Anomaly Detection on Spatio-Temporal Pollution Data

Simon Gonzalez Matthew Gustafson Hyrum Bailey

Key Idea

Applying spatial and spatio-temporal scan statistics, typically used in fields such as epidemiology, for anomaly detection in spatio-temporal air pollution data, a domain where this method remains largely unexplored. Scan statistics, in its most basic form, tests the presence of clusters in a one-dimensional point process against the null hypothesis H0 of randomness [1]. In a point process defined over the interval [a,b], a sliding window of fixed size [t, t+w] moves along the interval. The number of points within the window is compared to its expected distribution under H0 using a discrepancy function as a test statistic. More generally, the Kulldorff scan statistic is defined as:

$$d(m_R, b_R) = m_R \log \frac{m_R}{b_R} + (1 - m_R) \log \frac{1-m_R}{1-b_R}$$

where m_R and b_R represent the proportion of the total value that falls inside the candidate region R, under the measurement and baseline distributions, respectively. This model extends to any collection of regions. Spatio-temporal scan statistics consider all possible 3D zones where the z-dimension corresponds to time. We investigated and used SaTScan [2] (https://www.satscan.org/) and pyscan [3] for anomaly detection.

Problem & Motivation

Air pollution is a significant public health issue, especially in Utah, where air quality worsens in the winter due to inversion. The primary measurable component of this air pollution is the levels of $PM_{2.5}$, which is fine particulate matter that is harmful to health. Understanding which factors drive $PM_{2.5}$ levels can help policymakers design better interventions to reduce pollution and protect health. Our goal is to identify anomalous behavior in the particulate matter concentrations to uncover potential factors, such as location or time of year, that may be most relevant to air quality.

Data Collection

Dr. Brenna Kelly from the University of Utah Population Health Sciences department in the School of Medicine provided us with the data originally derived from Joel Schwartz's ensemble model [4]. The dataset used consists of $PM_{2.5}$ particulates measurements, their location and time, in Utah over 2016. The final shape of the matrix (latitude, longitude, day) is $501 \times 530 \times 366$, and it contains about 82 million elements. We rasterized the data files into a regular grid ($1km \times 1km$) and transformed them into a single netCDF file to reduce size (from 752.4 MB to 299.4 MB) and improve reading speed.

Experiments with pyscan

For grid-based data, a practical approach is to define the range space as all possible subgrids. Then, using two horizontal and two vertical lines to sweep the space, each possible subgrid is defined, and its discrepancy is computed. The complexity of this algorithm is $O(g^4)$, where the grid has g×g cells. An approximate solution can be found using a linearization heuristic [5] to improve efficiency, reducing the complexity to $O(tg^3)$, where t depends on the precision of the approximation. This formulation is interesting since our data is structured as a grid. We extended the pyscan implementation with a faster grid construction method and an area-limited version of the Approx-Grid method. Figure 1 shows the results for one run with the original method (6.4 seconds) and the area-limited method (2.4 seconds) for a dataset containing ~250,000 data points. The area-limited version successfully identifies a more meaningful cluster. Finally, inspired by the elbow method, we experimented with several area limits to identify reasonable values for our dataset. The results are shown in Figure 2. For example, 450 and 700 seem suitable, as the maximum discrepancy subgrid appears more stable at those values.

While pyscan demonstrates promising efficiency, the linearization heuristic only works for convex discrepancy functions that can be written in terms of b_R and m_R. Our data, consisting of continuous values, better aligns with the Normal model [6], where the maximum log-likelihood ratio (LLR) is computed as:

$$\max_R \left[N \ln(\sigma) + \sum_i \frac{(x_i - \mu)^2}{2\sigma^2} - \frac{N}{2} - N \ln \sqrt{\sigma^2_R} \right]$$

Where x_i are the observed values, μ and σ^2 are the overall mean and variance, σ_R^2 is the variance within region R, and N is the total number of observations. While Kulldorff's scan statistic can be applied by weighting point counts with a continuous variable, we felt it did not suit our data well, and the interpretation of the results would be questionable. For this reason, we opted to use the Normal model in SaTScan for the remaining analysis.

