Urban Simulation Assessment

April 16, 2019

1 Part 1

1.1 Introduction

This is an analysis of resilience in the London tube network. The analysis uses a recursive function for node removal to calculate the marginal effect of a node's removal, pseudo-code for the function can be found in Appendix 1. Below, criteria for node removal and effect evaluation are discussed below.

1.2 Impact Evaluation

Because the node removal criteria was decided in the context of the network effect metric, the network effect metric will be discussed first.

Calculate community metric for network to justify focus on distance? *from igraph: cluster edge betweenness, cluster fast_greedy, cluster_label_prop, cluster_leading_eigen, cluster louvain, cluster optimal, cluster spinglass, cluster walktrap.*

Because London's tube doesn't demonstrate strong sub clusters,breaking the network into isolates was investigated but not pursued. Instead, the focus will be on forcing tube users to travel further for longer on their journeys.

This is investigated using shortest path and shortest topological path. Given the spatial nature of the london tubes, where edge attributes represent actual distances, true shortest path might be an attractive option. In the context of the London tube though, total time and effort are more important than total distance though. Trains can travel longer distances fairly rapidly while travelling through a high number of stations increases time dramatically because of the need to stop. Further, it is assumed that traveling through a higher number of stops implies a higher number of time and effort consuming train changes to switch. Thus by using geodesic, what is being maximized is the increase in stoppage time, and line change time for travelers in the network.

The statistics calculated to understand the effect of node removal include: % of connected nodes in the graph, average of the node specific statistics above,

The igraph package's mean_distance function was used to compute the average shortest path between nodes in the network. The unconnected parameter was used to specify that nodes that were not connected to the largest cluster were counted as 1 + the longest possible geodesic, the actual longest geodesic is much less than this. This demonstrates the incompatibility of different network measures. There isn't really a way to compare the effect of a longer trip with the effect of removing a trip possibility entirely. To do

that we would have to include alternate modes of transport like the bus network. The actual longest geo

1.3 Node Removal Criteria

The function was run for measures that include, degree, betweeness, topological betweenness, closeness, topological closeness and eigenvector centrality. The correlations for these values across stations can be reviewed in figure 1. It was noted that correlations between weighted and topological measures were high, indicating that the distances between tube stations are fairly consistent so that the number of stations between two stations is a decent approximation of the distance. The correlation of measures betweenness and degree is also fairly high, indicating that tube stations at the middle of a line, with higher betweenness, also tend to have multiple lines, high degree. Correlations between eigenvector centrality and the other measures was very low, indicating that this does not give the same information as other metrics. Lastly, correlation between weighted and topological eigenvector centrality was 0 **indicating a problem with how they were calculated!**.

In order to maximize the increase in travel time measured by the average length of geodesic paths, betweenness will be used to order node removals. This measure is the number of shortest paths between nodes that travel through a given node. Deleting the node with highest betweenness will force the highest number of trips to use an alternate, ostensibly longer, path through other stations.

One note about this process, deleting nodes in some places creates isolates. In this investigation **this was not an issue I hope because for the first 10 or so nodes by betweenness, they were not the only node connecting any other node to the network.**

