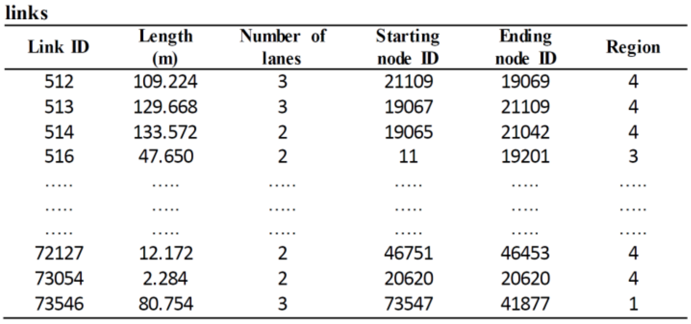
Traffic Data from Urban Inductive Loop Detectors and the Macroscopic Fundamental Diagram

Preface

The following document demonstrates an MFD (Macroscopic Fundamental Diagram) using a microsimulation environment that replicates real measurements of inductive loop detectors in cities. First, a few common traffic characteristics (flow, density, occupancy, speed) are estimated. Then, we plot relations between them. The end goal is to estimate the MFD of flow vs. density (production vs. occupancy) and analyze it with regional clustering.

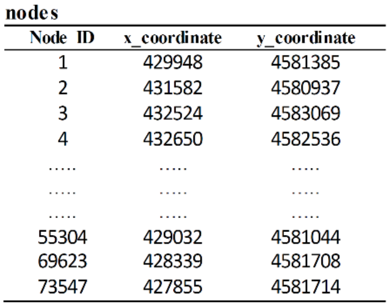
4 matrices are provided for data. All matrices are contained in the 'dataLab1.mat' file. The matrix representations are as shown in Tables 1 – 3.



links

Link ID	Length (m)	Number of lanes	Starting node ID	Ending node ID	Region
512	109.224	3	21109	19069	4
513	129.668	3	19067	21109	4
514	133.572	2	19065	21042	4
516	47.650	2	11	19201	3
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
72127	12.172	2	46751	46453	4
73054	2.284	2	20620	20620	4
73546	80.754	3	73547	41877	1

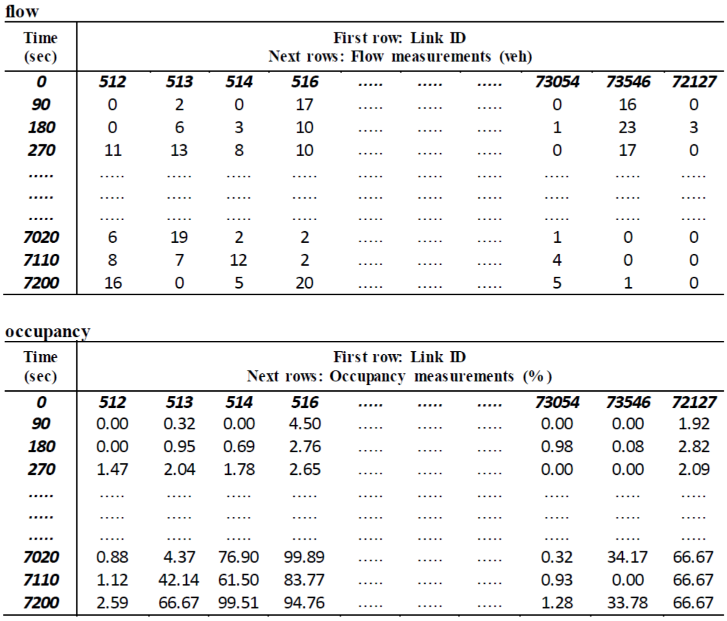
<Table 1.>



nodes

Node ID	x_coordinate	y_coordinate
1	429948	4581385
2	431582	4580937
3	432524	4583069
4	432650	4582536
.....	.....	.....
.....	.....	.....
.....	.....	.....
55304	429032	4581044
69623	428339	4581708
73547	427855	4581714

<Table 2.>



flow

Time (sec)	First row: Link ID Next rows: Flow measurements (veh)									
0	**512**	**513**	**514**	**516**	.....	.....	.....	**73054**	**73546**	**72127**
90	0	2	0	17	.....	.....	.....	0	16	0
180	0	6	3	10	.....	.....	.....	1	23	3
270	11	13	8	10	.....	.....	.....	0	17	0
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
7020	6	19	2	2	.....	.....	.....	1	0	0
7110	8	7	12	2	.....	.....	.....	4	0	0
7200	16	0	5	20	.....	.....	.....	5	1	0

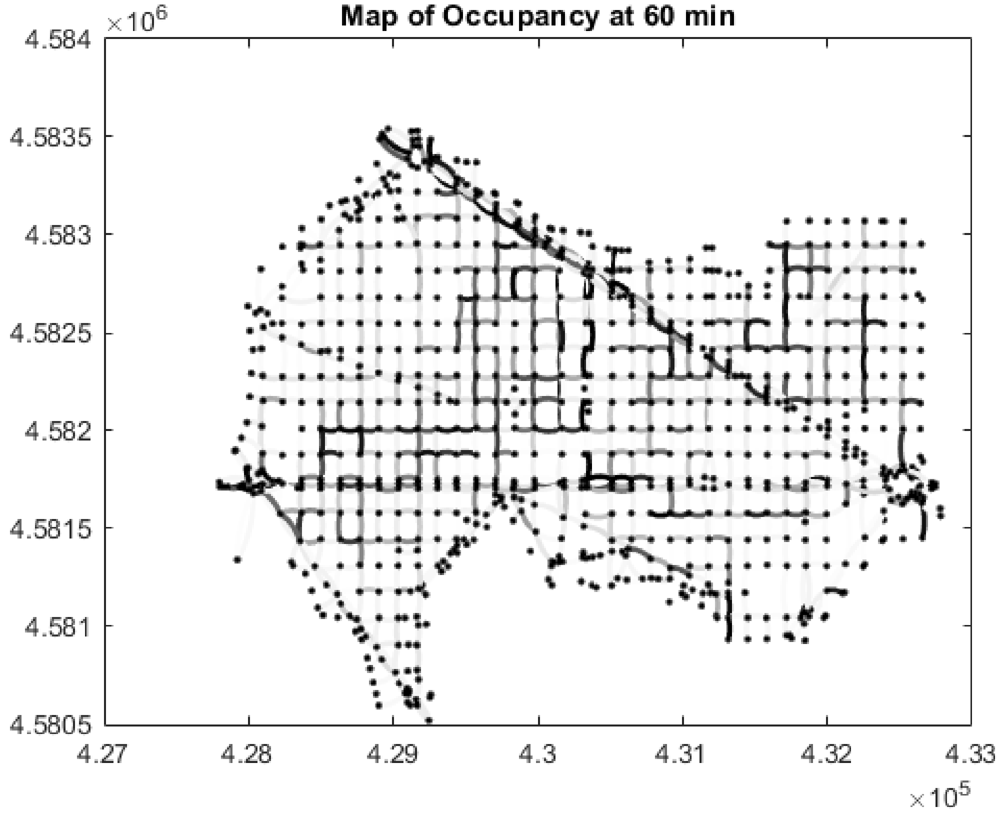
occupancy

Time (sec)	First row: Link ID Next rows: Occupancy measurements (%)									
0	**512**	**513**	**514**	**516**	.....	.....	.....	**73054**	**73546**	**72127**
90	0.00	0.32	0.00	4.50	.....	.....	.....	0.00	0.00	1.92
180	0.00	0.95	0.69	2.76	.....	.....	.....	0.98	0.08	2.82
270	1.47	2.04	1.78	2.65	.....	.....	.....	0.00	0.00	2.09
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
7020	0.88	4.37	76.90	99.89	.....	.....	.....	0.32	34.17	66.67
7110	1.12	42.14	61.50	83.77	.....	.....	.....	0.93	0.00	66.67
7200	2.59	66.67	99.51	94.76	.....	.....	.....	1.28	33.78	66.67

