Spatial Estimation of Chronic Respiratory Disease Based on Geospatial Learning Procedures – An Approach Using Earth-Engine-Based Remote Sensing Data and Air Quality Variables in the State of Pennsylvania

EMILY ZHOU, SHUAI WANG, MUSA6500-S24 Final Project

Original Study: Alvarez-Mendoza, C. I., Teodoro, A., Freitas, A., & Fonseca, J. (2020). Spatial estimation of chronic respiratory diseases based on machine learning procedures—An approach using remote sensing data and environmental variables in quito, Ecuador. Applied Geography, 123, 102273. https://doi.org/10.1016/j.apgeog.2020.102273

Replication Study: https://github.com/emilyzhou112/MUSA6500-Pennsylvania-CRD

Keywords: support vector machine, random forest, multiple layer perceptron, deep learning, bayesian information criteria, google earth engine, geospatial health

Remote Sensing and Machine Learning in Public Health Research

- **Enhance** the accuracy of disease prediction and early detection
- **Improve** the efficiency of healthcare resource allocation
- **Enable** the development of personalized treatment plans
- **Facilitate** the identification of social and environmental determinants of health.
- **Enable** the mapping and real-time monitoring of disease vectors.
- **Provide** consistent spatial and temporal coverage.

Using Machine Learning to Predict ICU Transfer in Hospitalized COVID-19 Patients

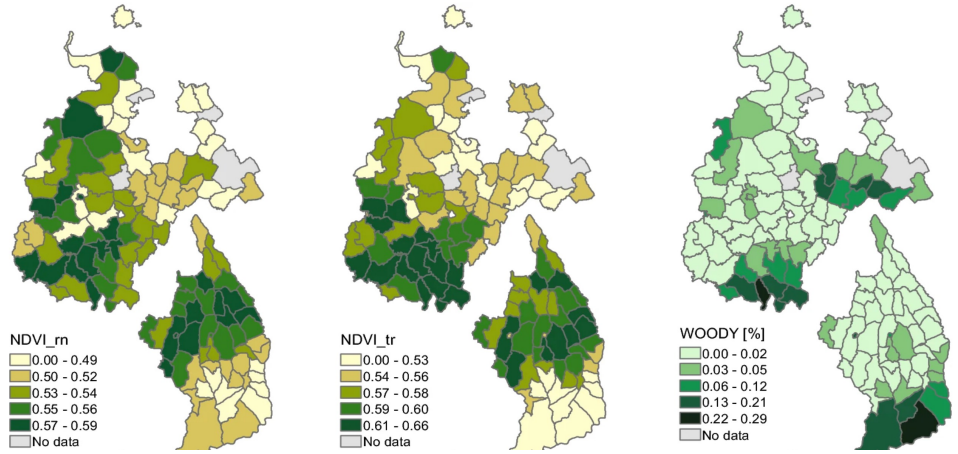
Fu-Yuan Cheng [1], Himanshu Joshi [1,2], Pranai Tandon [3], Robert Freeman [1,4], David L Reich [4,5], Madhu Mazumdar [1,2,*], Roopa Kohli-Seth [6], Matthew A. Levin [5,7], Prem Timsina [1,†] and Arash Kia [1,†]

1 Institute for Healthcare Delivery Science; Icahn School of Medicine at Mount Sinai, 1425 Madison Avenue, New York, NY 10029, USA; Fu-Yuan.Cheng@mountsinai.org (F.-Y.C.); Himanshu.Joshi@mountsinai.org (H.J.); Robert.Freeman@mountsinai.org (R.F.); Prem.Timsina@mountsinai.org (P.T.); Arash.Kia@mountsinai.org (A.K.)
2 Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, 1425 Madison Avenue, New York, NY 10029, USA
3 Respiratory Institute, Icahn School of Medicine at Mount Sinai, 10 E 102nd St, New York, NY 10029, USA; Pranai.Tandon@mountsinai.org
4 Hospital Administration; The Mount Sinai Hospital, 1 Gustave L. Levy Place, New York, NY 10029, USA;

Remote sensing of environmental risk factors for malaria in different geographic contexts

Andrea McMahon, Abere Mihretie, Adem Agmas Ahmed, Mastewal Lake, Worku Awoke & Michael Charles Wimberly

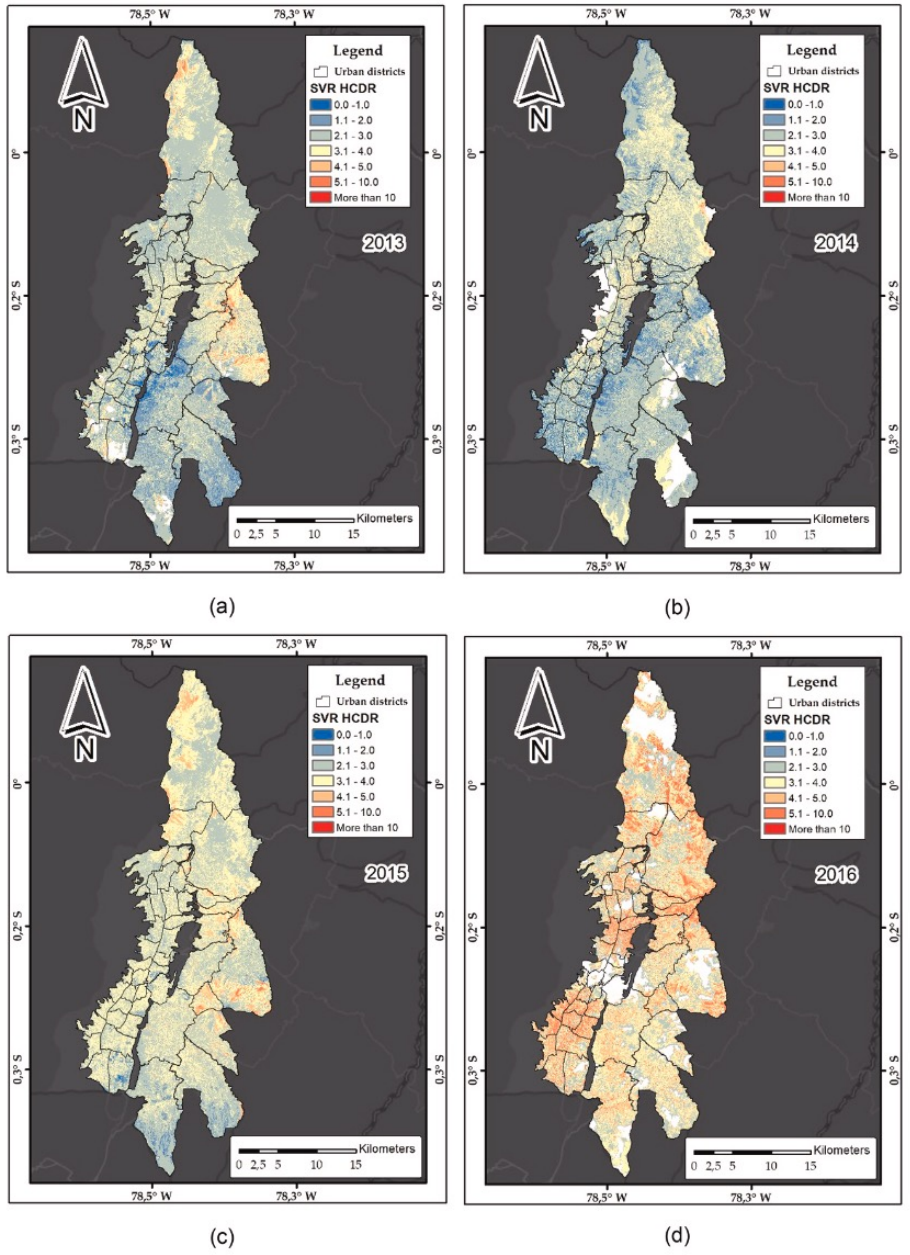
International Journal of Health Geographics **20**, Article number: 28 (2021) | Cite this article



Alvarez-Mendoza, et.al (2020)'s Study

Topic: Compare the effectiveness of several machine learning models in predicting the number of hospital discharge patients with chronical respiratory diseases using remote sensing and air quality data from 2013 to 2017.

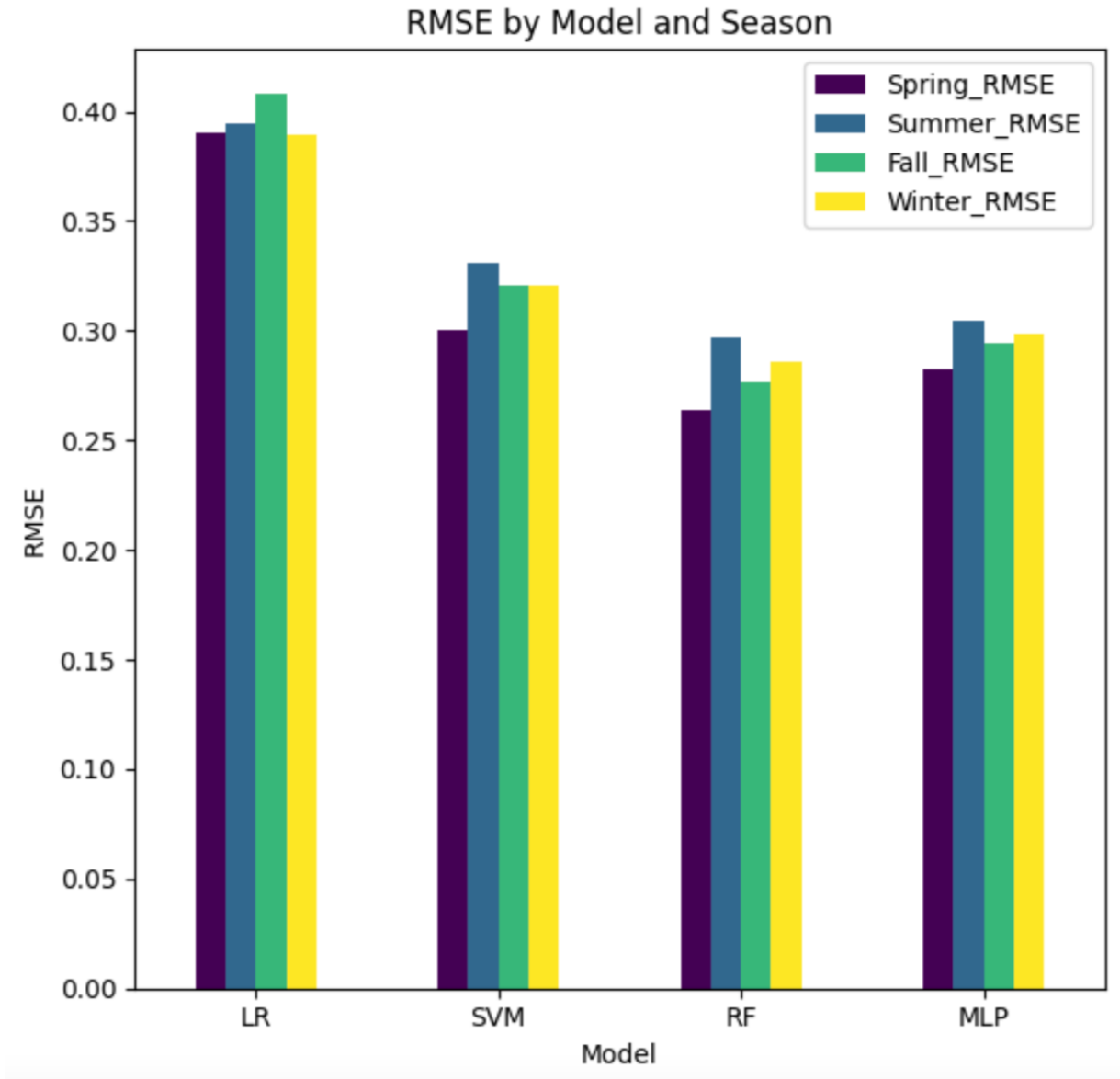
Goal: Understand the most significant spatial predictors and the spatial distribution of chronical respiratory disease in the city of Quito, Ecuador .