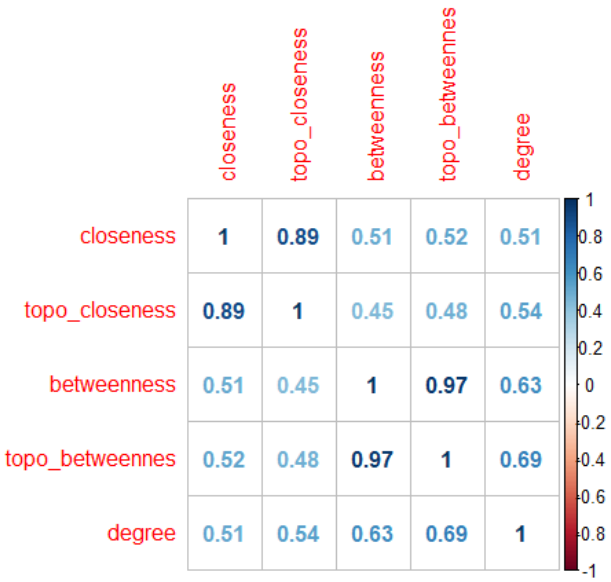

Figure 1: Correlation between station/node metrics

index	NodeDeleted	IncreaseGeodesic	Components
1	Green Park	0.464271515336074	1
2	King's Cross St. Pancras	30.9672711339296	2
3	Bank	1.01101854695752	2
4	Waterloo	1.13665048051966	2
5	Stockwell	17.7264111534033	4
6	Embankment	108.61606869706	5
7	Baker Street	2.77460558311449	6
8	Notting Hill Gate	30.6706717614217	7
9	Ealing Common	26.6151971068078	9
10	Stratford	10.5750849462714	10
11	Canning Town	5.34435028248589	12
12	Hammersmith	8.37777866974866	14
13	Shadwell	7.82942087022292	16
14	Harrow-on-the-Hill	2.89045216902605	18
15	Camden Town	2.37783237317933	20
16	Canary Wharf	1.95898061029061	23
17	Mile End	0.964285951026795	24
18	Paddington	0.182272933085244	28
19	Earl's Court	0.109301944667777	31
20	Oxford Circus	-0.440105209296462	33
21	Woodford	-0.425293708021684	34
22	Aldgate East	-0.627897731687995	35
23	Finsbury Park	-0.34940346537536	38
24	Northfields	-0.418389327214584	39
25	Wembley Park	-0.540097783642864	41
26	North Acton	-0.588994666477163	43
27	Upney	-0.5890868871723	44
28	Rayners Lane	-0.598069582065818	46
29	Liverpool Street	-0.601393755502556	48
30	London Bridge	-0.597733955470574	50

Betweenness unconn is false			
	Node Deleted	Increase in Avg Geodesic	Components
1	Green Park	0.464271515336074	1
2	King's Cross St. Pancras	30.9672711339296	2
3	Bank	1.01101854695752	2
4	Waterloo	1.13665048051966	2
5	Stockwell	17.7264111534033	4
6	Embankment	108.61606869706	5
7	Baker Street	2.77460558311449	6
8	Notting Hill Gate	30.6706717614217	7
9	Ealing Common	26.6151971068078	9
10	Stratford	10.5750849462714	10
11	Canning Town	5.34435028248589	12
12	Hammersmith	8.37777866974866	14
13	Shadwell	7.82942087022292	16
14	Harrow-on-the-Hill	2.89045216902605	18
15	Camden Town	2.37783237317933	20
16	Canary Wharf	1.95898061029061	23
17	Mile End	0.964285951026795	24
18	Paddington	0.182272933085244	28
19	Earl's Court	0.109301944667777	31
20	Oxford Circus	-0.440105209296462	33
21	Woodford	-0.425293708021684	34
22	Aldgate East	-0.627897731687995	35
23	Finsbury Park	-0.34940346537536	38
24	Northfields	-0.418389327214584	39
25	Wembley Park	-0.540097783642864	41
26	North Acton	-0.588994666477163	43
27	Upney	-0.5890868871723	44
28	Rayners Lane	-0.598069582065818	46
29	Liverpool Street	-0.601393755502556	48
30	London Bridge	-0.597733955470574	50

	By Betweeness		
	Node Deleted	Increase in Avg Geodesic	Components
1	Green Park	0.464271515336074	1
2	King's Cross St. Pancras	0.344894786795857	2
3	Bank	1.14339690184063	2
4	Waterloo	1.28494742196984	2
5	Stockwell	0.563343124925417	4
6	Embankment	-6.16772972588596	5
7	Baker Street	2.32031653445306	6
8	Notting Hill Gate	-1.73581862937826	7
9	Ealing Common	-2.85804986154228	9
10	Stratford	-0.357026106908554	10
11	Canning Town	0.336399630312155	12
12	Hammersmith	-1.11269563661233	14
13	Shadwell	-1.44335210818291	16
14	Harrow-on-the-Hill	-0.489621630559977	18
15	Camden Town	-0.565504186081095	20
16	Canary Wharf	-0.335125486278896	23
17	Mile End	-0.846532555502438	24
18	Paddington	-0.328322619604598	28
19	Earl's Court	-0.196290601211155	31
20	Oxford Circus	0.016453621763357	33
21	Woodford	-0.056258058828067	34
22	Aldgate East	-0.000392838426682	35
23	Finsbury Park	-0.101735403043495	38
24	Northfields	-0.213720988377057	39
25	Wembley Park	0.008126154915853	41
26	North Acton	-0.075498287049283	43
27	Upney	-0.183419855551612	44
28	Rayners Lane	-0.120649861972213	46
29	Liverpool Street	-0.09859581775937	48
30	London Bridge	-0.079075958422429	50

