Optimal Base Station Network Based on Topological Data Analysis

Minhao Lyu

***Abstract*—The decision of which base stations need to be removed due to the cost is always a difficult problem, because the influence on the cover rate of the network caused by the removal should be kept to a minimum. However, the common methods to solve this problem such as K-means Clustering show a low accuracy. Barcode, which belongs to TDA, has the possibility to show the result by identifying the Persistent Homology of base station network. This essay mainly illustrates the specific problem of optimal base station network, which applies the TDA(Topological Data Analysis) methods to find which base stations need removing due to the cost K-means Clustering and Topological Data Analysis methods were mainly used. With the simulated distribution of telecommunication users, K-means Clustering algorithm was used to locate 30 best base stations. By comparing the minimum distance between the results (K=25 and K=30), K-means Clustering was used again to decide base station points to be removed. Then TDA was used to select which 5 base stations should be removed through observing barcode. By repeating above steps five times, Finally the average and variance of cover area in original network, K-means Clustering and TDA were compared. The experiment showed that the average cover rate of original network was 81.20% while the result of TDA and K-means Clustering were 92.13% and 89.87%. It was proved by simulation that it is more efficient to use TDA methods to construct the optimal base station network.**

***Index Terms*—Topological data analysis, optimal base station network, cluster.**

I. Introduction

With the rapid development of information, more and more users are in demand of high quality of internet connection. However, due to the cost and low-efficient methods used before, some base stations need to be removed. It becomes important to choose which base station to be removed, so that the base station network will be influenced least. K-means Clustering method is widely-used to solve the problem. TDA (Topological Data Analysis) is a method combing topological methods with statistical skills. Now, TDA has been widely used to process with the data in high dimensions [1]. When facing the problem of removing the base station and not to low the network efficiency, the essay is aimed to solve the specific base station problem mentioned above with TDA. Compared with the result of K-means Clustering, flexibility and accuracy of TDA are proved.

Manuscript received October 18, 2021; revised December 22, 2021.

Minhao Lyu is with Hefei University of Techhnology, China (e-mail: lmh0918@mail.hfut.edu.cn).

II. Related Work

Problems of base station network have been discussed for a long time. Related work on base station placement studied how to place base station so that the network flow is proportionally maximized subject to link capacity [2]. In the problems of base station construction, Clustering methods based on distance are often used. It is shown that the use of clustering in conjunction with a mobile base station for data gathering can significantly prolong network lifetime and balance energy consumption of sensor nodes [3]. In fact, optimizing the base station network is a kind of topology problem. Coverage problems in sensor networks with minimal sensing capabilities [4]. The large literature on coverage problems for networks rests on the computational geometry approach [5]. In the site selections of base stations in reality, base stations often had to be rebuilt and re-sited. There is a situation where a certain number of optimal base stations are known, considering the production cost to reduce base station construction, how to choose the base stations that needs to be removed in order to affect the cover area least. While the common algorithms such as K-means Clustering is more complicated to operate in such problems, the process is cumbersome and the results have a certain deviation.

III. Related Methods

A. K-means Clustering

Clustering algorithm is a kind of algorithms that can be used to process the problems with the data which are not marked. By taking advantage of clustering, more properties of data can be found.

means Clustering is one of the classical and widely-used clustering algorithms. Given a sample set D= $\{x_1, x_2, \ldots, x_m\}$, for the clustering $C = \{C_1, C_2, \ldots, C_k\}$, K-means algorithm minimize the square error

$$\sum_{i=1}^{k} \sum_{x \in C_i} \left\| x - \mu_i \right\|_2^2 \quad (1)$$

It is a NP-complete problem to find the optimal solution of the formula, all possible clustering of sample set should be considered [6]. K-means algorithm takes the greedy strategy to get approximate solution of this formula by

DOI: 10.7763/IJMO.2022.V12.792

iteration. The specific algorithm is shown as follow.