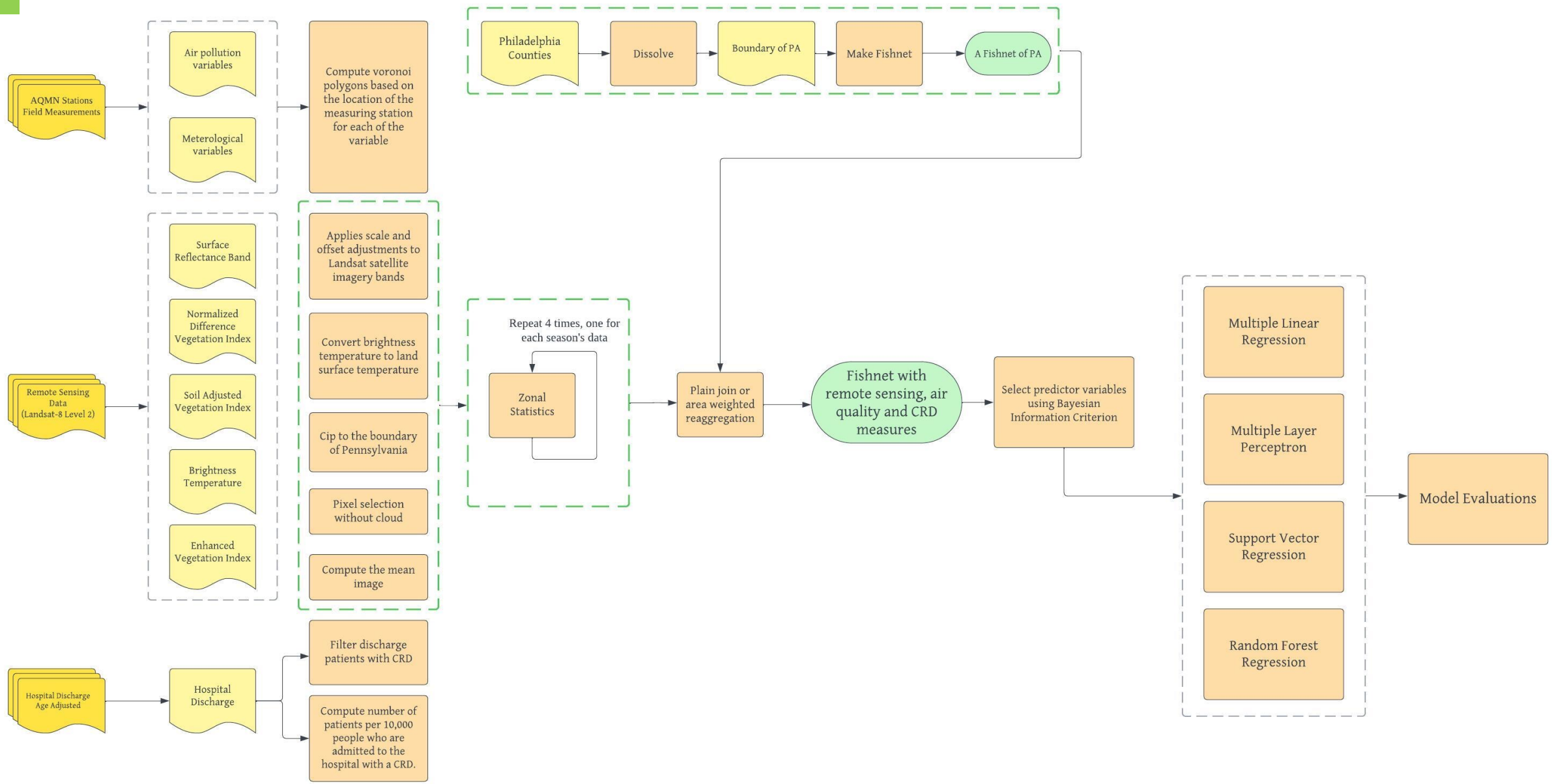
Our Study

Topic: replicate and improve upon Alvarez-Mendoza et al.'s method to investigate the effectiveness of several machine learning models in predicting the number of hospital discharge patients with CRD in the state of Pennsylvania by different seasons.

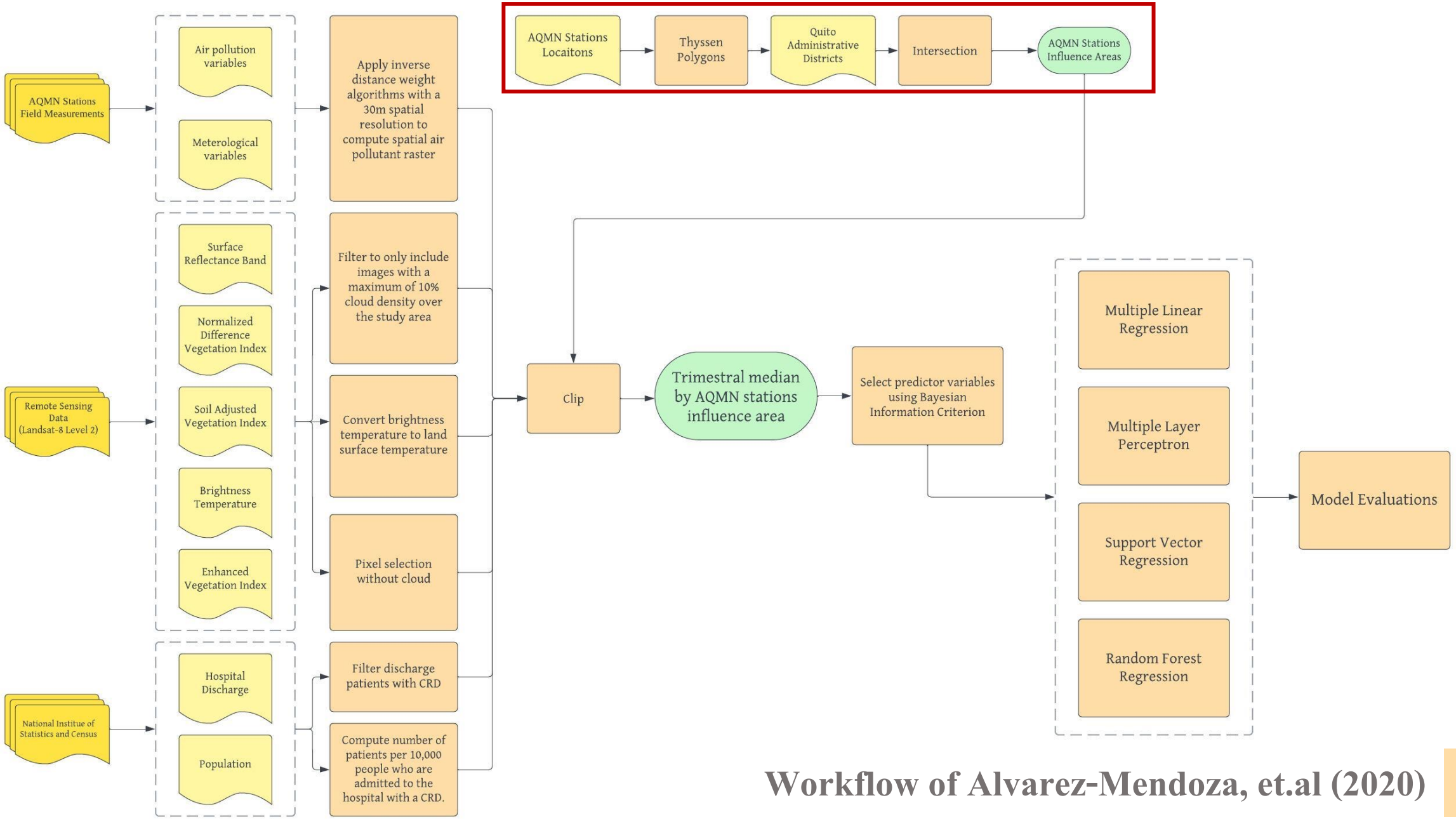
Goal: understand the most significant environmental and atmospheric factors leading to higher CRD risk in Pennsylvania, compare the performance of different machine learning models, compare the influence of seasonality on CRD discharge rate.



Our Workflow

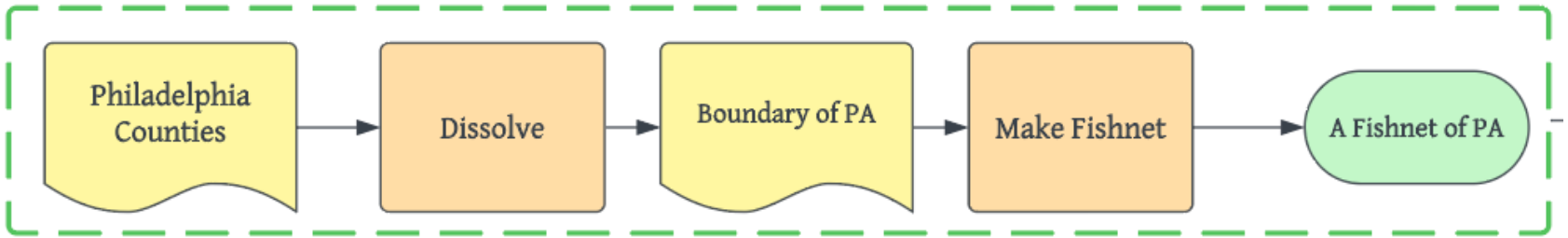


1. Geographic Transformation



Workflow of Alvarez-Mendoza, et.al (2020)

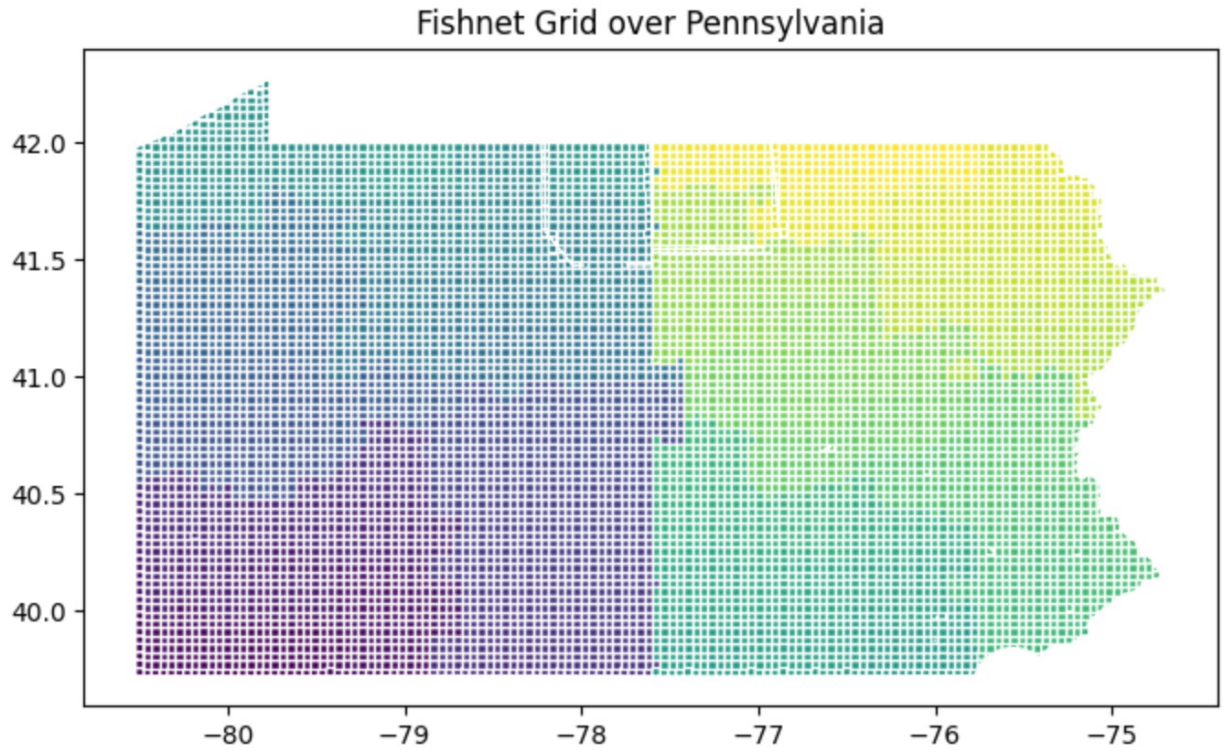
Geographic Transformations



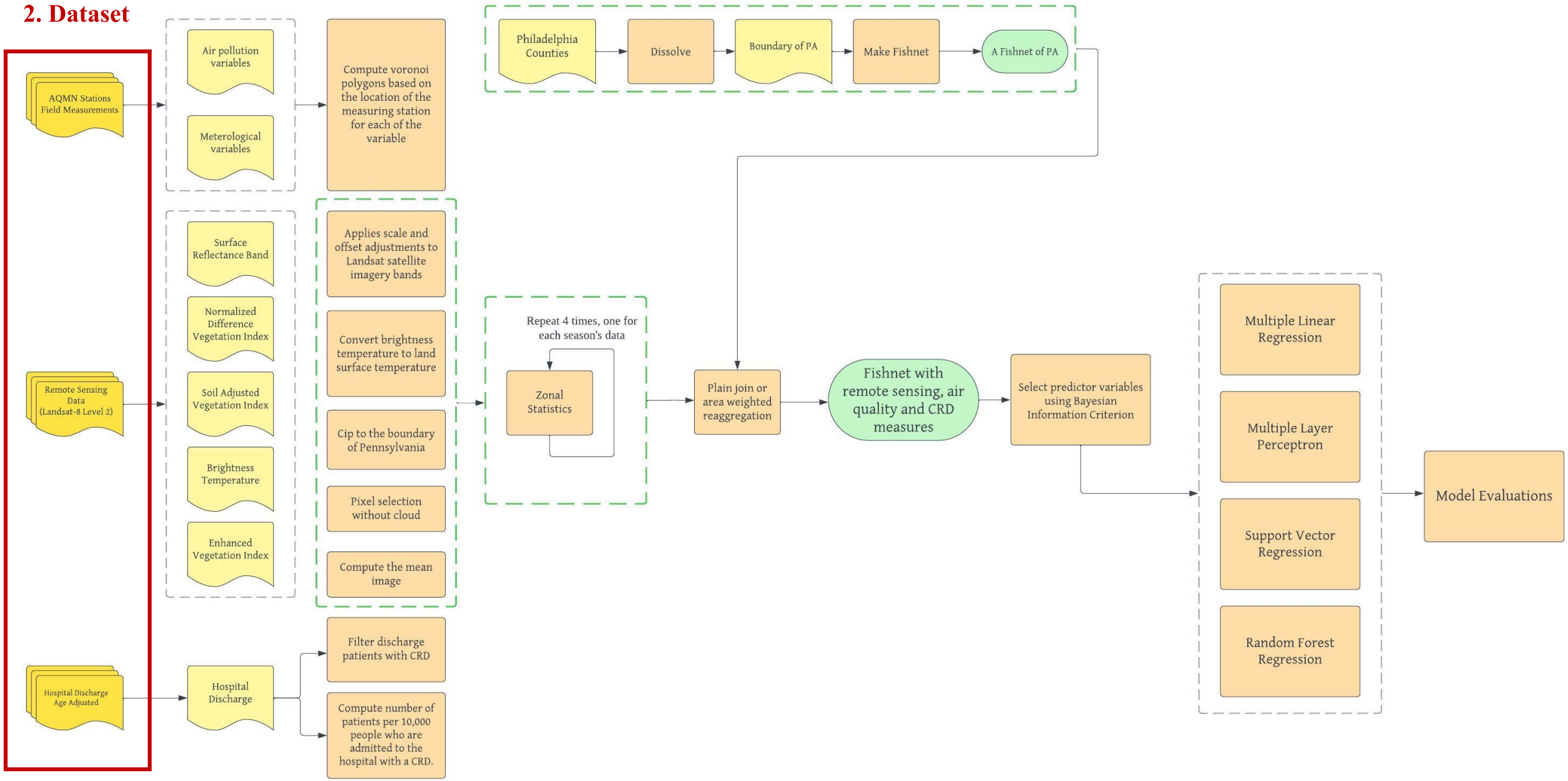
Using a nested loop, iterates over the X and Y coordinates within the bounds of PA boundary. Within each iteration, creates a square polygon geometry representing a 5000 * 5000 grid cell.