	By Eigenvector Centrality		
	Node Deleted	Increase in Avg Geodesic	Components
1	Embankment	0.124797831125548	1
2	Cannon Street	-0.020785011623886	2
3	Moorgate	0.221126748710891	2
4	West India Quay	0.0217058916729	2
5	Great Portland Street	0.114088504753353	2
6	Farringdon	0.032692113174246	3
7	Paddington	-0.069066969738513	4
8	Leicester Square	0.031071992674539	4
9	Heron Quays	-0.213633720509144	5
10	Gloucester Road	0.404000167032718	5
11	Euston	-0.232471150322526	6
12	Aldgate	0.028208599848918	6
13	St. James's Park	0.034699837369548	6
14	Mile End	0.776660400757532	6
15	Oxford Circus	0.508933395700687	6
16	Notting Hill Gate	2.80231756571204	7
17	Rotherhithe	0.044124240677434	7
18	Blackfriars	0.001798634055188	8
19	Baker Street	-6.40337680481831	12
20	Barons Court	-0.217740790392554	13
21	Aldgate East	0.046883136212973	14
22	Ruislip Manor	-0.0251731067831	15
23	Blackwall	0.07893822023436	15
24	King's Cross St. Pancras	-0.928512463963177	18
25	Island Gardens	0.016339374551825	19
26	West Ham	0.366915223830574	21
27	Holborn	0.014000633210076	24
28	Waterloo	0.86009000361571	24
29	Victoria	-0.609084466226621	25
30	Custom House	-0.385377841072922	26

	By Eigenvector Centrality		
	Node Deleted	Increase in Avg geodesic	Components
1	Embankment	0.124797831125548	1
2	Cannon Street	5.66059665069182	2
3	Moorgate	0.215731180123939	2
4	West India Quay	0.02021050051145	2
5	Great Portland Street	0.110753998811909	2
6	Farringdon	1.91439512290877	3
7	Paddington	23.3379220818089	4
8	Leicester Square	0.017700366401854	4
9	Heron Quays	15.4821648982723	5
10	Gloucester Road	0.319473942355302	5
11	Euston	34.9321449605953	6
12	Aldgate	-0.027020815846441	6
13	St. James's Park	-0.022791930367916	6
14	Mile End	0.500638234735106	6
15	Oxford Circus	0.310169191636348	6
16	Notting Hill Gate	3.48423397547472	7
17	Rotherhithe	-0.022978065465537	7
18	Blackfriars	-1.59050475761802	8
19	Baker Street	90.790225864921	12
20	Barons Court	5.840627408535	13
21	Aldgate East	1.495445128136	14
22	Ruislip Manor	1.32642824899307	15
23	Blackwall	-0.342224789480696	15
24	King's Cross St. Pancras	15.4170099145231	18
25	Island Gardens	-1.07286413770476	19
26	West Ham	9.42385307392772	21
27	Holborn	2.77304884203235	24
28	Waterloo	-0.370341481560303	24
29	Victoria	5.9845029956258	25
30	Custom House	2.3713875848423	26

1.4 Analysis

When Kings Cross is deleted, it creates a new unconnected component out of the 11 stations on the north east end of the Picadilly line. In igraph, the two ways to handle this for the mean_distance() function are either to exclude distances between those unconnected nodes and the rest of the network or to assume that the distance is one greater than the longest possible geodesic in the network, that is ,the number of nodes on the network. Data is included for both options.

Looking at the effect data it seems reasonable to say that betweenness did a better job than eigenvector centrality of prioritizing nodes to remove. Using betweenness created

more isolates. It's difficult to judge which method lengthened average shortest path the most because of the options for dealing with disconnected networks. This will be discussed in the conclusion.

1.5 Conclusions

To improve this work, it would be good to add data about transportation networks besides the underground. In particular information about bus routes connected nodes would be useful because it would allow for a better estimate of average shortest path when subway stations become disconnected as the shortest path could then go through a bus route instead. Similarly, it would be good to include more granular data about where a rider would have to change trains. The current network assumes there's no cost to switch trains relative to staying on the same train passing through a station. Anyone who has walked from the Picadilly line to the Northern line at Kings Cross knows that there is a big difference.

