Outlier Detection for Traffic Measurement

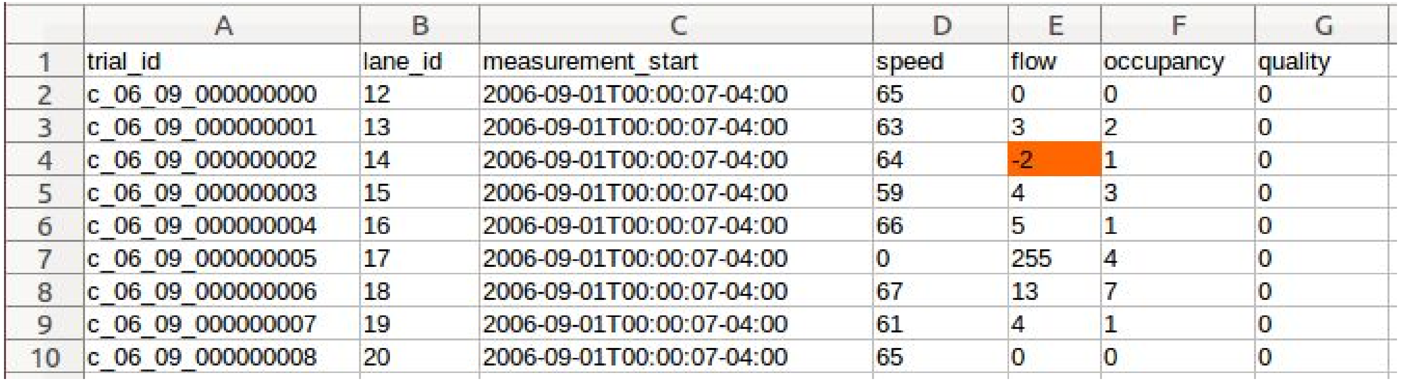
Cleaning I

Task Overview

- Description

  Given a collection of erroneous measurement data (e.g. flow, speed, occupancy), you are asked to predict the probability that a specific measurement is correct.

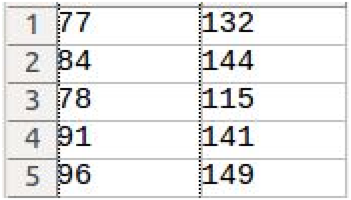
- Example Erroneous Measurement



	A	B	C	D	E	F	G
1	trial_id	lane_id	measurement_start	speed	flow	occupancy	quality
2	c_06_09_000000000	12	2006-09-01T00:00:07-04:00	65	0	0	0
3	c_06_09_000000001	13	2006-09-01T00:00:07-04:00	63	3	2	0
4	c_06_09_000000002	14	2006-09-01T00:00:07-04:00	64	-2	1	0
5	c_06_09_000000003	15	2006-09-01T00:00:07-04:00	59	4	3	0
6	c_06_09_000000004	16	2006-09-01T00:00:07-04:00	66	5	1	0
7	c_06_09_000000005	17	2006-09-01T00:00:07-04:00	0	255	4	0
8	c_06_09_000000006	18	2006-09-01T00:00:07-04:00	67	13	7	0
9	c_06_09_000000007	19	2006-09-01T00:00:07-04:00	61	4	1	0
10	c_06_09_000000008	20	2006-09-01T00:00:07-04:00	65	0	0	0

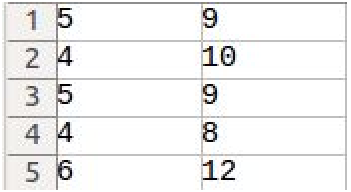
Data

Measurements are divided by zones, where each zone can have one or more detectors. Detectors in the same zone are geographically next to each other. For each zone, you are given the following data:



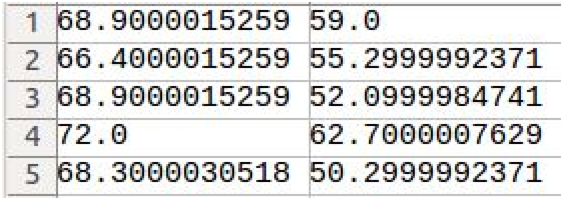
77	132
84	144
78	115
91	141
96	149

flow.tsv



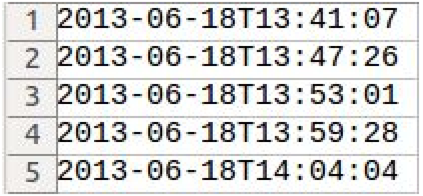
5	9
4	10
5	9
4	8
6	12

occupancy.tsv



68.9000015259	59.0
66.4000015259	55.2999992371
68.9000015259	52.0999984741
72.0	62.7000007629
68.3000030518	50.2999992371