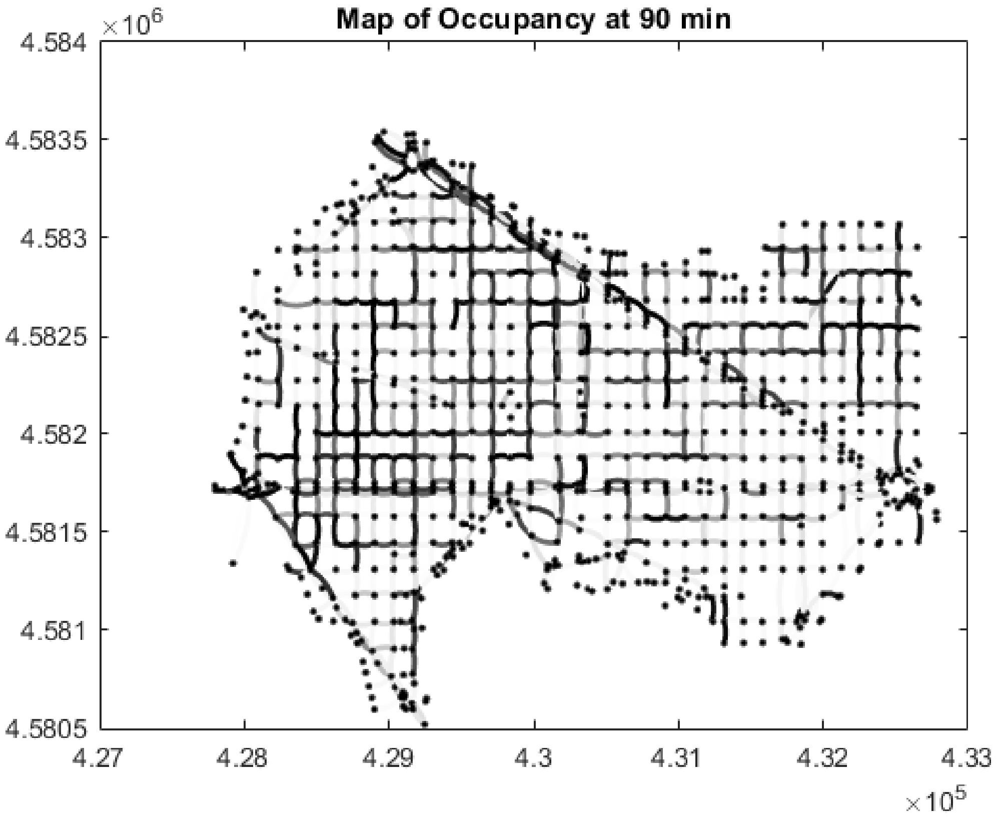
<Table 3.>

Step 1. Congestion of links

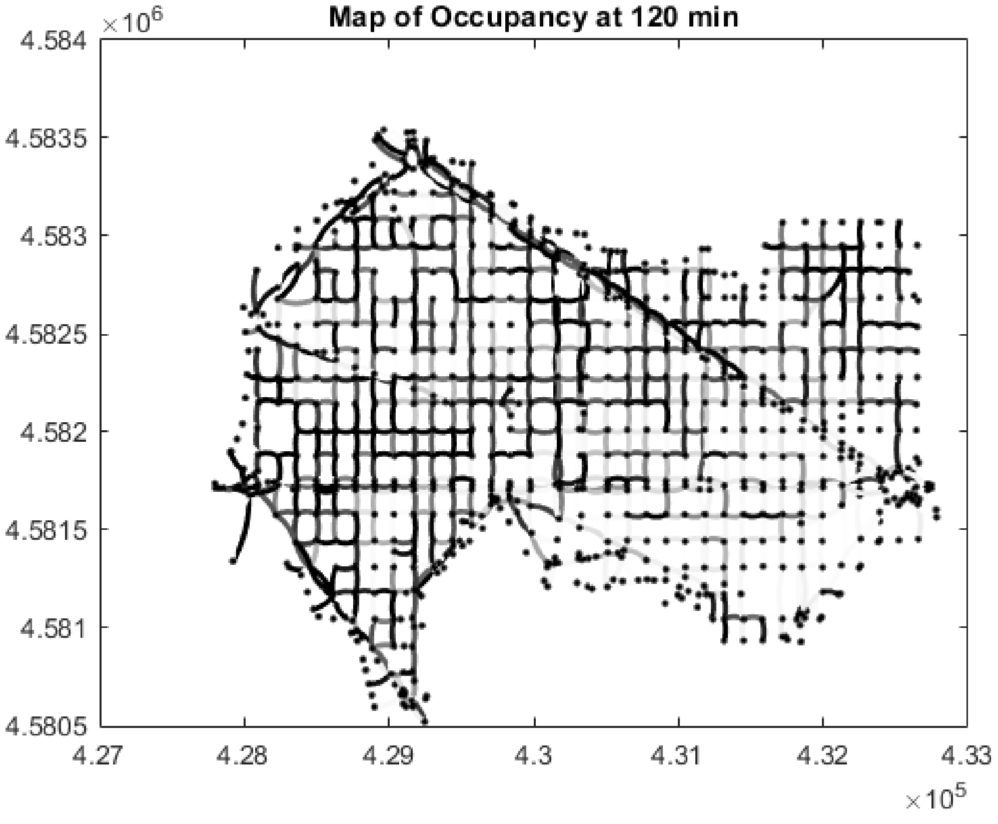
Figure 1, Figure 2, and Figure 3 depict the congestion of links at time 60min, 90min, and 120min respectively. Congestion is depicted in grayscale with 100% occupancy being black and 0% occupancy being white.



<Figure 1.>


Map of Occupancy at 90 min
×10^6
4.584
4.5835
4.583
4.5825
4.582
4.5815
4.581
4.5805
4.27
4.28
4.29
4.3
4.31
4.32
4.33
×10^5

<Figure 2.>

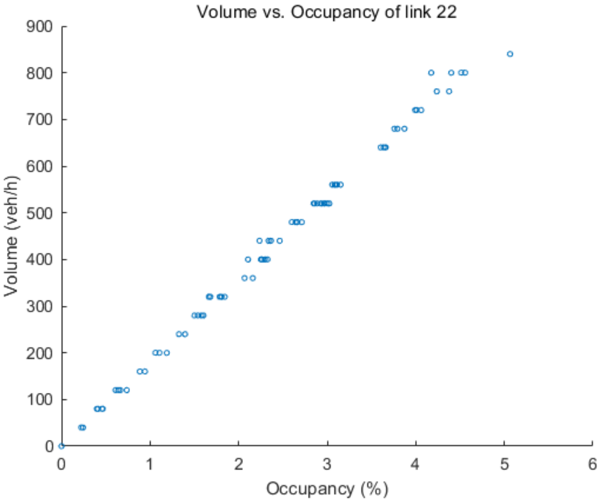

Map of Occupancy at 120 min
×10^6
4.584
4.5835
4.583
4.5825
4.582
4.5815
4.581
4.5805
4.27
4.28
4.29
4.3
4.31
4.32
4.33
×10^5

<Figure 3.>

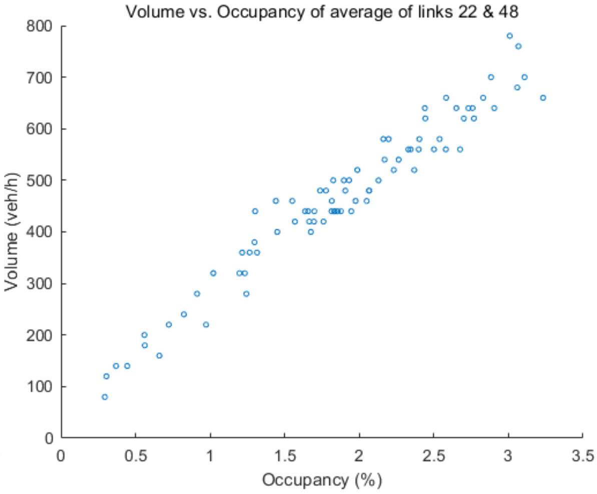
Overall, we can see that traffic occupancy is increasing overtime. However, the bottom right area of the map stays low on occupancy.

Step 2. Volume vs. Occupancy

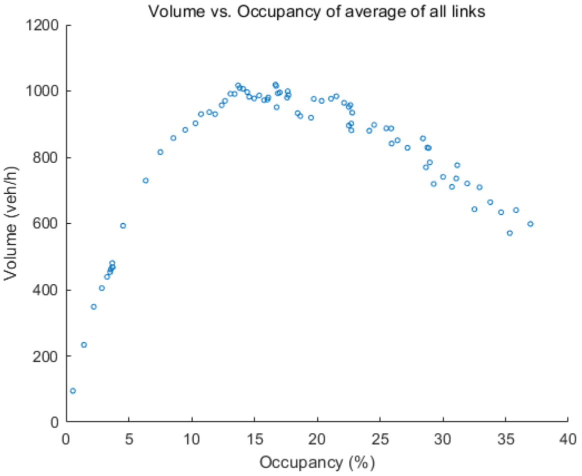
First, the flow measurements for 90 second intervals are converted to vehicles per hour, which is the volume measurement. Figure 4 shows the volume vs. occupancy scatter plot of a randomly selected link of 22. Figure 5 shows the volume vs. occupancy scatter plot of the average of two randomly selected links 22 and 48. Figure 6 shows the volume vs. occupancy scatter plot of the average of all links. Figure 7 does the same with the average for each of the 4 region clusters.



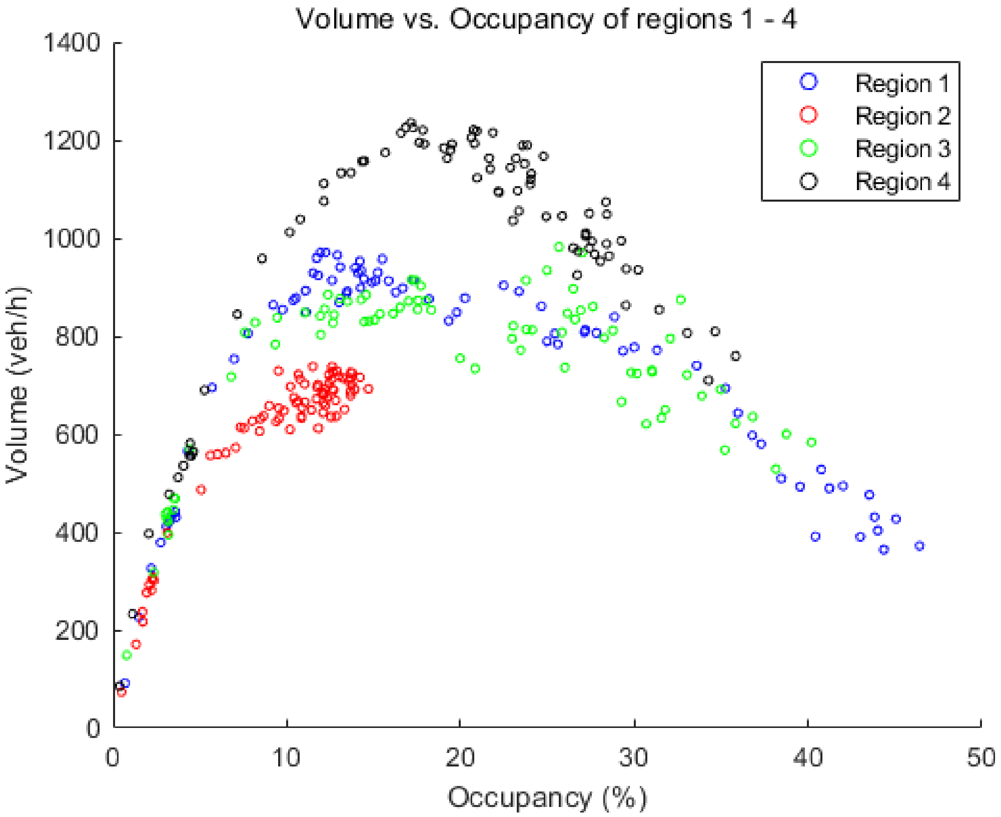
<Figure 4.>



<Figure 5.>



<Figure 6.>



<Figure 7.>

Figure 7 shows that regions 1&3 show almost identical patterns with region 1 having few examples of higher occupancy. Region 4 shows considerably higher capacity for the same amount of occupancy. Interestingly, most cases in region 2 are stuck at a low occupancy. Also, the trend of the MFD would suggest that region 2 has less capacity compared to all other regions. Looking at this figure only by itself would suggest that the partitioning of regions 4 and 2 is meaningful, whereas regions 1 and 3 could perhaps be a single region.

