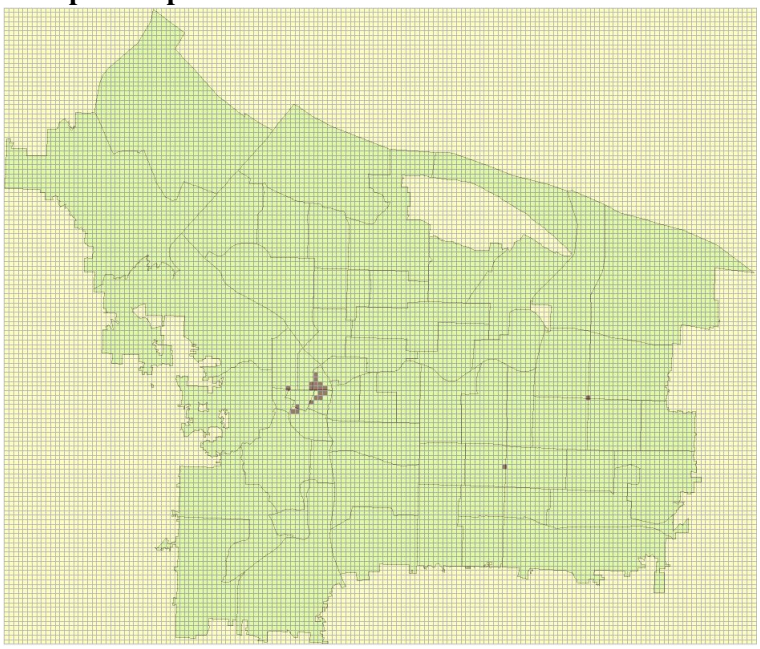
Real-Time Crime Forecasting – Phase I

UMass-Boston 3/2/2017

1 Educational Goal

Apply data mining techniques to real-world crime data.

- **Goal:** Choose best model parameters - cell size, cell shape and training time window - for crime prediction, based on historical Portland crime data and NIJ competition success measures.
- **Due Date:** 4:00 pm, March 21, 2017
- **Programming language:** python.
- **Hot spots map:**



- **Evaluation:** Use PAI and PEI* to evaluate your results, the definitions of PAI and PEI* are as follows:

  Prediction Accuracy Index (PAI): The PAI will measure the effectiveness of the forecasts with the following equation:

$$PAI = \frac{\frac{n}{N}}{\frac{a}{A}} \tag{1}$$

  where n equals the number of crimes that occur in the forecasted area, N equals the total number of crimes, a equals the forecasted area, and A equals the area of the entire study area.

  Prediction Efficiency Index* (PEI*): The PEI* will measure the efficiency of the forecasts with the following equation:

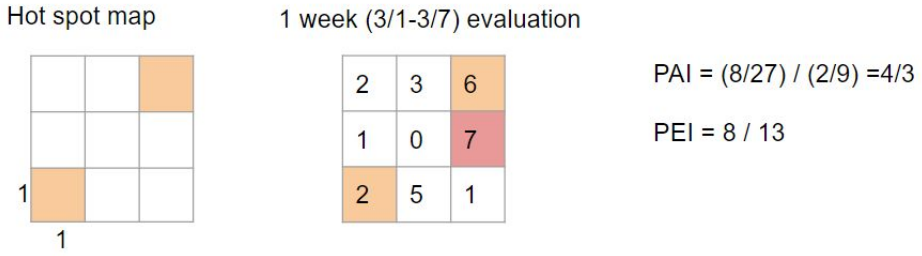
$$PEI^* = \frac{PAI}{PAI^*} \tag{2}$$

where PEI* equals the maximum obtainable PAI value for the amount of area forecasted, a. As such:

$$PEI^* = \frac{n}{n^*} \tag{3}$$

where $n*$ equals the maximum obtainable n^* for the amount of area forecasted, a.

- **A example of how to do evaluation:**



- **Dataset:** The dataset includes the calls-for-service (CFS) records provided by the Portland Police Bureau (PPB) for the period of March 1, 2012 through January 31, 2017.
  It contains four crime categories: all calls-for-service; burglary; street crime; and motor vehicle thief, by date and GPS coordinates.