1) Input: sample set D= { $x_1, x_2, \ldots, x_m$ }, the number of clusters K

2) Process:

1: choose k samples as initial mean vector { $\mu_1, \mu_2, \ldots, \mu_k$ } from D randomly

2: repeat

3: let $C_i = \varnothing \ (1 \le i \le k)$

4: for j=1,2,...,m do

5: calculate the distance between x_j and $\mu_i (1 \le i \le k)$: $d_{ji} = \|x_j - \mu_i\|_2$

6: based on shortest distance, determine the cluster x_j : $\lambda_j = \arg\min_{i \in \{1,2,\ldots,k\}} d_{ji}$

7: put x_j into the cluster: $C_{\lambda_j} = C_{\lambda_j} \cup \{x_j\}$

8: end for

9: for i=1, 2..., k does

10: calculate latest mean vector $\mu_i' = \frac{1}{|C_i|} \sum_{x \in C_i} x$

11: if $\mu_i' \neq \mu_i$ then

12: replace the mean vector μ_i into μ_i'

13: else

14: keep the mean vector

15: end if

16: end for

17: until no update of the mean vector

3) Output: Cluster $C = \{C_1, C_2, \ldots, C_k\}$

B. Topological Theorem

Topological Data Analysis (TDA) is noticed and put into dealing with the problems which are not easily solved by clustering methods. Now, TDA methods are widely and frequently used to cope with problems in high dimensions. To enter the TDA, there are some basic points that need to be known.

1). Metric space

A data set D can be taken as a finite set of observation vectors $x_1, x_2, \ldots, x_n$, each of length $d \in Z_+$. A function $\phi: D \times D \to R_+$ is a metric on D. The pair (D, ϕ) is called a metric space.

2). Simplicial complex

A k-simplex is a set of k+1 vertices which will be denoted as $[p_1, p_2, \ldots, p_k]$. For example, the 2-simplex $[p_1, p_2, p_3]$ is a triangular disc and a 3-simplex $[p_1, p_2, p_3, p_4]$ is a regular tetrahedron. To construct a simplicial complex from data, it is intuitive to use a generalization of an ε-neighboring graph. The result is called a Vietoris-Rips complex:

$$v_\varepsilon(D) = \{\sigma \subseteq S \mid \phi(u, v) \le \varepsilon, \forall u \neq v \in \sigma\} \quad (2)$$

where ϕ is the Euclidean metric.

4) Persistent Homology

Persistent Homology is an important part of TDA. An example is shown that the notion of α-neighboring graph immediately extends. For a given set of points X in a metric space (M, ρ) and a real number $\alpha \ge 0$, the Vietoris-Rips complex $Rips_\alpha(X)$ is the set of simplices $[x_0, \ldots, x_k]$ such that $d_X(x_i, x_j) \le \alpha$ for all (i, j)[7].

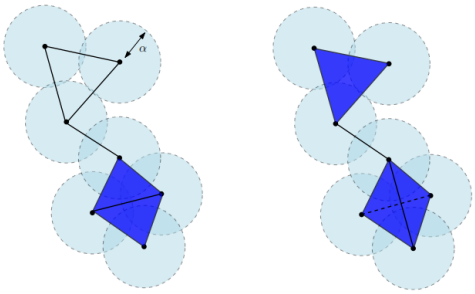

Fig. 1. The Cech complex $Cech_\alpha(X)$ [left] and the Vietoris-Rips $Rips_{2\alpha}(X)$ [right] of a finite point cloud in the plane R^2

3). Barcode

A barcode graphically represents $H_k(C; F)$ as a collection of horizontal line segments in a plane. While in the plane, horizontal axis corresponds to the parameter and vertical axis represents an ordering of homology generators [8]. Figure 2 gives an example of barcode representations of the homology of the sampling of points in an annulus.

A barcode best represents the persistence analogue of a Betti number. The k-th Betti number of a complex, $\beta_k = rank(H_k)$, acts as a coarse numerical measure of H_k. As with β_k, the barcode for H_k does not give any information about the finer structure of the homology, but merely a continuously parameterized rank. A barcode representation has the ability to qualitatively filter out topological noise and capture significant features [9].