Step 3. Space-mean Speed

First, we calculate the following three values with the given formulas.

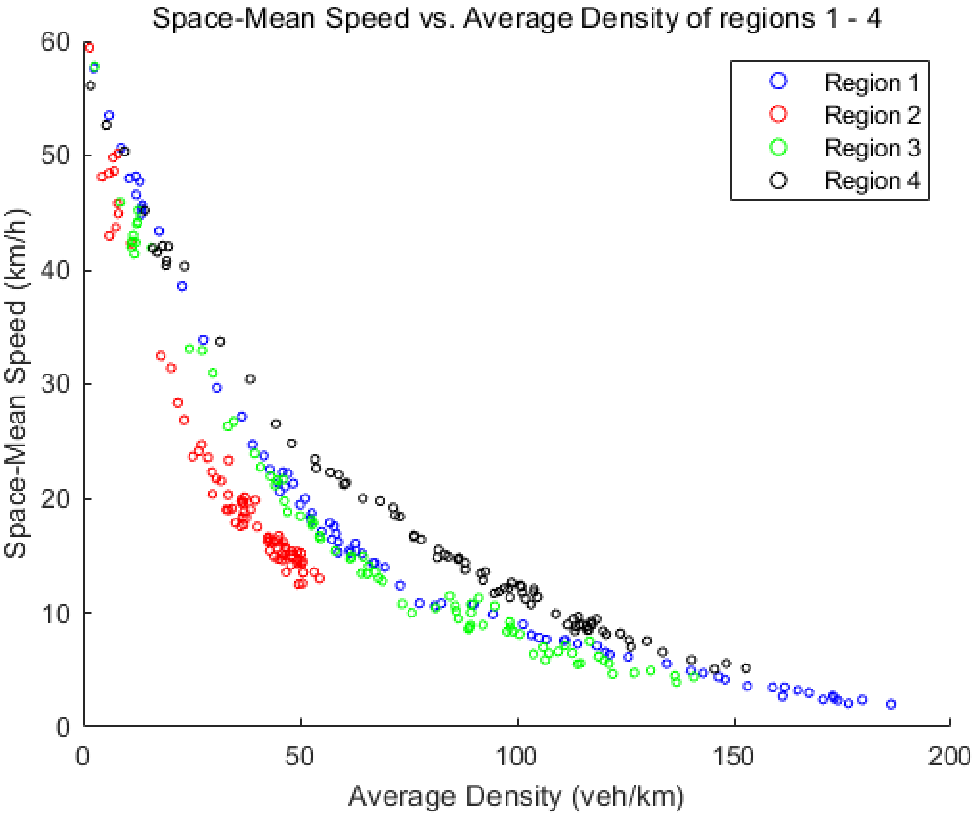
$$Density(k) = \frac{\frac{occupancy}{100} * \mu}{L + L_D} \left({}^{veh}/_{km} \right)$$

$$\mu: Number\ of\ lanes\ in\ a\ link$$
$$L_D = 2m\ (length\ of\ detector)$$
$$L = 5m\ (average\ vehicle\ length)$$

$$Linkspeed\left({}^{km}/_{h} \right) = \frac{Volume(V)\ ({}^{veh}/_{h})}{Density(k)\ ({}^{veh}/_{km})}$$

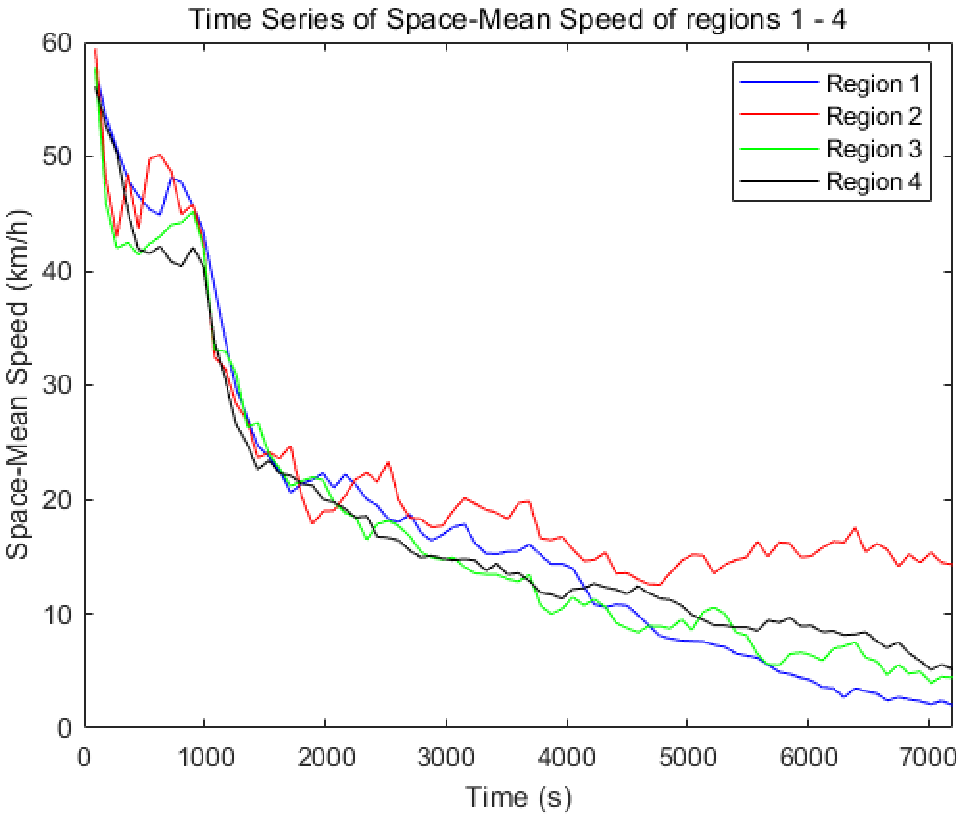
$$MeanSpeed(u)\left({}^{km}/_{h}\right) = \frac{\sum_{\forall link}(V * length)}{\sum_{\forall link}(k * length)}$$

Figure 8 shows the scatter plot of the space-mean speed as a function of average density for each of the regions. Again, region 2 is stuck at a low density and shows less speed for the same density. Regions 1 and 3 show almost identical behaviors again, and region 4 has the highest speed per density overall.



<Figure 8.>

Figure 9 depicts the time series of the space-mean speed of each region over the 2-hour period. For all regions, the speed starts out high, but drops significantly around the 15 min mark. Interestingly, region 2 is the only on that maintains a speed of 16 km/h while all other regions drop considerably. This is mostly since the occupancy of region 2 never passes the critical point and the MFD of region 2 never enters the congested phase (which does happen with other regions, bringing the speed down).



<Figure 9.>

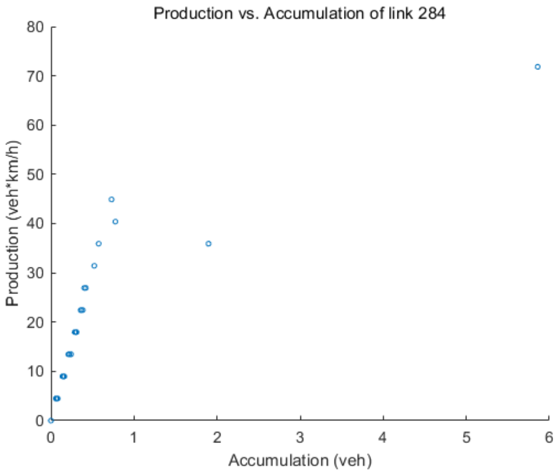
Overall, Figure 9 would suggest that this time interval is the start of a rush-hour and that the rush-hour started at around the 16 min (960 seconds) mark.

Step 4. Production vs. Accumulation

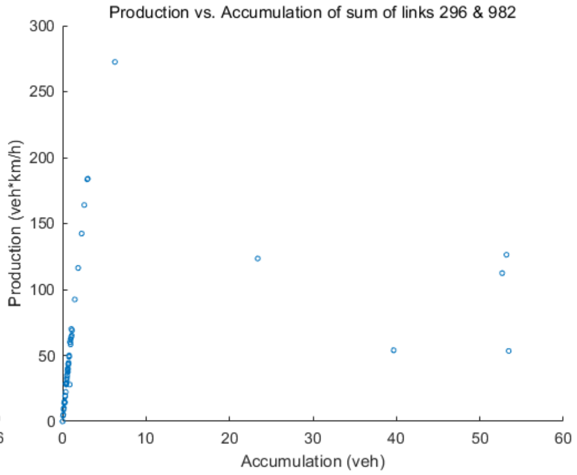
We first calculate the accumulation and the production based on the following formulas.

$$Accumulation\ (N) = density * length\ (veh)$$
$$Production\ (P) = volume * length\ ({}^{veh\, *\, km}/_{h})$$

Figure 10 depicts a scatter plot of production vs. accumulation for a randomly chosen link of 284. Figure 11 depicts the same plot for the summation of two randomly chosen links 296 and 982. Figure 12 depicts the same plot for the sum of each region cluster.