Clip fishnet grids to the extent of study area.

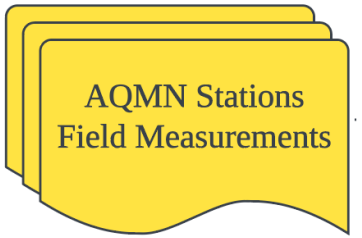
In total, there are 8379 grids.



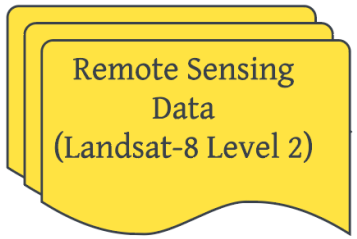
2. Dataset


AQMN Stations Field Measurements
Air pollution variables
Meterological variables
Compute voronoi polygons based on the location of the measuring station for each of the variable
Philadelphia Counties
Dissolve
Boundary of PA
Make Fishnet
A Fishnet of PA
Remote Sensing Data (Landsat-8 Level 2)
Surface Reflectance Band
Normalized Difference Vegetation Index
Soil Adjusted Vegetation Index
Brightness Temperature
Enhanced Vegetation Index
Applies scale and offset adjustments to Landsat satellite imagery bands
Convert brightness temperature to land surface temperature
Cip to the boundary of Pennsylvania
Pixel selection without cloud
Compute the mean image
Repeat 4 times, one for each season's data
Zonal Statistics
Plain join or area weighted reaggregation
Fishnet with remote sensing, air quality and CRD measures
Select predictor variables using Bayesian Information Criterion
Multiple Linear Regression
Multiple Layer Perceptron
Support Vector Regression
Random Forest Regression
Model Evaluations
Hospital Discharge Age Adjusted
Hospital Discharge
Filter discharge patients with CRD
Compute number of patients per 10,000 people who are admitted to the hospital with a CRD.

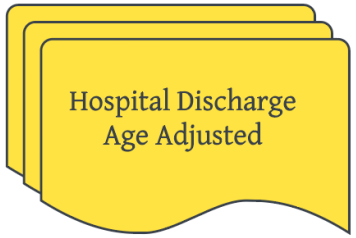
Datasets



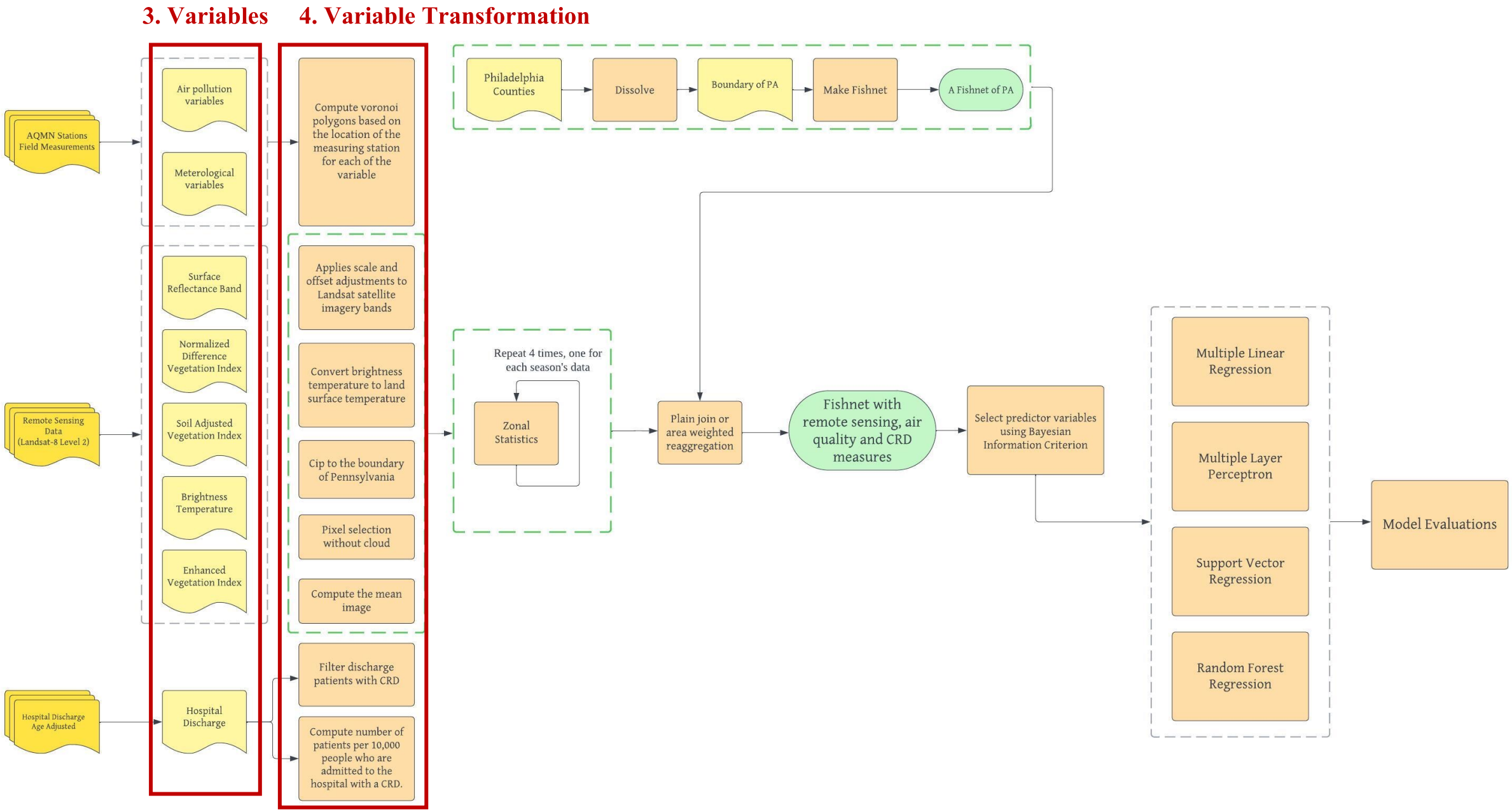
Field Data from Pennsylvania Department of Air Quality Monitoring Network (AQMN) that provides hourly field measurements of air pollutants and meteorological variables.



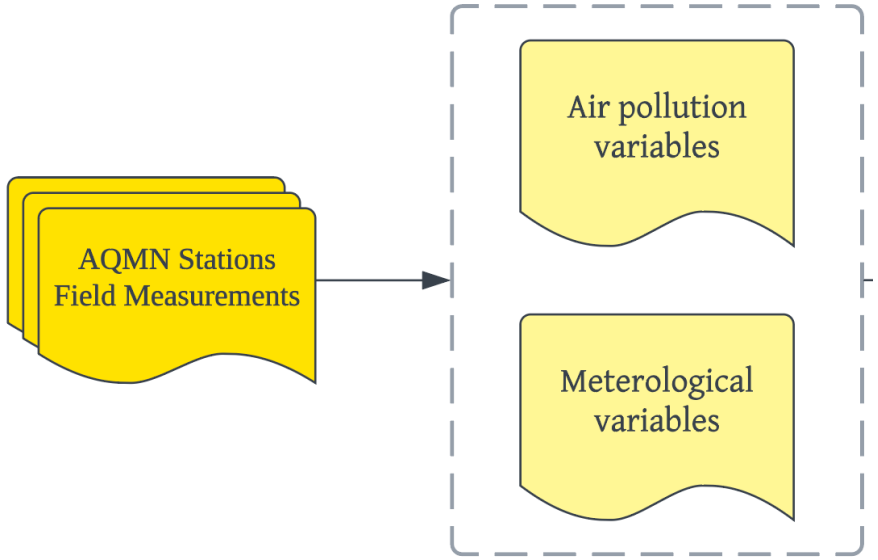
Landsat 8 Level 2 Collection 2 Tier 1 images between Spring 2022 and Spring 2023 retrieved from the Google Earth Engine API with *30m of spatial resolution*. The images were acquired by season.



Hospital discharge of patients with CRD collected by Pennsylvania's Department of Health.


3. Variables
4. Variable Transformation
AQMN Stations Field Measurements
Air pollution variables
Meterological variables
Compute voronoi polygons based on the location of the measuring station for each of the variable
Philadelphia Counties
Dissolve
Boundary of PA
Make Fishnet
A Fishnet of PA
Remote Sensing Data (Landsat-8 Level 2)
Surface Reflectance Band
Normalized Difference Vegetation Index
Soil Adjusted Vegetation Index
Brightness Temperature
Enhanced Vegetation Index
Applies scale and offset adjustments to Landsat satellite imagery bands
Convert brightness temperature to land surface temperature
Cip to the boundary of Pennsylvania
Pixel selection without cloud
Compute the mean image
Repeat 4 times, one for each season's data
Zonal Statistics
Plain join or area weighted reaggregation
Fishnet with remote sensing, air quality and CRD measures
Select predictor variables using Bayesian Information Criterion
Multiple Linear Regression
Multiple Layer Perceptron
Support Vector Regression
Random Forest Regression
Model Evaluations
Hospital Discharge Age Adjusted
Hospital Discharge
Filter discharge patients with CRD
Compute number of patients per 10,000 people who are admitted to the hospital with a CRD.

