Representative Polygons for Idaho Air Quality Monitors

For Machine Learning Forecast Areas and Real-Time Monitor Areas

Idaho Department of Environmental Quality January 25, 2021

Objective

The aim of the analysis is to create polygons that represent two types of monitor reporting areas in Idaho.

The first type is *machine learning forecast areas*. A daily forecast is prepared for each air quality monitor in Idaho, using machine learning algorithms and forecasted meteorology. Polygons for each monitor are required to represent the forecast area for each monitor.

The second type is needed to report the real-time Air Quality Index (AQI) for monitor sites in Idaho. The *real-time monitor areas* will be used in a mobile application that returns the current AQI based on a user's location.

A polygon representing a daily forecast area for a monitor should capture a stable airshed that is well-mixed during the forecast period. Topography, local meteorology, emissions, and atmospheric chemistry all contribute to the formation of a recognizable airshed. Diurnal cycles of emissions, meteorology, and chemistry are evident and predictable within airsheds. The spatial relationship between these factors, when combined with local topography and location relative to major landforms, delineates the area of a cohesive airshed.

The area that represents the air measured at a monitor in real-time is not the same as a well-mixed, stable airshed. An instantaneous measurement does not have the time to average out the daily fluctuations in emissions, wind speed, or solar radiation that affect the concentrations at the monitor. Real-time monitor representativeness may change quickly, depending on the circumstances. An area for one monitor may be the extent of a wildfire smoke plume that blows through for an hour, or it may constitute the region of a week-long wintertime inversion. Area representativeness can also vary by pollutant, since emissions rates, transport time and distance, and chemistry differ between pollutants. Despite the inherent variability of the area represented by a real-time monitor, the extent must still be defined. Idaho DEQ will use the polygons developed for the *machine learning forecast areas* for the *real-time monitor areas* also, while noting the gap between desirability and achievability.

Airshed boundaries are not like political boundaries or the demarcation between an earthen bank and a river. The Earth's atmosphere is a single airshed; smaller tesselations of the ultimate airshed are necessarily indistinct. The exchange of air does not stop at a border. Therefore, defining an area that represents a part of the atmosphere that is similar to the air measured at a single monitor is difficult, and the result is ill-defined. But attempts must be made; this analysis represents one such attempt.

Background

Machine Learning Forecast Areas

A new machine learning air quality tool developed by Washington

State University (WSU) and Washington Ecology provides daily forecasts for monitors in the Pacific Northwest. Representative polygons for Idaho's monitors are needed to define the forecast areas.

The new air quality forecasting tool is based on machine learning algorithms which are a form of artificial intelligence. By looking for and discovering patterns and relationships in recently observed air quality data, along with forecast meteorological parameters from the University of Washington's Weather Research and Forecast model (WRF), these machine learning algorithms create air quality forecasts for permanent monitoring sites with more than three years of data. DEQ is currently developing a similar tool based on a more advanced machine learning structure that will complement the tool developed by WSU.

Real-time mobile app

The design of a real-time mobile app that returns an AQI based on a person's location in Idaho (determined by mobile phone coordinates), prompts a number of questions. What are the representative areas for each monitor? Should the provided AQI, depending on location, be derived from one or multiple monitors? Are there areas in Idaho that have no representative monitors and therefore no AQI? The development of monitor representativeness polygons aims to answer these questions.

Methods

The central method used to construct monitor representativeness polygons is *Spatially Constrained Multivariate Clustering*. It is a spatial statistics tool available in the latest version of ESRI's ArcPro desktop GIS software. Spatially Constrained Multivariate Clustering finds spatially contiguous clusters of features based on a set of feature attribute values and a given number of clusters. The method looks for a solution where cluster members are as similar as possible and the difference between clusters is maximized.

More information about *Spatially Constrained Multivariate Clustering* can be found here:

How Spatially Constrained Multivariate Clus...

Spatially Constrained Multivariate Clustering uses unsupervised machine learning methods to determine...

https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/how-spatially-constrained-multivariate-clustering-works.htm

Inputs

The main input to the clustering analysis used to develop monitor representativeness polygons is a grid of PM2.5 24-hour design value *background concentrations*. These PM2.5 concentrations values (in µg/m3) were developed from model and monitoring data for the period of July 2014 through June 2017.

Model data are 2014 –2017 4 km gridded CMAQ design values from AIRPACT v4 and v5. A photochemical air quality Eulerian grid modeling system (https://www.epa.gov/scram/photochemical-air-quality-modeling), AIRPACT incorporates emissions from land use, mobile, industrial, and biogenic sectors combined with predicted meteorological fields. Model data accounts for effects of terrain, meteorology, emissions, and photochemistry at a 4 km spatial scale.

Monitor data includes all permanent PM2.5 air quality monitors (including IMPROVE monitors) within the AIRPACT modeling domain. PM2.5 monitoring sites with less than one year of data, seasonal sites that did not run all three years, and temporary monitors were omitted.