speed.tsv



2013-06-18T13:41:07
2013-06-18T13:47:26
2013-06-18T13:53:01
2013-06-18T13:59:28
2013-06-18T14:04:04

timestamp.tsv

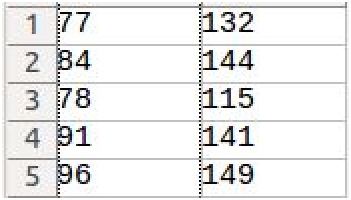
- #columns = #lanes: Each column is corresponding to one lane (e.g. data by the same detector).
- #rows = #timestamp: Each row represents measurement at specific time given by timestamp.tsv
- Missing data: flow, occupancy and speed can have missing data. If a measurement of specific lane at specific timestamp is missing, then that corresponding field is empty.
- Discontinuous timestamps: most of the time, the timestamp increases with fixed interval. But, this is not guaranteed. You should NOT assume nearby rows are measured in nearby time intervals. Always check the timestamp to see if they are continuous or not.

Task Description

- Step #1: Construct measurement vectors

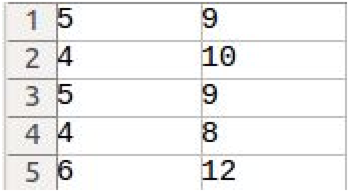
Construct a set of measurement vectors MV = {(flow$_k$, speed$_k$, occupancy$_k$) | k = 1 ,..., R*C}, where R is #rows, and C is #columns. Each vector should be measurement of the same detector at the same timestamp. Because we have #rows timestamps and #columns detectors, so |MV| = R*C.

- Example



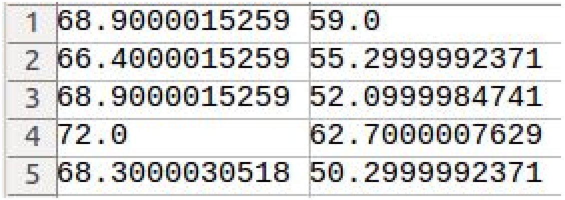
1	77	132
2	84	144
3	78	115
4	91	141
5	96	149

flow.tsv



1	5	9
2	4	10
3	5	9
4	4	8
5	6	12

occupancy.tsv



1	68.9000015259	59.0
2	66.4000015259	55.2999992371
3	68.9000015259	52.0999984741
4	72.0	62.7000007629
5	68.3000030518	50.2999992371

speed.tsv

This should come to 5*2 = 10 vectors. The first two vector are (77, 5, 68.9) and (132, 9, 59.0).

Task Description

- Step #2: Model probability distribution of vectors

  Given MV = {(flow_k, speed_k, occupancy_k) | k = 1 ,..., R*C}, construct a probability model, e.g. M, to estimate the probability density at each vector in MV.

- Example Approach

  Divide flow, speed and occupancy into N continuous intervals (e.g. N ~ 20), such that number of data points falling into any interval is mostly the same and maximum difference between any pair of values in the interval is no more than a threshold (e.g. flow: 10, speed: 10, occupancy: 5). Then for each 3D box defined by three intervals (flow, speed, occupancy), count the number of vectors falling into it. Next, for each box we have probability density estimated by: p = #vectors_in_box / (#total_num_vectors * 3D_box_volume). Finally, the probability density of a vector is given by the p value of the box it belongs to. For those measurement (e.g. negative values) that we know for sure is incorrect, it may be useful to just assign 0 probability density to them.

Submission

For each of the zones (3445, 3532, 3451, 3232, 1160), submit one file named “zone_id.txt”. The zone_id is the name of folder that contains the measurements. Each row of “zone_id.txt” file is separated by TAB, like this (order is important):

```
flow	speed	occupancy	probability
```

- Rows should be SORTED by probability in ascending order.
- You should NOT submit all R*C rows, instead you should sample 1% of the rows by taking every 100th values in the sorted list, like: [1st, 100th, 200th, 300th, …, end]. So you should only have around (0.01*R*C) rows for each zone.
- Flow, speed, and occupancy should be rounded to integers, and probability (between 0 and 1) should preserve 8 decimals.
- Only one member need to submit the results.
- **Also submit a report (pdf file) describing all details.**