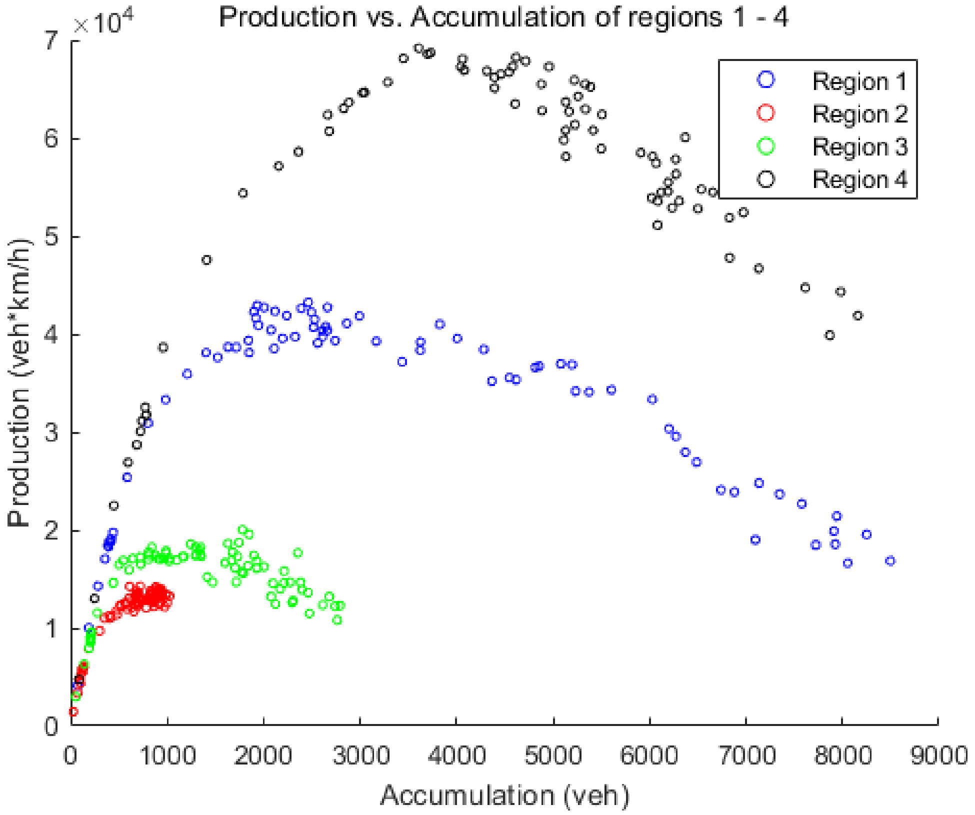


<Figure 10.>



<Figure 11.>

Figure 10 and 11 both show an almost linear line at first with some outliers to the right side of the graph. Considering the low number of examples, the outlier in Figure 10 can be just considered as random noise. On the other hand, for Figure 11, it can be assumed that the right side of the MFD is in its congested stage and therefore has lower production even with higher accumulation.

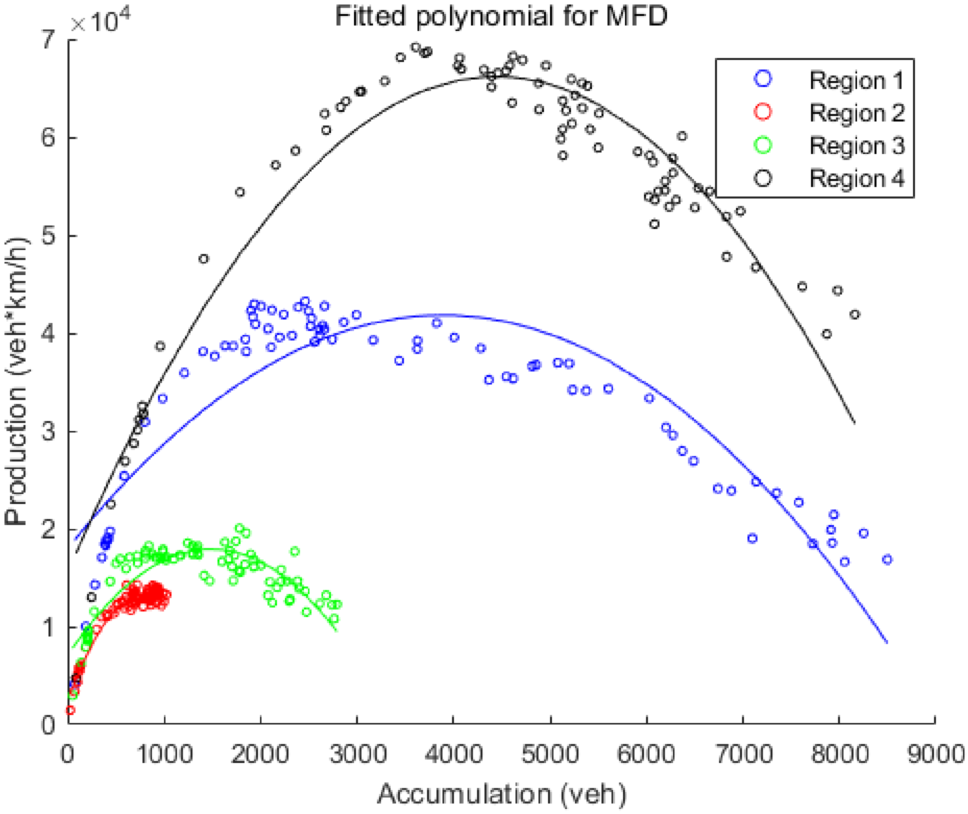


<Figure 12.>

Figure 12 shows a similar pattern to Figure 7. However, since the total accumulation varies greatly between regions 1 and 3, they are not overlapped as they were in Figure 7. Also, region 4 shows an even greater discrepancy due to the lack of scaling as well.

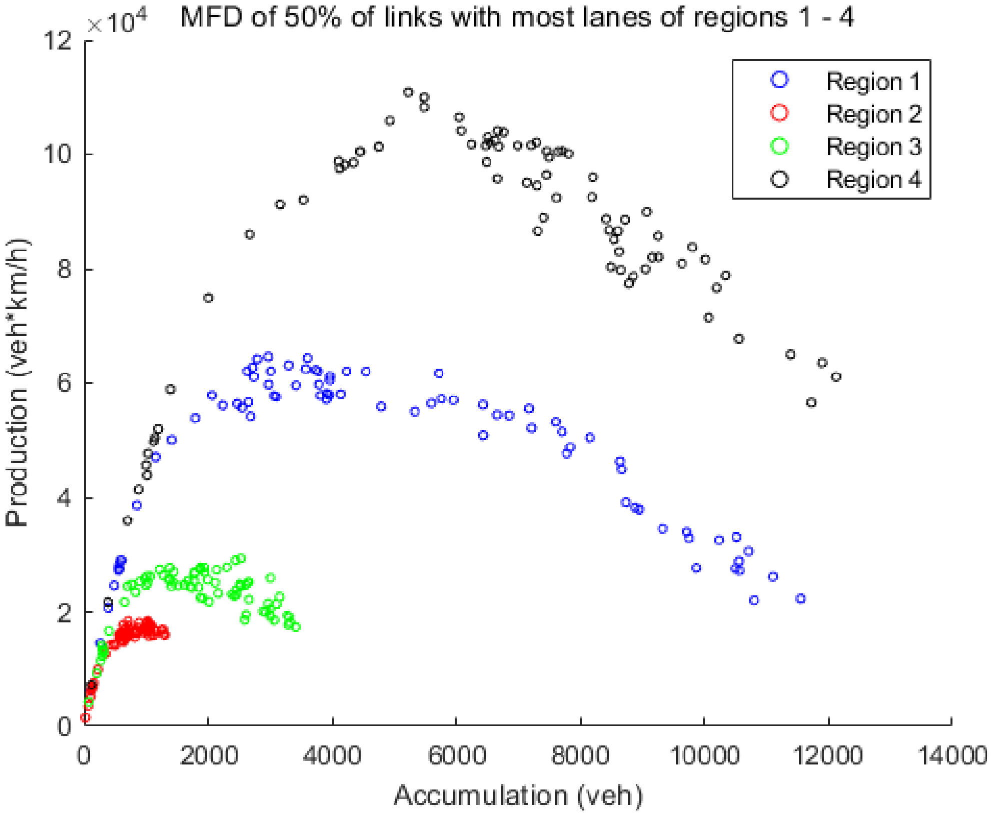
Step 5. Selecting 50% of the links

In this section we first fit polynomials for the MFD found in step 4. Then, we select only the top 50% of the links based on certain criteria and estimate the MFD of each region. Next, we compare the MFDs with the one obtained in step 4 using the fitted polynomials of each MFD. Figure 13 depicts the polynomial that was fitted to the MFD in step 4.

