Pre-Pilot Evaluation Guideline

Dihong Gong
Dept. of CISE

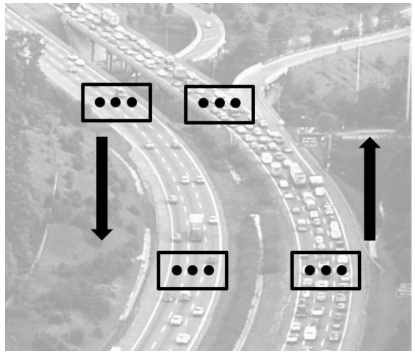
Overview

1. **Cleaning Task**
    a. clean traffic lane detector measurements containing incorrect flow values, providing correct traffic flow values for the erroneous traffic flow measurements.
2. Alignment Task
    a. analyze video from camera feeds to detect an event and match it to a separate inventory of traffic events (disabled car, accidents, etc).
3. **Prediction Task**
    a. develop a system that can predict the number and types of traffic events by type for a given (geographical bounding, interval of time) pair.
4. Forecasting Task
    a. leverage past traffic information and current conditions (weather, maps) to forecast vehicle flows on major roads.

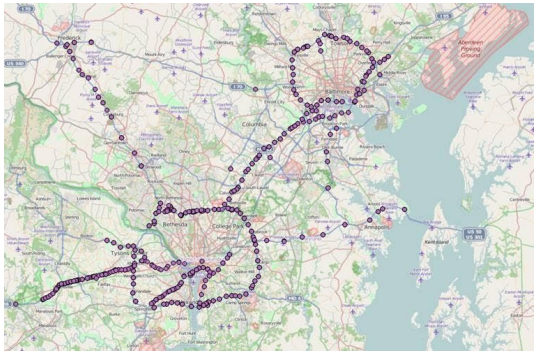
* You only need to complete the task #1 and #3.

Data

1. Lane_measurements
    a. detector_lane_inventory.csv
        i. lane_id: uniquely identify a detector (totally 2,135).
        ii. zone_id: identifier of a zone in a road.
        iii. road: on which road, e.g. I-66.
        iv. location_description: e.g. I-66 NEAR Sudley Rd @ MM 49.02
        v. Geographical coordinate: (latitude, longitude)
        vi. There are 11 other less important fields.



lane and zone illustration (courtesy by NIST)



Detector distribution (courtesy by Sreten Cvetojevic)

Data

1. Lane_measurements
    b. test
        i. lane_id: identifier of a detector that this record is collected from.
        ii. measurement_start: timestamp when measurement starts, e.g. 2007-04-09 14:04:12-04
        iii. speed: measured average speed (mph) of the last interval, e.g. 70.
        iv. flow: number of vehicles passed through the lane detector in the last interval, e.g. 9.
        v. occupancy: the average percent of time a vehicle was in front of detector in the last interval, e.g. 2.
        vi. There are totally 108 csv files, with file size ranging from 100 MB to 2GB, totally hundreds of gigabytes.

Data

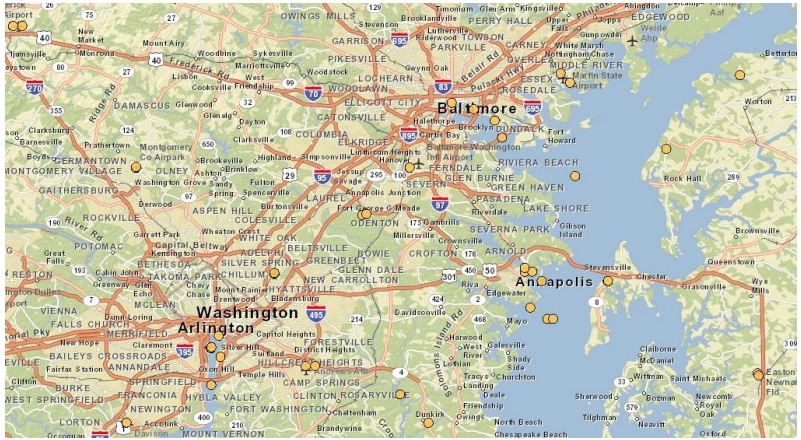
2. traffic_events
    a. events_train.csv
        i. event_id: uniquely identify an event, e.g. “MDOT_CHART_4aff02b300110095003f0be8b3035daa”
        ii. event_description: a text description about an event, e.g. “Disabled Vehicle Event @ I-495 AT MD 187”
        iii. Timestamps: times the event was created, confirmed, and closed (some are missing).
        iv. event_type: the type of an event, e.g. “accidentsAndIncidents”.
        v. geographical location: (latitude, longitude)
        vi. There are 9 other less important fields.