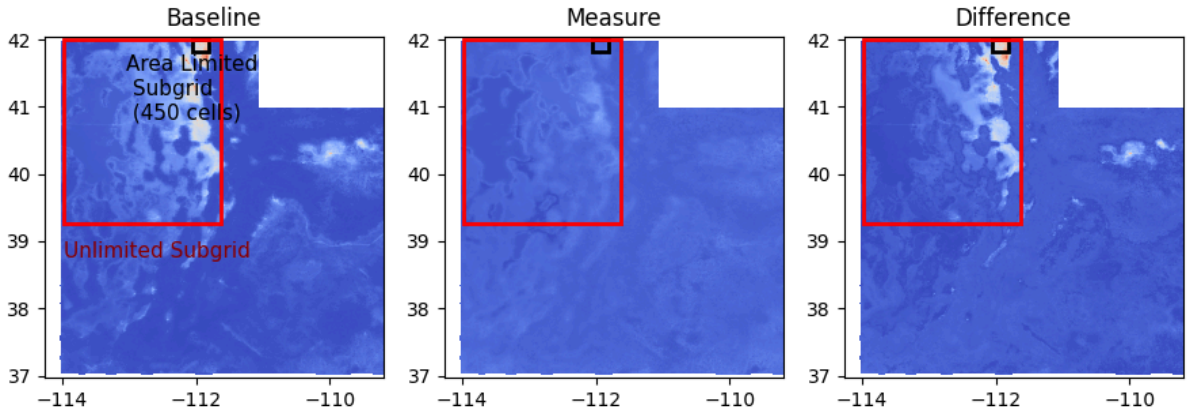

Figure 1: Maximum discrepancy subgrids obtained with the Approx-Grid algorithm and its area-limited version over the state of Utah $PM_{2.5}$ values.

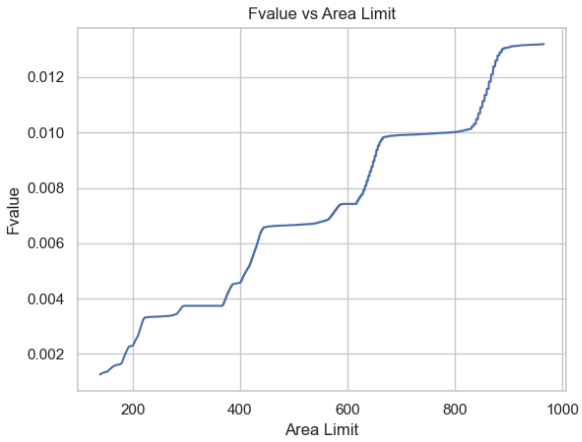

Figure 2: Discrepancy for the maximum discrepancy subgrid with different area limits given in number of cells.

Experiments with SatScan

even though it meant that scalability would be a problem.

Differing Timescales & Case Count Thresholds

Due to the lack of temporal variants to pyscan algorithms, we settled on using the SaTScan library. We defined our area of interest to include Salt Lake City and smaller suburban population centers such as Bountiful, Herriman, Midvale, etc. (Latitudes = (39.954, 42.000), Longitudes = (-112.475, -111.357)), and experimented with different configurations for SaTScan (e.g., setting maximum cluster size from 5% – 50%, time aggregates from 1-day to 1-month, and 5 km – 50 km circle diameter limits). We settled on the following configurations:

- Configuration 1: Temporal-only Normal model, 1-day (no time aggregate), 10% of the data as the maximum cluster size.
- Configuration 2: Space-Time Normal model, 1-week time aggregate, 10% of the data as the maximum cluster size.
- Configuration 3: Space-Time Normal model, 1-month time aggregate, 10% of the data as the maximum cluster size.

