**Railroad Network Mapping and Analysis**

**Final Project Report**

**CSE 519 – Data Science Fundamentals**

**Team Members:**

Abhinav Jain (111495982) abjain@cs.stonybrook.edu

Chaitanya Kalantri (111446728) ckalantri@cs.stonybrook.edu

Jitendra Savanur (111491676) jsavanur@cs.stonybrook.edu

**Objectives:**

1. To analyze the railroad network of the United States in order to study the railroad crossing frequencies across various segments of the railroad.
2. To speed up computations for embedding crossings on railroad segments i.e. to form a mapping between the segments and crossings occurring on them and vice-versa.
3. To study the railroad crossing frequencies on paths without branches and to check if it is consistent
4. To study the distribution of crossing frequencies
5. To come up with a list of 1000 places on the network where sensors can be installed to monitor maximum railroad traffic

**Tasks:**

1. **Computational speed up for embedding crossings data on railroad segments**

The general overview of how we were doing this is in the previous phase:

* Read the list of all railroad crossings
* For each of the crossings in this list

-Loop through the segments in the same *state*

-Compute distance between each point on the segment with the crossing coordinate

-Get the segments whose any point is at the least distance from the crossing

Pros:

-Accuracy of this approach was decent since we were comparing each coordinate on the segment with every crossing coordinate.

Cons:

-Since there are thousands of crossings in a state and thousands of railroad segments in the same state, the loop would run many times and take up a lot of computational time.

Optimizations:

The optimizations that we implemented helped reduce the computational time considerably and let us play with the graph for the railroad network of the entire United States.

The new approach for embedding the crossings data on the segments is as follows:

* Read the list of all railroad crossings.
* For each of the crossings in this list

-Loop through the segments in the same *county*.

-Instead of getting the distance between the crossing coordinate and every coordinate on each segment, we consider only three points for our distance calculation.

1. The starting point of the segment

2. The middle point of the segment (this is not same as the mid-point. For instance, if there are 10 coordinates given for a segment, we consider the 5th coordinate for our distance calculation)

3. The ending point of the segment

-Map the crossing to that segment which gives the least distance from any of the above points with that crossing.

Pros:

This approach is very time optimal since:

-For every crossing, we are considering the segments occurring in that *county* only. That reduces our search space greatly.

-For every segment, we are considering only three points on the segment. Thus, for every segment, however there are many points on the segment, we shall compute distance of the crossing coordinate from only the three points.

Cons:

-The gain in time comes with a loss in accuracy. Some crossings do not get embedded with the actual segment they lie on, since this is an approximate approach to solve the problem.

**2. Study of Railroad Crossing Frequencies**

This is the main objective of this project i.e. to study the consistencies in the railroad crossing frequencies across segments. The objective here is to study the number of trains passing through a crossing during a given time period.

For this, we have considered the columns *DayThru* and *NghtThru* in the crossing data set, which indicate the number of train movements through the crossings during the day and night respectively.

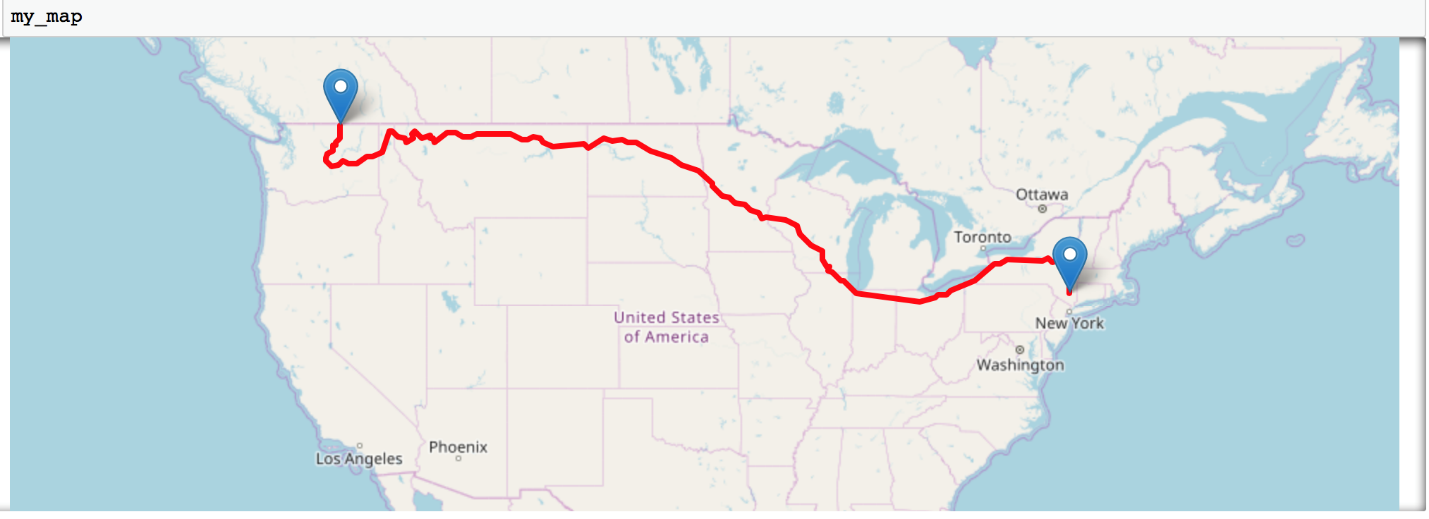
Intuitively, we can say that if there is a path which has no branches i.e. there is no alternate segment, that a train can choose on that path, the crossing frequency in that segment has to be similar.

For example, consider that 10 trains were to go from point A to point B in the network, and there is no branching in the path between A and B. Further, consider that there are 5 crossings on this path from A and B. Now if 10 trains are going from A to B in a given time frame, then the number of trains going through the 5 crossings is 10. We can also say that if there were gates on these 5 crossings, then all those gates will go up and down 10 times, since there are 10 trains going though those crossings. This is true, unless some of the 10 trains might go in parallel, if there are multiple parallel tracks.