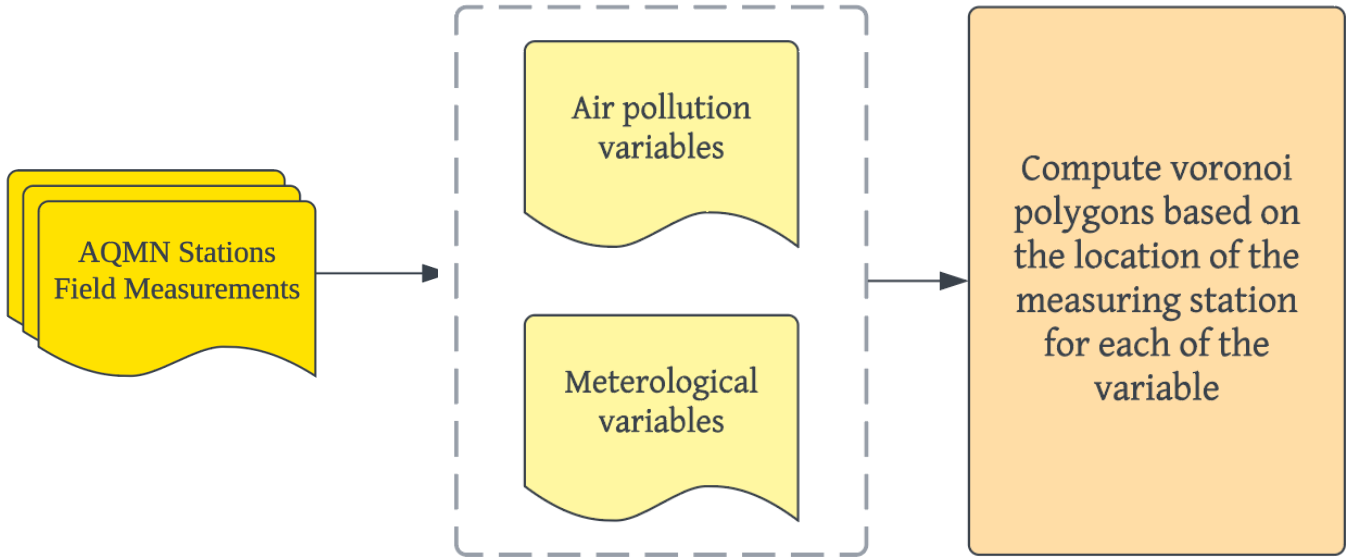
Input Variables and Transformation

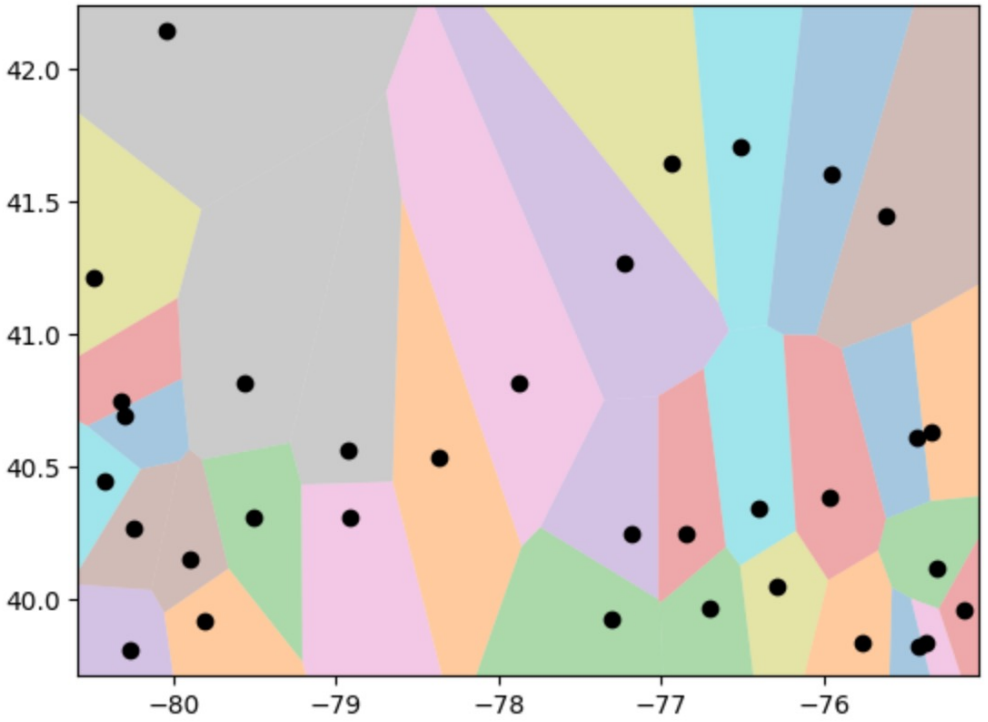


PM2.5 and O_3

wind speed and solar radiation

Input Variables and Transformation

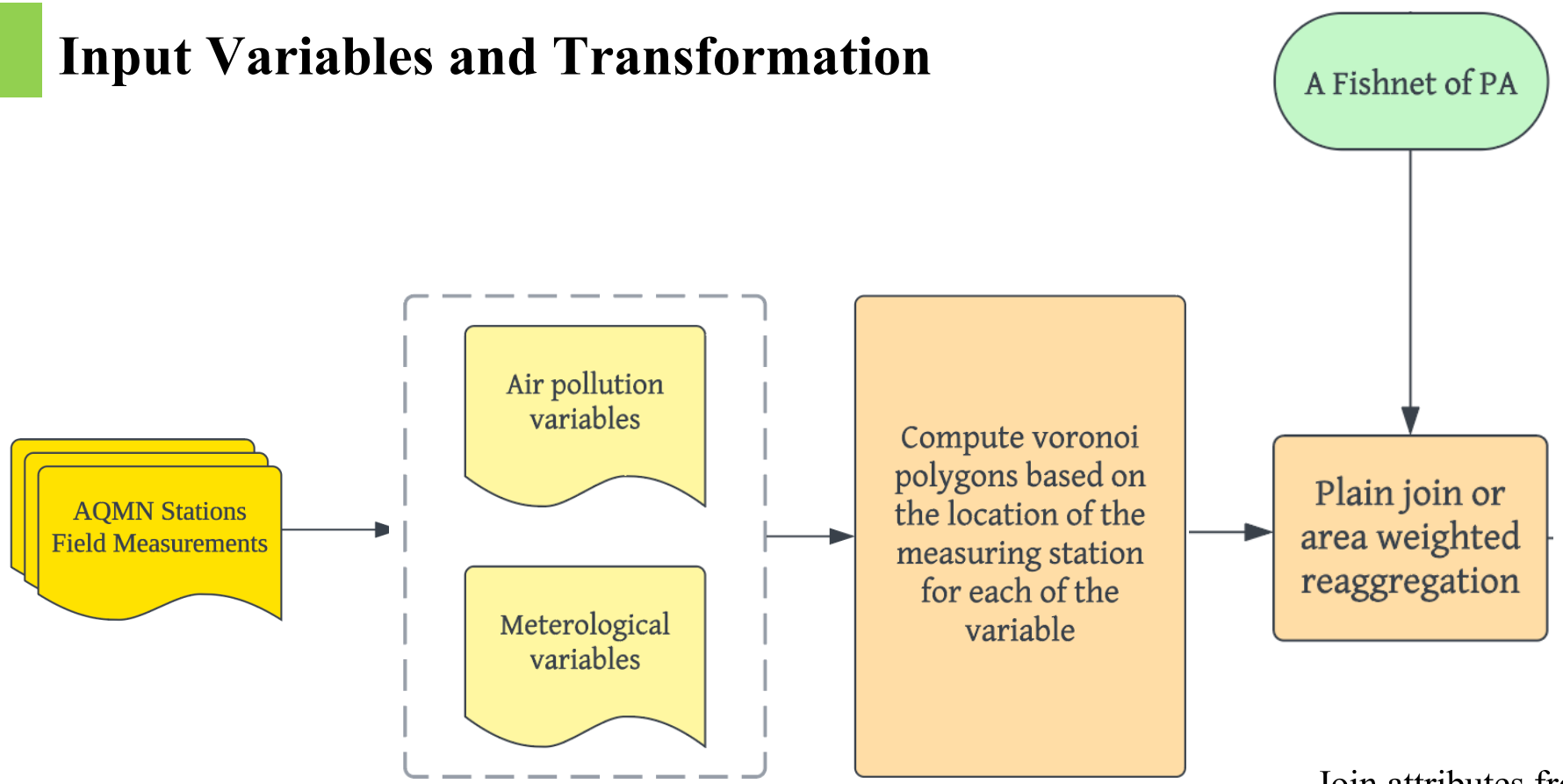




Generate polygons from a set of sample points such that any location inside the polygon is closer to that point than any of the other sample points.

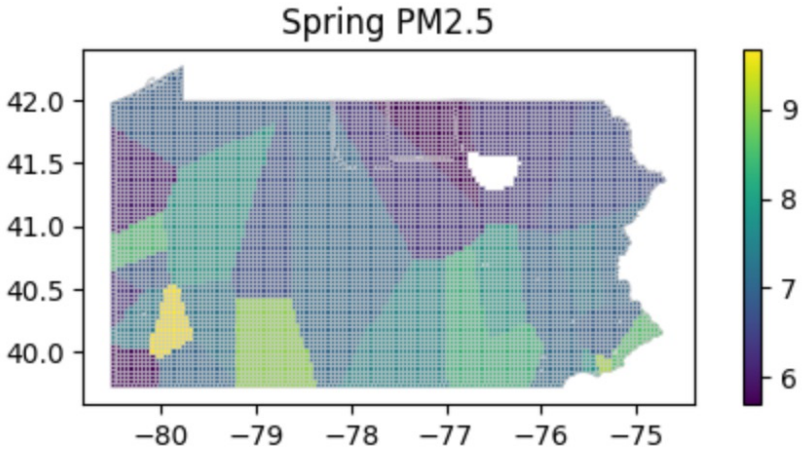
Each station is associated with a polygon representing the area where it has the closet proximity.

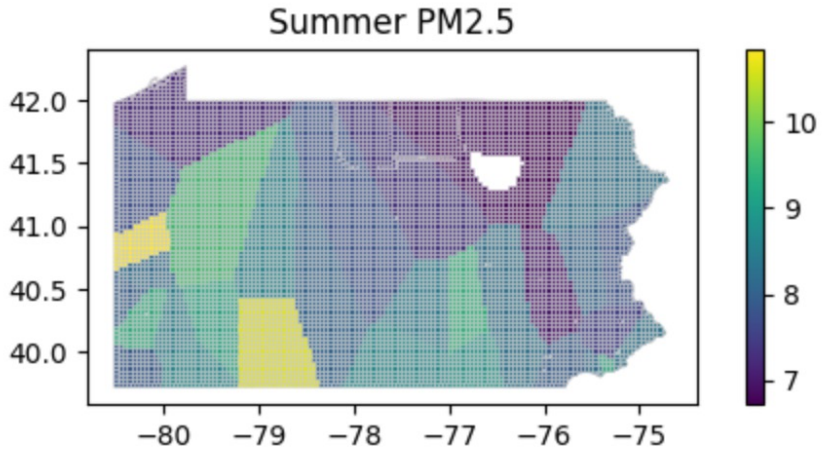
Input Variables and Transformation

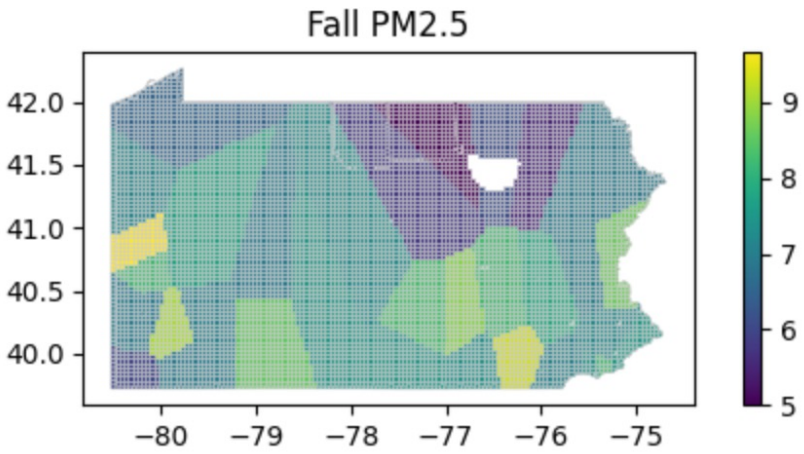


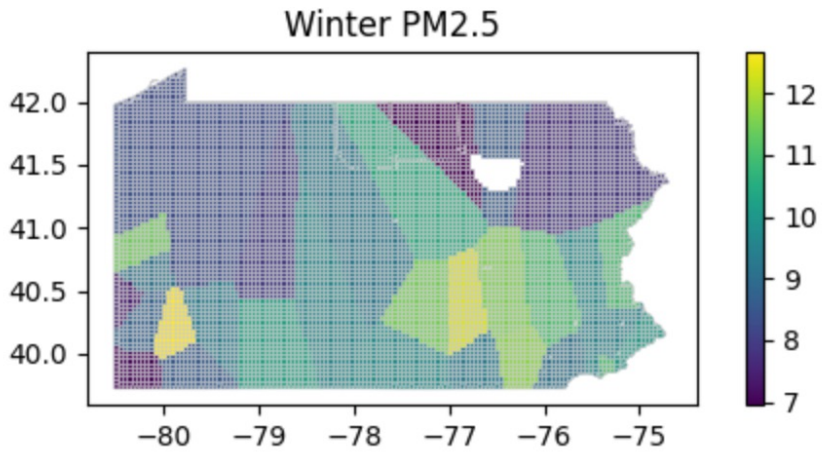
Join attributes from Voronoi polygons into the fishnet by comparing the centroid of the fishnet to the polygon. If the centroid intersects with the polygon, then assign the attributes of that polygon to the fishnet grid.

Intermediaries



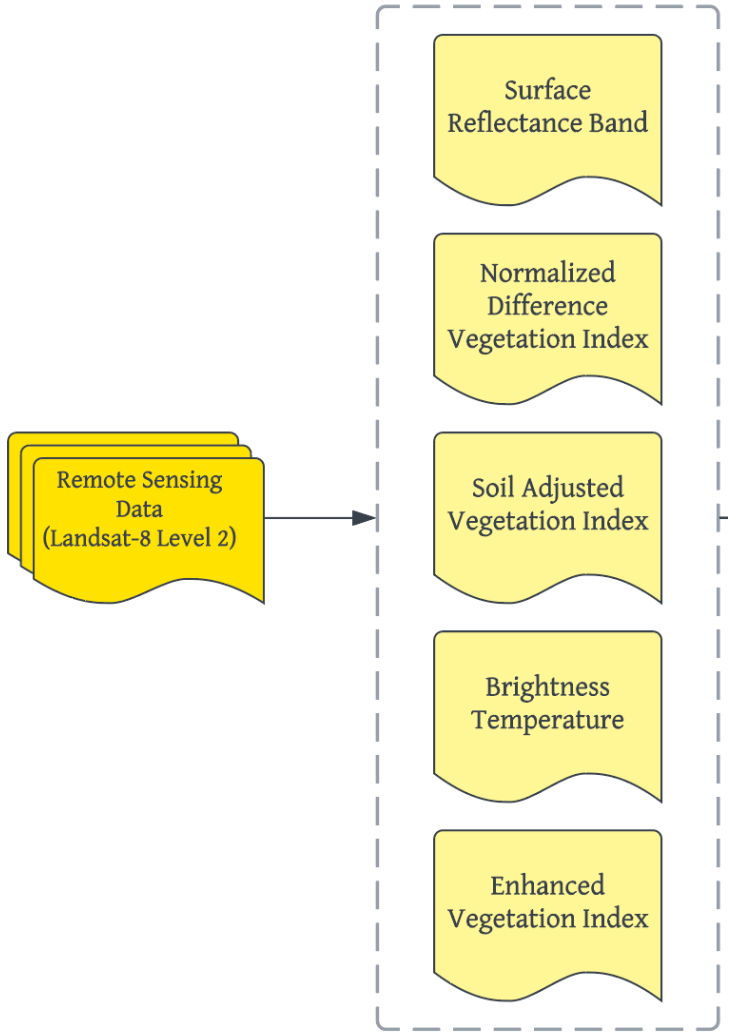






PM2.5 tends to be higher in winter and summer while lower in spring and fall. Some counties in particular, stand out as having much higher PM2.5 than others.

Input Variables and Transformations



Coastal aerosol band (B1), blue band (B2), green band(B3), red band (B4), near-infrared (B5), short-wave infrared 1(B6), short-wave infrared 2(B7)

Quantifying the health and density of vegetation.

NDVI = (Band 5 – Band 4) / (Band 5 + Band 4)

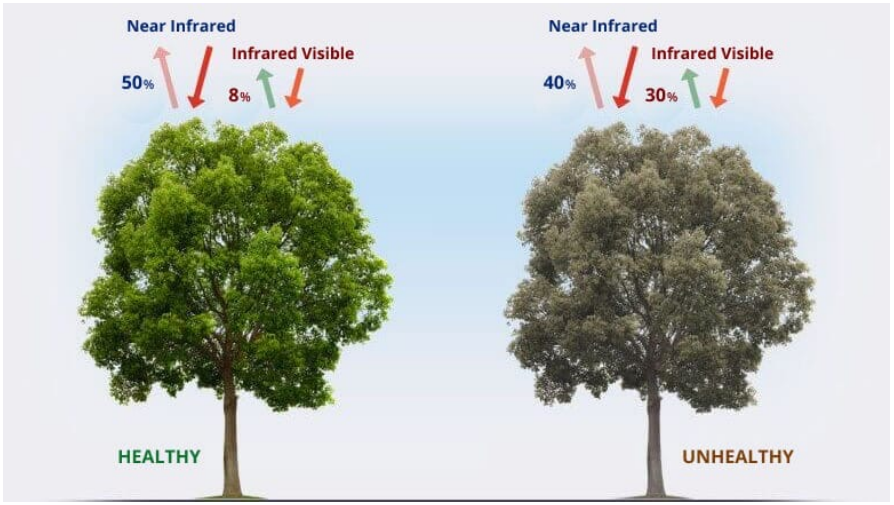
Correct NDVI for the influence of soil brightness in areas where vegetative cover is low

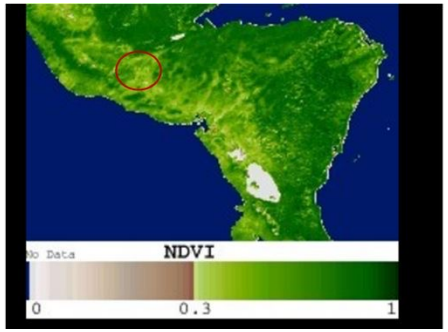
SAVI = ((Band 5 – Band 4) / (Band 5 + Band 4 + 0.5)) * (1.5)

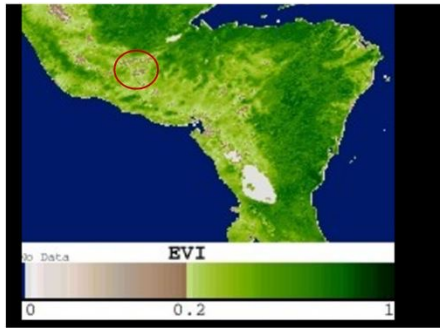
Derived from thermal band 10, and 11.