An improvement to the data would be to use travel time data instead of using distance as an approximation.

Lastly, it would be interesting to build an igraph function that can compute average shortest path using edge weights since the current function cannot. This could confirm or reject the thought that tube stations are spaced fairly regularly based on the high correlation between weighted and topological centrality measures.

2 Part 2

2.1 The Models

2.1.1 Unconstrained

The model is constrained to the total flows of the system but flows out of an origin and into a destination can be any value between 0 and total system flows.

This is useful for studying the change in connectivity between regions, for instance if a new transportation link was built. In particular, it is useful for studying long term effects of a change where residence and employment are more flexible.

2.1.2 Production

The direction of flows can change but the total flows from each origin will remain constant. This is useful for studying the effect of a new employment or consumption location that changes the destinations of people going to work or to spend money. In terms of the matrix, it implies that the sums of the rows of the matrix are constant.

2.1.3 Attraction Constrained

The source of flows into a region can change but the total flows into a region will remain constant. That is, any reduction in flows into a destination from another region will be fully replaced by flows into the destination from another region. This could be used to study a new housing development that pulls people into residence in a different part of an area or a natural disaster that forces residents out of an area. Employers outside the area still need workers but will not be able to draw them from the same places after a housing change or natural disaster. In terms of the matrix the sums of the columns are held constant. Additionally, it can be used to study the effect of a specific change to employment where the model can be constrained to the values that result from that change.

2.1.4 Doubly Constrained,

Doubly constrained models could be used to test the short term effects of a change to transportation networks given that homes and businesses won't relocate but flexible behavior patterns like shopping could change almost immediately due to the change in accessibility or travel times between locations. In this model, the sums of both the columns and rows are held constant.

2.2 The Parameters

Sensitivities to origin attributes, destination attributes, and linkage attributes

2.3 A Scenario

The scenario used for this assessment will be: What if teleportation was invented and dramatically reduced travel times but could only be used in the outer boroughs because of the need to avoid teleporting through tall buildings. Thus origin and destination attributes remain constant but the travel costs change dramatically in the most peripheral boroughs: Hilingdon, Harrow, Barnet, Enfield, Waltham forest, Redbridge, Havering, Bexley, Bromley, Croydon, Sutton, Kingston, Richmond, Hounslow.

This will be investigated using a doubly constrained model to estimate the short term effects where residences and businesses cannot relocate and a total constrained model to see the long term effects on business and residence locations as a result of the incredible new discovery.

I selected a subset of the london Boroughs that included the main business areas, WEstminister, the city, and Camden, as well as the outermost boroughs

idea: pick out the main business areas, city, westminister, camden

and

Then do a production constrained model about what would change if salaries in one of the urban boroughs went up, or if a new transit line went in.

select a scenario and explore the consequences of varying model parameters and inputs on interaction flows and the origin/destination estimates

One thing noticed is that as the number of constraints goes up the R^2 goes up.

3 Part 3

define, compare and contrast Cellular Automata and Agent Based models Compare CA and ABM

While Cellular automata models are often viewed as separate modeling techniques, a more modern view of these methods considers them to be cousins or related in the sense that all CA are a subset of ABM.

Expanding a CA model usually results in the creation of an ABM. Cellular automata models are defined by a set of homogeneous cells that interact only with each other according to a defined set of input/output functions, e.g., if a cell with value 1 is surrounded by other cells f value 1 it's value becomes 0. The models can be extended to "n" states but all cells should be capable of reaching all n states to maintain the homogeneity of the cells. This can be contrasted with ABM where cells can have infinite heterogeneity and future states can be functions of the "environment" other cells or agents" and the cells own state.

The simplicity of CA models make them very useful for studying mathematical processes

whereas Agent Based Modeling flexibility make them useful for modeling more complex "real world" phenomena and make them more accessible to non-technical audiences. Often the value of ABM comes from the ability to conduct parameter sweeps, to study combination rules that apply to multiple conceptual processes with different parameter values. The value of cellular automata models tends to be focused on the effect of initialization states on the long term outcomes and equillibria of the model e.g. for the same model, outcomes are a function of the rule set and initialization values, where as agent based models tend to be calculated for a large number of initialization values in order to study the effect of model dynamics independent of initialization values that may not accurately reflect the real world.

vary model parameters to construct 3 scenarios, describe them, find time required to reach steady state, minimum runs required for statistically meaningful results.

