ANOMALY DETECTION ON SPATIO-TEMPORAL POLLUTION DATA

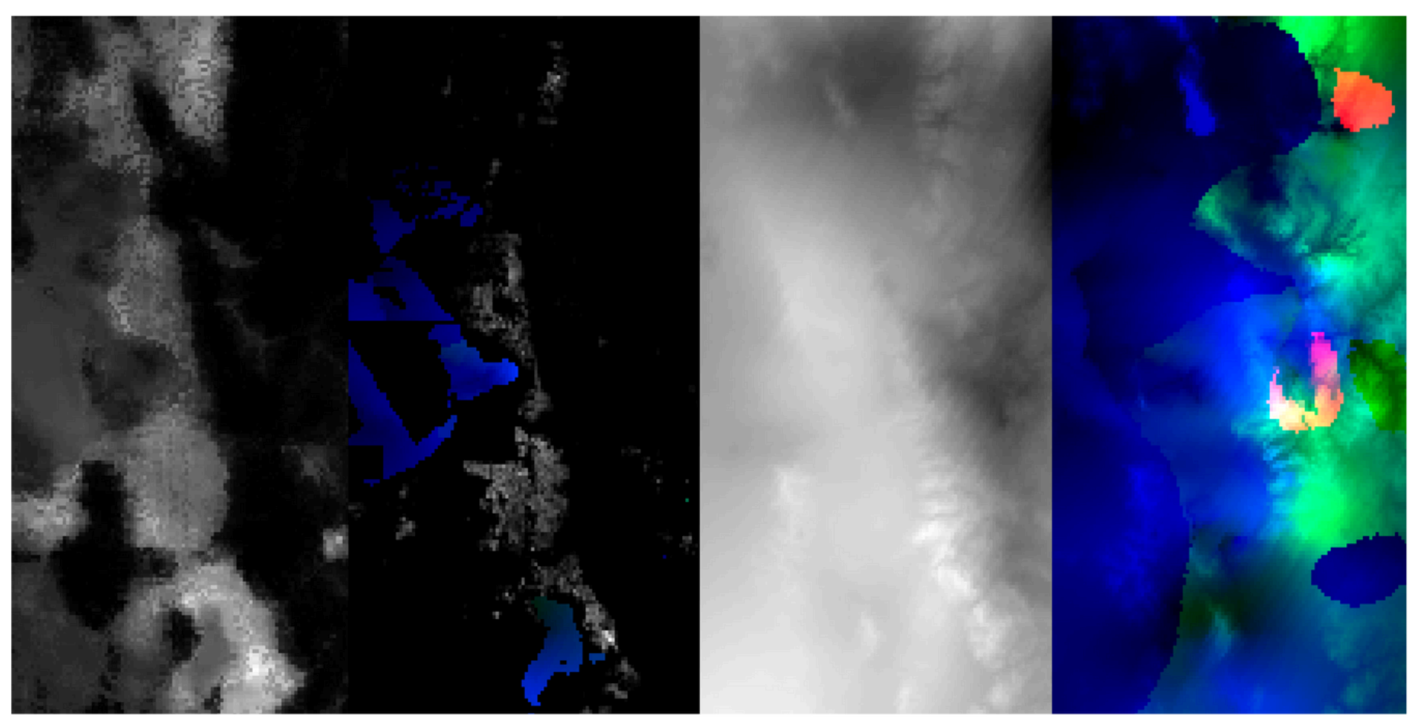
SIMON GONZALEZ MATTHEW GUSTAFSON HYRUM BAILEY
GROUP 7: DATA MUNCHERS

MOTIVATION

- Air pollution is a serious public health concern in Utah, particularly during winter inversions and summer wildfires.
- **Focus**: We analyze $PM_{2.5}$, a fine particulate pollutant known to impact respiratory and cardiovascular health.
- **Goal**: Detect spatio-temporal anomalies in PM2.5 levels to identify patterns related to location, time of year, or other contributing factors.

DATA

- **Daily PM2.5 concentrations** measured in a 1km grid over Utah.
- **Covariate data**: precipitation, temperature, windspeed and population density.
- All datasets were regridded, interpolated, and aligned to a common grid.



