

## Proposal of Implementation for Satellite Data + Ground Measurement Fusion – Part 1

### Objective

One major problem in our remote sensing applications is that the ground sensor measurements are much sparser than the satellite data. Hence, the main objective of this work is to find a proper way to combine the satellite data with the ground sensor measurements and hopefully raise the predictive performance of target variables such as ground-level  $PM_{2.5}$  values.

### Current Solution

In the current setting, we define areas of interest (AOI) around the ground sensors and only match the satellite images within that AOI to corresponding ground measurements (see details in Section 2.3 of [1] and Section 2.1-2.2 of [2], Section 2.1 of [2] also elaborates on the reason why we use daily averaged measurements as our target). As a result, all unlabeled images outside the AOI are discarded. [3] proposes to run contrastive learning on these unlabeled images to enhance the supervised learning performance, but the performance lift is only obvious in terms of capturing the spatial correlation instead of the overall predictions.

### Proposed Solution

In this project, we propose to manually fill in the missing labels by *spatial interpolation* over a certain time period. To be more specific, say we have an unlabeled satellite image, and we know the spatial location (i.e. latitude and longitude) of its geometric center (marked as green dot) as shown in Figure 1 below. Also, we have two nearby ground sensors (marked as red dots) where we know the ground measurements of  $PM_{2.5}$  values on a certain day. Our goal is to estimate the  $PM_{2.5}$  values at the location where the unlabeled satellite image is located based on the known ground measurements. There are several ways to do this estimation including:

- Kriging (simple kriging, ordinary kriging, regression kriging, etc.)
- Inverse distance weighted interpolation
- Nearest neighbor
- Density estimation

Here we should focus on kriging and inverse distance weighting.

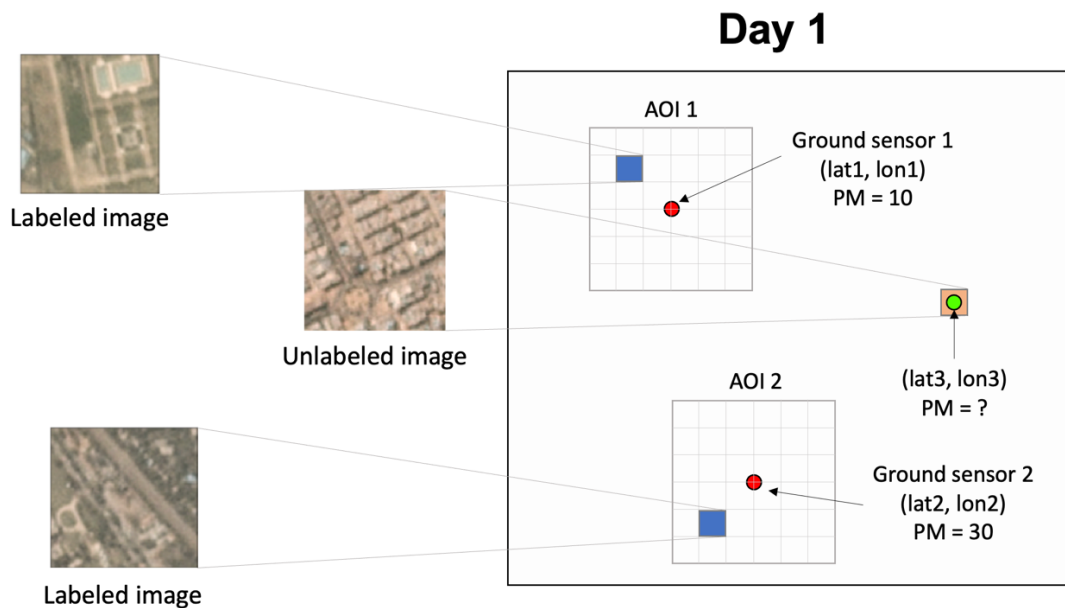



Figure 1

### Steps of Implementation

- Collect PM values from different ground sensors in one city over a certain time period. For example, Figure 2 below shows the PM values collected from Alipur air quality monitoring (AQM) station in Delhi starting from Nov 1, 2018.



CENTRAL POLLUTION CONTROL BOARD

CONTINUOUS AMBIENT AIR QUALITY

Date: Thursday, Jul 23 2020

Time: 03:41:30 AM

City	Delhi				
Station	Alipur, Delhi - DPCC				
Parameter	PM10,PM2.5				
AvgPeriod	24 Hours				
From	01-11-2018T00:00:00Z 00:00				
To	01-07-2020T18:09:59Z 00:00				
Alipur, Delhi - DPCC					
Prescribed Standards		0-100	0-60		
Exceeding Standards		NA	NA		
Remarks					
From Date	To Date	PM10	PM2.5		
01-11-2018 00:00	02-11-2018 00:00	None	None		
02-11-2018 00:00	03-11-2018 00:00	None	None		
03-11-2018 00:00	04-11-2018 00:00	None	None		
04-11-2018 00:00	05-11-2018 00:00	None	None		
05-11-2018 00:00	06-11-2018 00:00	None	None		
06-11-2018 00:00	07-11-2018 00:00	None	None		
07-11-2018 00:00	08-11-2018 00:00	None	None		
08-11-2018 00:00	09-11-2018 00:00	None	None		
09-11-2018 00:00	10-11-2018 00:00	None	None		
10-11-2018 00:00	11-11-2018 00:00	None	None		
11-11-2018 00:00	12-11-2018 00:00		431.63	247.97	
12-11-2018 00:00	13-11-2018 00:00		343.36	236.08	
13-11-2018 00:00	14-11-2018 00:00		428.75	284.73	
14-11-2018 00:00	15-11-2018 00:00		205.45	144.52	
15-11-2018 00:00	16-11-2018 00:00		163.7	114.01	
16-11-2018 00:00	17-11-2018 00:00		204.73	140.85	
17-11-2018 00:00	18-11-2018 00:00		278.21	170.78	
18-11-2018 00:00	19-11-2018 00:00		348.95	218.71	
19-11-2018 00:00	20-11-2018 00:00		351.96	533.12	

Figure 2

The Beijing PM data can be downloaded from: <https://data.mendeley.com/datasets/n3ywbm3y2t/4> and the corresponding latitudes longitudes of ground sensors (AQM stations) are listed in Table 1 in [2].