Data

3. OpenStreetMap (OSM)
    a. Description: map data from from OpenStreetMap, describing the road network in the DC-MD-VA area as well as locations including airports, public transportation stations, and buildings that host large events. These maps also support lookup by latitude and longitude coordinates.
    b. More information: https://wiki.openstreetmap.org/wiki/Main_Page

Data

4. Integrated Surface (ISD) Weather Data
    a. A dataset of measurements from weather stations in the DC-MD-VA area with a variable number of measurements. Measurements include station information, temperature, air pressure, weather condition, **precipitation**, and other elements
    b. More information: http://www.ncdc.noaa.gov/data-access



Weather Station Distribution

Cleaning Task

Description: participants are asked to clean traffic lane detector measurements containing incorrect flow values, providing correct traffic flow values for the erroneous traffic flow measurements.

Input: lane_measurements (**nist-prepilot-ufl/core/lane_measurements/test/*.csv**)

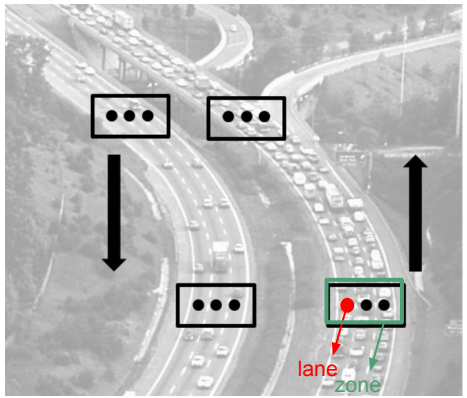
Output: cleaned lane_measurements (with flow values corrected)

Training/Development data: detector_lane_inventory.csv and OpenStreetMap, in addition to input lane_measurements (*Note: no other data is allowed to be used for this task).

Cleaning Task

Hints

1. Extraneous flow values usually violate some constraints*.
   a. In a period of time, flows in the same zone of different lanes should have similar values.
   b. In a period of time, flows of nearby zones should be similar.
   c. Flow values must be nonnegative numbers.
   d. Measurements of (flow, speed, occupancy) should be consistent.
   e. Your own constraints.
2. How to correct extraneous flow values.
   a. Correct the values as values of nearby location or time.
   b. Correct the values as output of a regression model F(location,time)$\rightarrow$flows**. The F is learned from the dirty data.



* Some flow values may be correct even they violate constraints, when unusual event such as accidents occur. How you tackle this problem?

** The model F should have very low VC dimension for better smoothness.

Cleaning Task

Hints

3. Dive into the data and come up with your own constraints/models.

*Note: we highly encourage you to think creatively, and come up with extra improvements.

Prediction Task

Description: participants will develop a system that predicts the number and types of traffic events by type for a given (geographical bounding, interval of time) pair.

*Note: 1. The interval of time is around 30 days.

2. All possible roads within given bounding box should be counted, not just limited to roads in the given training data.

Input: geographical bounding boxes and time intervals.

Output: predicted counts for each specified type of traffic event.

Training/Development data: Lane_measurements, traffic_events, OpenStreetMap, and Integrated Surface (ISD) Weather Data.

Prediction Task

List of events* to predict

- ❏ Accidents and Incidents
- ❏ Roadwork
- ❏ **Precipitation**
- ❏ Device Status
- ❏ Obstruction
- ❏ Traffic Conditions