The model and monitor data were combined using a geostatistical interpolation method, Empirical Bayesian Kriging with Regression Prediction (EBKRP). The ratio of the monitor design value to the median model value at each monitoring site was interpolated and then multiplied by the AIRPACT-5 median model values across the domain.

More information about *background concentrations* input data is available here:

Background Concentrations 2014 - 2017

Estimated background concentrations of criteria air pollutant design values for use in air permit engineering.

https://idahodeq.maps.arcgis.com/apps/MapSeries/index.html?appid=0c8a006e11fe4ec5939804b873098dfe

Secondary inputs to the monitor representativeness polygons development are previously established Idaho *airshed boundaries*. In 2007, Idaho DEQ generated airshed boundary polygons for the major metropolitan centers in Idaho. These areas include Coeur d'Alene, Idaho Falls, Lewiston, Pocatello, Treasure Valley, and Twin Falls.

The *airshed boundaries* were developed by combining conventional photochemical atmospheric dispersion modeling, reverse modeling, meteorology, emissions, topography, and population parameters in a weighted overlay.

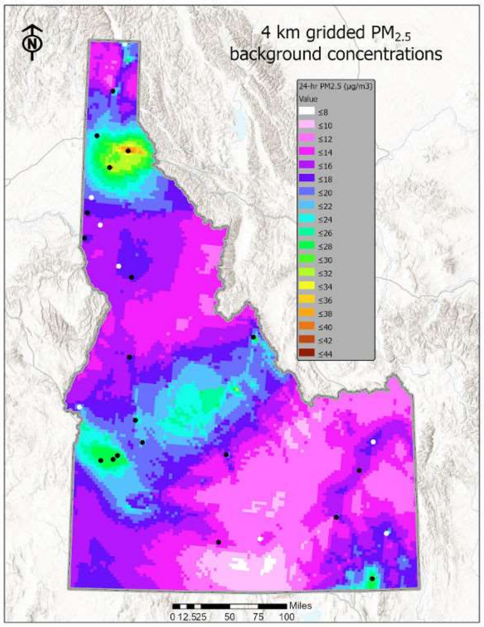
More information about *airshed boundaries* input data is available here:

Airshed Definition Presentation 2007

A Powerpoint presentation describing the development process of airshed boundaries for Idaho.

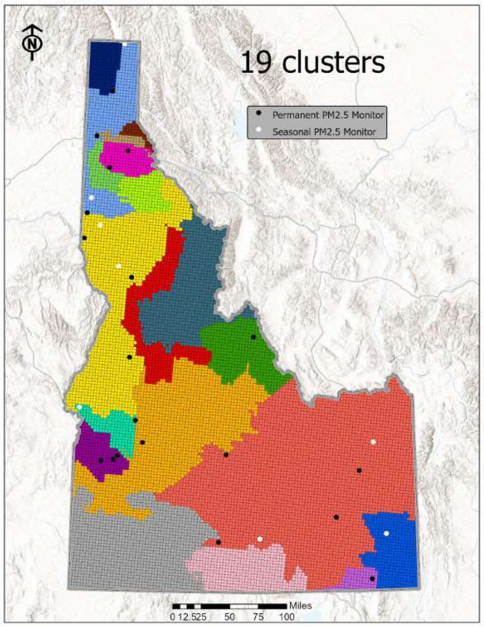
https://idahodeq.maps.arcgis.com/sharing/rest/content/items/398ad840a66b4fc0a07081f6a0e7f9cc/data

Spatially Constrained Multivariate Clustering was applied to 4 km gridded 24-hour PM2.5 background concentrations. Multiple iterations were run while varying the Number of Clusters parameter. At each step, statistical diagnostics were examined to identify and explore clustering optimization. The clustering results delineate known geographic regions in Idaho. The following examples show the input data and example cluster iterations.



Input data: 4 km gridded PM2.5 background concentrations

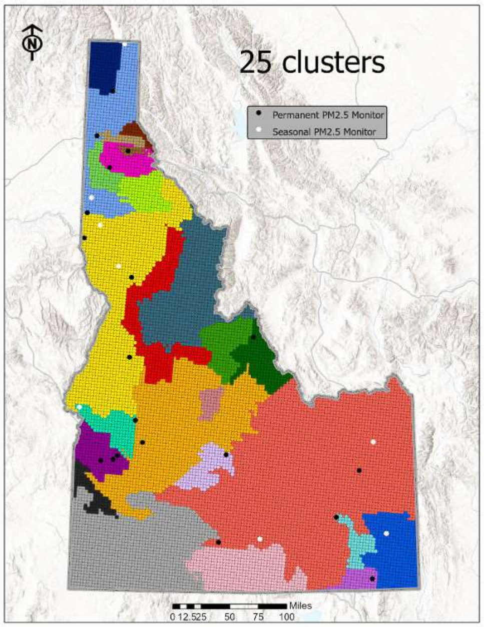
The *Spatially Constrained Multivariate Clustering* algorithm takes the input data and Number of Clusters parameter and creates spatially contiguous clusters where features within each cluster are as similar as possible and all clusters are as different from each other as possible.