Correct NDVI for some atmospheric conditions and canopy background noise

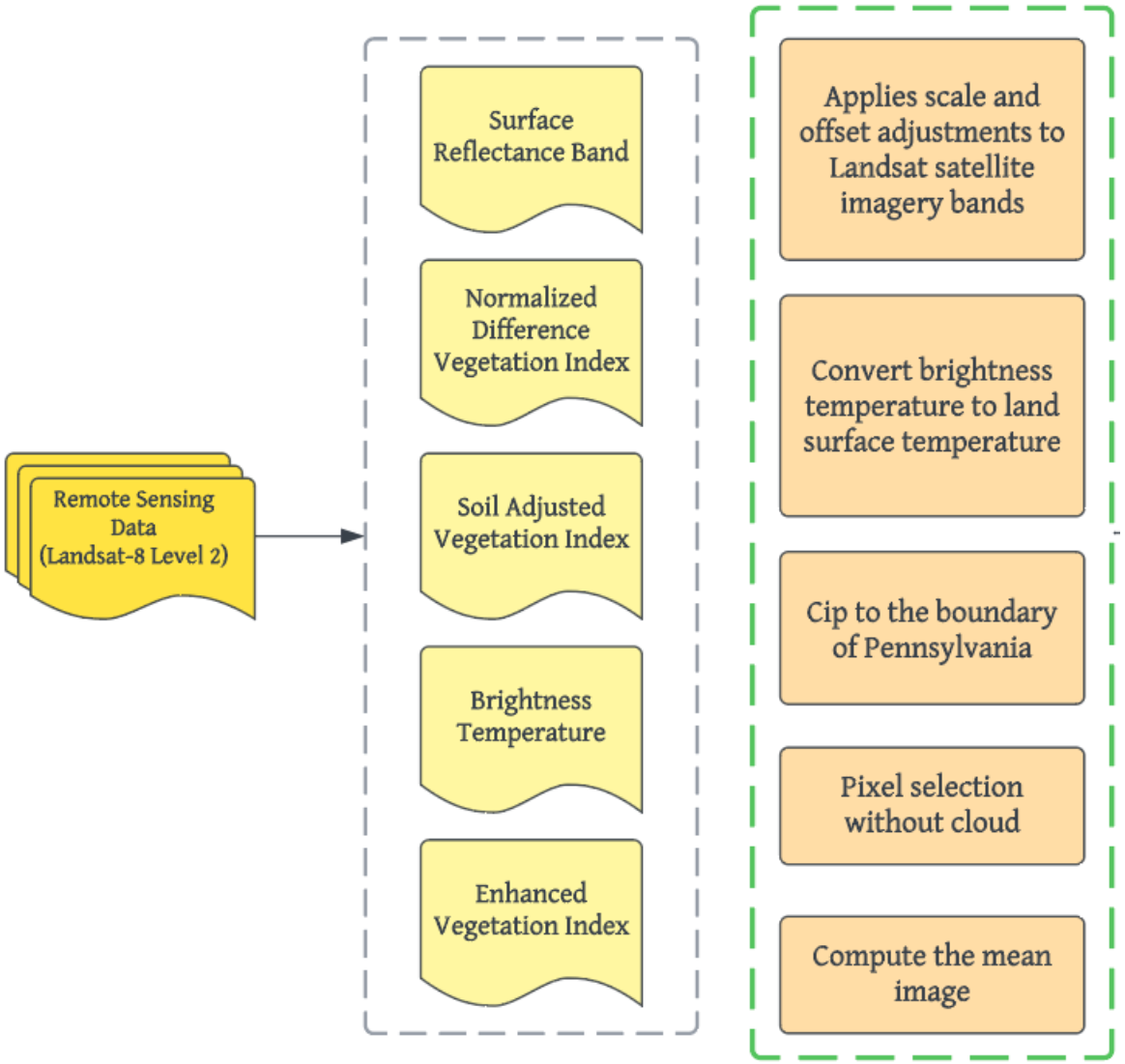
EVI = 2.5 * ((Band 5 – Band 4) / (Band 5 + 6 * Band 4 – 7.5 * Band 2 + 1))







Input Variables and Transformations

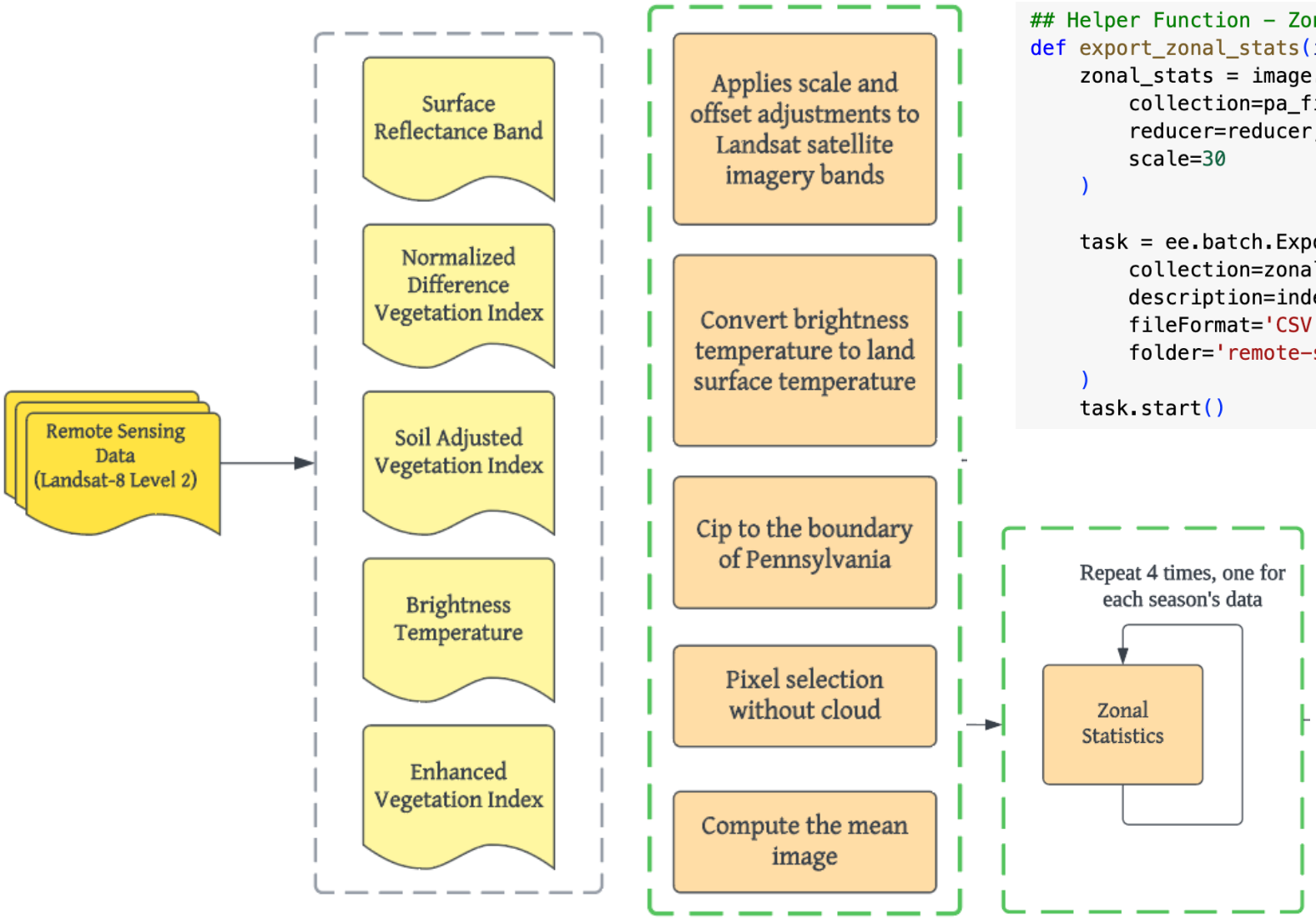



```
## Load Spring Image Collection
imageSpring = ee.ImageCollection("LANDSAT/LC08/C02/T1_L2") \
              .filterBounds(aoi) \
              .filterDate(startSpring, endSpring) \
              .map(apply_scale_factors) \
              .map(cloud_mask) \
              .median() \
              .clip(aoi)
```

Translate the brightness temperature measurements from remote sensing instruments into meaningful temperature values for the land surface.

$$LST = \frac{BT}{\left(1 + \left(\frac{\lambda * BT}{\rho}\right) ln\varepsilon\right)} - 273.15$$

Input Variables and Transformations

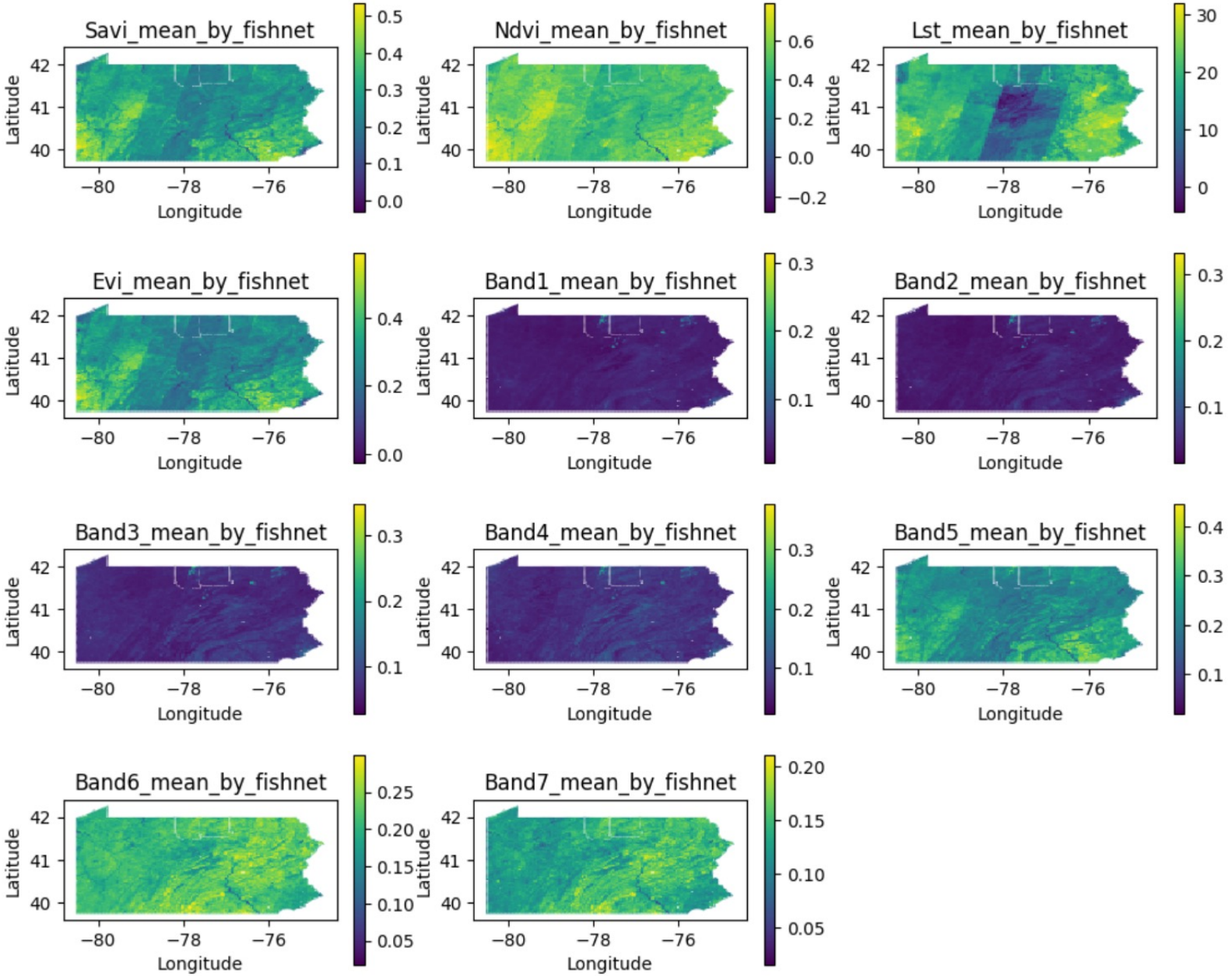



```python
## Helper Function - Zonal Statistics
def export_zonal_stats(image, reducer, index_name):
    zonal_stats = image.reduceRegions(
        collection=pa_fishnet_ee,
        reducer=reducer,
        scale=30
    )

    task = ee.batch.Export.table.toDrive(
        collection=zonal_stats,
        description=index_name,
        fileFormat='CSV',
        folder='remote-sensing' # default to our shared google drive for the project
    )
    task.start()
```

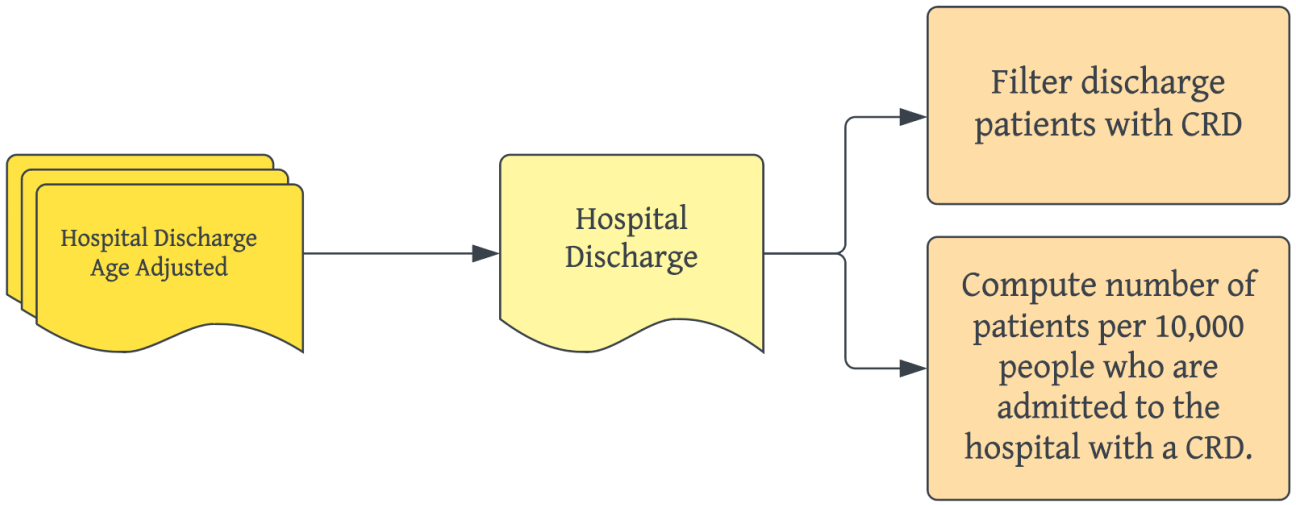
Zonal Statistics: one of the easiest method of brining raster information into vector geometries. In Earth Engine, zonal statistics can be calculated by aggregating pixel values within geometries defined by feature collections (in our case, the fishnet).