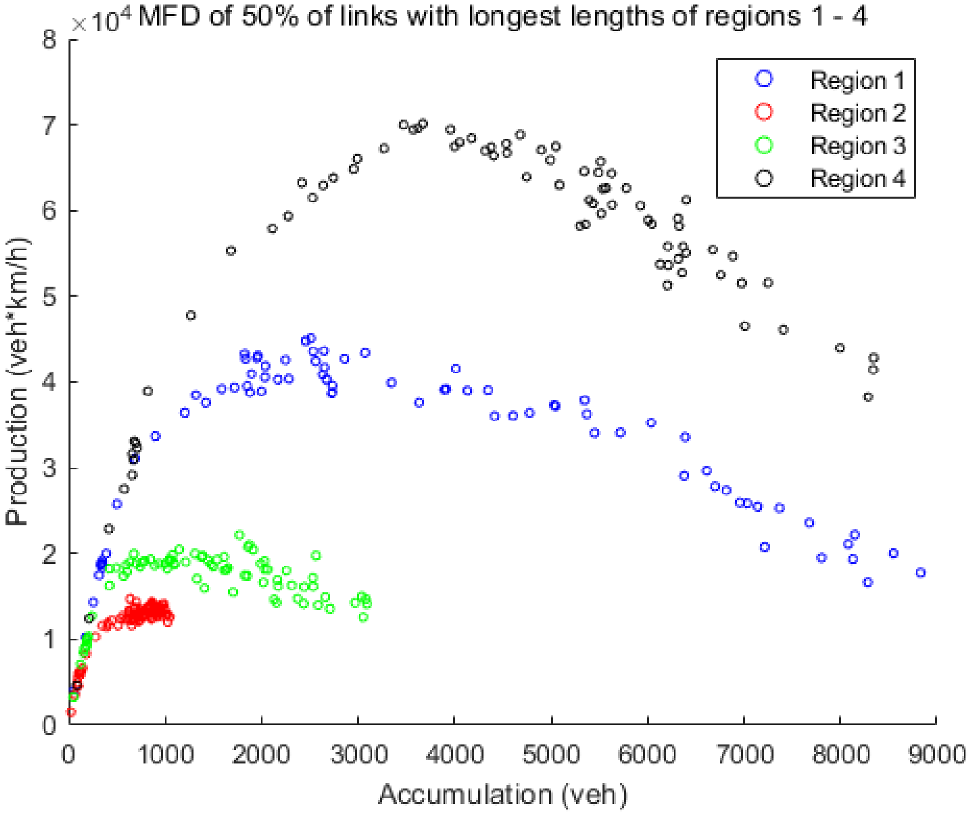


<Figure 13.>

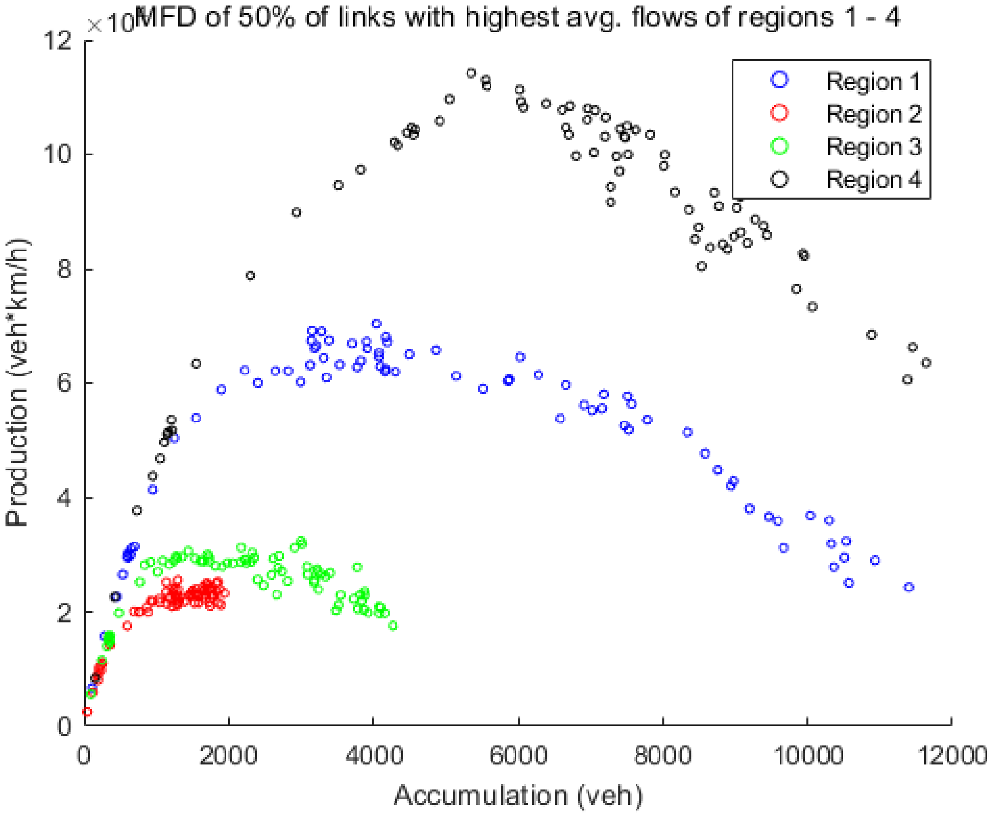
The criteria of choosing the links are as follows: i) Number of lanes; ii) Link length; iii) Average flow. The MFDs with production vs. accumulation of each region (clusters). Also, both axes of all MFDs are scaled by the following ratio: [total link length of the region / total link length of the selected sample]. Figures 14 to 16 show the MFD based on each link selection criteria. Figures 17 to 20 depict the fitted polynomial for the MFDs at each region for different link selection methods.


MFD of 50% of links with most lanes of regions 1 - 4
×10^4
Production (veh*km/h)
0
2
4
6
8
10
12
0
2000
4000
6000
8000
10000
12000
14000
Accumulation (veh)
Region 1
Region 2
Region 3
Region 4

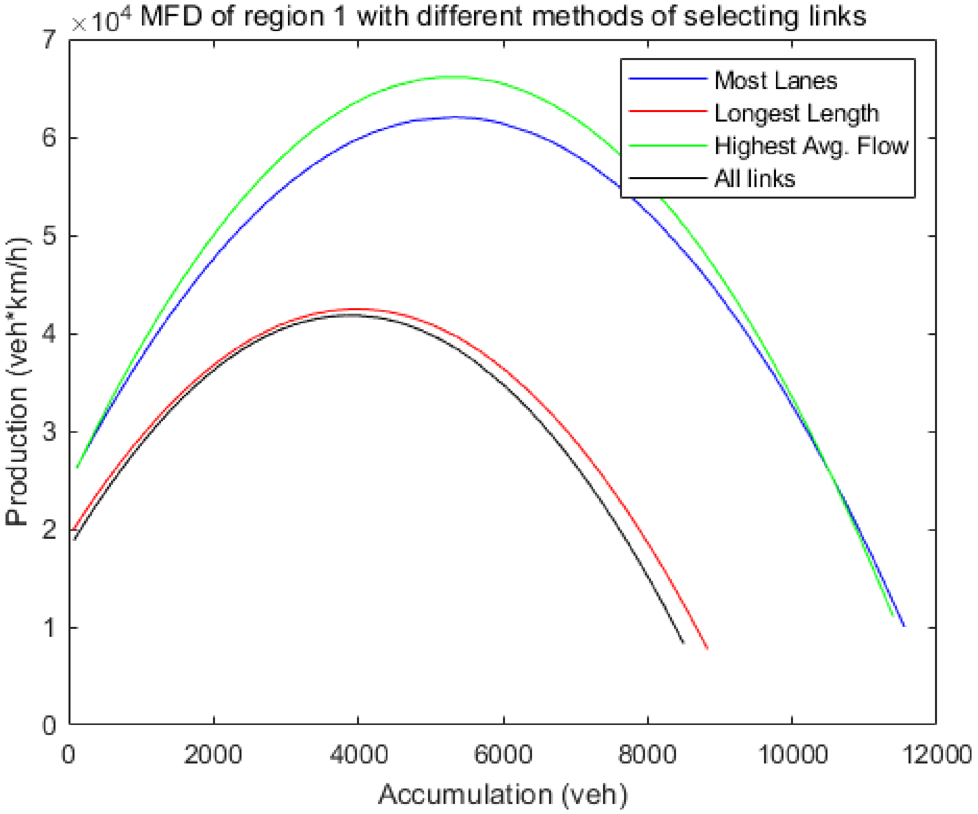
<Figure 14.>


×10^4 MFD of 50% of links with longest lengths of regions 1 - 4
Production (veh*km/h)
0
1
2
3
4
5
6
7
8
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
Accumulation (veh)
Region 1
Region 2
Region 3
Region 4

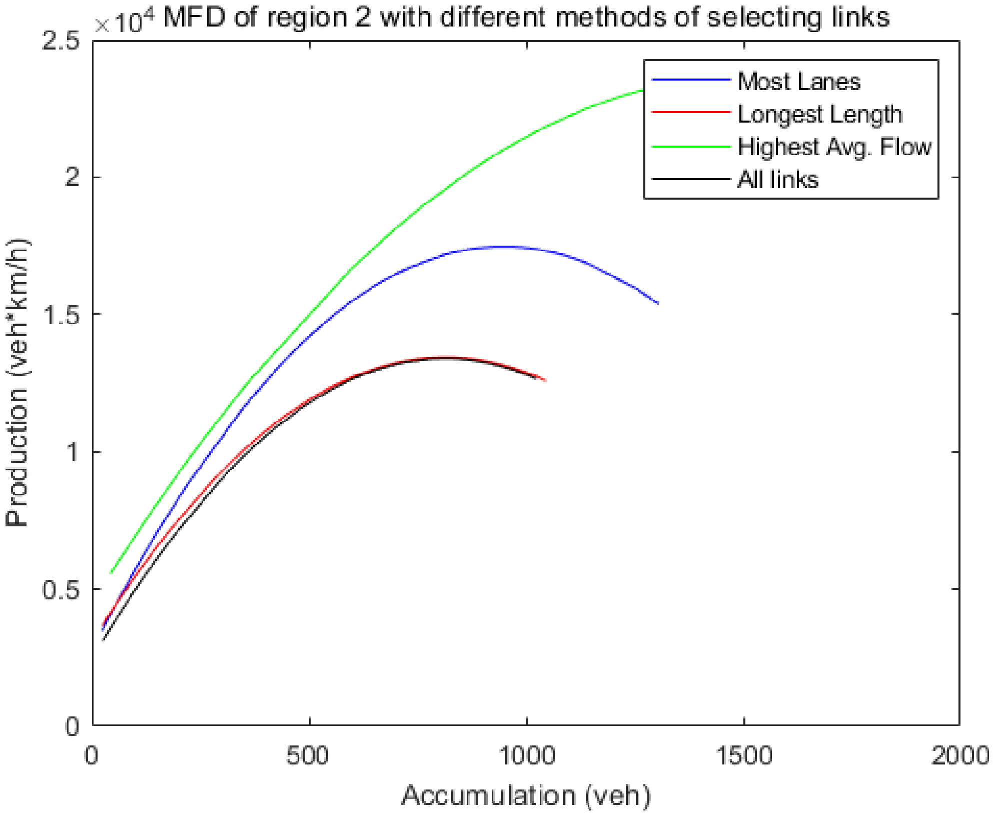
<Figure 15.>


×10
MFD of 50% of links with highest avg. flows of regions 1 - 4
Production (veh*km/h)
Accumulation (veh)
Region 1
Region 2
Region 3
Region 4
0
2
4
6
8
10
12
0
2000
4000
6000
8000
10000
12000

