Optimizing COVID-19 Sewage Surveillance by Mixed Integer Linear Programming

Kristy Luk

Received August 21, 2023
Accepted July 03, 2024
Electronic access July 15, 2024

Hong Kong researchers collected sewage samples from different districts to detect COVID-19 and identify risky areas. Thus, our objective is to build a math model using mixed-integer linear programming to find optimal pathways for sewage sample collectors to travel between districts to minimize the uncertainty that COVID-19 remains undetected in the city and the traveling time of the pathway chosen. Our method was to mathematically formulate the model with the two parameters discussed above using Shannon entropy and the traveling salesman problem. The Shannon entropy is the "uncertainty" that COVID-19 pathogens remain undetected in Hong Kong. Collectors must choose the districts to travel to that minimize the collective Shannon entropy of the city the most. The Traveling Salesman Problem calculates the optimal pathway for collectors to travel to visit the maximum number of districts in minimal time. We ran two tests: the first to determine the effectiveness of Shannon entropy as a parameter compared to other optimization methods, and the second to determine the optimal pathways provided that Shannon entropy was optimal. The results of the tests and conclusions drawn from the tests are discussed below. First, we proved the effectiveness of Shannon entropy as a parameter to be optimized. We conducted multiple tests on various trees, first using Shannon entropy to choose the building in the district that would be most worth optimizing, and then using other methods and computing the final entropies of each for comparison. We find that the novel application of Shannon entropy is the most consistent and effective parameter for optimizing models compared to other previously proposed parameters, including the number of siblings a building has and the distance between the building and the sewage treatment work. The second result is that four sewage sample collectors are most effective for traveling between ten districts collectively. For four sewage sample collectors, the four optimized pathways are 1) HKU, North West Kowloon, Kwun Tong, Tseung Kwan O, HKU, 0.5834 hrs, 11.719 km; 2) HKU, To Kwa Wan, Aberdeen, Wan Chai East, HKU, 0.7907 hrs, 26.519 km; 3) HKU, Kwai Chung, Central, HKU, 0.5830 hrs, 11.696 km; 4) HKU, Tsing Yi, North Point, HKU, 0.6396 hrs, 12.262 km. For future work, we can elevate our math model from testing one building per district to testing multiple buildings per district to further minimize entropy and optimize the number of buildings tested per district. To achieve this, we can use a piecewise function that models the minimized entropy of a tree after every test with mixed integer linear programming and consider multi-dimensional functions to model all entropies for every combination of buildings tested within the district.

Introduction

When the COVID-19 pandemic was very serious in Hong Kong, researchers collected sewage samples from buildings to detect COVID-19 and identify risky areas in the city. This data collection method was efficient and effective for COVID-19 detection[1]. However, Hong Kong is a city with over seven million people living in 400 square kilometers. It is impossible for researchers to collect sewage samples from every building in every area to produce timely results[2]. Thus, the objective is to reduce collection time by building a math model that finds optimal pathways for sewage sample collectors to travel between districts. The pathways minimize two parameters: 1. the uncertainty that COVID-19 remains undetected in the city and 2. the traveling time of the pathways. These two parameters can be represented by Shannon entropy and the traveling salesman problem.

The first parameter is the "uncertainty" that COVID-19 remains undetected in Hong Kong, and is quantified by Shannon entropy[3]. Researchers must choose the districts to travel to that minimize the collective Shannon entropy of the city the most. Take Figure 1, a simplified model of buildings in a district. Each transparent circle represents a building and has two numbers - the first is the probability of COVID-19 presence in the building, which is calculated by population density and other factors. These probabilities are used to calculate the second number: the Shannon entropy of the whole district if that building is chosen for sewage sampling. In this example, building 3 results in the lowest Shannon entropy, making it the optimal building to sample from.