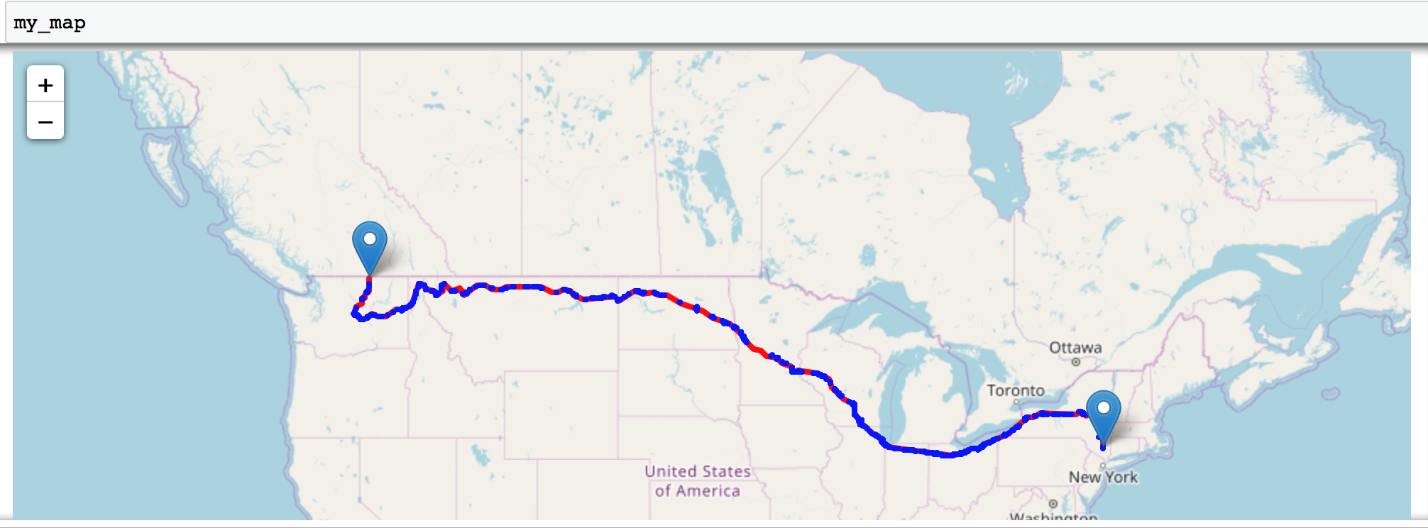
We have considered a number of runs in different locations and plotted the graph for the frequency of crossings for that path.

**1. Path from Newburgh, NY to Oroville, WA**

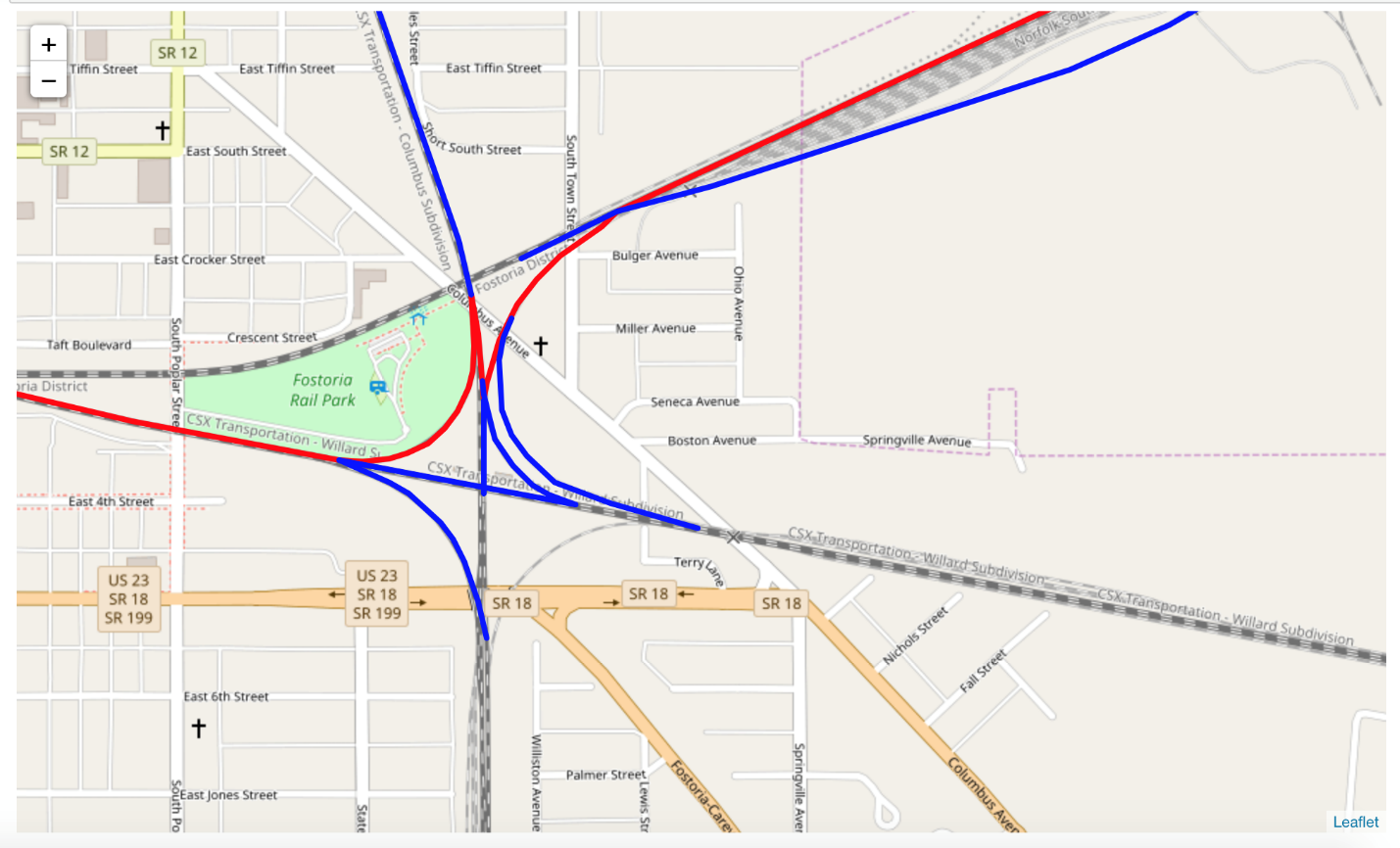
The shortest path from Newburgh, NY to Oroville, WA looks like this:



*Figure 1: Shortest path from Newburgh, NY to Oroville, WA*

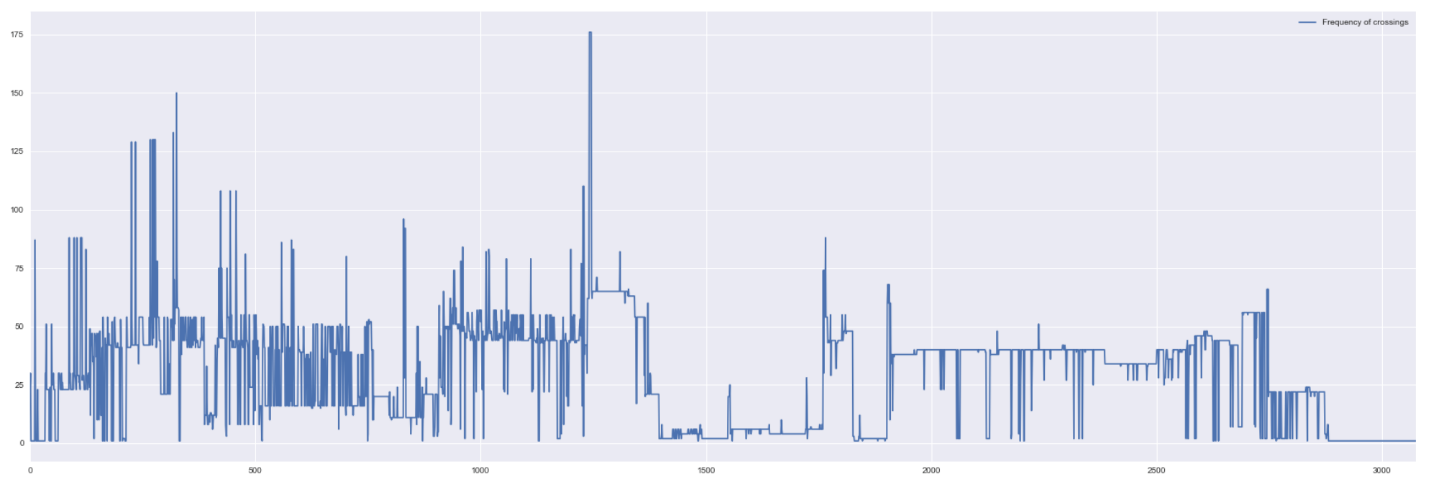


*Figure 2: Shortest path from Newburgh, NY to Oroville, WA, with branching. Red lines indicate the path and blue lines indicate the branching. The blue lines are very dense here because of the dense branching.*



*Figure 3: Shortest path from Newburgh, NY to Oroville, WA, with branching zoomed-in*

The distribution of crossing frequencies on the above path are plotted as a line chart:

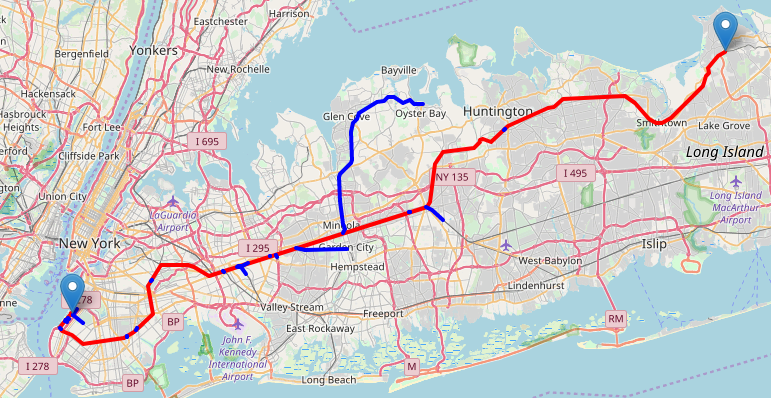


*Figure 4: Crossing frequency distribution for path from Newburgh, NY to Oroville, WA.*

The above graph is not very easy to read since there are many crossings (~2500) on this path. The following graphs for shorter paths are easier to understand.

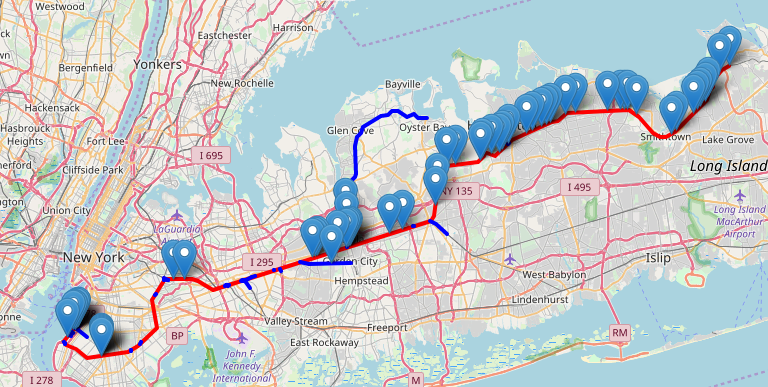
**2. Path from East Setauket to Brooklyn**