3.1 Scenarios

This will be an investigation of the behavior of an epidemic across different population types. The baseline population scenario will have mid-levels of immunity and recovery chances(50%). The high immunity scenario will have immunity chance at 90% and recovery chance at 10% while the high recovery chance scenario will set recovery chance to 90% and immunity chance to 10%.

3.2 Assumptions and Preliminary Investigation

To understand a reasonable context for the scenarios above, behavior space was used to simulate a variety of initial-people and the number of people infected. Focusing on the three scenarios allowed the investigation to avoid a full parameter search of the effect of the 4 dimensions (initial population, number of turtles infected, immune-chance, and recovery-chance) on the final percentage of turtles infected.

The way the model is written, the actual number of people infected is the % population immunity multiplied by the number of people infected because if the model selects a turtle at random to infect that is immune, there is no infection. This for a population half immune, on average infecting 50 turtles with result in 25 turtles actually becoming infected.

Figures 2, 3, and 4 show the percentage of population infected over time for the three scenarios. The different lines represent the averages of 5 runs for 60 different combinations of initial population and number infected. Population size ranged from 100 to 1000 incremented by 100. Number infected ranged from 1 to 51 incremented by 10. It can be seen below that results are consistent within scenarios. Roughly the same trend and final steady state is reached by all lines within a chart. It was checked that this was true for individual runs as well with no information lost by averaging the five trials. There were no "outlier" trials.

It is notable that while the baseline and high recovery scenarios reached clear steady

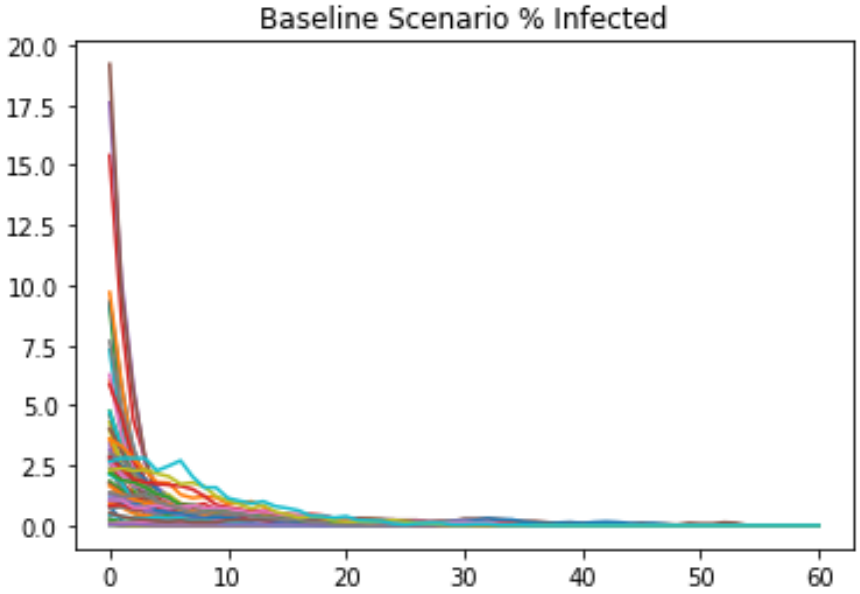

Figure 2: % of Population Infected Over Time

states in all cases, there were cases in the high Immunity scenario where the percentage of people infected over time continued to fluctuate after step 6 by which time all trials in the other scenarios were 0. There were not trials were the percentage of people infected increased meaningfully though.

Now add charts of distribution of outcomes for Immunity scenario.

Considering figures 2 through 7, it's clear that the baseline and high recovery scenarios reach steady states where no agent is infected. This is the only true steady possible in the model since no additional infections can occur without an infected agent present. In the case of the high immunity scenario, at 60 ticks, there were still trails where a meaningful proportion of the population was infected.

Also it is notable that all steady states identified in the trials are reached comfortably before 60 ticks, so this time limit will be used in the investigation.

