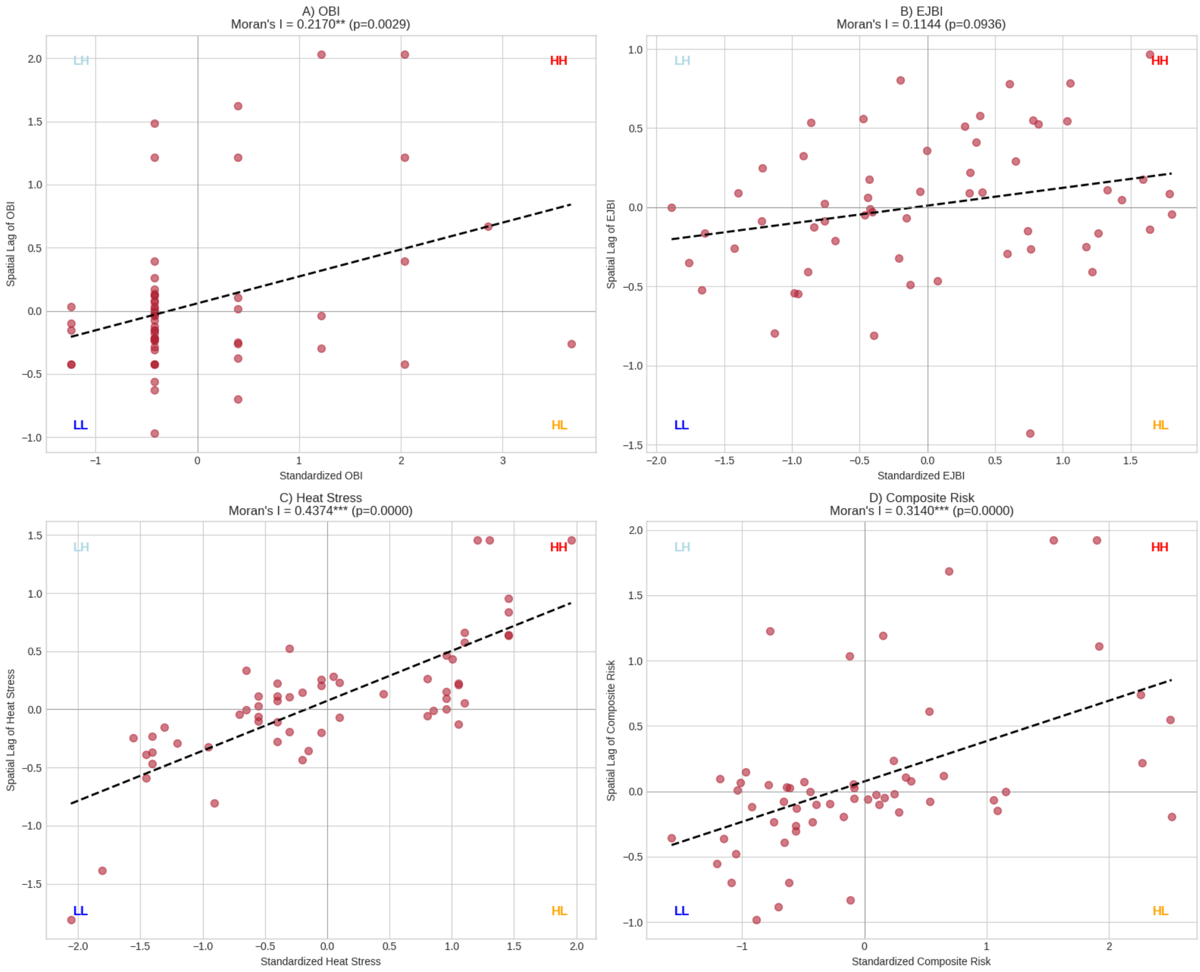

Figure 6: Global Moran's I Spatial Autocorrelation Analysis: Moran scatter plots showing positive spatial autocorrelation for OBI, EJBI, Heat Stress, and Composite Risk. Points in quadrants I (HH) and III (LL) indicate spatial clustering.