The shortest path from East Setauket to 39 St/2 Av Brooklyn looks like this:



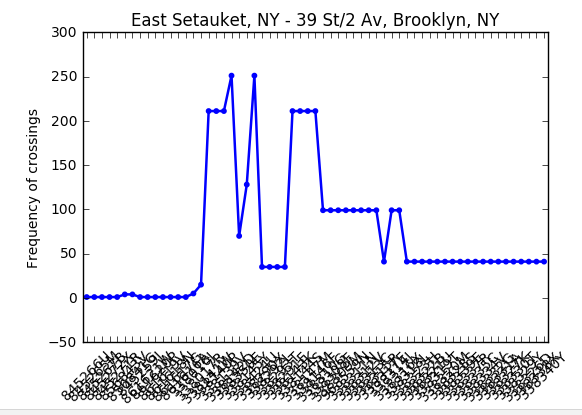
*Figure 5: Shortest path from East Setauket to 39 St/2 Av Brooklyn. The red line indicates the actual shortest path and the blue lines indicates the branching on the path.*

If we plot the crossings which come along the shortest path above, we have:



*Figure 6: Path from East Setauket to Brooklyn with the crossings on the path*

We plotted the frequency of crossing by considering the columns DayThru and NghtThru of the crossings on the above path and plotted the same from the source to destination.

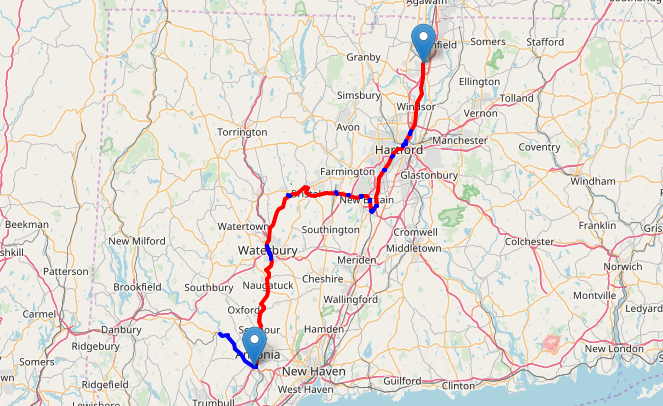


*Figure 7: Distribution of crossing frequencies for the path from East Setauket to Brooklyn*

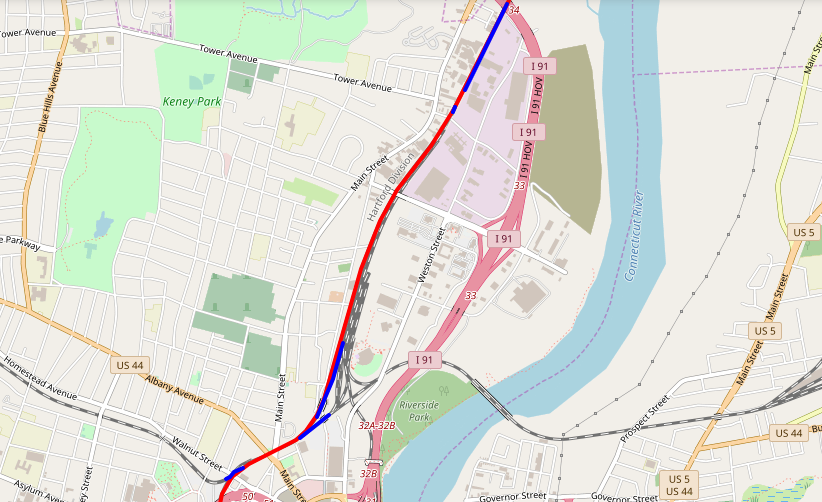
Here, we observe that for the initial 15 crossings, the crossing frequency is consistent. Also, wherever there is no branching, the crossing frequency remains consistent. The crossing frequency shoots up after that and the variation in the crossing frequencies implies branching.

**3. Path from Windsor, CT to Shelton CT**

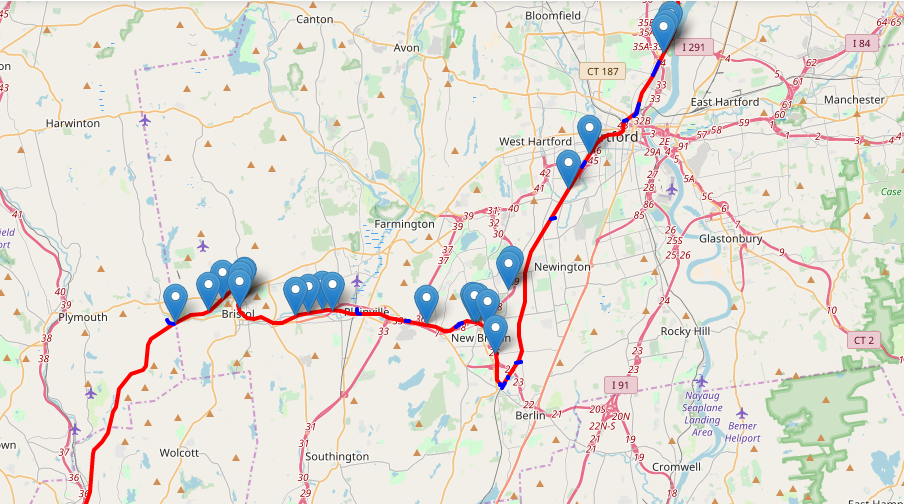
The path from Windsor, CT to Shelton CT is:



*Figure 8: Path from Windsor Locks, CT to Shelton, CT. Red lines indicate the path and blue ones indicate branching.*

