Note: You’ll need tslearn to run my code. I have added a requirements.txt file. You can get the required libraries using pip install -r requirements.txt.

1. The number of clusters and the cluster label for each sensor location (0.5pt)

Number of clusters - 4

{1128: 3, 132561: 0, 1650: 0, 1852: 0, 2504: 1, 2749: 2, 3487: 2, 5222: 0, 6164: 2, 6472: 2, 6836: 1, 9076: 0, 9108: 1, 9168: 0, 17583: 0, 23263: 0, 23295: 0, 23347: 0, 27109: 2, 27113: 0, 28457: 1, 28515: 2, 28973: 0, 29295: 0, 30737: 0, 32339: 0, 33235: 1, 33855: 0, 34517: 2, 34519: 2, 35027: 2, 37193: 1, 37217: 0, 38399: 0, 39155: 0, 40639: 0, 42251: 2, 42857: 0, 52071: 1, 53825: 0, 54977: 0, 54981: 1, 54985: 1, 54993: 0, 55251: 0, 55351: 2, 56073: 2, 56141: 2, 63387: 2, 69201: 2, 71813: 2, 72459: 0, 74069: 0, 75839: 0, 76465: 1, 76469: 2, 76473: 2, 76627: 0, 76867: 2, 77427: 2, 78587: 0, 80069: 2, 80657: 1, 80807: 2, 81219: 2, 81363: 2, 81851: 2, 82271: 0, 83041: 2, 83081: 2, 83647: 2, 83895: 2, 84069: 2, 84083: 2, 84489: 2, 85253: 2, 85359: 2, 85373: 2, 85829: 2, 85985: 2, 86185: 0, 86367: 2, 86951: 0, 87315: 0, 87323: 2, 87413: 0, 87415: 2, 88895: 2, 90005: 0, 94397: 0, 96487: 2, 96793: 1, 99139: 2, 102032: 2, 104892: 2, 104900: 0, 104940: 1, 108518: 0, 110602: 0, 6228: 1, 6244: 1, 6886: 1, 14859: 2, 22675: 0, 34495: 0, 36143: 2, 36717: 0, 37885: 0, 40449: 2, 40721: 1, 40725: 1, 51323: 0, 52277: 1, 53043: 0, 61565: 0, 70793: 0, 78637: 1, 81379: 0, 84101: 2, 87129: 0, 87477: 1, 87567: 2, 90819: 1, 96243: 0, 102064: 0, 107960: 0}

1. The total number of features in the geographic abstraction vector (0.5pt)

5854

1. The top 20 important features and their importance (1pt)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Geo Feature** | **Feature Type** | **Buffer Size** | **Importance \(%\)** |
| **0** | waterway | stream | 3000 | 0.00756072051263904 |
| **1** | highway | track | 2500 | 0.007112029006588840 |
| **2** | building | residential | 2900 | 0.006960398676567070 |
| **3** | highway | service | 2900 | 0.006844928821713640 |
| **4** | building | residential | 3000 | 0.006620001845966780 |
| **5** | building | warehouse | 2800 | 0.006512322532633790 |
| **6** | amenity | fast\_food | 3000 | 0.005817743169094100 |
| **7** | highway | traffic\_signals | 2400 | 0.005757050047354120 |
| **8** | highway | track | 2800 | 0.005705449143858510 |
| **9** | waterway | stream | 2800 | 0.005125496643271650 |
| **10** | building | residential | 400 | 0.0050996318638686800 |
| **11** | natural | scrub | 2800 | 0.005010624762139430 |
| **12** | highway | stop | 1800 | 0.004998259132458570 |
| **13** | tourism | motel | 2600 | 0.004942587129746120 |
| **14** | highway | crossing | 2000 | 0.00467896299269153 |
| **15** | highway | traffic\_signals | 2500 | 0.004572951456910940 |
| **16** | building | apartments | 2500 | 0.004396996442702900 |
| **17** | highway | stop | 2200 | 0.004340243880995160 |
| **18** | building | warehouse | 3000 | 0.004284636253924170 |
| **19** | building | industrial | 2800 | 0.004260775892842770 |

1. Overall MSE and R2 on test samples and the plot showing hourly MSE and R2 scores (1pt)

Hourly R2 Score:

Chart, line chart

Description automatically generated

Hourly MSE Score:

Chart, line chart

Description automatically generated

R2 Score for the entire dataset: 0.6164421802520904

Mean Squared Error Score for the entire dataset: 146.37517723504774

5. Four plots of fine-grained prediction results (screen shots) (1pt)

Plot of 6 am to 9 am:Map

Description automatically generated

Plot of 4 pm to 7pm:

Map

Description automatically generated

Plot of weekdays:

Map

Description automatically generated

Plot of weekends:

Map

Description automatically generated

6. Your findings and discussion on the selected features, MSE and R2 scores, and plots of finegrained predictions in several sentences (no more than 300 words) (1pt)

Selected features:

I have selected the top 60 features based on the importance scores. They intuitively make a lot of sense because highway occurs quite often in those features. 25 out of the 60 features have highway. So does building which occurs in 18 out of the 60 features. Waterway occurs in 6 of those features. Highway and building will negatively impact the pm2.5 concentration whereas waterway will have the opposite effect.

MSE and R2 scores:

The overall MSE and R2 scores are good and show that the model did learn something and was able to make decently accurate predictions. The R2 scores peak during the daytime. I think that can be attributed to lower variability in pm2.5 scores as opposed to the nighttime.

Plots of fine-grained predictions:

It is clearly visible that pm2.5 concentrations are shown to be higher along the roads especially when the roads are not surrounded by vegetation. This is beautifully illustrated in the La Puento area where the pollution concentration takes the shape of the curved road. The area around Malibu always has lower concentration of pollution primarily because of abundance of waterway and vegetation. Roads that directly lead to downtown always have higher concentration when compared to other roads. Weekdays have higher concetration of pm2.5 around the downtown area compared to weekends. The area near vegetation has really low levels of pm2.5.

Appendix:

Number of clusters vs Inertia

Chart, line chart

Description automatically generated

TimeSeries Clusters:

Chart, line chart, histogram

Description automatically generated

Chart, line chart, histogram

Description automatically generated