SCAN STATISTICS

- Scan statistics provide a rigorous statistical framework for detecting spatial and spatio-temporal anomalies.
- **Normal model** for continuous data:

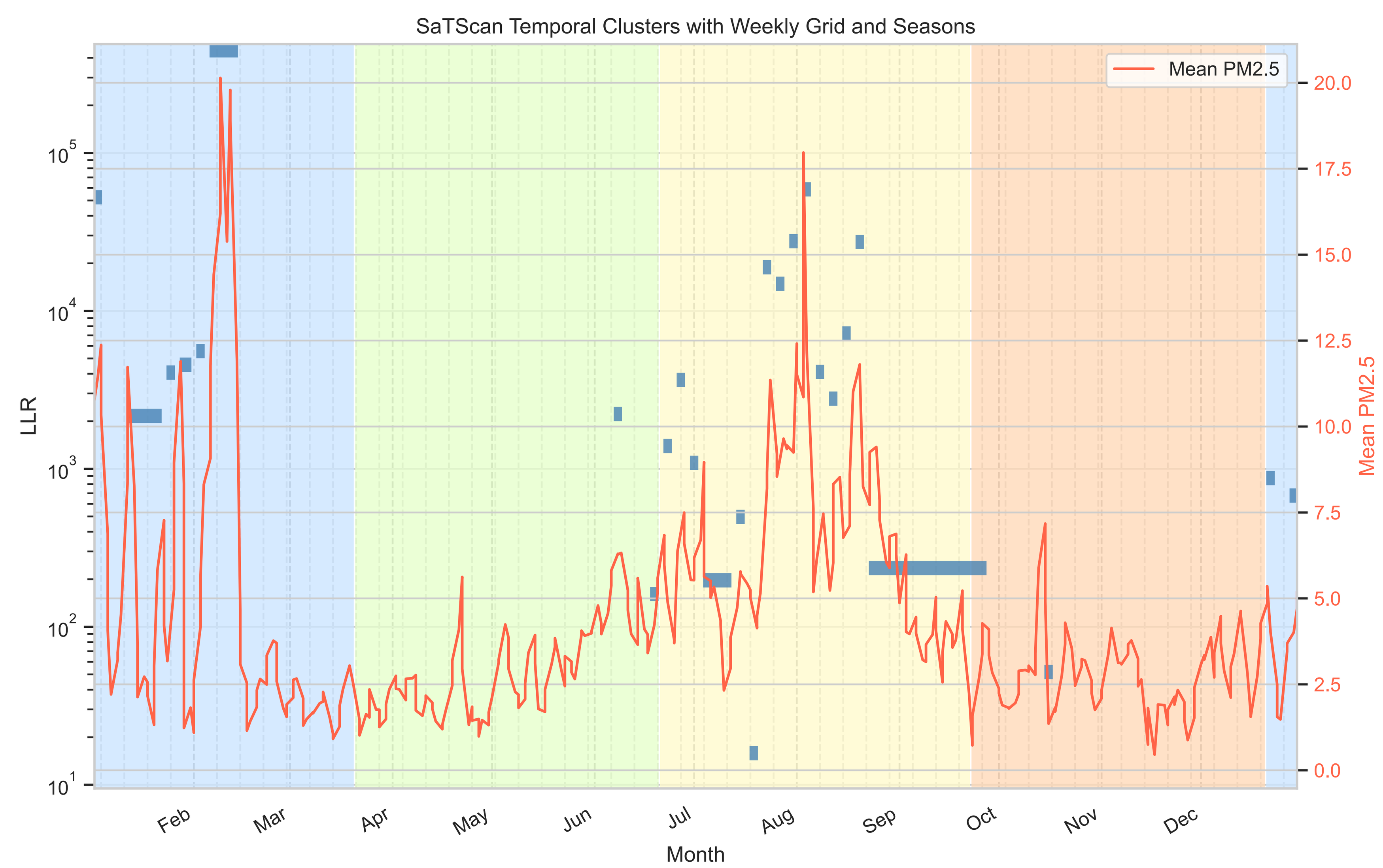
$$\max_z \frac{\ln L_z}{\ln L_0} = \max_z \left(N\ln(\sigma) + \sum_i \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{N}{2} - N\ln\left(\sqrt{\sigma_z^2}\right)\right)$$

- To isolate unexplained variation, anomaly detection can be performed on the residuals of a **regression model**:

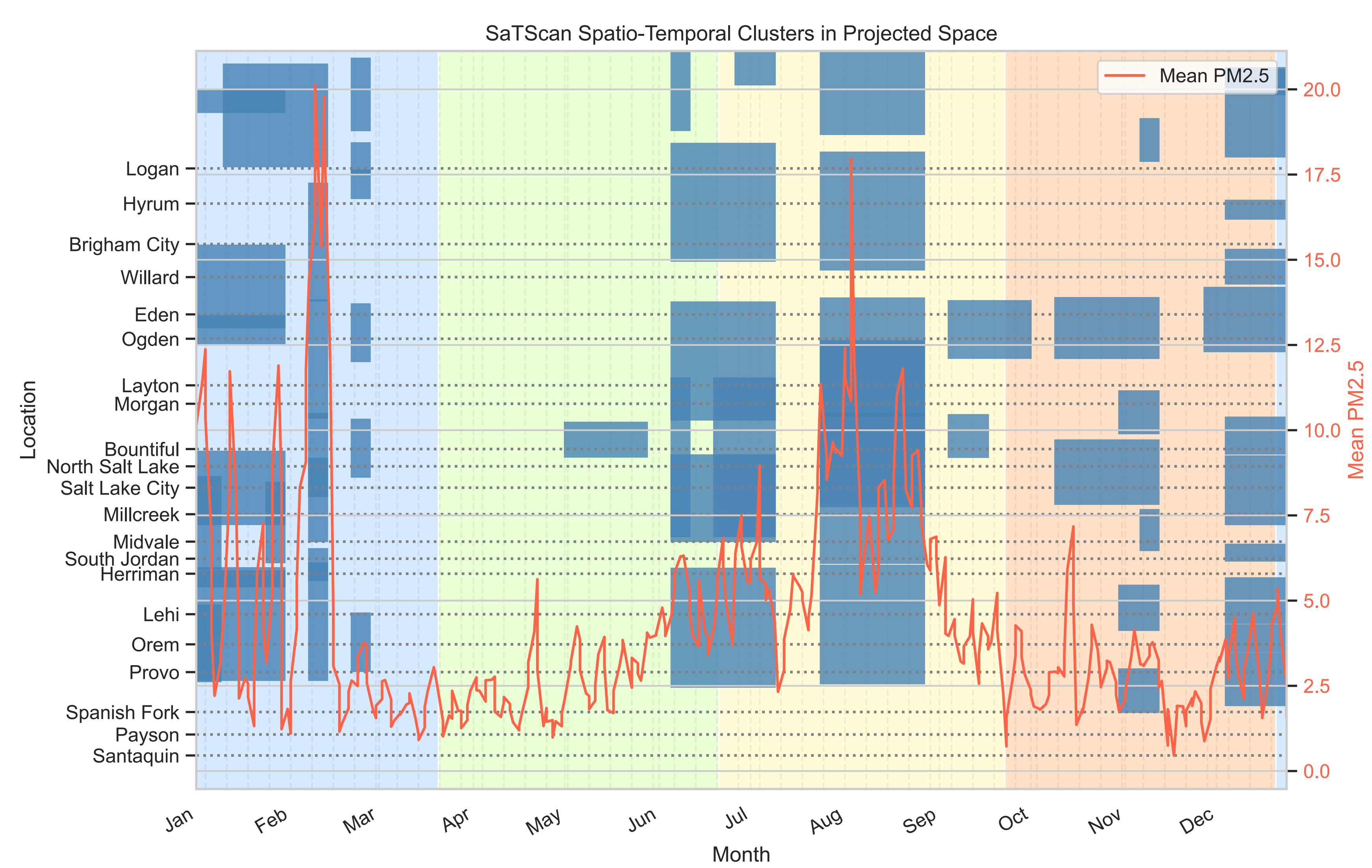
$$\log(PM_{2.5}) = \beta_0 + \beta_1 Wind + \beta_2 Prec + \beta_3 Temp + \beta_4 \log(Pop) + \epsilon$$