These configurations allowed us to obtain small enough non-trivial clusters in a relatively manageable runtime (up to 17 hours was the worst-case scenario).

Adjusting for Covariates

We considered two potential covariates influencing pollution levels: immediate weather-related factors and population density. Accounting for these factors could isolate interesting, unexplained anomalous concentrations. We obtained daily gridded data for 2016 from the following sources:

- Precipitation and mean temperature from Daymet [7], at 1 km resolution.
- Windspeed, calculated as daily means from hourly values in a 9 km resolution grid, sourced from ERA5 [8].
- Population density from the WorldPop [9] 2016 census dataset, at 1 km resolution.

The covariate layers were interpolated to match the $PM_{2.5}$ grid. To remove the influence of these covariates, we initially applied a standard linear regression, as recommended in the SaTScan manual; however, the model showed a poor fit. As an alternative, we implemented a daily linear regression model on the log $PM_{2.5}$ values using OLS, as done by Wrightson et al. [10]:

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

where β are the parameters of the linear model, and ϵ is the residual. Theoretically representing pollution not explained by weather or population, these residuals were then used as input for SaTScan's Normal model. The fitted model showed:

- A relatively low coefficient of determination $R^2 = 26.6\%$, meaning that the covariates explain only a limited portion of the variance.
- A moderate Mean Absolute Error $MAE = 0.276$.
- A residual mean of 0, meaning that the model is not biased.

Results

Non-Covariate-Adjusted Results

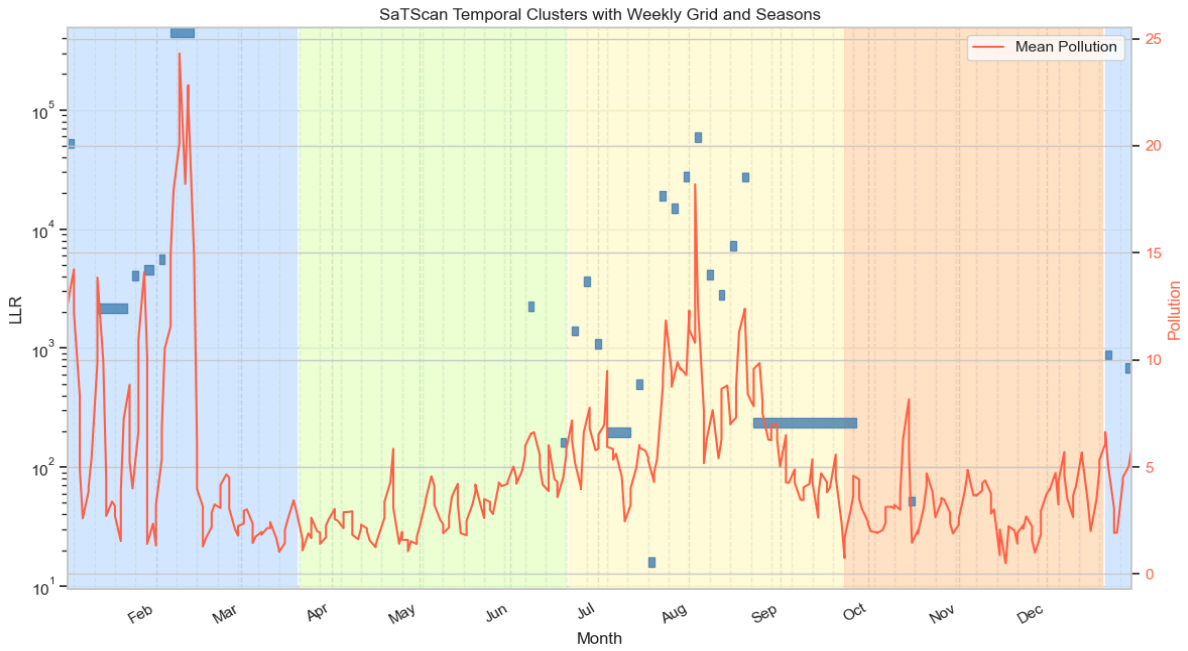

Figure 6: Mean $PM_{2.5}$ concentrations in the Salt Lake Valley with clusters shown in blue,

Model Configuration 1, 2016.