Looking at figure 8 it seems that at 60 ticks. a higher proportion of turtles were infected at higher population sizes. This indicates that population density of turtles without immunity is a key consideration in infection rates and that immune turtles perhaps should be thought of simply as not existing for the purposes of the mode.

Thus in light of the findings above, the actual investigation uses a population of 1000, across 60 ticks.

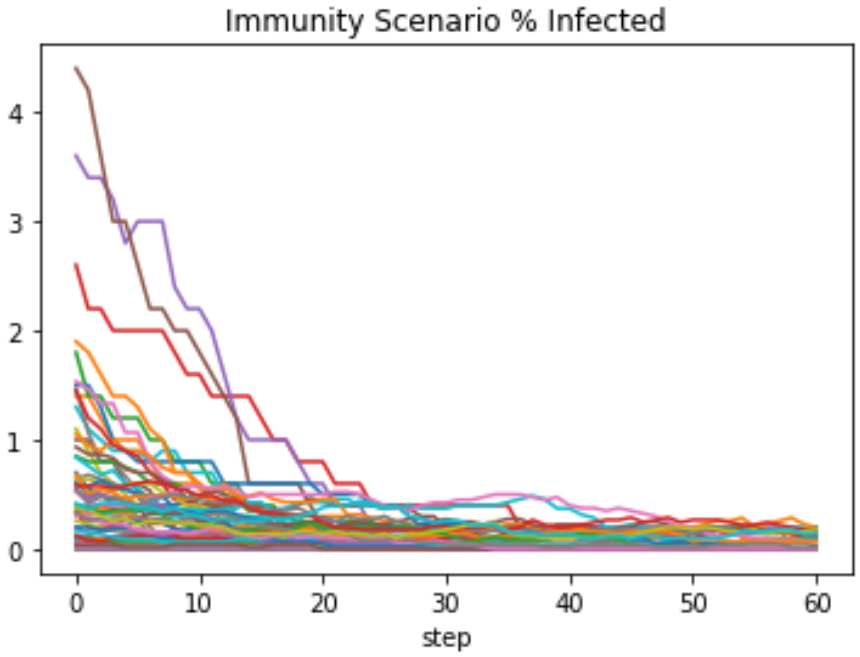

Figure 3: % of Population Infected Over Time

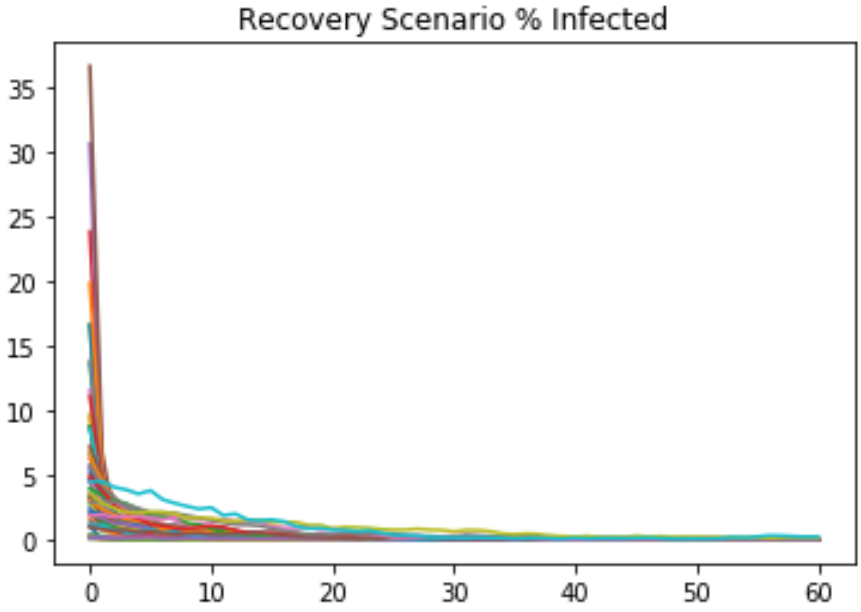

Figure 4: % of Population Infected Over Time

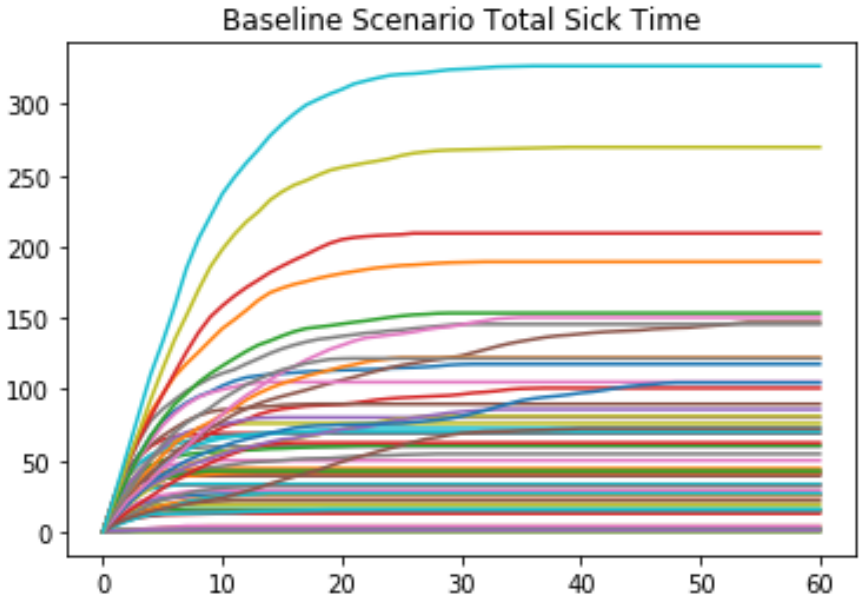

Figure 5: Total Time Sick Normalized for size of Turtle Population

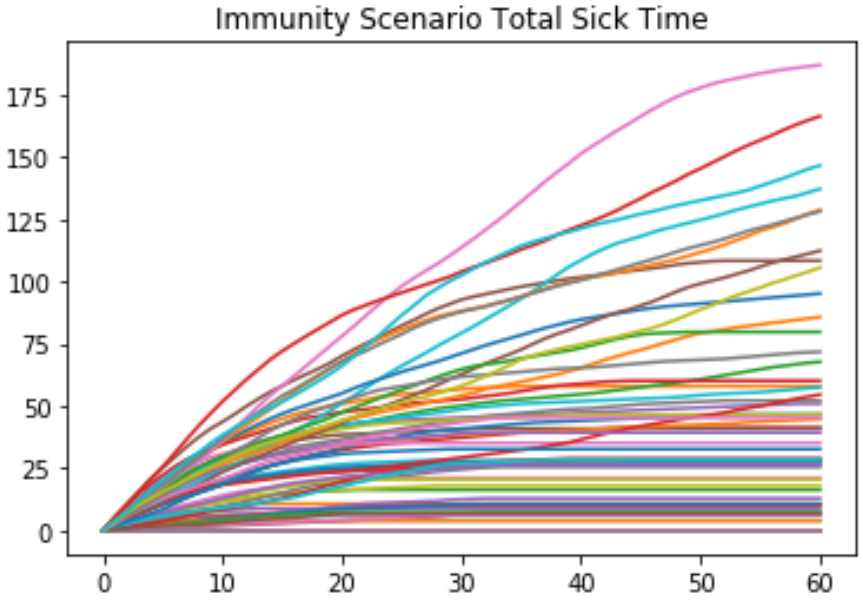

Figure 6: Total Time Sick Normalized for size of Turtle Population

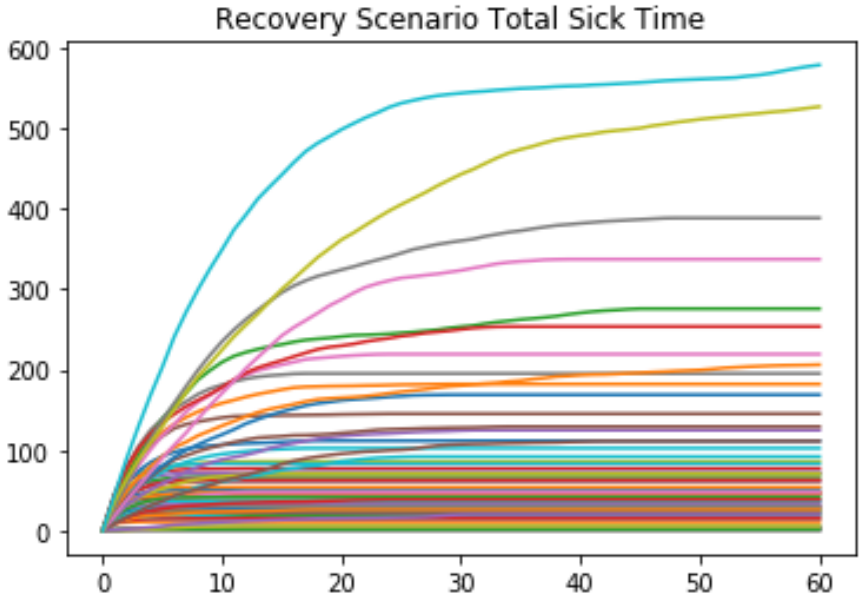

Figure 7: Total Time Sick Normalized for size of Turtle Population

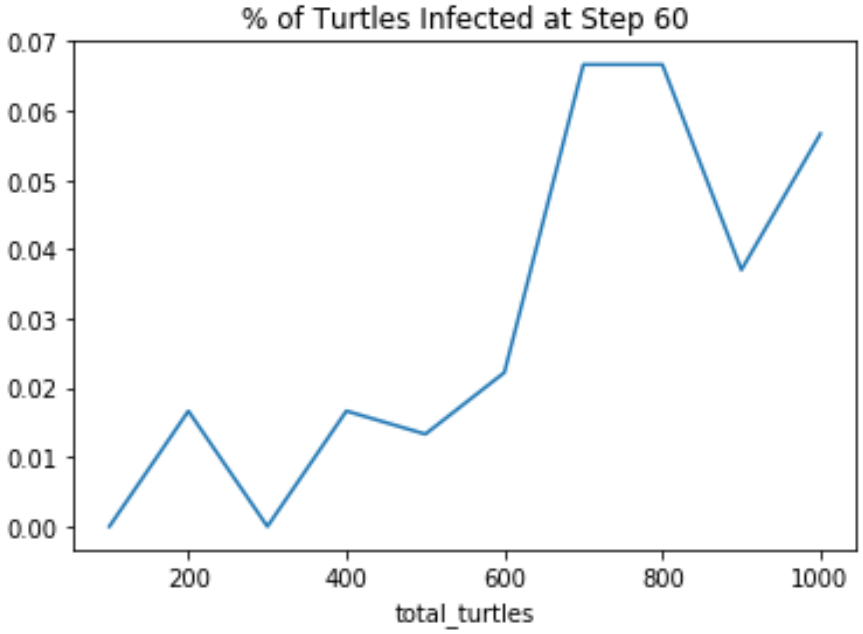