<Figure 16.>

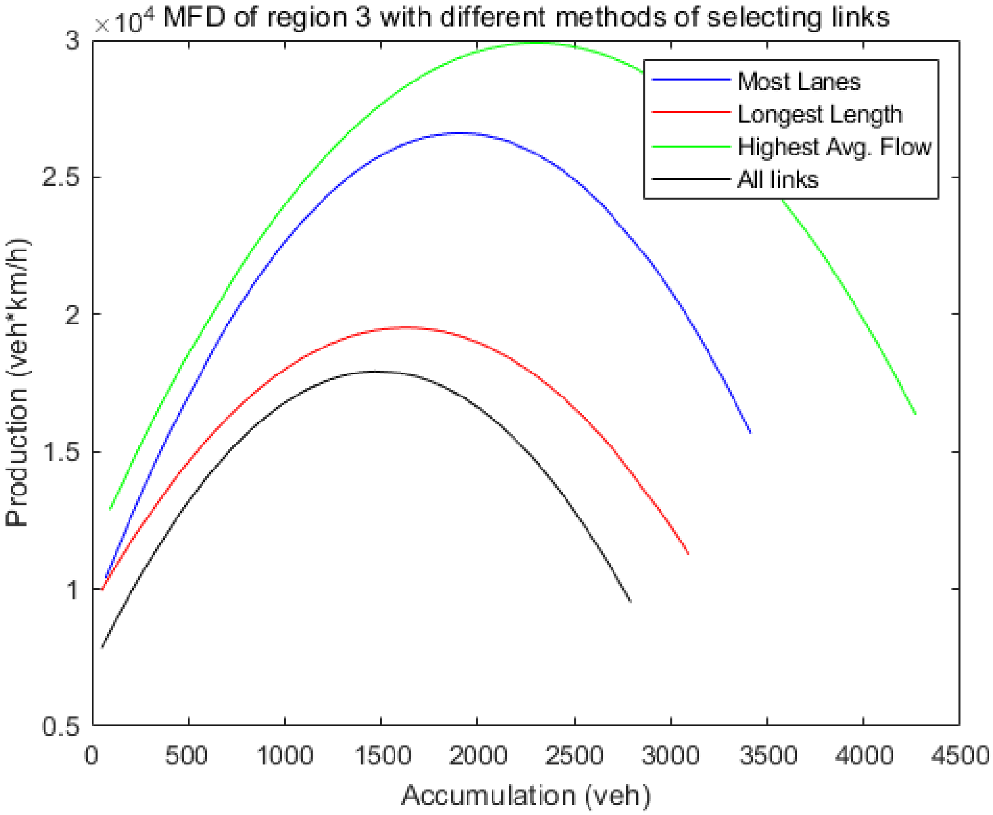

×10⁴
MFD of region 1 with different methods of selecting links
Production (veh*km/h)
Accumulation (veh)
Most Lanes
Longest Length
Highest Avg. Flow
All links
0
1
2
3
4
5
6
7
0
2000
4000
6000
8000
10000
12000

<Figure 17.>



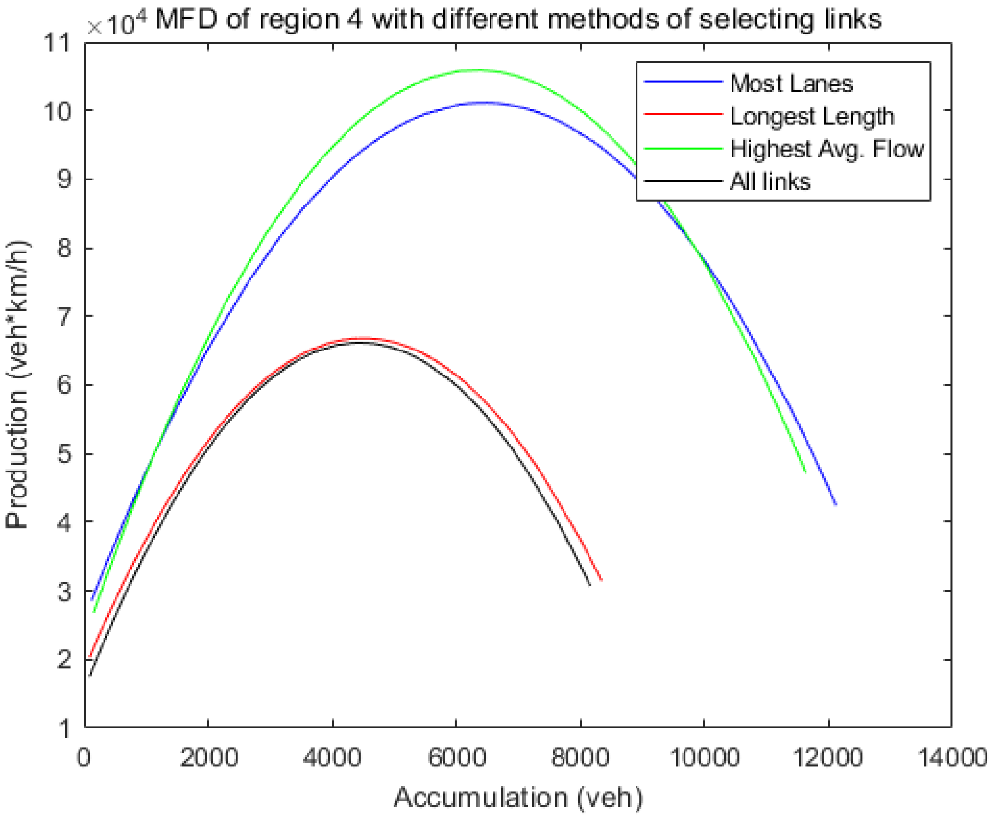
$\times 10^4$ MFD of region 2 with different methods of selecting links
Most Lanes
Longest Length
Highest Avg. Flow
All links
Production (veh*km/h)
Accumulation (veh)
0
0.5
1
1.5
2
2.5
0
500
1000
1500
2000

<Figure 18.>



$\times 10^4$ MFD of region 3 with different methods of selecting links
Most Lanes
Longest Length
Highest Avg. Flow
All links
Production (veh*km/h)
Accumulation (veh)
0.5
1
1.5
2
2.5
3
0
500
1000
1500
2000
2500
3000
3500
4000
4500

<Figure 19.>



<Figure 20.>

Throughout all regions, we can clearly see that Longest Length method of selection generally has the most similar MFD to that of the original data. Selecting links with the Most Lanes method increases the critical accumulation as well as the max capacity of the MFD since the links have more lanes to take in more vehicles at any given time.

Also, the critical accumulation for the Highest Avg. Flow method can be a bit unstable and not necessarily either lower than higher than that of the Most Lanes method. However, the capacity of the Highest Avg. Flow method is consistently higher than that of all other methods including the Most Lanes method. Reasoning is quite simple. Since the selection is done based on flow, it is a more direct way of selecting the links with the highest production (flow*travel length), even more so than selecting by Most Lanes. Therefore, it can pick up some outlier links where it has higher flow (consequently higher production) than other links with more lanes. The steeper curve of the MFD shows that the speed on these links is higher as well.

In order to measure the error of each selection method, we use the Frechet distance and measure the similarity of the fitted MFD polynomials. The Frechet distance is a measure of similarity between curves that considers the location and ordering of the points along the curves (https://en.wikipedia.org/wiki/Fr%C3%A9chet_distance). The code for calculating the distance was provided by Zachary Danziger (https://www.mathworks.com/matlabcentral/fileexchange/31922-discrete-frechet-distance). All link selection methods were tested against the original MFD of step 4 (all links) at every region for the Frechet distance. Table 4 depicts the results of the calculation.

Frechet Disance	Most Lanes	Longest Length	Highest Avg. Flow
Region 1	20290.14	1053.086	24354.58
Region 2	4074.287	571.9451	10283.02
Region 3	8684.965	2087.415	11997.7
Region 4	35031.4	2791.435	39791.59

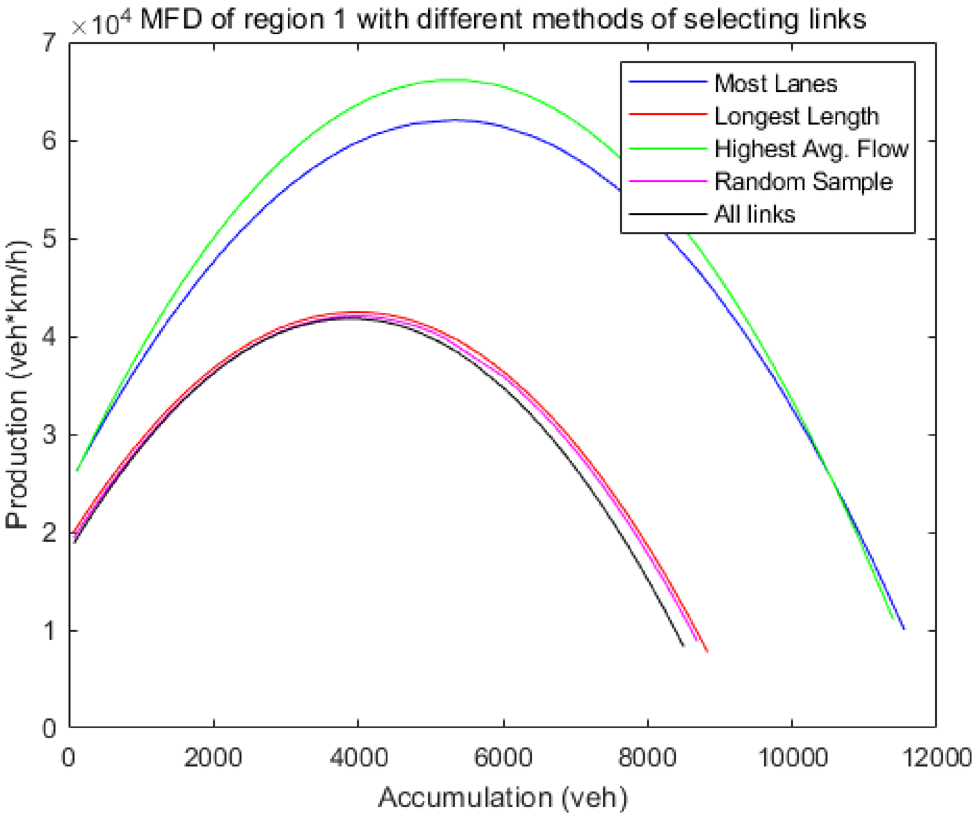
<Table 4.>

The numbers in Table 4 matches what we can tell intuitively by looking at the fitted MFD polynomial figures. The Longest Length method has by far the lowest distance and is followed by the Most Lanes method and the Highest Avg. Flow method, in that order.