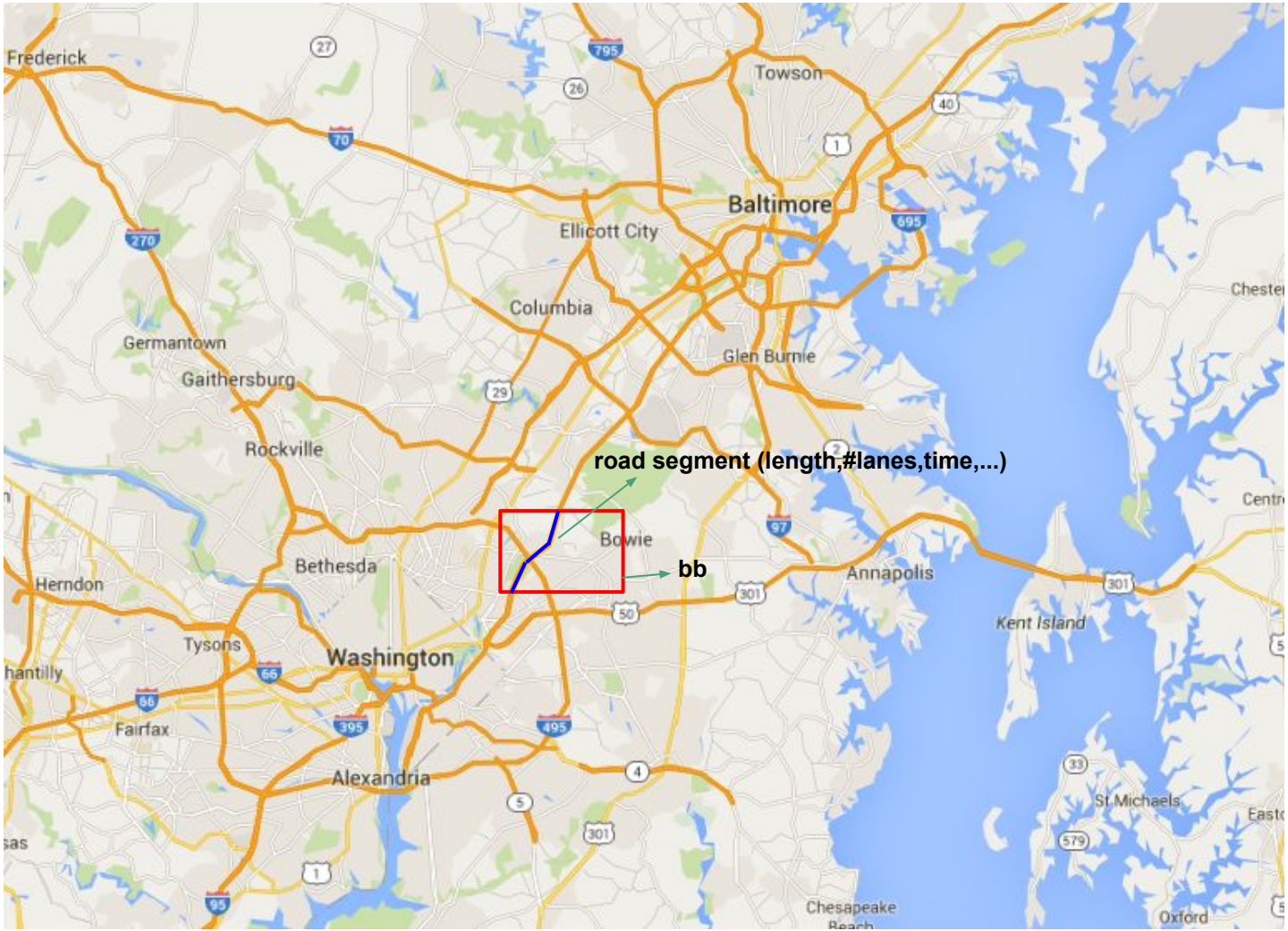
* The events consist a subset of all possible events in the training data.

Prediction Task

Hints

1. For a given geographical bounding box *bb*, identify all major road segments contained within *bb*, using OpenStreetMap. Denote set of road segments as S.
2. For the *k-th* segment s_k in S, predict the total number of events for each event type will occur within the given time interval *ti* (interval is always 1 month). Denote predicted number of events will occur for the *i-th* event type e_i as n_{ik}.
3. Predict the total number of occurrences for events of the i-th event type within given *bb* and *ti* as: $y_i = sum\_k (n_{ik})$.

The crux is in the 2nd step: how to predict #events for a road segment?


road segment (length,#lanes,time,...)
bb
Frederick
Towson
Baltimore
Ellicott City
Columbia
Germantown
Gaithersburg
Glen Burnie
Rockville
Bowie
Bethesda
Herndon
Annapolis
Kent Island
Tysons
Washington
Fairfax
Alexandria
St Michaels
Chesapeake
Oxford

Prediction Task

Hints

Predict #events of type event type e for a road segment s within time interval t.