Intermediaries

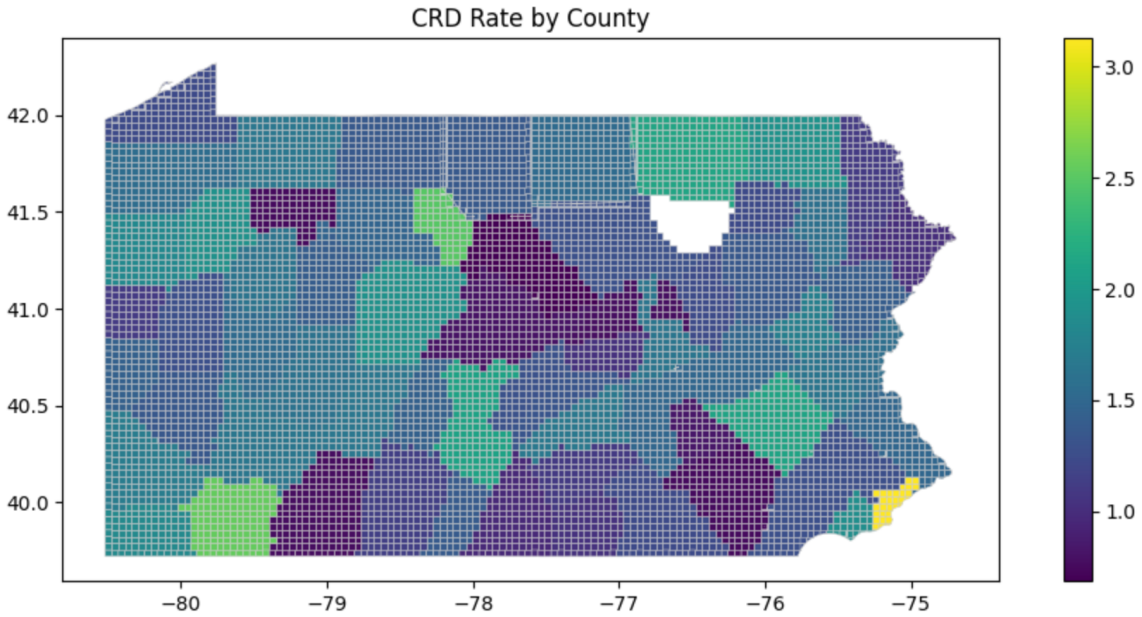
We may see that the NDVI (as well as two other vegetation indices) are higher in the southeastern and southwestern part of the state but lower in the central part of the state.



Input Variables and Transformations

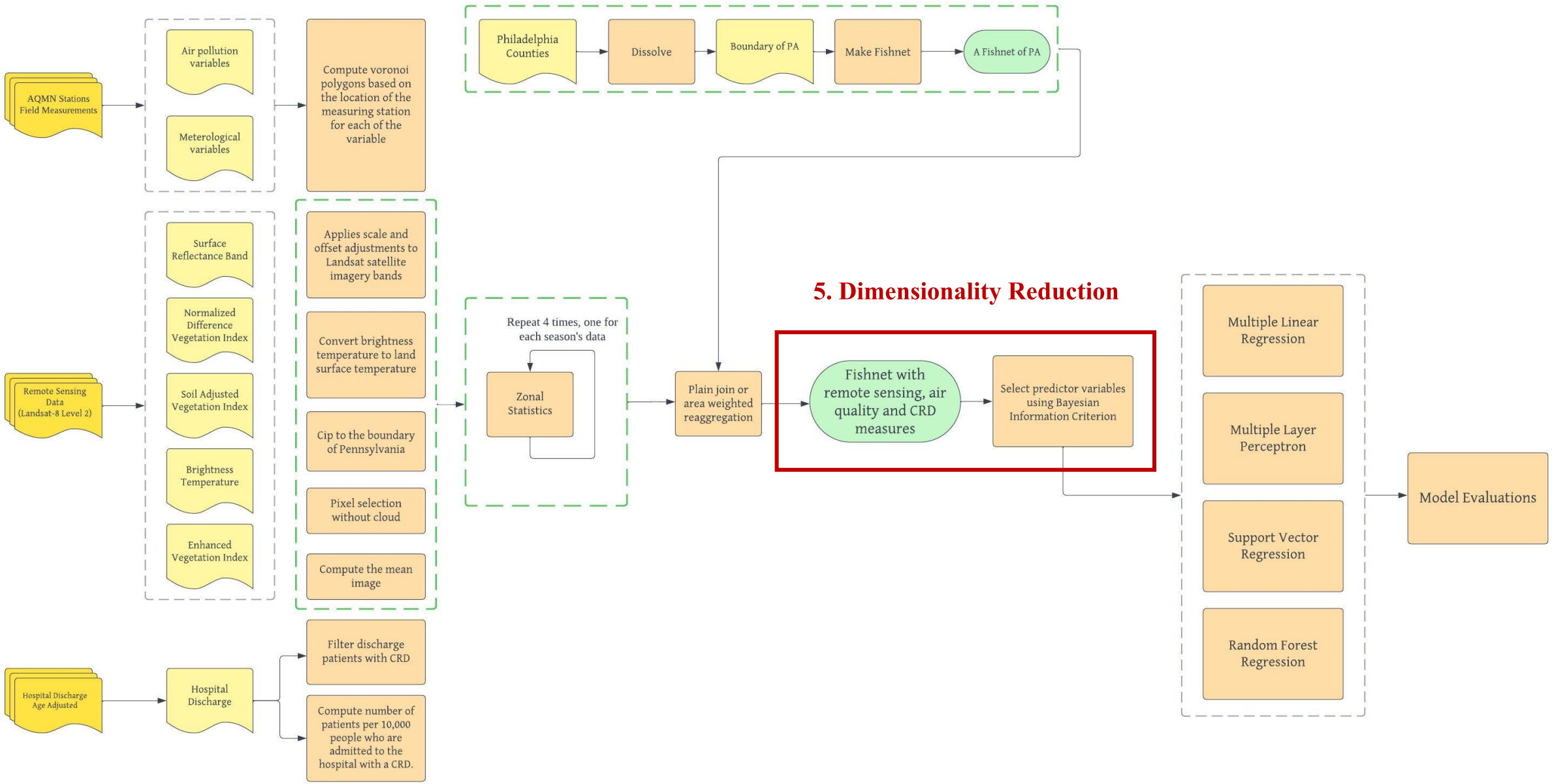


Normalize the data by total population.

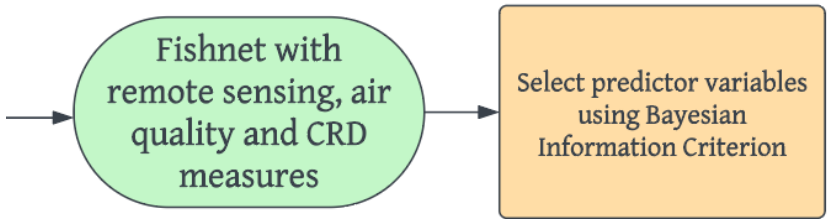


Major Deviations from the Original Study (Summary)

	Original Study	Replication Study
Variable Selection	24 predictor variables in total	15 predictor variables in total due to the limitations in air quality dataset
Satellite Image Selection	46 Landsat 8 Level 2 images from 2013 to 2017 grouped by trimester	all Landsat 8 level 2 images from 2022 to 2023 grouped into 4 by the median of each season.
Unit of Analysis	air quality monitoring network stations influence area	divide the state of Pennsylvania into about 8000 fishnet grids
Spatial Interpolation	compute Voronoi polygon for air quality	use zonal statistics to summarize pixels into each grid in addition to computing Voronoi polygon for each air quality monitoring station
Modeling	4 machine learning models	16 machine learning models, 4 on each season's data

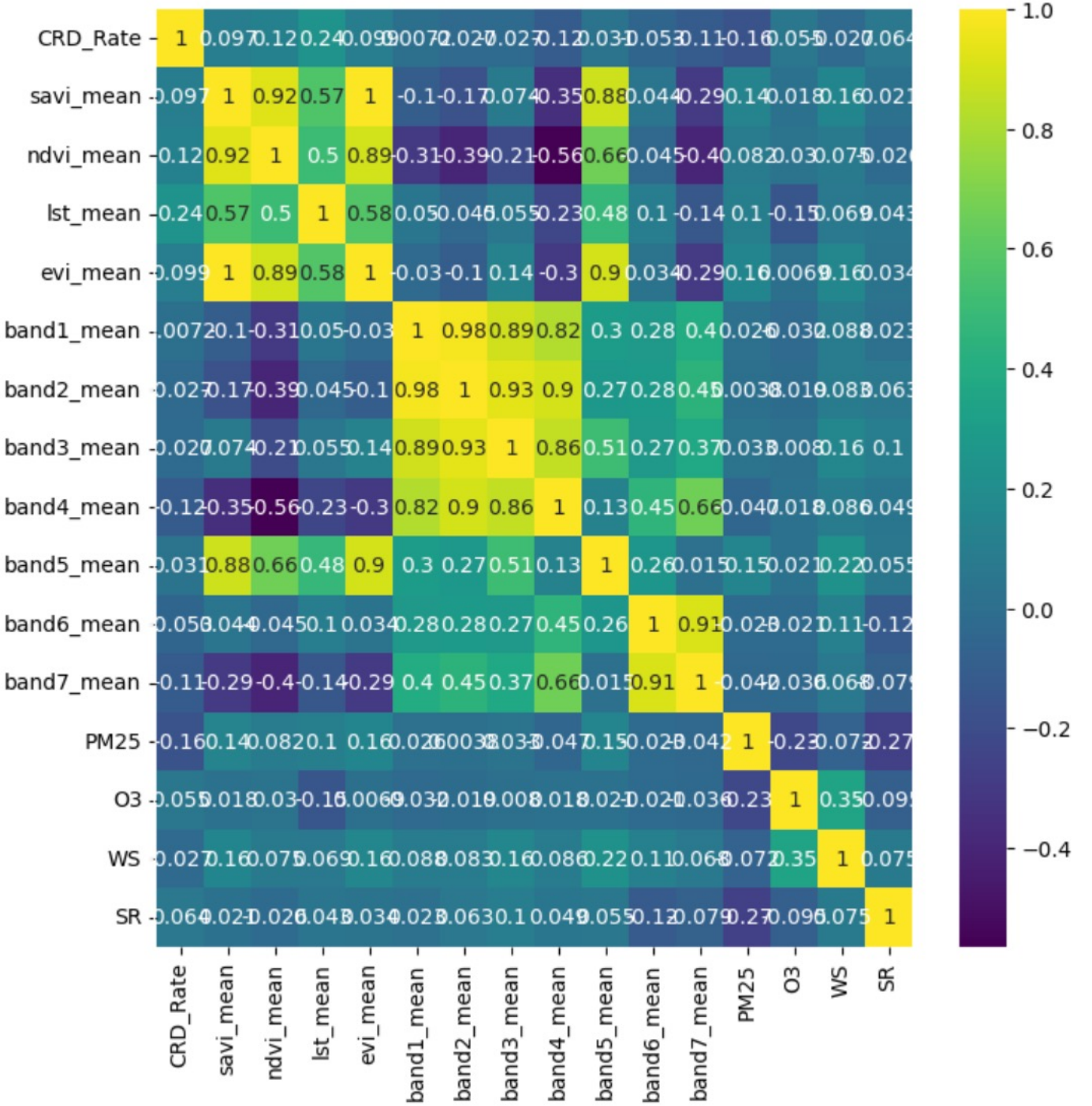

AQMN Stations Field Measurements
Air pollution variables
Meterological variables
Compute voronoi polygons based on the location of the measuring station for each of the variable
Philadelphia Counties
Dissolve
Boundary of PA
Make Fishnet
A Fishnet of PA
Remote Sensing Data (Landsat-8 Level 2)
Surface Reflectance Band
Normalized Difference Vegetation Index
Soil Adjusted Vegetation Index
Brightness Temperature
Enhanced Vegetation Index
Applies scale and offset adjustments to Landsat satellite imagery bands
Convert brightness temperature to land surface temperature
Cip to the boundary of Pennsylvania
Pixel selection without cloud
Compute the mean image
Repeat 4 times, one for each season's data
Zonal Statistics
Plain join or area weighted reaggregation
5. Dimensionality Reduction
Fishnet with remote sensing, air quality and CRD measures
Select predictor variables using Bayesian Information Criterion
Multiple Linear Regression
Multiple Layer Perceptron
Support Vector Regression
Random Forest Regression
Model Evaluations
Hospital Discharge Age Adjusted
Hospital Discharge
Filter discharge patients with CRD
Compute number of patients per 10,000 people who are admitted to the hospital with a CRD.