The second parameter is minimizing the traveling time between districts and is quantified by the Traveling Salesman Problem[4]. Consider Figure 2, a simplified model with num-

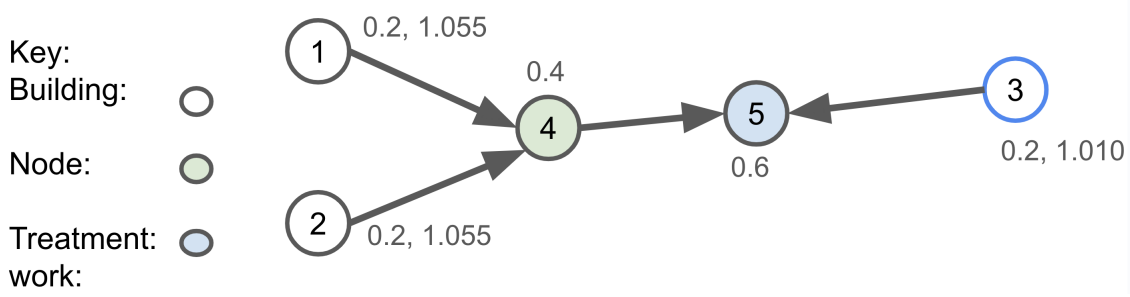

Fig. 1 Example of an entropy tree with buildings in white, nodes in green, and optimal building for testing in blue

bered circles representing districts and the numbers on lines between the circles representing the traveling time between the districts. The traveling salesman problem returns the pathway between all districts, which takes the least amount of time.

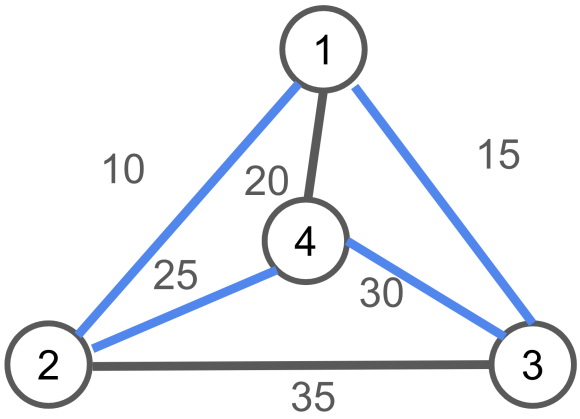

Fig. 2 Example of Traveling Salesman Problem with Optimal Path in Blue

Methodology

The objective function mathematically expresses the two optimized parameters, Shannon entropy and traveling time, along with all other constraints for the algorithm to process the data and calculate the optimized sewage pathways.

First, we aim to minimize the Shannon entropy of the system, which can be expressed as

$$-\sum_{i=1}^{n} p_i \log p_i$$

where p_i is the probability that node i contains COVID-19, based on various real-world parameters such as population density, and n is the number of nodes in the district.

The solver can only test one node and return one node to test, which is expressed by

$$\sum_{i=1}^{n} x_i = 1.$$

Recall that $x_i = 1$ when a node is visited and 0 when it isn't visited. Thus, if the sum of all x_i of all buildings in a district equals 1, only one of the x_i can equal one. In other words, only one building can be visited and tested.

The probability is assumed to be uniformly distributed among nodes with the same immediate parent. This assumption must have been made to make the model feasible because data on the population densities of each building are not published. However, this assumption is reasonable as the buildings that share the same immediate sewage intersection are relatively close to each other, so population density and thus COVID-19 uncertainty would not fluctuate dramatically among these buildings. The initial probability values are determined by the relative population density of the area.

The objective is to minimize the collective time taken by all collectors to complete their tours or minimize the objective function below, which is the weighted sum of the combined entropy of the system and the time taken for all salesmen to finish their tours. This can be expressed by

$$0.001 \sum_{k=1}^{4} u_{1,k} + \sum_{i=1}^{n} f_i(x).$$

Since traveling times are by nature larger than Shannon entropies, the sum is weighed by a scaling factor of 0.001. In the Shannon entropy expression above, p is the probability that Covid-19 is present in a district, and probability has a range between 0 and 1, making Shannon entropy have a value of around 1. Thus, traveling times will naturally be much greater numerically than the Shannon entropies. Thus, by multiplying the traveling time by the scale factor 0.001, we ensure that both the Shannon entropy and the traveling time have equal priorities in the optimization function. The use of the scaling factor of 0.001 is also effective because it allows policymakers to adjust

it depending on the structure of their cities and how they relatively proritize the uncertainty of Covid-19 remaining versus the traveling time between districts.