ANALYSIS

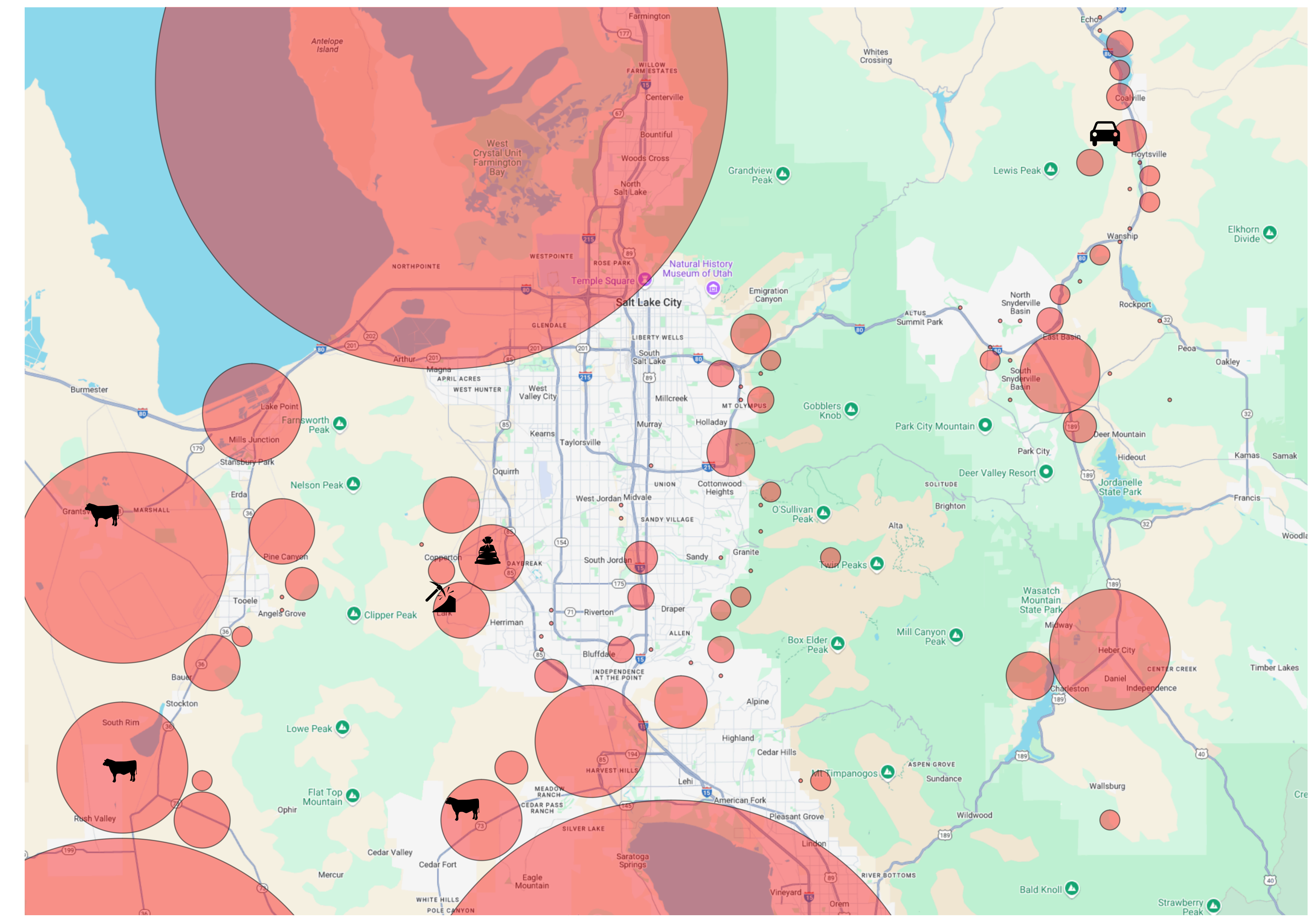
- **SaTScan** and **pyscan** were tested as anomaly detection tools, SaTScan was used for supporting the normal model.
- Multiple configurations were tested, varying spatial and temporal scopes, temporal resolution, and cluster size limits.



Results for **temporal-only** normal model with daily resolution analysis.



Results for **spatio-temporal** normal model with weekly resolution.



Results for spatial-only **covariate adjusted** normal model.

RESULTS

- Temporal spikes in February and August align with known pollution events:
  - February: Severe **inversion** and stagnant air.
  - August: Wildfire smoke elevated $PM_{2.5}$ levels.
- Spatiotemporal clusters reveal:
  - Anomalies extend beyond major cities (e.g., Provo, Orem).
- Spatial-only covariate adjusted clusters suggest:
  - High **traffic** emissions along highways.
  - **Farming** activity.
  - Bingham Canyon **Mine** activity
  - Emissions originating from the Trans-Jordan **Landfill**.

CONCLUSIONS AND FUTURE RESEARCH

- Spatial scan statistics are relatively effective in identifying spatiotemporal anomalies in $PM_{2.5}$ concentrations.
- Some clusters aligned with winter inversion and summer wildfires.
- Covariate adjustment helped explore anomalies from more localized sources.
- Future work could apply these methods to larger regions or longer time periods.
- Scalability needs to be improved for the normal model.