PROGRESS REPORT ON STUDY OF NYC TRAFFIC

VAIBHAV KARVE

1. Restricting to a block

We restrict our attention to the block between W. 44th to 45th St. and 8th to 9th Ave. This block was selected at random. On extracting the data from the data file for 2011, we see that these streets are all one-way access only.

1.1. Naming convention.

1.1.1. *Intersections.*

- A = W. 44th St. and 8th Ave. (Node_id = 42435671)
- B = W. 44th St. and 9th Ave. (Node_id = 42443561)
- C = W. 45th St. and 8th Ave. (Node_id = 42432700)
- D = W. 45th St. and 9th Ave. (Node_id = 42432703)

1.1.2. *Roads.*

- BA (Link_id = 128255)
- AC (Link_id = 169017)
- CD (Link_id = 182993)
- DB (Link_id = 181188)

We restrict our attention to only these links, one at a time. For each link, from the database, we extract 2 separate arrays: one giving the average travel time in seconds, for every hour of the year; and the other giving the number of trips on that used that link, for every hour of the year.

1.2. Periodicity Analysis.

1.2.1. *Stratifying the data.* Intuitively, one may expect that traffic patterns repeat themselves every 7 days (weekly) or maybe every 30 days (monthly). Or perhaps, there is no such periodicity. Whatever be the case, the periodicity should not be imposed, but rather should be inferred from the data itself. To do so, we we assume a particular period in days, call it `period` and divide the entire data into $24 \times$ `period` number of bins. This converts our data from flat lists of trips and traveltimes to something that may be viewed as stratified data.

Stratified data now looks like:

$$\begin{pmatrix} \textbf{Stratum\_1} & \textbf{Stratum\_2} & \textbf{Stratum\_3} & \cdots & \textbf{Stratum\_N} \\ 5 & 2 & 32 & \cdots & 10 \\ ? & 7 & 5 & \cdots & ? \\ 33 & ? & 12 & \cdots & ? \\ \vdots & \vdots & \vdots & \cdots & \vdots \end{pmatrix}$$

Here, $N = 24 \times$ `period`. We let period range from 2 to 38 for our periodicity analysis.

Date: May 20, 2016.

1.2.2. *Missing values.* The question marks (?) in the above matrix correspond to those hours for which we have no data on our link. This ofcourse does not mean that there is no traffic on that link at that time, just that we don't know what it it. These are to be treated as missing values in our data and need to be somehow inferred.

The links BA, AC, CD and DB have 20%, 0%, 0% and 1% of their values missing, respectively.

The most natural first approximation for these missing values is the mean value of the corresponding stratum to which each missing value belongs.

The Inferred Stratified data looks like:

$$\begin{pmatrix} \textbf{Stratum\_1} & \textbf{Stratum\_2} & \textbf{Stratum\_3} & \cdots & \textbf{Stratum\_N} \\ 5 & 2 & 32 & \cdots & 10 \\ Mean_1 & 7 & 5 & \cdots & Mean_N \\ 33 & Mean_2 & 12 & \cdots & Mean_N \\ \vdots & \vdots & \vdots & \cdots & \vdots \end{pmatrix}$$

Here, we calculated the mean for each stratum by ignoring the missing values.

1.2.3. *Stratified variances.* To establish the optimal choice of `period` for the data, we calculate the variance of the inferred stratified data we obtained in the previous subsection.

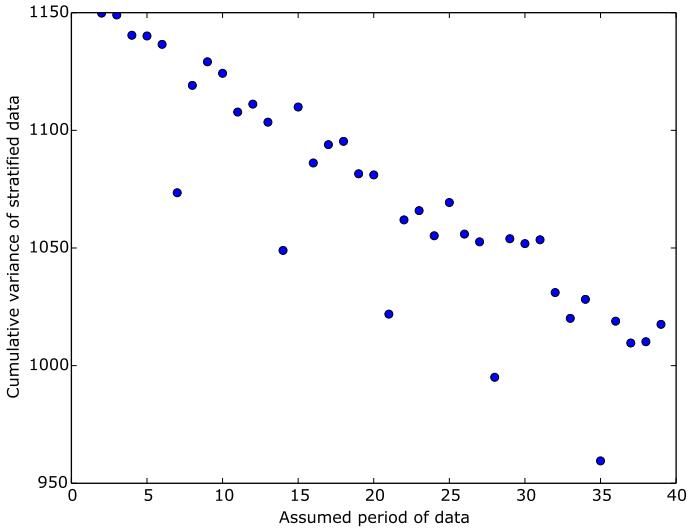
$$\text{Variance of stratified data} \approx \frac{1}{n} \sum_{n} \text{Variance of each stratum}$$

where $n = 24 \times$ `period` is the total number of strata.

If there truly does exist a periodicity in the data, for the correct `period`, the values within each stratum will lie close to each other and hence, the variance will be minimized. Hence, we look for dips in the variance.

1.3. **Conclusion of periodicity analysis.** We calculate the variance for each value of `period` from 2 to 38. The graph of the variances for link BA for the No. of Trips during 2011 is given below:

FIGURE 1. The dips in variance correspond to `period` $= 7, 14, 21, 35$.



The other three links show similar dips in variance at `period` values which are multiples of 7, in both data – Travel times as well as No. of Trips. **Conclusion:** NYC traffic data has a 7-day periodicity.