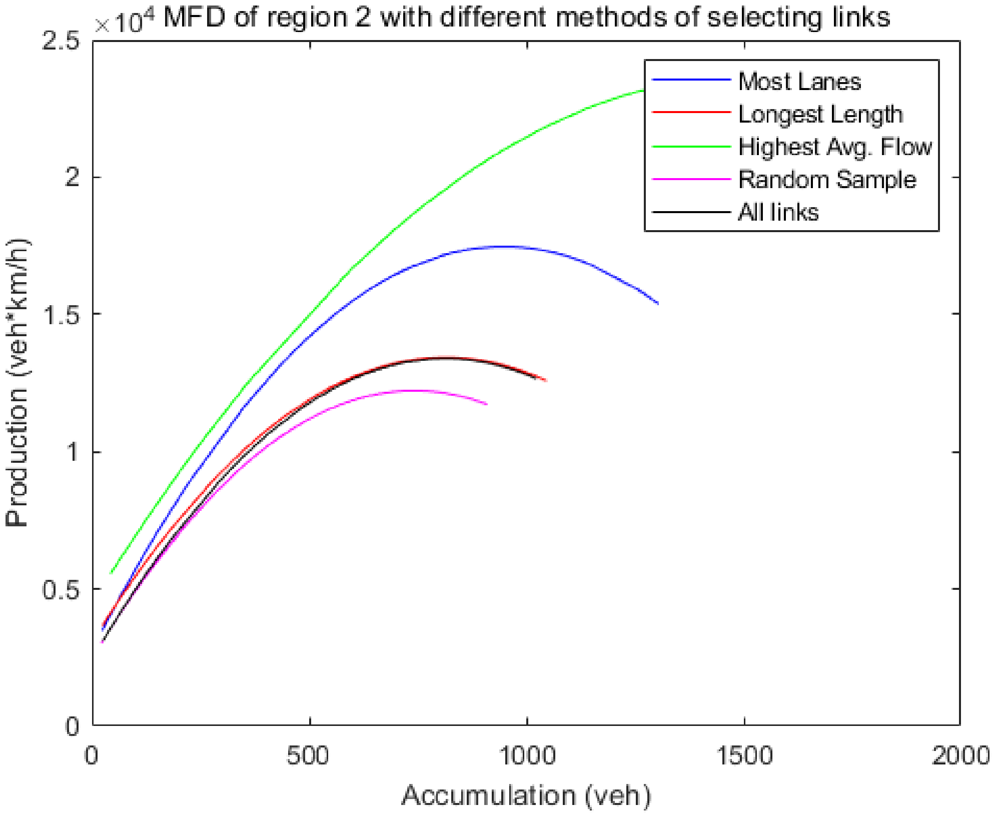
Step 6. Method of selecting 50% of the links.

The purpose of this step is to find a way to select only 50% of the links and multiplying all measurements by a factor of 2 whilst having the minimum difference in terms of the MFD with the original links data.

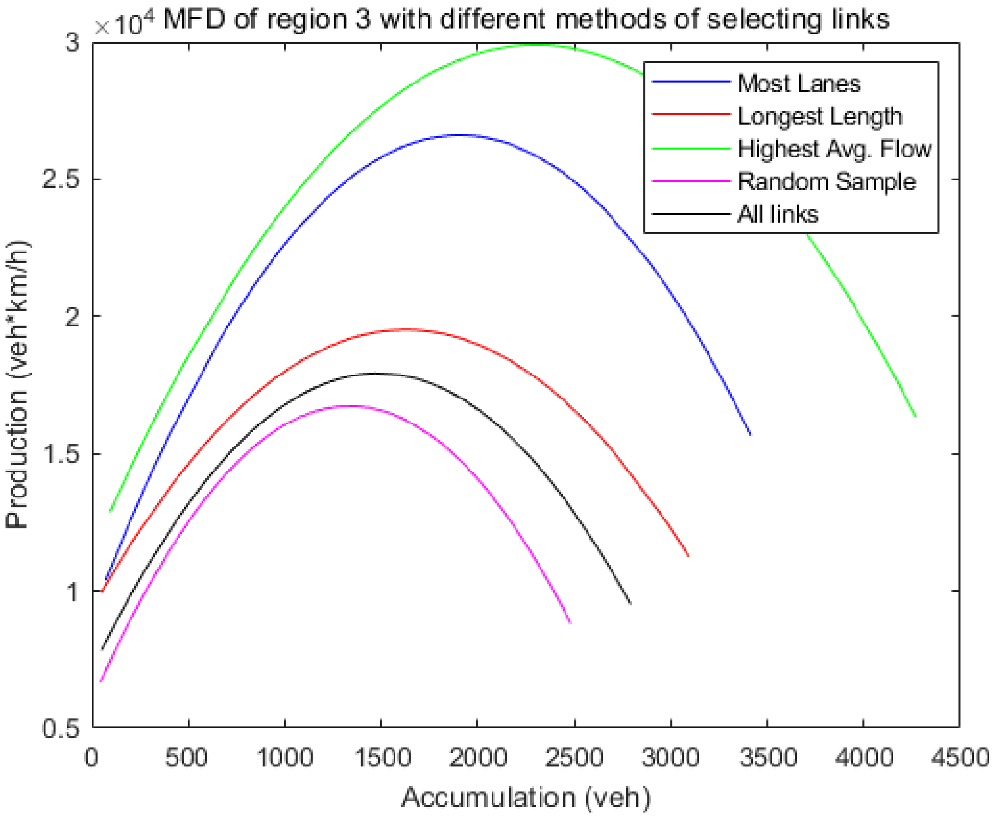
To do so, we propose the Random Selection method. As the name suggests, this method randomly selects 50% of the links with equal probabilities from the original data. Figures 21 to 24 show the results of this method. Naturally, since this method involves randomness, the results might vary per iteration. However, we have found that the variation does not meaningfully affect the conclusion we will draw.


×10^4 MFD of region 1 with different methods of selecting links
Production (veh*km/h)
Accumulation (veh)
0
1
2
3
4
5
6
7
0
2000
4000
6000
8000
10000
12000
Most Lanes
Longest Length
Highest Avg. Flow
Random Sample
All links

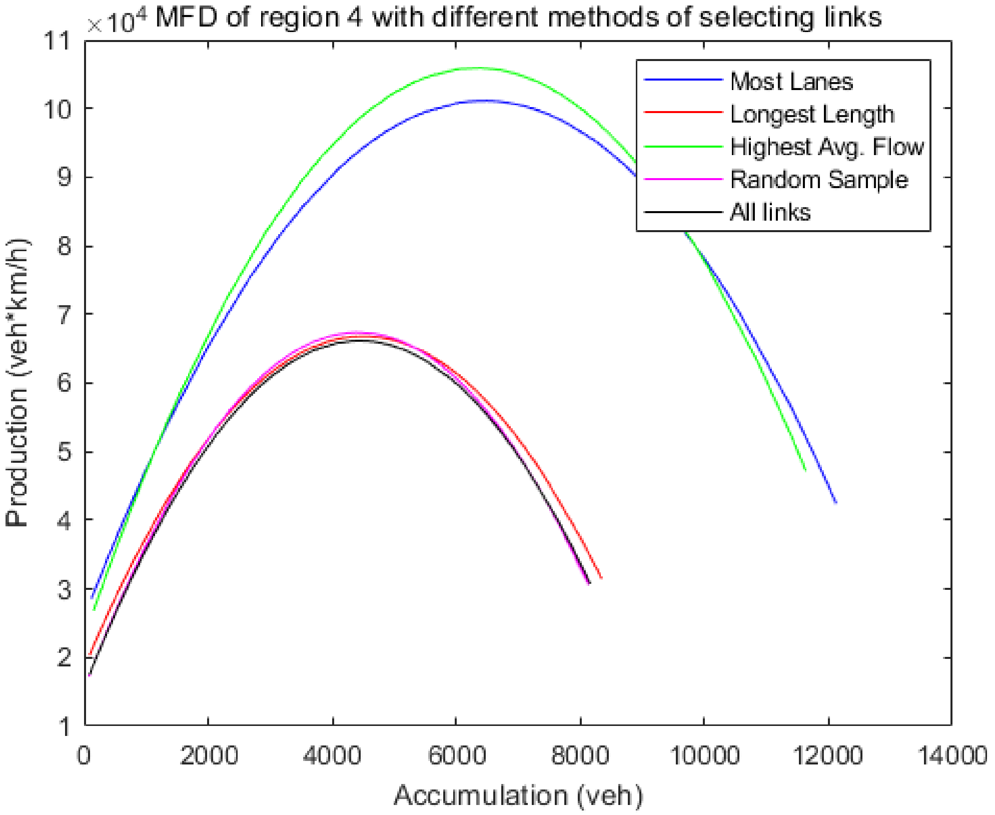
<Figure 21.>


×10^4 MFD of region 2 with different methods of selecting links
Production (veh*km/h)
Accumulation (veh)
0
0.5
1
1.5
2
2.5
0
500
1000
1500
2000
Most Lanes
Longest Length
Highest Avg. Flow
Random Sample
All links

<Figure 22.>



<Figure 23.>



<Figure 24.>

Also, Table 5 shows the Frechet distance of the Random Sample method.

Frechet Disance	Most Lanes	Longest Length	Highest Avg. Flow	Random Selection
Region 1	20290.14	1053.086	24354.58	1528.801
Region 2	4074.287	571.9451	10283.02	1172.004
Region 3	8684.965	2087.415	11997.7	1204.2
Region 4	35031.4	2791.435	39791.59	3393.016

<Table 5.>

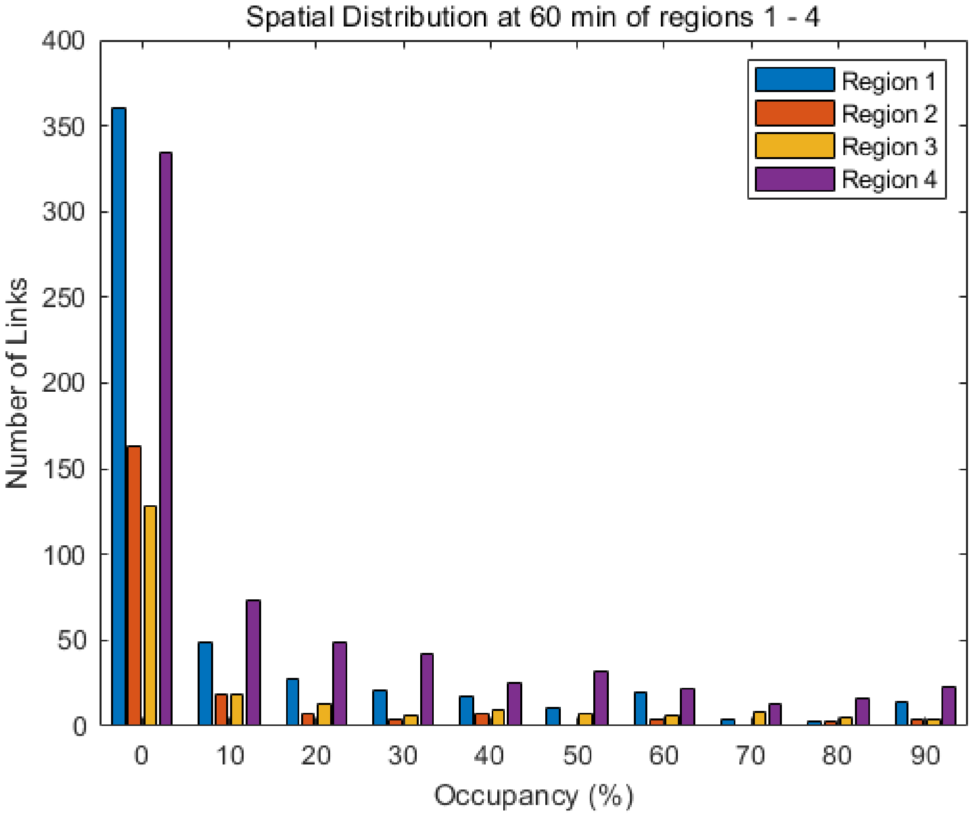
Intuitively, the random selection method should be one of the best ways to impartially select 50% of the links as it is not correlated with any other factors. And therefore, it should be the closest to the original. However, even in this case, the Longest Length method beats it out in regions 1, 2, and 4. And other iterations have shown that the Random Selection method loses at all regions quite often (and never the other way around).