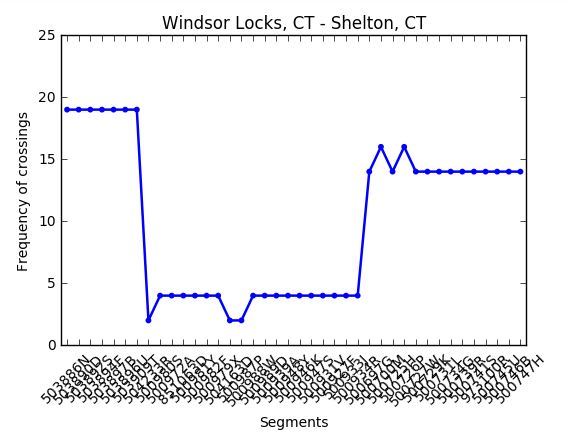


*Figure 9: Zoomed-in version to get a better view of the branching*



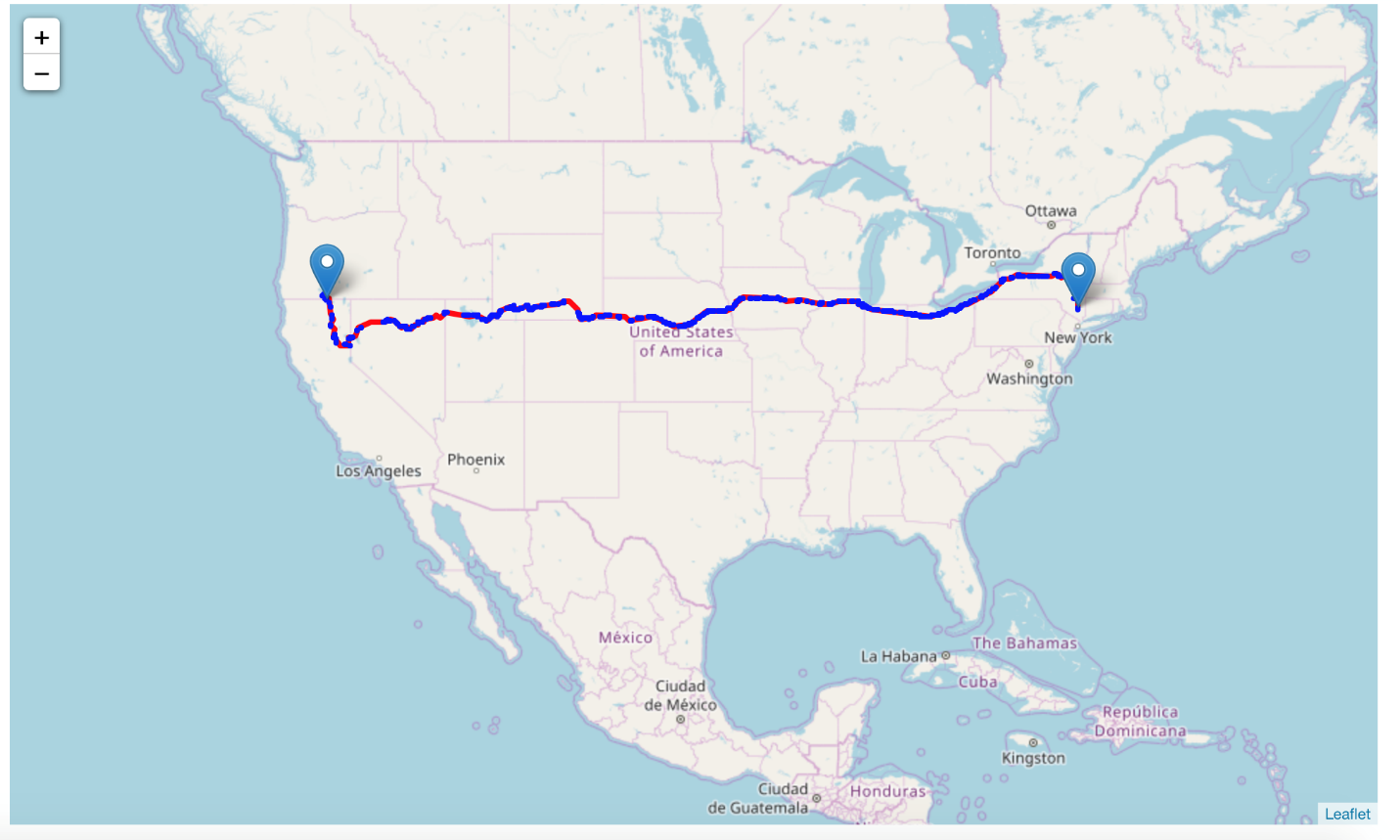
*Figure 10: Crossings present on the path from Windsor Locks to Shelton*

Distribution of crossing frequencies on the above path as a line chart:

