Benchmarking Low-cost Air Quality (PM2.5) Sensors-Examining Their Potential to Complement Existing Pollution-Measurement Frameworks in Pittsburgh

Abhishek Viswanathan, Vasco Xu
Instructor: Dr. Amy Babay
CS 3551: Advanced Topics in Distributed Information Systems
Spring 2020
Problem

- Not enough air quality data on local scales
- Uncertainty about accuracy of low-cost sensors
  - If the data is inaccurate, how can we adjust/calibrate its values?
- How to incorporate low-cost sensor data into environmental models
- How to get citizens to participate with easy-to-use devices
- How to automatically visualize citizen-science data
Related Work

PurpleAir

PurpleAir PA-II
$229.00
Related Work

SmellPGH
Approach

- Built prototype with inexpensive PMS5003 sensor to automatically post data to web server.
Approach

- Built web-server hosted on Heroku to receive and store sensor data.

<table>
<thead>
<tr>
<th>Device Name</th>
<th>Time</th>
<th>PM0.3</th>
<th>PM0.5</th>
<th>PM1.0</th>
<th>PM2.5</th>
<th>PM10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maple</td>
<td>20/04/2020 19:54:18</td>
<td>351</td>
<td>102</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Pine</td>
<td>20/04/2020 19:49:03</td>
<td>348</td>
<td>104</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Maple</td>
<td>20/04/2020 19:44:10</td>
<td>378</td>
<td>104</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Pine</td>
<td>20/04/2020 19:38:54</td>
<td>282</td>
<td>84</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Maple</td>
<td>20/04/2020 19:34:00</td>
<td>264</td>
<td>75</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pine</td>
<td>20/04/2020 19:28:53</td>
<td>288</td>
<td>76</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Maple</td>
<td>20/04/2020 19:23:59</td>
<td>357</td>
<td>95</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Pine</td>
<td>20/04/2020 19:18:44</td>
<td>288</td>
<td>90</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Maple</td>
<td>20/04/2020 19:13:50</td>
<td>267</td>
<td>83</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Pine</td>
<td>20/04/2020 19:08:42</td>
<td>330</td>
<td>98</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Approach

- Visualize data from other projects - SmellPGH, PurpleAir, Allegheny County Official Data
Approach

- ML + Data Analysis
Analysis
Analysis

- Random Forest Regression (random forests are run in parallel)
  - Attempts to avoid overfitting.
  - Useful for non-linear data.
Results/Analysis

- Left figure: PittAir sensor values (placed inside a house) vs. PurpleAir sensor values (presumably also inside).
- Right figure: Calibrated PittAir sensor with Random Forest Regression ML (of 77% accuracy).
- Total values used: 1132.
- 75% of values used for training and 25% used for testing.
Results/Analysis

- Tested the accuracy of the model by using values not used in building or testing the ML model.
- Left figure: PittAir sensor values vs. PurpleAir sensor values.
- Right figure: Calibrated PittAir sensor values with ML model from previous slide (141 values used).
Future Work

- Improve outdoor models by calibrating near existing sensors hourly and analyzing its accuracy.
- Making collected data available in different formats for analysis.
- Automatically visualize the data on a graph on the website.
- Posting analysis/data automatically to Slack via an API.
Questions?
Thank you!