Figure 8: Trend in infection rates across population size

3.3 Analysis

The overall questions investigated here is, "Is there a meaningful difference in disease transmission between populations with high immunity, high recovery rates, and populations with moderate levels for each.

To investigate this, a t-test is used to determine whether the difference between % infected at tick 6 for each of the difference scenarios is statistically significant.

Looking at the Baseline scenario with 50% chance of immunity and 50% chance of recovering, trails consistently found the 0 infection steady state and it can be said that there isn't a meaningful difference between outcomes in this state.

Comparing those two scenarios with the high immunity scenario, there is a noticeable difference. A t-test confirms that there is a significant difference between the two distributions.

t test

Important to note is that the distribution of outcomes is not close to normal, mainly because there cannot be negative infection rates, trimming the distribution at 0 and indicating that more sophisticated statistical methods are necessary.

possibilities for t test

- test for difference between average % infected between scenarios
- test for difference between averages of % infected at different ticks
- test for difference in outcomes between initial population sizes or number infected.

3.4 Questions

Scenarios

- Baseline
- high immunity
- high recovery

Record the run by using Behavior space ? Change script so it infects people automatically?

How is this going to be 1000 words?

`use pseudocode`

XXXX words excluding headings, figures, and references.

Do I need to cite anything?

References

Grimm, Volker et al. (2010). "The ODD protocol: A review and first update". eng. In: *Ecological Modelling* 221.23, pp. 2760–2768. ISSN: 0304-3800.

Tasseron, G. and K. Martens (2017). "Urban parking space reservation through bottom-up information provision: An agent-based analysis". In: *Computers, Environment and Urban Systems* 64, pp. 30–41. ISSN: 01989715.