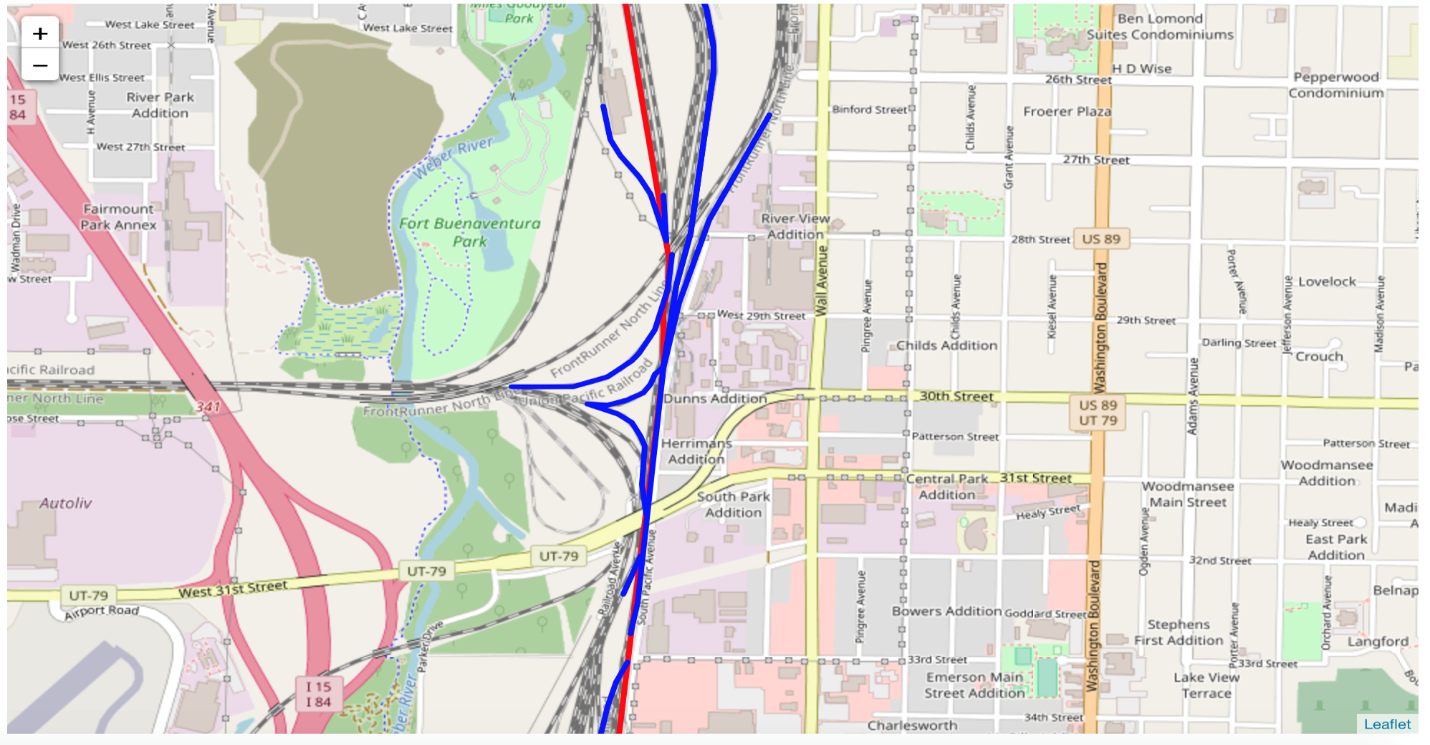


*Figure 11: Crossing frequencies from Windsor Locks to Shelton*

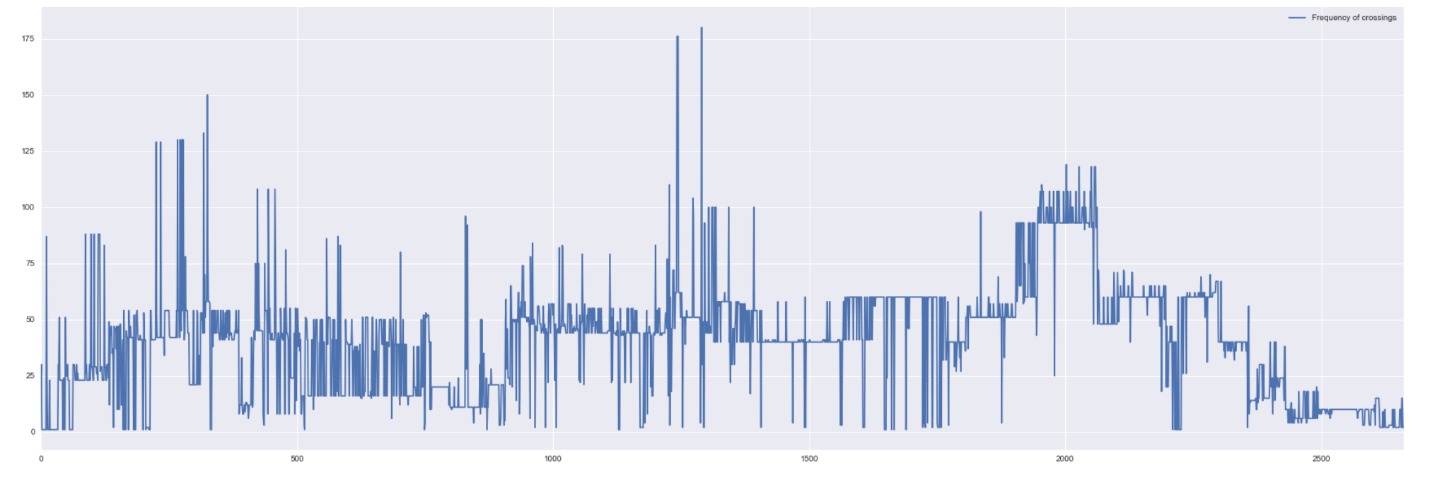
**4. Path from Newburgh, NY to Tulelake, CA**



*Figure 12: Path from Newburgh, NY to Tulelake, CA. Red lines represent the path and blue ones are branches.*