The scaling factor of 0.001 is beneficial because it makes the model flexible for policymakers who can adjust the scale depending on the structure of their cities and how they relatively prioritize the uncertainty of COVID-19 remaining versus the traveling time between districts. At the same time, it ensures all solutions lie on the Pareto Frontier[5], ensuring that both time and entropy are minimized optimally regardless of given conditions.

The constraint of the problem is that only one node in each district's tree may be tested at a time.

How COVID-19 is spreading is outside the scope of this paper. Our math model is only concerned with measuring the uncertainty of COVID-19 being present in buildings, regardless of its source.

Each collector must depart from and return to node 1. Exactly one tour at a time can enter and exit each node other than the first node. The number of collectors entering each node is equal to the number of collectors exiting it. No subtours permitted (Miller Tucker Zemlin constraint).

We used the Julia programming language to model and optimize. OpenStreetMapX.jl package provided real world data on population density and traveling time of various areas. HK drainage services provided information on the sewage pipe relationship within districts.

Results

After supplying the model with data, the objective function, and the constraints, the model solved for the optimal pathway for each of the four sample collectors shown in Figure 2, as well as the predicted time for completion and the length of the pathways as shown in Table 1.

Simulating the research conducted by the University of Hong Kong (HKU)[6], the algorithm used the GPS coordinates of HKU as node one, and the coordinates of each preliminary sewage treatment work as each node. Table 1 defines each node by its respective sewage treatment work. Table 2 describes the optimal solution for 4 salesmen, including the node paths, times, and distances, and this is visually represented in Figure 3.

Consider the same data set used in the Variants of Traveling Salesman Problem optimization problem. Table 4 presents the optimal solution, and this is visually represented in Figure 3.

Also, we proved the effectiveness of Shannon entropy as a parameter to be optimized. We conducted multiple tests on various trees, first using Shannon entropy to choose the leaf that would be most worth optimizing, and then choosing the leaf with the least or most number of siblings or the leaf closest to or farthest from the root node. Finally, we compute the final entropies of each for comparison. The results are shown in Fig 4.

Node	Sewage Treatment Work
1	The University of Hong Kong
2	North West Kowloon Preliminary Treatment Works
3	To Kwa Wan Preliminary Treatment Works
4	Kwun Tong Preliminary Treatment Works
5	Tseung Kwan O Preliminary Treatment Works
6	Kwai Chung Preliminary Treatment Works
7	Tsing Yi Preliminary Treatment Works
8	Central Preliminary Treatment Works
9	Wan Chai East Preliminary Treatment Works
10	North Point Preliminary Treatment Works
11	Aberdeen Preliminary Treatment Works

Table 1 Node Numbers and Respective Hong Kong Sewage Treatment Works

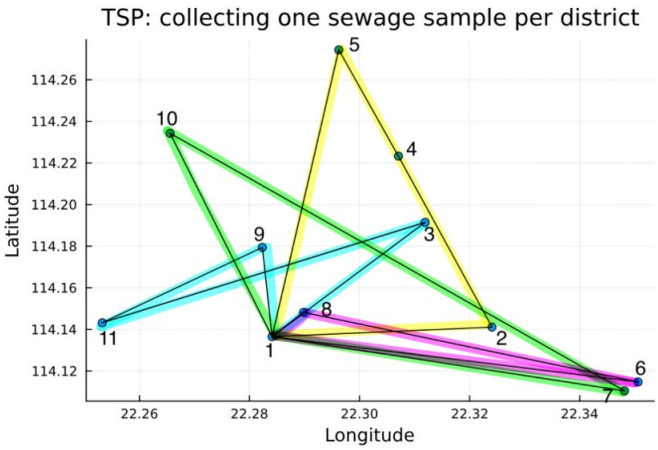

Fig. 3 Solution to mTSP with 4 Salesmen and 11 Nodes: Minimizing Time and Entropy

Salesman	Node Path, Time, and Distance
1	[1, 2, 4, 5, 1], 0.58358 hrs, 11.719 km
2	[1, 3, 11, 9, 1], 0.79071 hrs, 26.519 km
3	[1, 6, 8, 1], 0.58301 hrs, 11.696 km
4	[1, 7, 10, 1], 0.6396 hrs, 12.262 km

Table 2 mTSP Minimizing Time and Entropy with 4 Salesmen and 11 Nodes

As shown in Table 2, while other parameters such as the number of "siblings" and the distance between the building and the treatment work can be maximized or minimized depending on the entropy tree, minimizing Shannon entropy is the only consistent parameter for quantifying effectiveness.