In Figure 6, a visualization of Model Configuration 1 (our temporal-only Normal model), we can see the clusters ordered by LLR; the higher the LLR, the more interesting the anomaly. The clusters' starting dates roughly correspond to peaks in mean pollution values across all regions, and the end date roughly corresponds to the effect of the increase in mean no longer being a factor.

As we can see, there are two major spikes of LLR at two points of the year, one in February and one in August. In February 2016, it was reported that poor inversion and a lack of a significant storm to push the bad air out led to significantly higher levels of $PM_{2.5}$ than usual, a spike unseen since 2013 [11]. Meanwhile, in August 2016, summer wildfires were the likely cause of the spike in $PM_{2.5}$ concentrations, as smoke-filled skies were reported and led to concerns regarding health [12].

We also created a 2D space-time visualization of Model Configuration 2 (spatiotemporal Normal, 1-week analysis). We utilized PCA to compress latitude and longitude to a single dimension and added PCA-projected city locations to indicate roughly where each cluster is spatially located. Cluster height corresponds to cluster diameter. Figure 6 represents this projected space-time visualization:

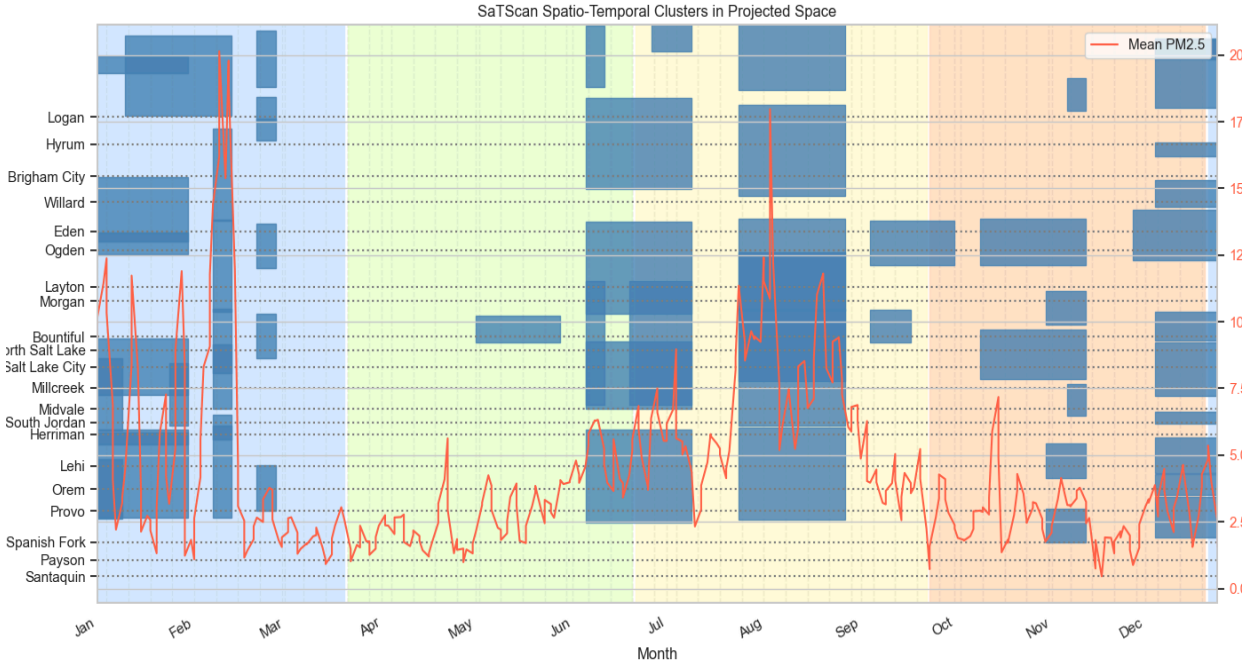

Figure 7: Spatiotemporal clusters of $PM_{2.5}$ concentrations in a PCA-projected space, Model Configuration 2, 2016.