- Download mosaiced satellite images in the same city using PlanetScope API over the same time period. Rasterize the images into desired dimension (e.g.  $100\text{ m} \times 100\text{ m}$ ) and discard the images outside the boundary of the city.
- Categorize the satellite images into labeled (inside the AOIs of ground sensors) and unlabeled ones (outside the AOIs of ground sensors). Pair the images with PM values.
- Since there will be a much larger number of unlabeled images than labeled images, some filtering might be necessary.
  - For example, as shown in Figure 2 above, some PM values are missing from 11/01/2018 to 11/10/2018 at Alipur station. If the ground measurements on a certain day are too sparse, we can simply skip that day.
  - We can also set up some metrics to determine whether we want to estimate the PM values for an unlabeled image. For example, we can calculate the mean distance to all available ground sensors on a certain day and determine whether we want to accept/keep or reject/discard the image as illustrated in Figure 3.

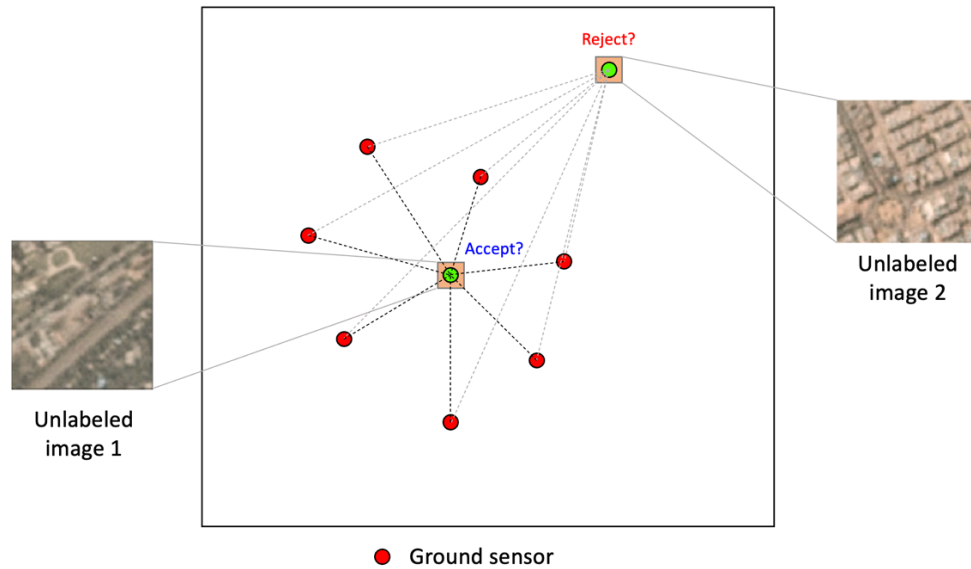


Figure 3

- Estimate the PM values for the filtered unlabeled satellite images by spatial interpolation based on the latitude and longitude of the center of the image. Pair the estimated PM values with unlabeled images.
- Finally, we can use these satellite images with estimated PM labels either for *pre-training* or *augmenting our labeled dataset* and see if we can enhance the supervised predictive performance of PM values. In fact, this step is quite *open-ended*. For example, we can also consider taking these estimated PM values as another model input, combining them with satellite data, and predicting whether the satellite data is labeled or unlabeled (just like DANN [5] in domain adaptation) as a pre-training step.

### ***Deliverables and Stretch Goals***

One thing we want to include in the final deliverables is a comparison of overall/spatial prediction errors/accuracies for the cases with and without including the estimated PM values by spatial interpolation. If the results look promising for one city, we can apply this pipeline to other cities on the earth and see if it generalizes well. Also, as we observed before that satellite images mainly contain information about the *spatial variation* of PM instead of the actual PM values, we can also consider including other PM-related information and try out the ICKy [4] framework.

### ***References***

- [1]. Zheng, Tongshu, et al. "Local PM2.5 Hotspot Detector at 300 m Resolution: A Random Forest–Convolutional Neural Network Joint Model Jointly Trained on Satellite Images and Meteorology." *Remote Sensing* 13.7 (2021): 1356.
- [2]. Zheng, Tongshu, et al. "Estimating ground-level PM2.5 using micro-satellite images by a convolutional neural network and random forest approach." *Atmospheric Environment* 230 (2020): 117451.
- [3]. Jiang, Ziyang, et al. "Improving spatial variation of ground-level PM2.5 prediction with contrastive learning from satellite imagery." *Science of Remote Sensing* 5 (2022): 100052.
- [4]. Jiang, Ziyang, Tongshu Zheng, and David Carlson. "Incorporating Prior Knowledge into Neural Networks through an Implicit Composite Kernel." *arXiv preprint arXiv:2205.07384*(2022).
- [5]. Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." *The journal of machine learning research* 17.1 (2016): 2096-2030.