Fig. 4 Lack of consistency of alternative parameters number of siblings (buildings that share the same immediate sewage intersection) and distance between a building and a treatment work for Shannon entropy with optimal building in blue

Conclusions & Future Work

The novel application of Shannon entropy is the most consistent and effective parameter for optimizing models compared to other previously proposed parameters, including the number of siblings a building has and the distance between the building and the sewage treatment work. These alternatives fail as the prediction of which building should be tested for COVID-19 depends on whether the probability that the building contains COVID-19 is less than that of its siblings, which is only accounted for by the mathematical expression for Shannon entropy, as shown in the first expression of the Methodology. The scope of the use of Shannon entropy can be expanded to many similar mathematical optimization problems defined by variables, constraints, and an objective function.

For ten selected treatment works on Hong Kong Island, it is optimal for four sewage sample collectors and each pathway between sewages is under one hour. For four sewage sample collectors, the optimized pathways are 1) HKU, North West Kowloon, Kwun Tong, Tseung Kwan O, HKU, 0.5834 hrs, 11.719 km; 2) HKU, To Kwa Wan, Aberdeen, Wan Chai East, HKU, 0.7907 hrs, 26.519 km; 3) HKU, Kwai Chung, Central, HKU, 0.5830 hrs, 11.696 km; 4) HKU, Tsing Yi, North Point, HKU, 0.6396 hrs, 12.262 km. These pathways could be used to track the spread of viruses in future epidemics, or the spread of any pathogens in any scenario of interest, making it exceedingly useful.

The math model has some limitations and challenges when implemented in a real-world setting. The computational complexity exponentially increases by n! with each increase in sewage district to test. The data is accurate and consistently updated by the OpenStreetMap Julia package, but this takes a great toll on the computational complexity and makes the solving time relatively slow. This issue is inevitable due to the size of the dataset, and this is already mitigated by using Julia as a programming language, which is the most cutting-edge in mathematical optimization compared to Matlab and other alternatives.

Another current limitation is that although the model is validated using the most powerful optimizer available, JuMP, and is backed by real-time updated data from OpenStreetMap, it has yet to be tested in reality. The next step is to use data from demographers on COVID-19 cases and population densities to test the model, collect data, and evaluate the accuracy and power for further improvements.

We can elevate our math model from testing one building per district to testing multiple buildings per district to further minimize entropy and optimize the number of buildings tested per district. To achieve this, we can use a piecewise function that models the minimized entropy of a tree after every test with mixed integer linear programming and consider multi-dimensional functions to model all entropies for every combination of buildings tested within the district[7,8].

Acknowledgments

Thank you to Dr. Benoit Legat for elevating the concept of the paper and being so gracious with his time and teaching. Thank you to Mrs. Mary Dee Mulligan for advising in the presentation process.

References

1 *Government follows up on positive results of sewage surveillance and appeals to residents to undergo virus testing*, https://www.info.gov.hk/gia/general/202206/05/P2022060300390.htm, 2022.

2 D. Yu, *Science of The Total Environment*, 2022, **821**, year.

3 C. E. Shannon, *The Bell System Technical Journal*, 1948, **27**, 379–423.

4 G. B. Dantzig and M. N. Thapa, *Linear Programming : 2: Theory and Extensions*, Springer, 2003.

5 M. Mesquita-Cunha, J. R. Figueira and A. P. Barbosa-Póvoa, *European Journal of Operational Research*, 2023, **306**, 286–307.

6 T. Zhang, *Grid monitoring of SARS-CoV-2 in Sewage For An Early-Warning Sign of Community Outbreak*, https://rfs1.healthbureau.gov.hk/images/jsn_is_thumbs/images/past_event/Health_Research_Symposium_2021/Materials/HRS2021_T3a_powerpoint.pdf, 2021.

7 S. U. Ngueveu, *European Journal of Operational Research*, 2019, **275**, 1058–1071.

8 M. Bernreuther, *Solving mixed-integer programming problems using piece-wise linearization methods*, Unpublished work.