Like Figure 6, the blue areas represent the clusters of particulate concentrations; in Figure 7, these correspond to different major cities across the Salt Lake Valley. An interesting pattern to note is that clusters emerge for nearly all cities during the previously mentioned spikes of February and August, indicating that in those periods of anomalous $PM_{2.5}$ concentrations, more than just the largest population centers (e.g., Salt Lake City, Provo, Orem, etc.) experienced poorer air quality.

Figure 8 shows a similar 2D space-time visualization, but this time for Model Configuration 3 (spatiotemporal Normal, 1-month analysis):

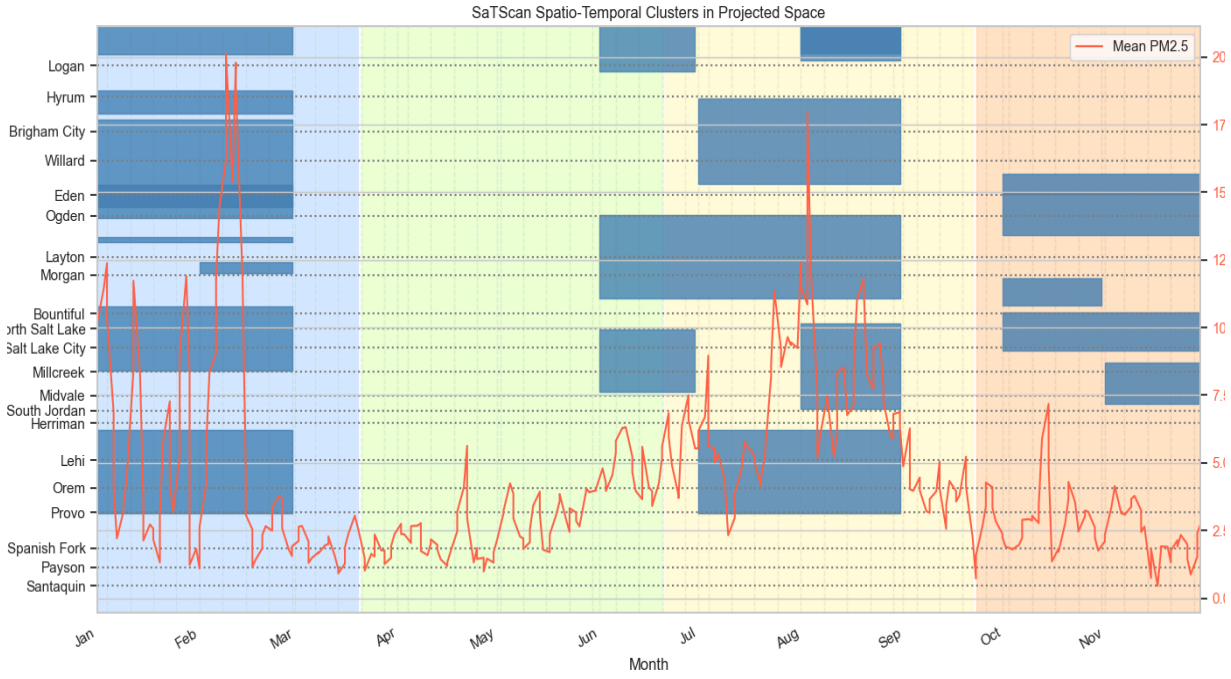

Figure 8: Spatiotemporal clusters of $PM_{2.5}$ concentrations in a PCA-projected space, Model Configuration 3, 2016.