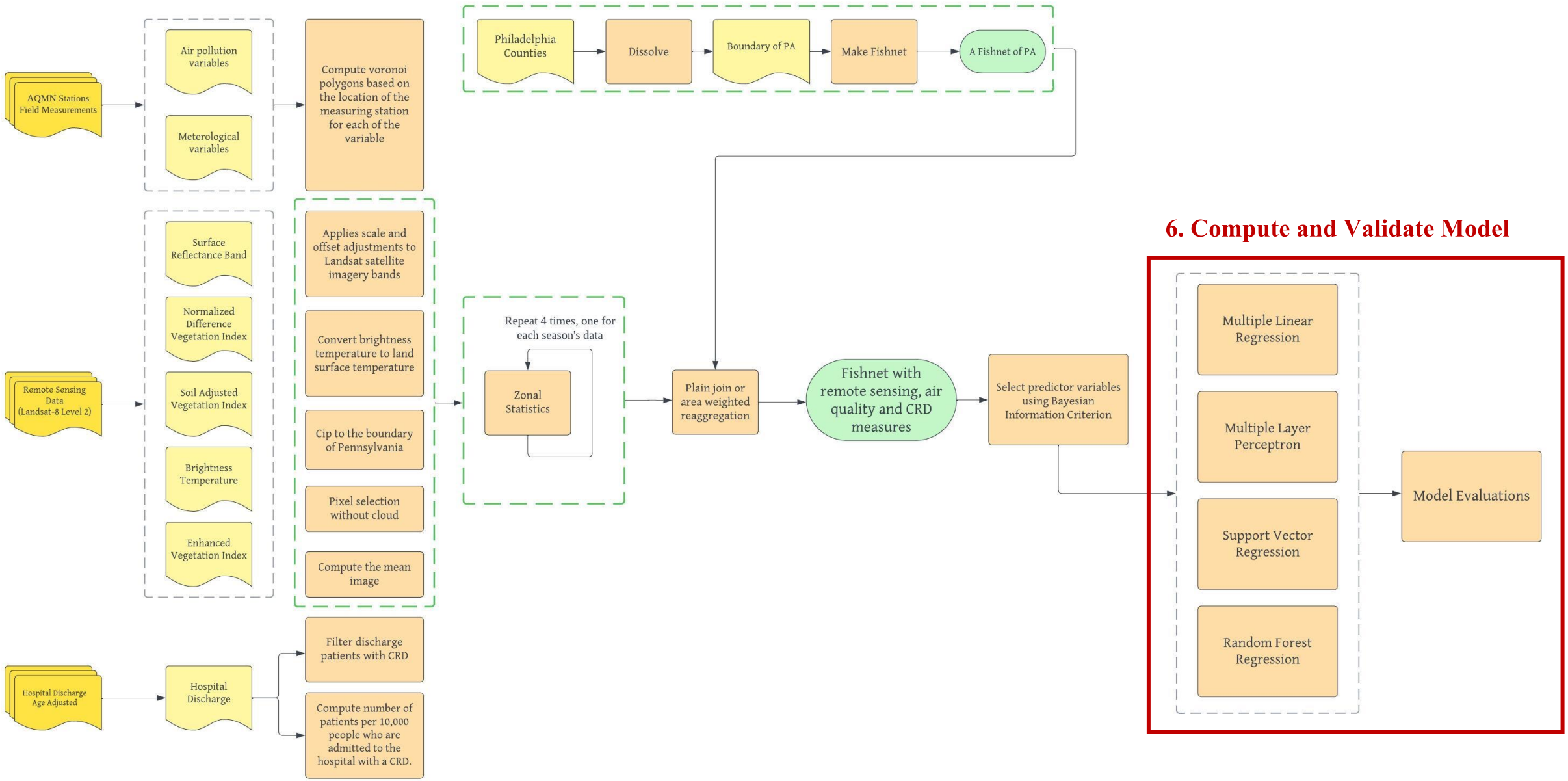
Dimensionality Reduction



We choose the Bayesian information criterion (BIC) was considered to conduct backward elimination, by which the lowest BIC values were used to choose the predictors.

Eventually, we selected the SAVI, NDVI, Land Surface Temperature, EVI, band2, band3, band5, band6, PM2.5, O3, and wind speed as our predictor variables.




AQMN Stations Field Measurements
Air pollution variables
Meterological variables
Compute voronoi polygons based on the location of the measuring station for each of the variable
Philadelphia Counties
Dissolve
Boundary of PA
Make Fishnet
A Fishnet of PA
Remote Sensing Data (Landsat-8 Level 2)
Surface Reflectance Band
Normalized Difference Vegetation Index
Soil Adjusted Vegetation Index
Brightness Temperature
Enhanced Vegetation Index
Applies scale and offset adjustments to Landsat satellite imagery bands
Convert brightness temperature to land surface temperature
Cip to the boundary of Pennsylvania
Pixel selection without cloud
Compute the mean image
Repeat 4 times, one for each season's data
Zonal Statistics
Plain join or area weighted reaggregation
Fishnet with remote sensing, air quality and CRD measures
Select predictor variables using Bayesian Information Criterion
Hospital Discharge Age Adjusted
Hospital Discharge
Filter discharge patients with CRD
Compute number of patients per 10,000 people who are admitted to the hospital with a CRD.
6. Compute and Validate Model
Multiple Linear Regression
Multiple Layer Perceptron
Support Vector Regression
Random Forest Regression
Model Evaluations

Machine Learning Models

80% Training

20% Testing

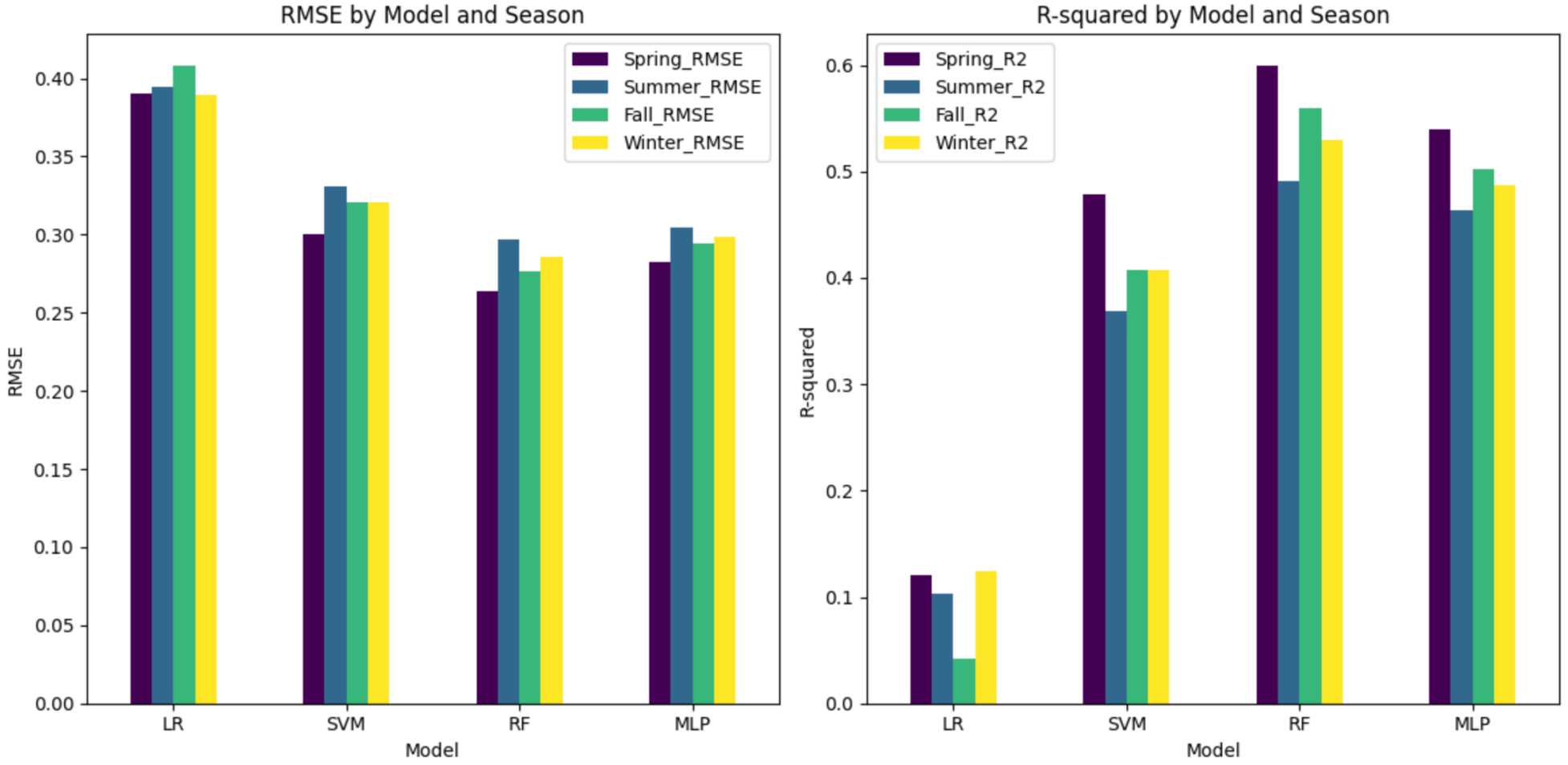
- Multiple Linear Regression
- Multiple Layer Perceptron
- Support Vector Regression
- Random Forest Regression

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 128)               1408

 dropout (Dropout)           (None, 128)               0

 dense_1 (Dense)             (None, 64)                8256

 dropout_1 (Dropout)         (None, 64)                0

 dense_2 (Dense)             (None, 1)                 65

=================================================================
Total params: 9729 (38.00 KB)
Trainable params: 9729 (38.00 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Special notes on the MLP model:

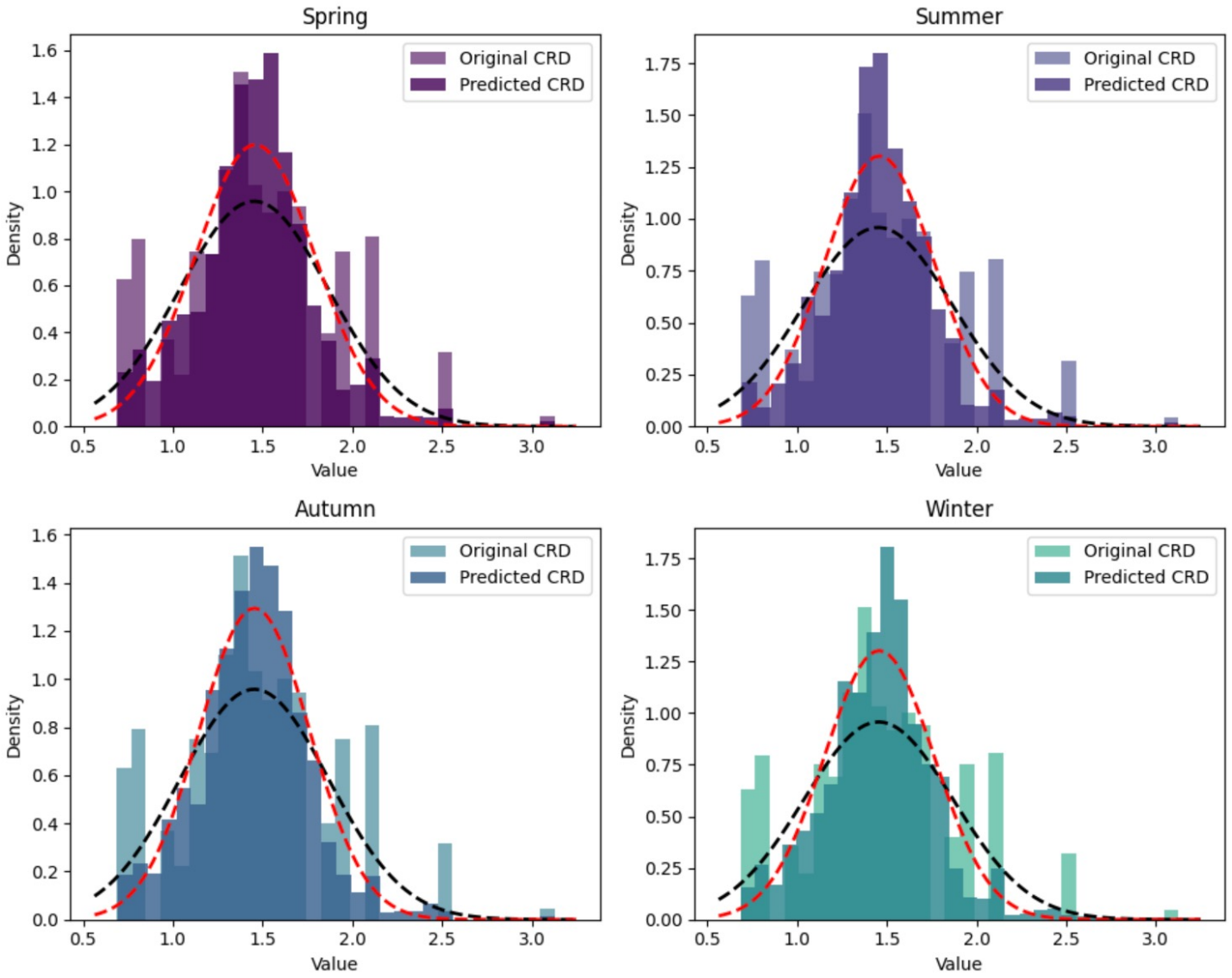
Our architecture consists of two dense layers with ReLU activation functions followed by dropout layers to regularize the network and prevent overfitting.