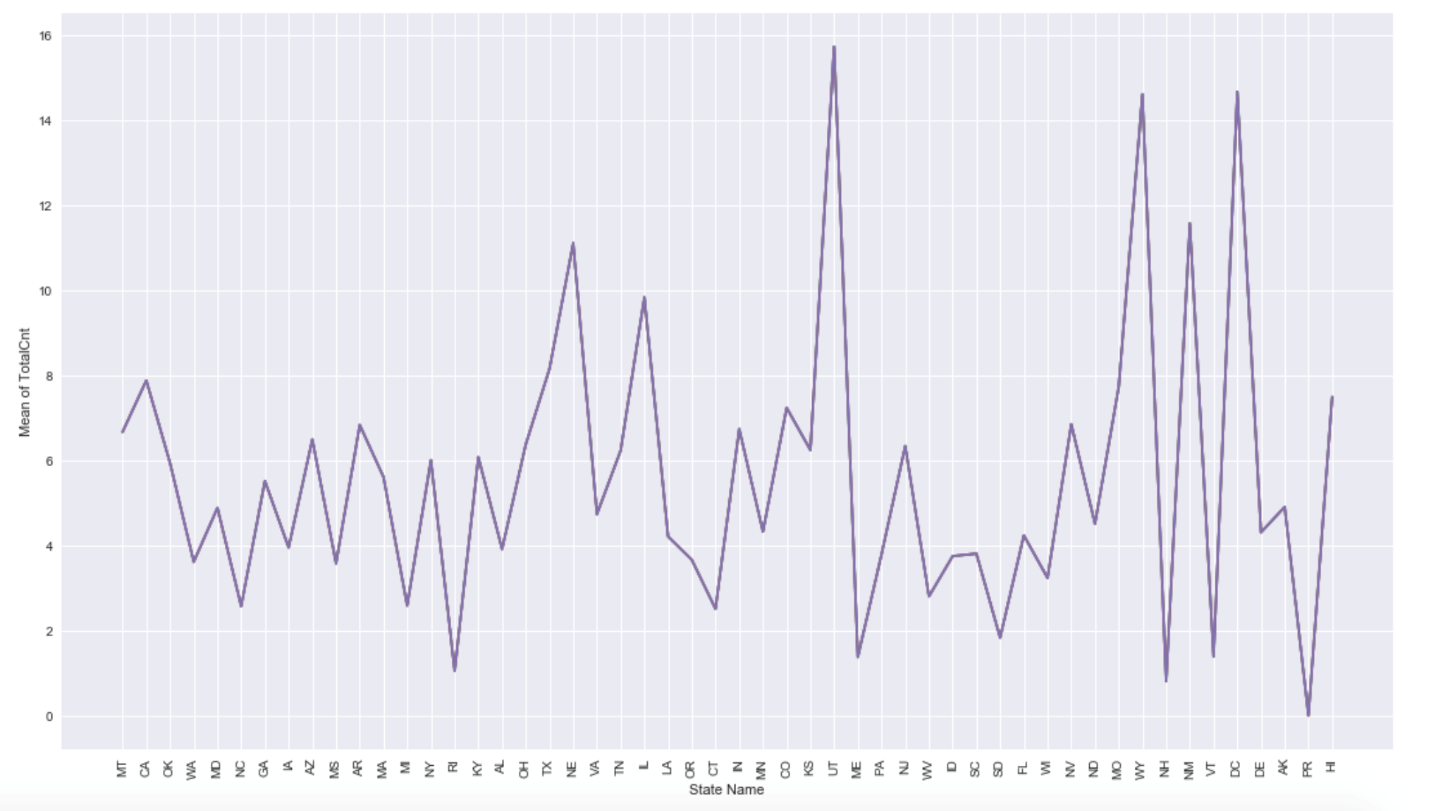
*Figure 13: Zoomed-in path from Newburgh, NY to Tulelake, CA. Red lines represent the path and blue ones are branches.*



*Figure 14: Crossing frequencies for path from Newburgh, NY to Tulelake, CA.*

**3. Distribution of Crossing Frequencies.**

Here, we study the distribution of crossing frequencies across states in the US. This will give us an idea of how dense the railroad network is in various states of the US.



*Figure 15: State wise distribution of railroad crossing frequencies.*

We observe that the state of Utah has the highest crossing frequency, followed by Wyoming, which is followed District of Columbia

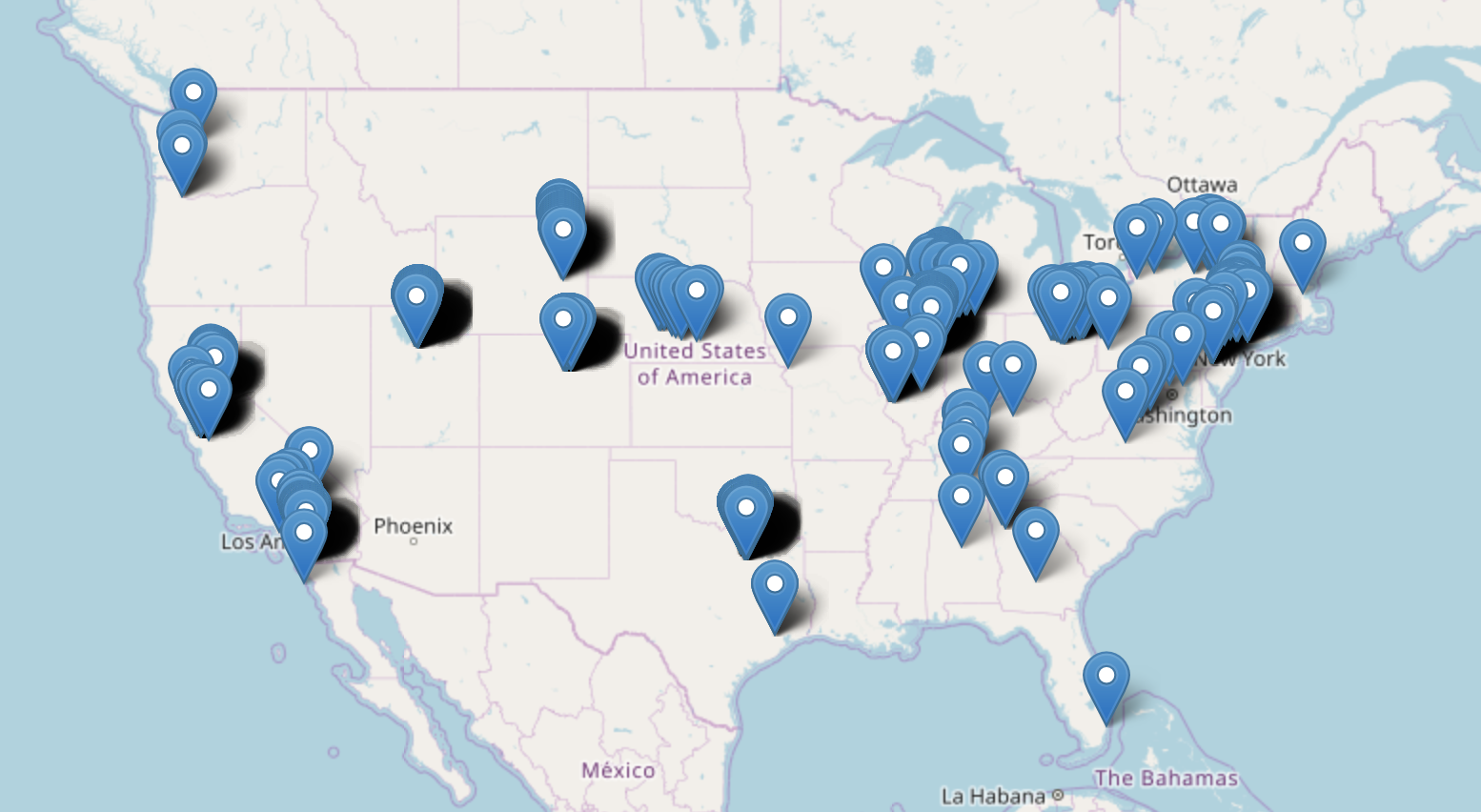
**4. Top Thousand places for placing sensors**