Output: 19 clusters iteration

The 19 clusters solution divides Idaho into geographically recognizable areas. Separate clusters contain permanent PM2.5 monitors:

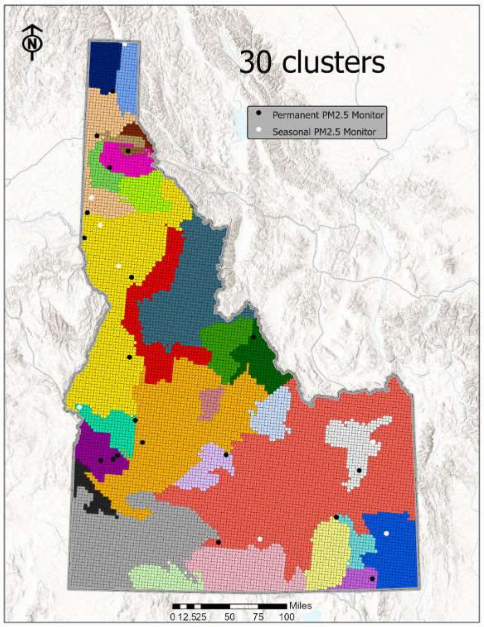
1. Sandpoint/Coeur d'Alene/Moscow
2. Pinehurst/St. Maries
3. Lewiston/Grangeville/McCall
4. Salmon
5. Idaho City
6. Garden Valley
7. Treasure Valley monitors
8. Twin Falls/Ketchum/Pocatello/Idaho Falls
9. Preston



Output: 25 clusters iteration

The 25 clusters solution separates Pinehurst, Pocatello, and Ketchum into their own clusters. The Salmon cluster is further divided. Individual clusters containing permanent PM2.5 monitors now include:

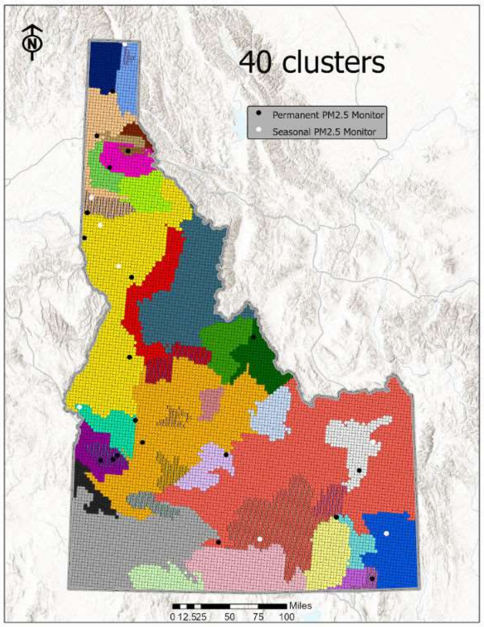
1. Sandpoint/Coeur d'Alene/Moscow
2. Pinehurst
3. St. Maries
4. Lewiston/Grangeville/McCall
5. Salmon
6. Idaho City
7. Garden Valley
8. Treasure Valley monitors
9. Twin Falls/Idaho Falls
10. Ketchum
11. Pocatello
12. Preston



Output: 30 clusters iteration

The 30 clusters solution divides Sandpoint from Coeur d'Alene and Moscow and pulls Idaho Falls out into its own cluster. The list of clusters containing permanent PM2.5 monitors now includes:

1. Sandpoint
2. Coeur d'Alene/Moscow
3. Pinehurst
4. St. Maries
5. Lewiston/Grangeville/McCall
6. Salmon
7. Idaho City
8. Garden Valley
9. Treasure Valley monitors
10. Twin Falls
11. Idaho Falls
12. Ketchum
13. Pocatello
14. Preston



Output: 40 clusters iteration

The 40 clusters solution creates a cluster for the Moscow monitor, separate from Coeur d'Alene. The list of clusters containing permanent PM2.5 monitors now includes:

1. Sandpoint
2. Coeur d'Alene
3. Moscow
4. Pinehurst
5. St. Maries
6. Lewiston/Grangeville/McCall
7. Salmon
8. Idaho City
9. Garden Valley
10. Treasure Valley monitors
11. Twin Falls
12. Idaho Falls
13. Ketchum
14. Pocatello
15. Preston

Further clustering iterations do not separate Lewiston, Grangeville, and McCall because the input data for those regions are too similar.

If these three monitors are to have their own representative polygons, other methods must be applied to separate them.

The second input dataset, *airshed boundaries*, was used to further modify cluster boundaries to include areas of high population or for management considerations.

Results

Idaho Monitor Representativeness Polygons...

Click here to view final polygons on an interactive map.

https://idahodeq.maps.arcgis.com/apps/View/index.html?appid=c7c026ef0aeb41a987a88b9200efc91e

Credits

StoryMap	Sara Strachan (Technical Services, IDEQ)
Analysis	Sara Strachan, Brian Himes, Wei Zheng (Technical Services, IDEQ)

Powered by ArcGIS StoryMaps