1. Extract set of road features from segment s. The road features can be*:
   a. length of the road segment.
   b. #lanes of the road segment.
   c. month when the measurement starts (e.g. Jan, Feb, ..., Dec).
2. Predict #events as output of H_e (*set of road features*) → #events.
   a. The H_e is a regression model learnt from training data.
   b. The subscript e denotes “event”, which implies we should train one model for per event.

Note: The road features are for your reference only. You are encouraged to come up with new features. Useful features usually have noticeable impact on occurrence of events.

Prediction Task

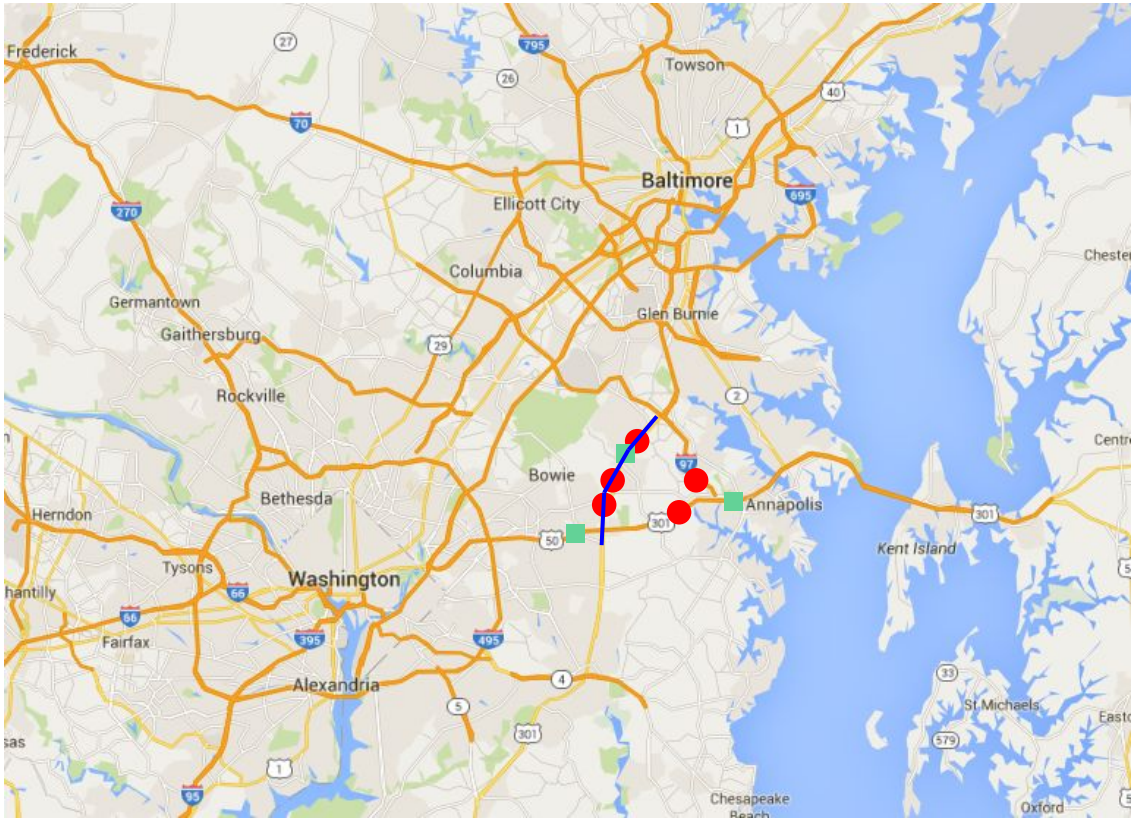
Hints

Train a regression model.

1. Construct training data.
    a. Identify a set of reasonable road features that your can obtain during both training and testing*.
    b. Extract a series of (road features, #events) as training data, where #events is regression target.
2. Select a regression model.
    a. Polynomial regression
    b. Gaussian Process for regression

* In the testing stage, you are only provided with geographical bounding box (latitude, longitude) and time interval (start, end) of one month. Flows and speed values for some roads may **not** be available during the testing stage, which means they cannot be used as road features. However, you may be able to infer missing features which are available in training stage but not testing stage by various methods. For example, you can learn an extra model to predict flow values of a road based on the training data. In the testing stage, though flow values are not available, you can first predict flow values and then predict #events. You may also come up with latent analysis model which set flow values as latent variables.