Best 1000 locations to place sensors to monitor railway traffic frequency:

1. Initially, we cleaned the dataset. Moreover, we considered sum of DayThru and NghtThru columns as a new ‘Count’ column. These two columns indicate the number of trains passing in the day and night timings respectively.

2. We sorted the data based on the frequency of the 'Count' column. By this, we got the busiest stations of all over the USA railway network.

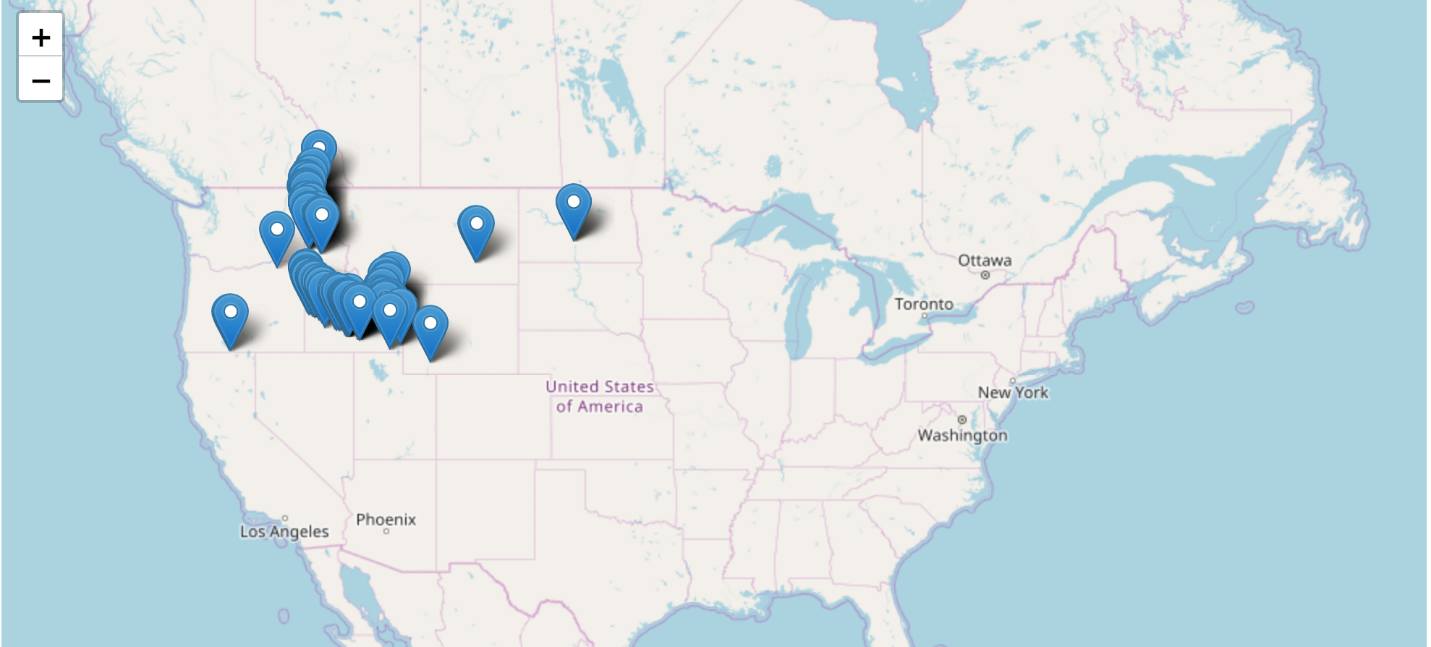
3. Finally, we extracted the top 1000 locations where the sensors should be placed in order to monitor the maximum railway network traffic.

****

*Figure 16: Busiest crossings in the United States.*

The above map shows the locations of the 1000 busiest crossings in the United States Railroad network. These can be the potential places where we can place sensors to monitor maximum railroad traffic.

**5. Top Thousand Nodes on the Railroad Network Graph**

****

*Figure 17: Top 1000 nodes in the railroad network graph. Note that these are the graph nodes and do not take into consideration the number of trains passing. Obtained by applying PageRank on the network graph.*

**Conclusion:**

We can draw the following conclusions from the analysis performed in this project:

* The railroad crossing frequencies appear to be consistent in a given segment.
* Based on the crossing frequencies, we can place sensors at a number of places to monitor real time railroad traffic.
* This helps us understand the branching of paths in the railroad network.

**Future Scope:**

* Analysis of train schedule data along with the crossing frequencies can give us a better idea as to how many trains pass through a given path and the trends in the opening and closings of gates.
* Separate analysis of freight and passenger trains might be useful in understanding further trends in the railroad network.