2 Approach

2.1 Step 1. Training and test data generation:

Using a few different values for cell size and shape, create grids of cells that cover the city of Portland. Transfer the CFS data (select a crime category) to the grids month-by-month, such that each month has its own copy of the grid and each cell within that grid contains the number of CFS records that occurred in that cell over that month. For example, you can use the burglary records in January 2016 to build a grid with $600ftx600ft$ cells, where each cell has the total number of burglary calls that occurred in the cell in January. Code examples will be available for converting GPS coordinates to grid cells.

Requirement: Create monthly cell grids for each month of a single calendar year. Populate those grids with call-for-service data from a chosen crime category.

1. Individual cell area: $62,500ft^2 - 360,000ft^2$
2. Minimum cell height: 125 ft
3. Cell shape: square or rectangle.

2.2 Step 2. Calculate PAI and PEI for a set of predicted hot-spot cells based on a single month of data (1month-based prediction):

For a given month, choose a number of hot spots, n, and find the n cells with the highest concentrations of crimes and use those cells as a prediction of the next month's hot spots, using PAI and PEI as an evaluation method. For example, if you choose $n = 20$ and January 2016 data for training, you should pick the top 20 cells in your January grid as hotspots and compare those 20 hotspots to the hotspots in the grid data for February 2016.

Requirement:

1. Total hotspots area: $0.25mi^2 - 0.75mi^2$.
2. Draw a plot (or multiple plots) to show all the results with different parameters (cell size, cell shape, category, number of hotspots).

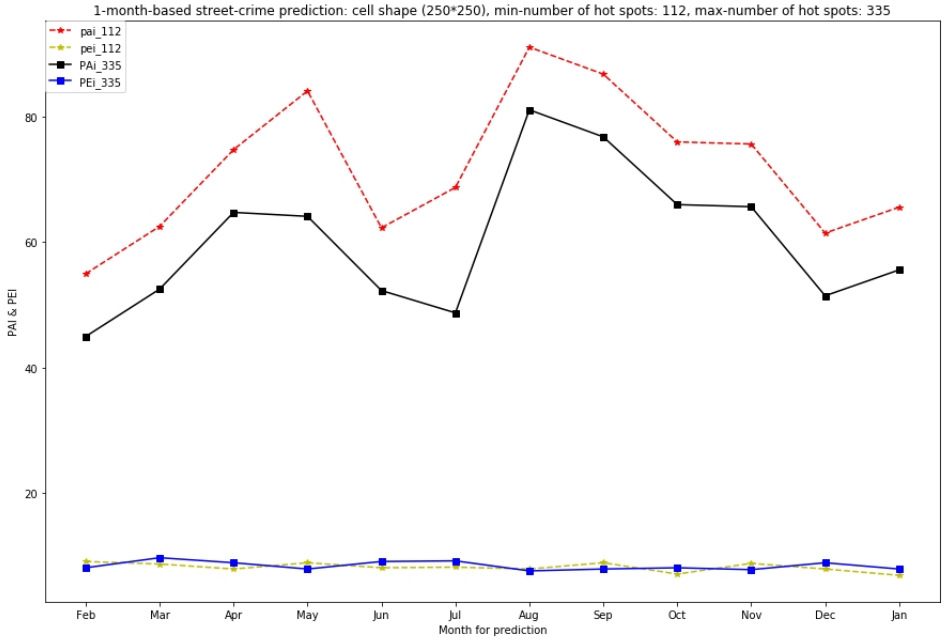
2.3 Step 3. Calculate PAI and PEI of multiple-month-based prediction:

Choose a number of months m, then using the best parameters you tried in part 2, perform a prediction based on an m-month time period, then evaluate that prediction against the month following the last month of the training window using PAI and PEI.