Properties of Persistent Homology can be understood and used in different situations. In these base stations

network problem, when the positions of base stations are abstracted as points in the coordinate axis, distances are mainly what should be focused on. Persistent Homology is related to the cover rate of base stations. More specifically, when R gets larger, it can be known which points can cover more areas. Barcode can describe the persistent Homology directly, which is continent to observe and decide which base stations need to be removed.

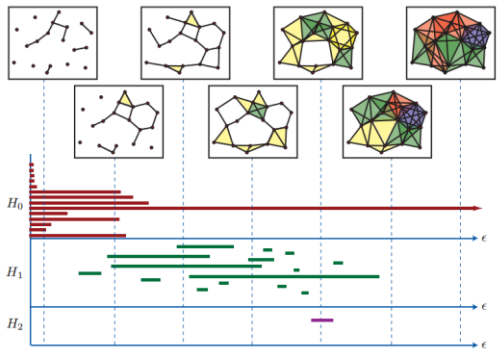

Fig. 2. [bottom] An example of the barcodes for $H^*(R)$. [top] The rank of $H_k(\varepsilon_i)$ equals the number of intervals in the barcode for $H_k(R)$ intersecting the (dashed) line $\varepsilon = \varepsilon_i$

IV. Experiment Premise

A. Problem Description

There were 5000 telecommunication users and 30 original base stations. Due to the cost, we had to keep 25 base stations. Among 30 points, 5 points should be removed and the users covered by base station network would get influenced least.

B. Assumption

Before the experiment, there are some assumptions to make the methods run possibly.

1) **A1:** All base stations have the same lifetime and their belonging equipment is normal.
2) **A2:** When the positions of users and base stations are considered, the natural effect should be ignored.
3) **A3:** The cost of the construction of base station is not influenced.

For **A1**, the state of equipment in the base station may influence the performance of the base station. In the experiment, the location of base station should be considered more. **A2** means that in specific occasions, due to some natural characters like geographical factors, the chosen points cannot be used to construct the base station. **A3** shows that many factors that may have effects on the cost of constructions should be ignored in this problem.

In this experiment, the data of 5000 telecommunication users and 30 original base station was generated and used.

V. Process of the Simulation

A. Original Location of base Station by K-mean Clustering

First, 5000 users and 30 base station networks were randomly simulated as follows.

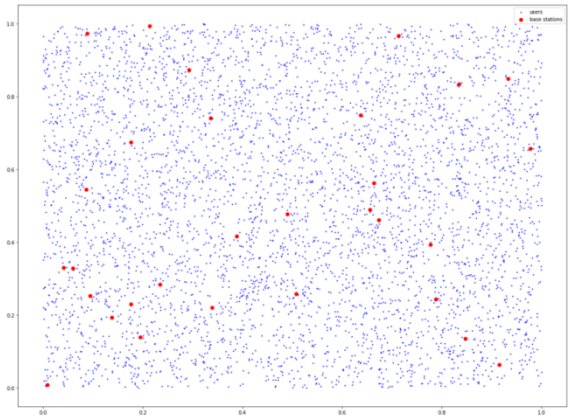

Fig. 3. The simulation of distribution of telecommunication users and base stations

Considering of the distance, the optimal base station network can be realized by K-means Clustering. 30 points were chosen as the base station randomly at first. Through ten times of iteration, 30 better base station locations were found which are closer to their neighbor user points. These 30 points are the center points of their user clusters.

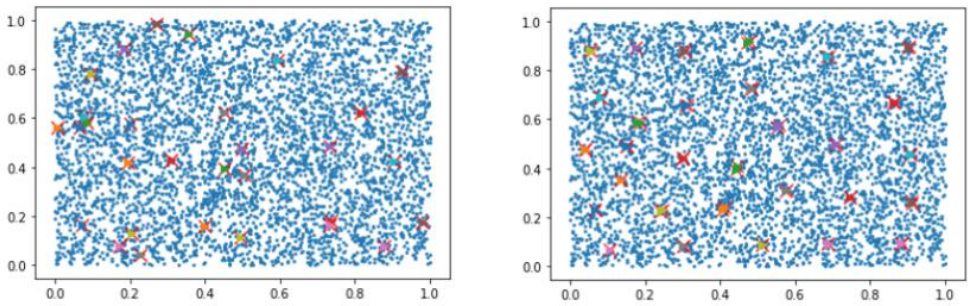

Fig. 4. 30 base stations randomly chosen[left]30 base stations determined by K-mean Clustering after ten times iterations[right]