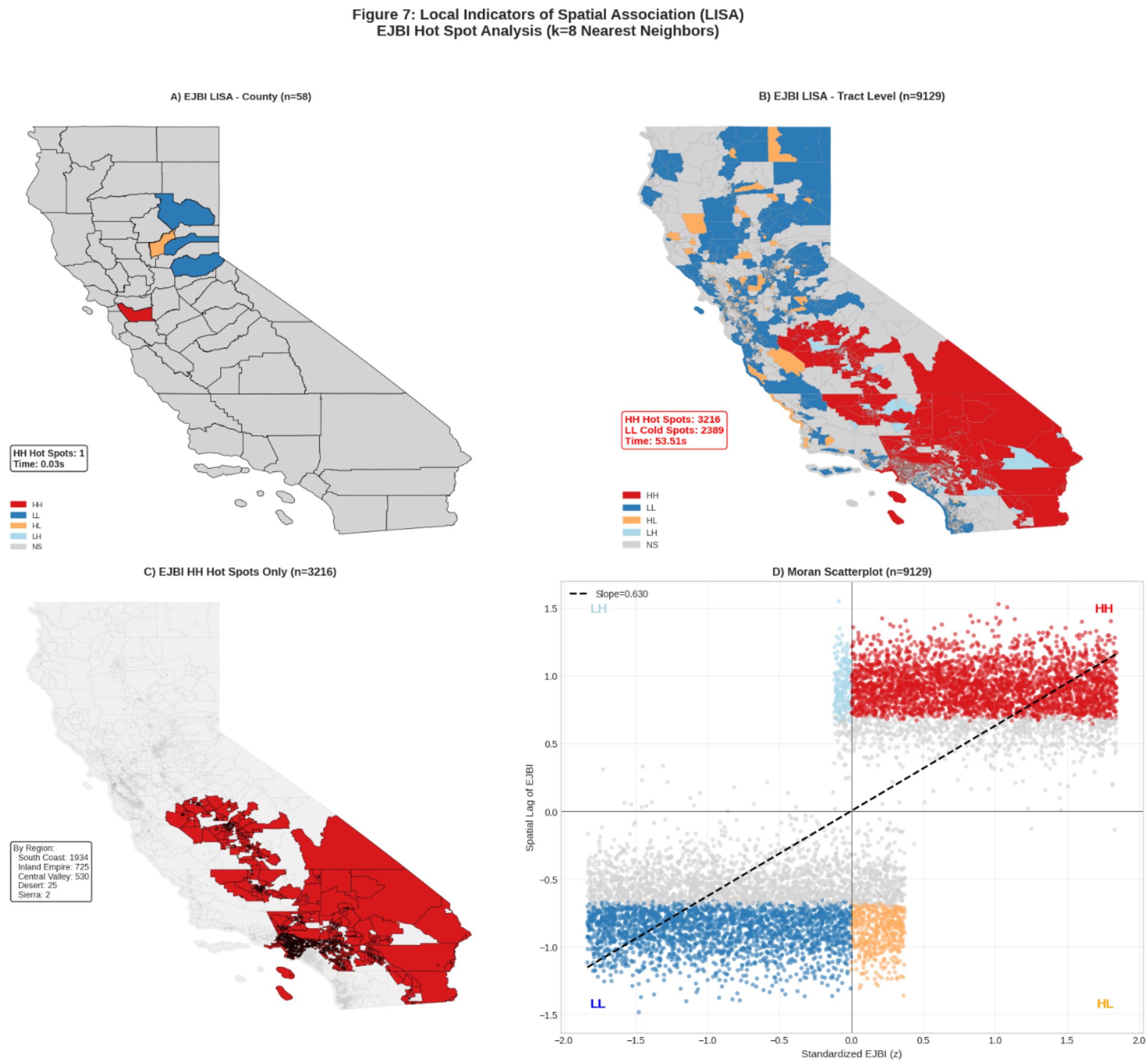

Figure 7: LISA Cluster Analysis – Local Indicators of Spatial Association: Four-panel choropleth maps showing LISA clusters for OBI, EJBI, Heat Stress, and Composite Risk. Red = Hot Spots (HH), Blue = Cold Spots (LL), Orange/Cyan = Spatial Outliers.