Covariate-Adjusted Results

After adjusting for covariates, we ran a temporal-only analysis and found no anomalies. One possible interpretation is that removing weather covariates, even if they capture only immediate effects and not complex temporal weather patterns, eliminates much of the temporal variability. As a result, we ran a spatial-only analysis, obtaining the results shown in Figure 9. Upon further investigation, we discovered that several small clusters were identified:

- Along transited routes such as I-15 and I-84, which can be explained by transit.
- In smaller and less populated towns and farmland. This could be explained by activities such as manure management and farm machinery, as well as the adjustment for population density.
- Around Bingham Canyon Mine and the Trans-Jordan Landfill, which could be explained by the presence of heavy machinery and waste producing small particulate matter.

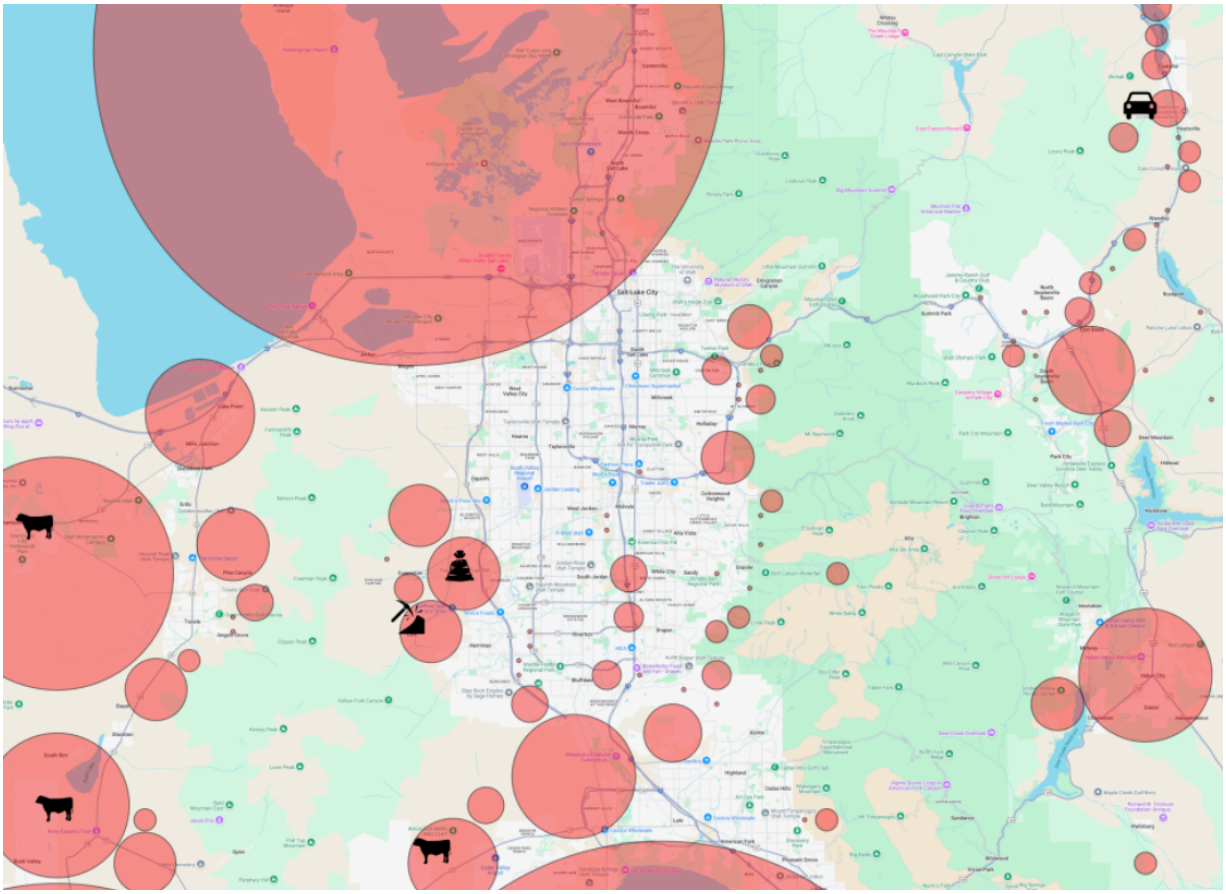

Figure 9: Clusters over the residual distribution. Agricultural areas, interstate highways, the Bingham Canyon Mine and Trans-Jordan Landfill are highlighted.