This suggests that the Longest Length method is one of the best ways to select a smaller sample of the original links data. And that the links with the highest length are a great representation of the MFD for the whole region.

Additionally, few other methods have been tested. One example was taking the average of production per each region and choosing links with the minimum difference from the average in terms of production. However, none of these methods proved more effective that neither the Random Selection method nor the Longest Length method.

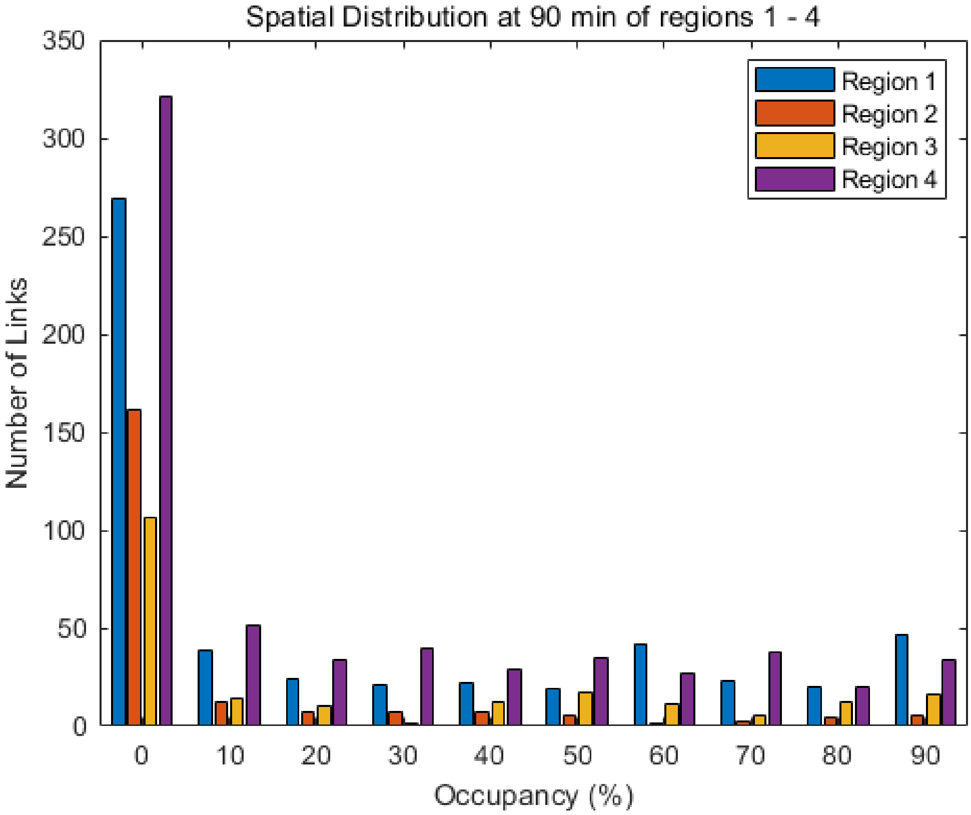
Step 7. Clustering of the links

Figure 25 and Figure 26 depict the histogram of each cluster based on their occupancy at times 60 min and 90 min. As shown on the figures, all clusters show similar patterns with most links at occupancy level of 0% - 10% and fewer links at higher occupancy levels. As for the changes from the 60 min mark to the 90 min mark, more links became more occupied. Regions 1 and 3 showed a very similar pattern where a reasonably large portion of links that were at the 0% - 10% occupancy level became more occupied and left the bin.



Spatial Distribution at 60 min of regions 1 - 4
Number of Links
Occupancy (%)
400
350
300
250
200
150
100
50
0
0
10
20
30
40
50
60
70
80
90
Region 1
Region 2
Region 3
Region 4

<Figure 25.>



Spatial Distribution at 90 min of regions 1 - 4
Number of Links
Occupancy (%)
350
300
250
200
150
100
50
0
0
10
20
30
40
50
60
70
80
90
Region 1
Region 2
Region 3
Region 4

<Figure 26.>

In order to check the effectiveness of the clustering, we calculated the total variance according to the following formula.

$$TV_n = \frac{\sum_{i=1}^{N_s}(N_{A_i} * var(A_i))}{N * var(A)}$$

As such, the total variation for no clustering would be 1. Table 6 shows the total variation of no clustering, 2 clusters (i. regions 1&4 together and regions 2&3 together, ii. regions 1&3 together and regions 2&4 together), 3 clusters (i. regions 1&4 together, ii. Regions 1&3 together), and 4 clusters. Note that the total variances in general are quite high. This is because all regions show similar behaviors and clustering does not make a huge difference.

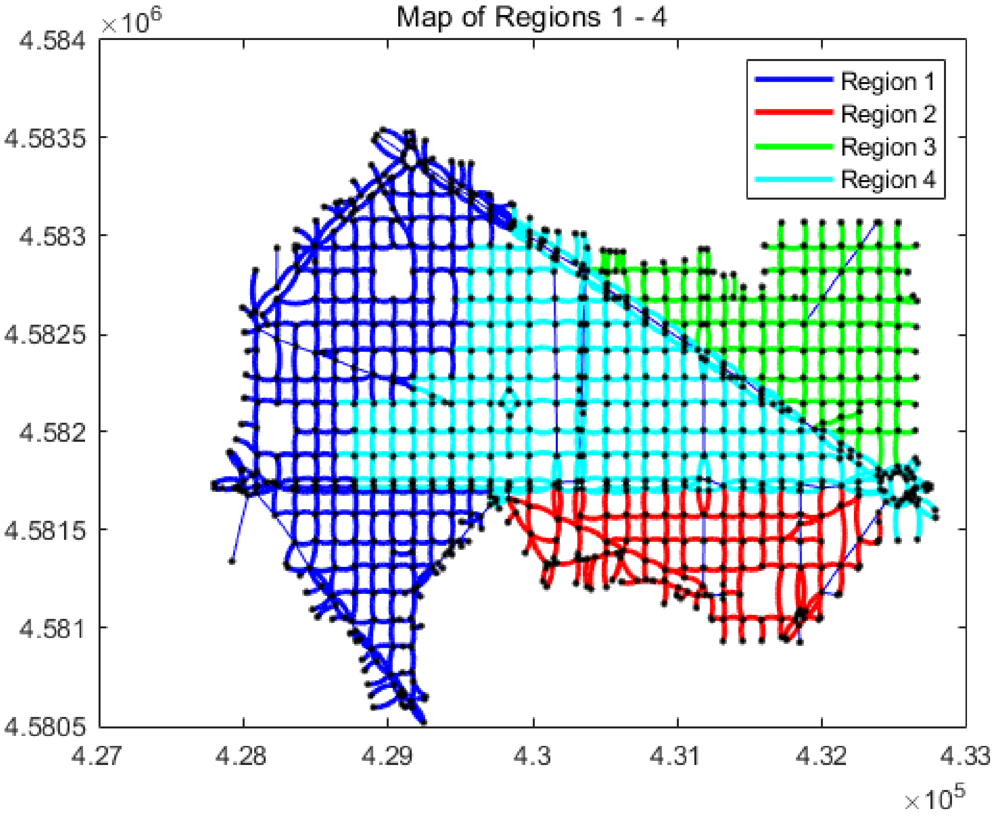
Total Variance	60 min	90 min
1 / 2 / 3 / 4 (4 clusters)	0.975736	0.971186
1 & 4 / 2 / 3 (3 clusters)	0.988875	0.97174
1 & 3 / 2 / 4 (3 clusters)	0.976674	0.970494
1 & 4 / 2 & 3 (2 clusters)	0.994587	0.989384
1 & 3 / 2 & 4 (2 clusters)	0.996766	0.991441
1 & 2 & 3 & 4 (1 cluster)	1	1

<Table 6.>

The results that the overall best clustering is either to have all 4 regions separate or to group regions 1 and 3 into one while having regions 2 and 4 be separate. Particularly, in the case of 90 min, the 3-cluster model outperformed the 4-cluster model in terms of the total variance. This confirms what we have suggested with Figure 7 in step 2, that regions 1 and 3 are so very similar that they can be merged into a single region. Additionally, as mentioned directly above, regions 1 and 3 showed similar behavior when it comes to the spatial distribution over time as well, further supporting this claim.

However, even though the calculation so far shows that regions 1 and 3 can be merged into a single region due to their similarities, this does not take into consideration the geography of these links. Figure 27 depicts the exact map of each region. According to the map, regions 1 and 3 are completely separated geographically and that they are divided by region 4. And since region 4 has sufficiently different characteristics from all other regions (highest capacity and traffic volume), we can conclude that this partitioning of regions is quite reasonable.

Additionally, this map explains why region 2 stayed on low occupancy at Figure 7 of step 2. It was because, as shown on Figure 1 – 3 on step 1, the area of region 2 (bottom right) stays low on occupancy within the simulation.


Map of Regions 1 - 4
×10^6
4.584
4.5835
4.583
4.5825
4.582
4.5815
4.581
4.5805
4.27
4.28
4.29
4.3
4.31
4.32
4.33
×10^5
Region 1
Region 2
Region 3
Region 4

<Figure 27.>