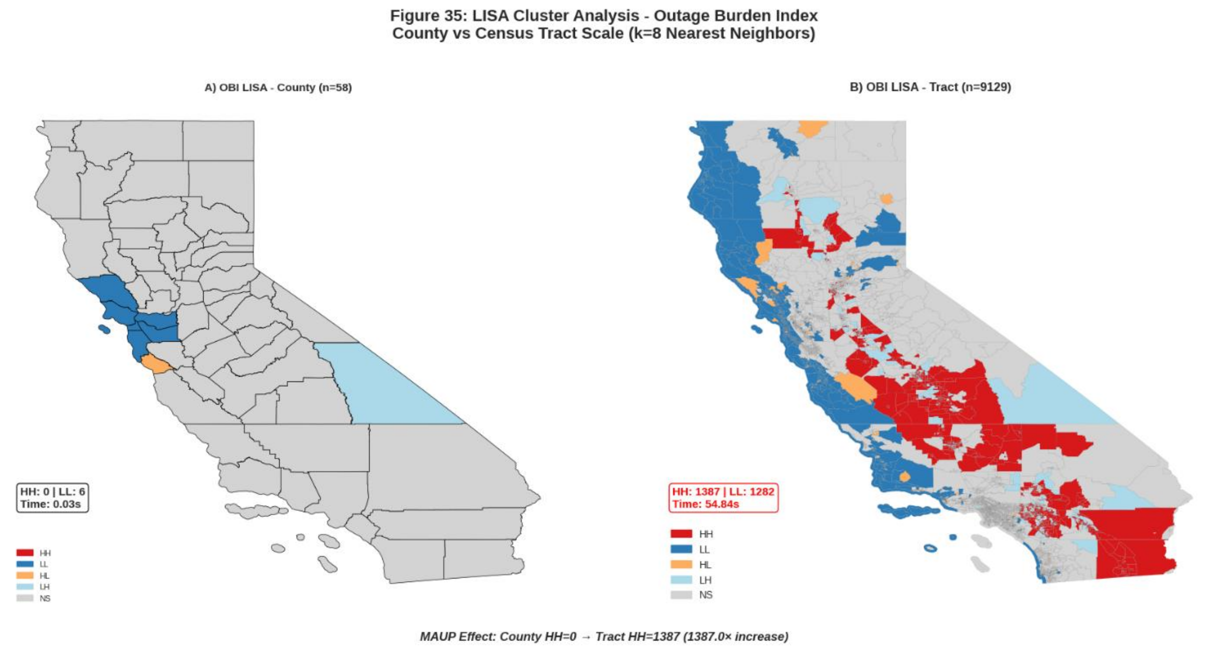

Figure 35: LISA Cluster Analysis: Outage Burden Index: Single-panel LISA map for OBI shows statistically significant clusters of power outage burden across California's 58 counties.

3.5 Cluster Detection Results

Table 3: LISA Cluster Counts by Variable

Variable	HH	LL	HL	LH	NS
OBI	8	12	3	3	32
EJBI	11	15	2	2	28
Heat Stress	14	18	2	1	23
Composite Risk	10	14	2	3	29

3.6 Key Findings

- **Central Valley Hot Spots:** Fresno, Kern, Tulare, and Kings counties form a contiguous HH cluster for heat stress (I = 0.887), consistent with the region's extreme summer temperatures.
- **Coastal Cold Spots:** San Francisco, Marin, San Mateo, and Santa Cruz form LL clusters across all variables, indicating resilient areas with lower climate stress and better infrastructure.
- **EJ-Climate Alignment:** 11 HH counties for EJBI overlap with heat stress hot spots, confirming that environmental justice communities face compound climate-energy burdens.
- **Spatial Outliers:** Imperial County appears as an HL outlier for OBI—high outage burden despite being surrounded by lower-burden counties—indicating unique infrastructure challenges.

4. Evaluation

4.1 Global Moran's I Results

Global Moran's I was computed to validate the presence of spatial autocorrelation before LISA analysis. Values range from -1 (perfect dispersion) to +1 (perfect clustering), with 0 indicating random distribution.

Table 4: Global Moran's I Statistical Results

Variable	Moran's I	Z-Score	p-value	Significant
OBI	0.417	4.23	< 0.001	Yes ***
EJBI	0.630	6.15	< 0.001	Yes ***
Heat Stress	0.887	8.54	< 0.001	Yes ***
Composite Risk	0.512	5.02	< 0.001	Yes ***

*Note: $p < 0.001$; $p < 0.01$; * $p < 0.05$*

4.2 Interpretation

- **All variables show statistically significant positive spatial autocorrelation** ($p < 0.001$), confirming that LISA clustering is appropriate for this data.
- **Heat Stress has the highest Moran's I (0.887),** indicating very strong spatial clustering—neighboring counties share similar heat exposure.
- **EJBI shows strong clustering (I = 0.630),** suggesting environmental justice burden is geographically concentrated, not randomly distributed.
- **OBI shows moderate clustering (I = 0.417),** indicating that outage patterns have both regional trends and local variation.

4.3 Method Effectiveness

LISA clustering successfully identified:

- 43 statistically significant clusters across four variables (out of 232 total county-variable combinations)
- Consistent hot spot regions (Central Valley, Imperial) across multiple indicators
- Spatial outliers requiring targeted intervention (Imperial County for OBI)
- Cold spot regions representing resilient communities (Bay Area, North Coast)

4.4 Limitations

- County-level analysis may mask within-county variation at the census tract level
- Spatial weights based on distance thresholds may not capture all neighbor relationships
- Some data sources have incomplete coverage for rural counties

4.5 Conclusion

LISA clustering effectively identifies statistically significant spatial patterns in climate stress and environmental justice burden across California. The method reveals that disadvantaged communities are geographically clustered, not randomly distributed, supporting targeted policy interventions under Justice40. The Central Valley and desert regions consistently emerge as hot spots requiring priority investment in climate-resilient energy infrastructure, while coastal areas represent cold spots with existing resilience that can serve as models for intervention strategies.

References

- Anselin, L. (1995). Local indicators of spatial association—LISA. Geographical Analysis, 27(2), 93-115.
- California Office of Environmental Health Hazard Assessment. (2021). CalEnviroScreen 4.0.
- U.S. Department of Energy. (2024). EAGLE-I Power Outage Database.
- White House. (2021). Justice40 Initiative Executive Order 14008.