Conclusions & Future Research

Overall, we found that using spatial scan statistics was a relatively effective tool in identifying anomalies in our data. We were able to draw connections between the clusters of high $PM_{2.5}$ concentrations throughout the year and environmental events that occurred during those times that suggest probable causes of these clusters, such as particularly poor inversion due to few winter storms or wildfires causing poorer air quality. Also, by adjusting for covariates, we were able, at least to some extent, to shift the anomaly detection toward identifying clusters explained by location.

Given more time and a tool with greater scalability, it would be beneficial to perform spatial scan statistics on data across the entire geography of Utah in future research. Additionally, our experimental data only encompassed a year's worth of values; more conclusions about anomalous patterns could be made on a longer time scale in future research. It would also be interesting to explore the development of approximate methods for the Normal model and to extend the pyscan approximate methods to spatio-temporal scan statistics.

References

[1] M. Kulldorff, “A spatial scan statistic,” *Communications in Statistics - Theory and Methods*, vol. 26, no. 6, pp. 1481–1496, Jan. 1997, doi: 10.1080/03610929708831995.

[2] “Software for the spatial, temporal, and space-time scan statistics,” SaTScan, https://www.satscan.org/ (accessed Apr. 16, 2025).

[3] M. Matheny, “Pyscan,” pyscan - pyscan 1.0 documentation, https://mmath.dev/pyscan/ (accessed Apr. 16, 2025).

[4] Q. Di *et al.*, “An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution,” *Environment International*, vol. 130, p. 104909, Sep. 2019, doi: 10.1016/j.envint.2019.104909.

[5] D. Agarwal, A. McGregor, J. M. Phillips, S. Venkatasubramanian, and Z. Zhu, “Spatial scan statistics, Approximations and Performance Study” in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA: ACM, Aug. 2006, pp. 24–33. Accessed: Mar. 17, 2025. [Online]. Available: https://doi.org/10.1145/1150402.1150410.

[6] M. Kulldorff, L. Huang, and K. Konty, “A scan statistic for continuous data based on the normal probability model,” *International Journal of Health Geographics*, vol. 8, no. 1, p. 58, 2009, doi: 10.1186/1476-072x-8-58.

[7] Daymet, https://daymet.ornl.gov/ (accessed Apr. 16, 2025).

[8] “ECMWF reanalysis V5,” ECMWF, https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5 (accessed Apr. 16, 2025).

[9] “Open spatial demographic data and Research,” WorldPop, https://www.worldpop.org/ (accessed Apr. 16, 2025).

[10] S. Wrightson, J. Hosking, and A. Woodward, “Higher population density is associated with worse air quality and Related Health Outcomes in tāmaki makaurau,” *Australian and New Zealand Journal of Public Health*, vol. 49, no. 1, p. 100213, Feb. 2025. doi:10.1016/j.anzjph.2024.100213

[11] E. Penrod, “Utah’s Bad Air is the worst it’s been in years — and it’s likely to stick around,” The Salt Lake Tribune, https://archive.sltrib.com/article.php?id=3524088&itype=CMSID (accessed Apr. 16, 2025).

[12] “Smoke filled skies cause health concerns,” University of Utah Health, https://healthcare.utah.edu/healthfeed/2016/08/smoke-filled-skies-cause-health-concerns#:~:text=Utahns%20are%20used%20to%20periods,the%20air%20is%20especially%20toxic (accessed Apr. 16, 2025).