Construct training data (road-dependent events)



- Accident And Incident Events (AAI)
- Roadwork Events
- Randomly Selected Road Segments

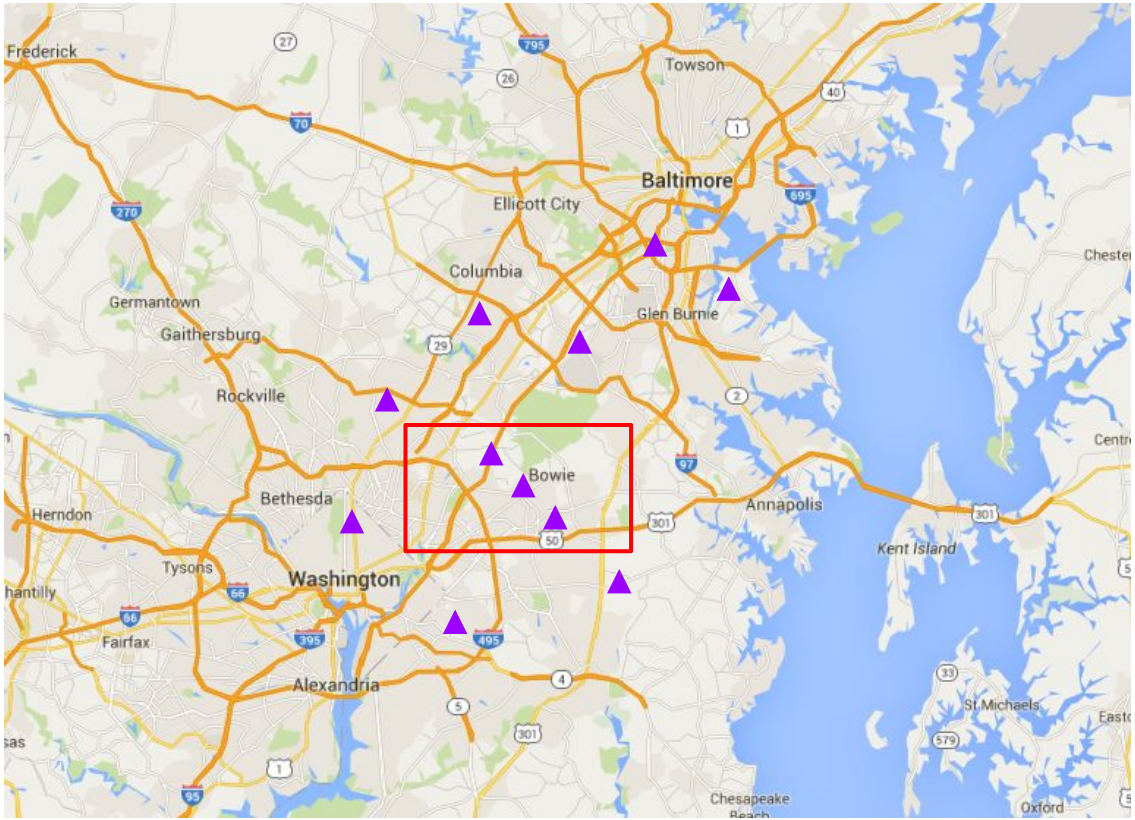
<u>Accident And Incident Events data construction</u>

1. Randomly select a road segment s.
2. Count #events of type AAI occurs within the segment in a random time interval of one month.
3. Obtain a single training entry: **(length,#lanes,time,...) → #events**
4. Repeat step 1-3 to generate more training entries.

Notes:

1. Selected road segments should have records in the training data.
2. Length of selected road segments should vary noticeably.
3. Time of event counts should vary noticeably (e.g. Jan, Feb, …, Dec).
4. The method is for your reference only. You are encouraged to come up with other solutions.

Construct training data (road-independent events)



Randomly Selected bounding boxes
Precipitation events

Precipitation data construction

1. Randomly select a region.
2. Count #events of type precipitation occurs within the region in a random time interval of one month.
3. Obtain a single training entry: **(longitude,latitude,w,h,t) → #events**
4. Repeat step 1-3 to generate more training entries.

Notes:

1. The shape of selected area should vary noticeably.
2. The method is for your reference only. You are encouraged to come up with other solutions.

Questions