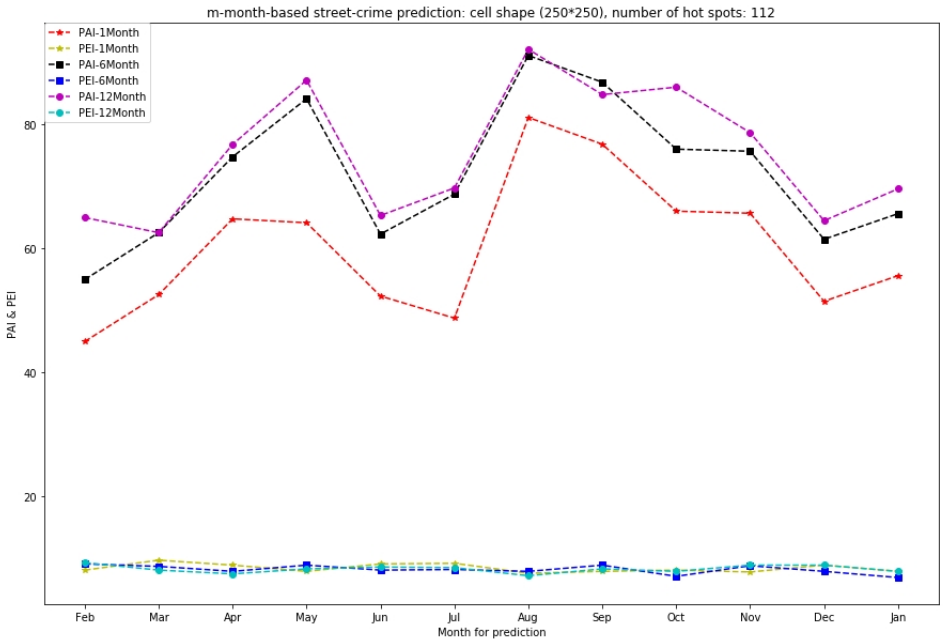
Requirement: Draw a plot to show the results with different values of m.

3 Team Task

1. Do 1month-based prediction for all months in 2016 (e.g. from "using 2016 JAN data to predict 2016 FEB data" to "using 2016 DEC data to predict 2017 JAN data").
2. Try at least 2 values for number of hot spots: minimal number of hot spots, maximal number of hot spots (the minimal and maximal number of hot spots are based on the individual cell area and the total hot spots area).
3. Draw plots, each plot corresponds a cell shape, the plot format is as follows:



4. Using the best parameters obtained from 1 month-based prediction to do multi-month-based prediction. Try at least 3 values for number of months m: m=1, m=6, m=12. Then draw a plot. The plot format is as follows:



5. Write a summary to discuss which parameters are the best for prediction, and what are the best PAI and PEI score you got.

3.1 Teams

1. **Team 1 :** Nam Nguyen, Zhi Cao
2. **Team 4 :** Nirmal Nepal, Nirajan Nepal, Kushal Khanal, Liru Hu
3. **Team 5 :** Han Nam Le, Aayush Choudhary
4. **Team 6 :** Saurav Paudyal, Justin Yang, John Allen
5. **Team 8 :** Luoyan Zhang, Hefei Qiu, Akshay Joshi
6. **Team 10:** Alex Millward, Tim Yip, Luke Chen, Chris Lo
7. **Team 11:** Shangyu Xie
8. **Team 14:** Abdul Fawzan

3.2 Team Tasks Table

Cell Shape	Burglary	Motor Thief	Street Crime	All CFS
250 * 250	Team 5	Team 4	Team 8	Team 1
600 * 600	Team 11	Team 14	Team 6	Team 10
125 * 2880	Team 5	Team 4	Team 8	Team 1
125 * 500	Team 11	Team 14	Team 6	Team 10

4 Server and Sample Code

4.1 python server:

create your own folder under the workspace folder. server link:

http://nanedi.cs.umb.edu:443/tree?token=e67697748ca0e12dbcfb53e85746c6954ab23d99755d4e93

4.2 sample code:

The "Sample-Code.ipynb" is under my folder at the python server. Do not change this file, you should modify your own copy in your folder.

5 Submission Requirements

1. Write an experiment report to discuss your experimental results, including detailed parameter settings, plots in step 2 and step 3, and the best parameter settings. Submit the paper copy of the report with the cover page in class. Only soft copy is required.
2. Submit the soft copy of the report through your UMassOnline account.

6 In Class Presentation

Each team should give a 6-minute presentation with any number of slides including all plots, the settings you used for each plot (cell shape, cell size, map size, number of hot spots), the corresponding PAI and PEI scores, and the conclusions you drew from your experiments (e.g. Which settings can help you get the best PAI and PEI? Were the best settings of PAI and PEI are the same, or different?). The date of presentation is as follows:

1. **Team 1 :** Thursday, Match 23th, 2017
2. **Team 4 :** Tuesday, Match 21th, 2017
3. **Team 5 :** Thursday, Match 23th, 2017
4. **Team 6 :** Thursday, Match 23th, 2017
5. **Team 8 :** Thursday, Match 23th, 2017
6. **Team 10:** Thursday, Match 23th, 2017
7. **Team 11:** Thursday, Match 23th, 2017
8. **Team 14:** Thursday, Match 23th, 2017