TABLE I. NETWORK CONTAINING 30 BASE STATIONS DETERMINED BY K-MEANS CLUSTERING

Base station number	x	y
1	0.06835781	0.22766937
2	0.41731278	0.24885483
3	0.48658042	0.92955139
4	0.74387416	0.28362089
5	0.71166166	0.50186038
6	0.30999498	0.8755727
7	0.68699622	0.09522051
8	0.30463839	0.08346989
9	0.05452252	0.88749054
10	0.90756333	0.46703017
11	0.317842	0.66180002
12	0.13234714	0.35477391
13	0.1946167	0.589382
14	0.86826205	0.67346873
15	0.17802362	0.89504101
16	0.90891361	0.27140247
17	0.8840804	0.0935934
8	0.58325384	0.31293485
19	0.25130456	0.23087939
20	0.69352689	0.85430563
21	0.15328295	0.48973193
22	0.04187901	0.48770122
23	0.45234844	0.41909839
24	0.30022577	0.44363468
25	0.56417122	0.58640208
26	0.90567346	0.89733275
27	0.10589404	0.06897774
28	0.48112495	0.74134079
29	0.5101992	0.08822913
30	0.08206836	0.69130229

B. Base Station Network after Removed by K-means Clustering

Due to the cost, now 5 base stations need to be removed. In this part, K-means Clustering algorithm was used again, while K=25.

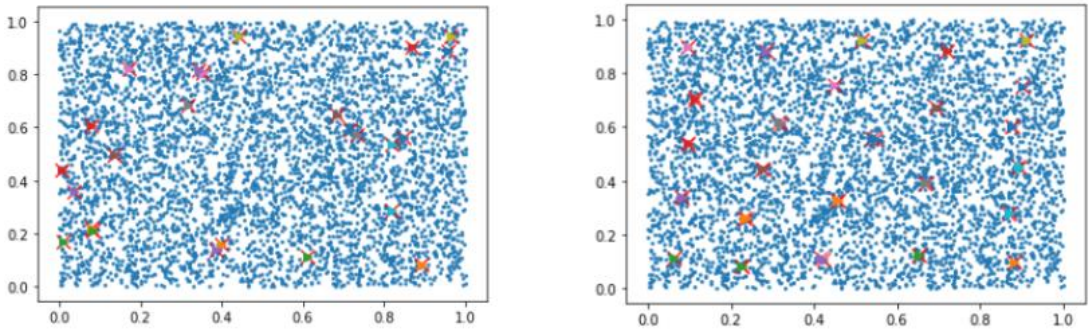

Fig. 5. 25 base stations randomly chosen[left]25 base stations determined by K-means Clustering after ten times iterations[right]

Then the minimum distance between 25 points (generated by K-means Clustering, K=25) and 30 points (generated by K-means Clustering, K=30) was calculated. By finding out 5 points with minimum distance, these 5 points which least influenced the cover area of base station will be removed while the structure is most similar to that before. The base stations judged by shortest distance were NO.6, NO.10, NO.17, NO.19, NO.24.

TABLE II. NETWORK CONTAINING 25 BASE STATIONS AFTER 5 STATIONS REMOVED BY K-MEANS CLUSTERING

Base station number	x	y
1	0.06835781	0.22766937
2	0.41731278	0.24885483
3	0.48658042	0.92955139
4	0.74387416	0.28362089
5	0.71166166	0.50186038
7	0.68699622	0.09522051
8	0.30463839	0.08346989
9	0.05452252	0.88749054
11	0.317842	0.66180002
12	0.13234714	0.35477391
13	0.1946167	0.589382
14	0.86826205	0.67346873
15	0.17802362	0.89504101