Results



The non-linear random forest model showed the best adjustment in the test data. The model presents the lowest **RMSE (0.263661 for spring, 0.296944 for summer, 0.276399 for fall, and 0.285896 for winter)** and the highest **R-Squared (0.598947 for spring, 0.491303 for summer, 0.559975 for fall, and 0.529131 for winter)** across all seasons. In particular, the RMSE is lowest for spring but highest for summer. The overall accuracy of random forest models is followed by multiple layer perceptron, support vector regression, and finally the linear regression model.

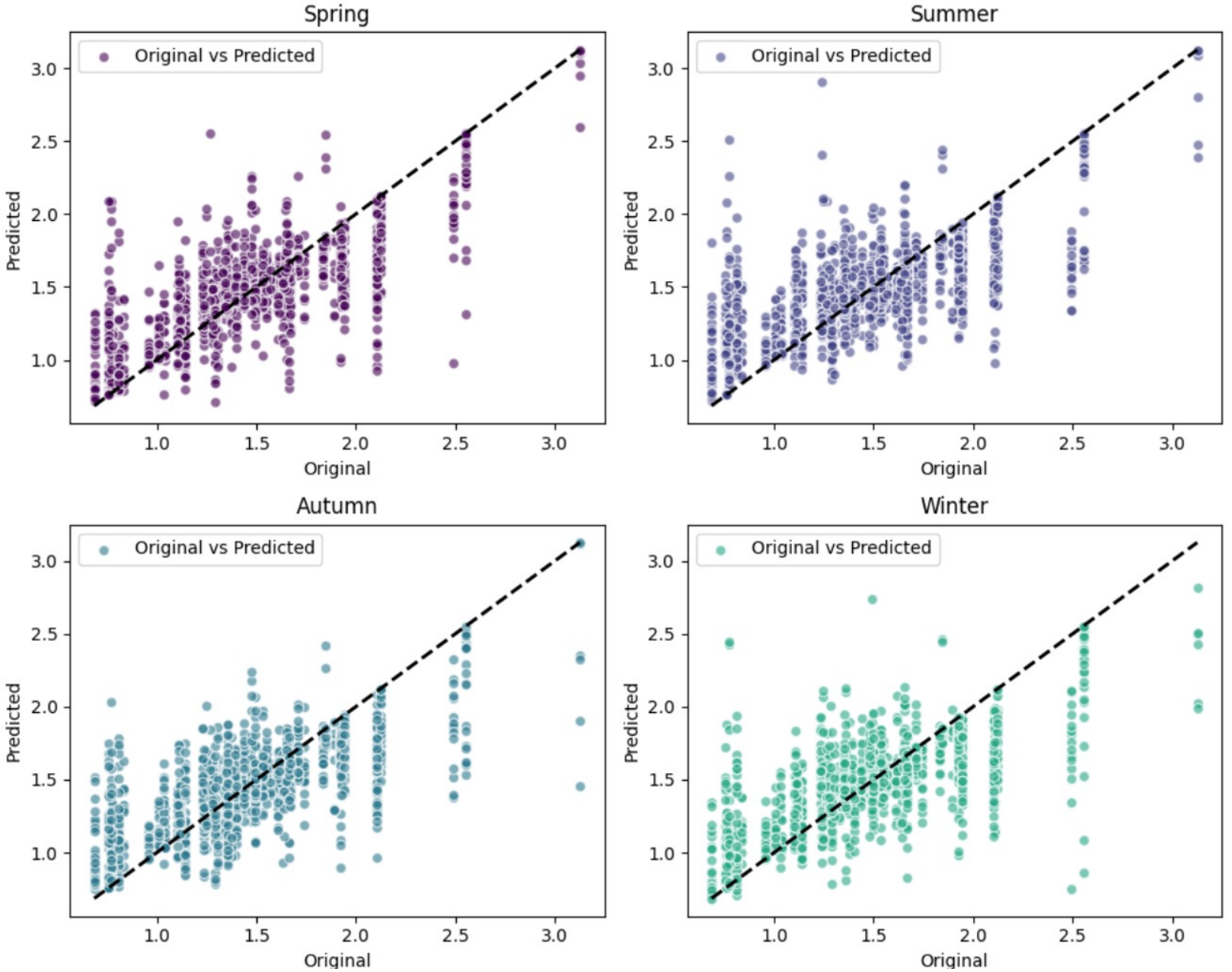
Results



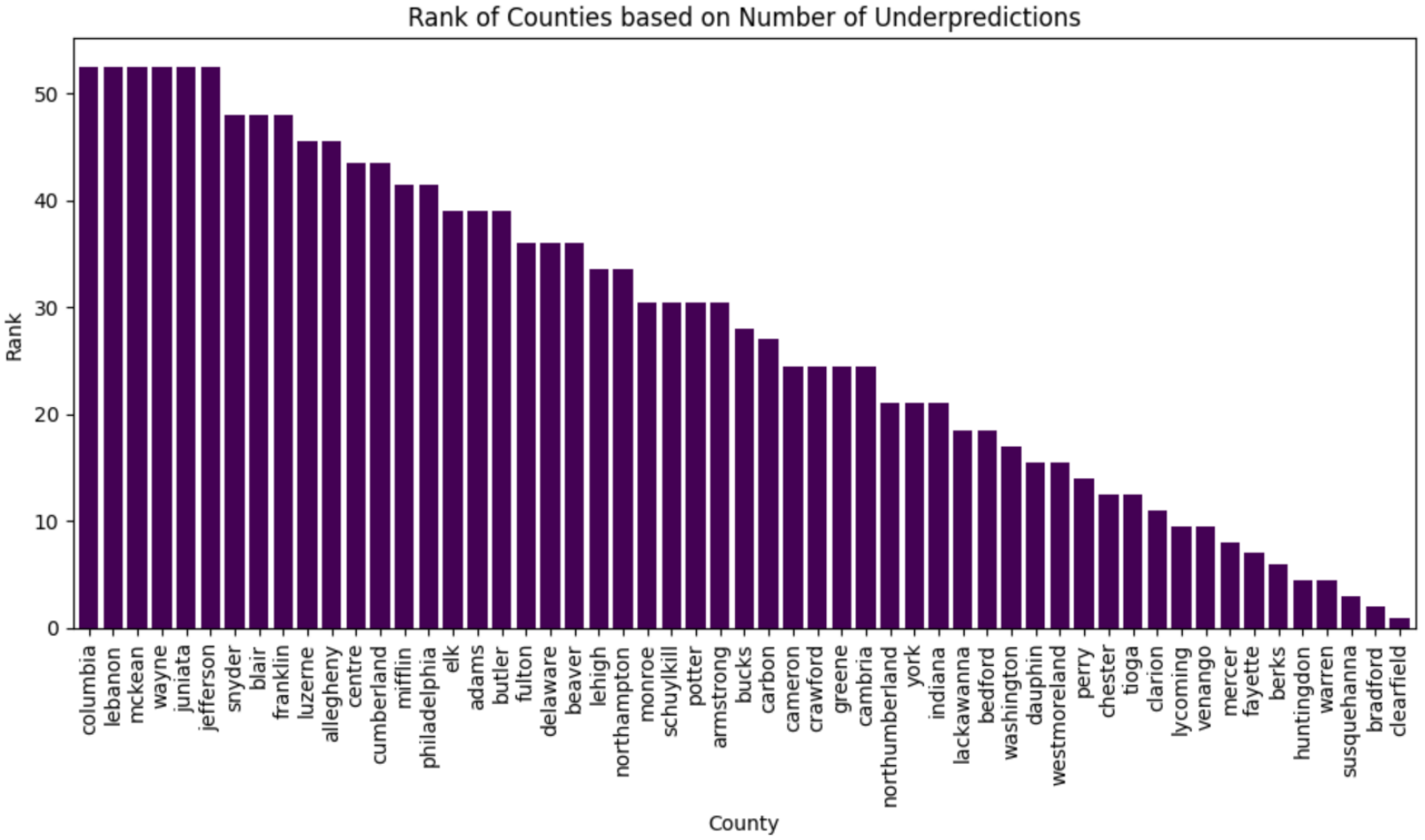
Comparing the distribution of the predicted and actual CRD hospital discharge values produced by the random forest model, we are able to learn the extent to which our model has over or under predict hospital discharge. For all four seasons, we can see that our model performs pretty well according to the normal distributions and indeed, the distribution curve of the spring model best matches the original curve.

Results

For each season, there are a few cases where CRD discharge has been significantly over/under predicted. The difference between each season is nuanced according to the visualization, but it seems like our models tend to underpredict CRD rates that were originally high.



Results



We selected those underpredicted values from the spring random forest model to take a closer look the counties at which it falls in. We would like to focus on underpredicted values because we want to be informed about the potential risks of CRD based on the local environment. We found that the following counties, especially **Columbia, Lebanon, McKean, Wayne, Junita, and Jefferson,** seems to contain a significant number of grids that have higher CRD rates but are underpredicted.

Limitations

Missing data is a concern in this project. There's one county in Pennsylvania with missing hospital discharge data. On top of that, different stations could be collecting completely different air quality measures. We also found that Philadelphia County's air quality data is maintained separately than other counties and are stored in a different format. This leads us to make a lot of compromises assumptions based on the available data and previous years' record in order to determine an approximate value for the month we need.

A1 | StationName

	A	B	C	D	E	F	J	K	L	M	N
1	StationName	CountyName	PM25-Spring-Avg	O3-Spring-Avg	WindSpeed-Spring-Avg	Solar-Radiation-Spring-Avg	Latitude	Longitude	NO2(1)	NO2(2)	NO2(3)
2	Allentown	Lehigh	7.27	31.10	5.93	178.4333333	40.611944	-75.4325			
3	Altoona	Blair	7.07	28.93	2.93	154.9666667	40.535278	-78.370833			
4	Arendtsville	Adams	7.57	38.20	9.10	162.6333333	39.9231	-77.3078	3.4	2.9	2.5
5	Beaver Falls	Beaver	8.50	28.50	3.00	155.9666667	40.747796	-80.316442	9.7	7.6	6.5
6	Brighton Township	Beaver		38.20	4.27	166.3666667	40.684722	-80.359722			
7	Bristol	Bucks		32.93	5.70	184.1	40.107222	-74.882222			
8	Carlisle	Cumberland	7.63		5.67	165.3666667	40.246528	-77.18675			
9	Charleroi	Washington	9.67		4.40	168.9	40.146667	-79.902222	8.8	7.2	6.5
10	Chester	Delaware	9.17	33.03	5.83	171.8	39.835556	-75.3725	10.5	8.5	7.6
11	Erie	Erie	6.90	31.93	6.03	166	42.14175	-80.038611	6.8	4.6	4.6
12	Farrell	Mercer	6.00	36.40	6.47	165.4333333	41.215014	-80.484779			
13	Florence	Washington	6.20	32.53	5.13	173.2666667	40.445278	-80.420833			
14	Fort McIntosh	Beaver	6.80	22.40	3.60	211.4	40.691857	-80.299211			
15	Freemansburg	Northampton	7.03	32.40	5.27	174.8	40.628056	-75.341111	9.6	6.4	5.8
16	Greensburg	Westmoreland	7.27	27.17	4.87	170.6	40.304694	-79.505667			
17	Harrisburg	Dauphin	8.30	30.07	4.83	168.9333333	40.246992	-76.846988			
18	Hershey	Dauphin		30.83	5.33	182.5666667	40.272222	-76.681389			
19	Holbrook	Greene	5.77	40.60	6.33	176.3666667	39.80933	-80.26567			
20	Houston	Washington	7.37	31.23	2.57	175.5333333	40.268963	-80.243995	3.8	3.4	2.9
21	Johnstown	Cambria	9.00	31.33	2.93	139.1333333	40.309722	-78.915	7.8	6	3.7
22	Kittanning	Armstrong	7.83	32.97	4.63	165.9666667	40.814183	-79.56475			
23	Kutztown	Berks		31.27	4.30	168.4666667	40.51408	-75.789721			
24	Lancaster	Lancaster	8.33	30.13	6.33	177	40.046667	-76.283333			
25	Lebanon	Lebanon	7.87	29.70	5.60	176.0666667	40.338516	-76.394788			
26	Marcus Hook	Delaware	7.80				39.818715	-75.413973			
27	Methodist Hill	Franklin		39.57	8.27	144.5666667	39.961111	-77.475556			
28	Montoursville	Lycoming		30.23	4.97	170.6333333	41.2508	-76.9238			
29	New Garden - Airport	Chester	7.10	32.57	7.00	181.5333333	39.834461	-75.768242			
30	Norristown	Montgomery	6.90	34.60	2.33	175.1	40.112222	-75.309167			
31	Reading Airport	Berks	7.00	35.73	6.63	180.5	40.38335	-75.9686			
32	Salladasburg	Lycoming	6.27		3.33	169.8666667	41.266263	-77.231189			
33	Scranton	Lackawanna	6.60	27.67	3.77	160.4	41.442778	-75.623056	11.2	8.2	8.2
34	State College	Centre	7.03	33.33	4.83	193.7666667	40.811389	-77.877028	4.5	3.6	2.7
35	Strongstown	Indiana	6.70	38.87	5.40		40.56333	-78.919972			
36	Swiftwater	Monroe		35.53	4.30	179.1666667	41.08306	-75.32328			

Limitations and Improvement

Using a standardize spatial unit makes the modeling process easier, but we have to do several rounds of area-weighted reaggregation to aggregate data at the county or measuring station influence area level into the fishnet.

Voronoi polygons is good at making spatial interpolations such that each station will be associated with a polygon representing the area where it has the closet proximity, it does not take into account local geographies

Improve upon the original paper by extending the temporal scale of analysis to include multiple years of data, rather than merely accounting for seasonal variations (especially if the hospital data are not at seasonal level).