16	0.90891361	0.27140247
8	0.58325384	0.31293485
20	0.69352689	0.85430563
21	0.15328295	0.48973193
22	0.04187901	0.48770122
23	0.45234844	0.41909839
25	0.56417122	0.58640208
26	0.90567346	0.89733275
27	0.10589404	0.06897774
28	0.48112495	0.74134079
29	0.5101992	0.08822913
30	0.08206836	0.69130229

C. Base Station Network after Removed by TDA Methods

In this section, TDA methods were used to find out which 5 base stations to be removed. From the related methods above, it is known that barcode can describe the persistent homology. By drawing the barcode of 30 base station points, ones which are persistent can be seen.

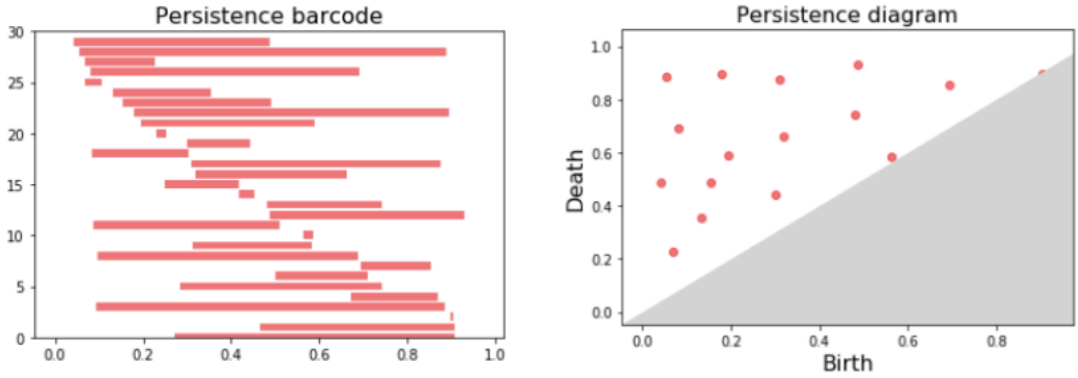

Fig. 6. The persistent barcode and diagram of networks containing 25 base stations

From Figure 6, it was found that the five base stations are NO.3, NO.11, NO.15, NO.21, NO.26. Then these 5 points which have shortest bar were removed to make the whole structure keep the persistent homology as long as possible. Finally, the network containing 25 base stations was decided as Table 3 shows.

TABLE III. NETWORK CONTAINING 25 BASE STATIONS AFTER 5 STATIONS REMOVED BY TDA METHODS

Base station number	x	y
1	0.06835781	0.22766937
2	0.41731278	0.24885483
4	0.74387416	0.28362089
5	0.71166166	0.50186038
6	0.30999498	0.8755727
7	0.68699622	0.09522051
8	0.30463839	0.08346989
9	0.05452252	0.88749054
10	0.90756333	0.46703017
12	0.13234714	0.35477391
13	0.1946167	0.589382
14	0.86826205	0.67346873
16	0.90891361	0.27140247
17	0.8840804	0.0935934
8	0.58325384	0.31293485
19	0.25130456	0.23087939
20	0.69352689	0.85430563
22	0.04187901	0.48770122
23	0.45234844	0.41909839
24	0.30022577	0.44363468
25	0.56417122	0.58640208
27	0.10589404	0.06897774
28	0.48112495	0.74134079
29	0.5101992	0.08822913
30	0.08206836	0.69130229

D. Result of the Simulation

To evaluate the validity of different designed base station network, the covered users of base station network was defined as cover rate.

Cover rate = the number of covered users/the number of all users (3)

The result shows that higher cover rate means the better performance of base station network.

Through this group of data, the cover rate of original network was 82.12% while the results for TDA methods and were separately 94.06% and 94.40%.

Due to the randomness of simulation, the experiment should be repeated to make the result closer to the real situations. Then 4 steps above were repeated five times. It was calculated that the average and variance of cover rate of original base station network, base station network after removed by K-means Clustering and base station network after removed by TDA methods.

TABLE IV. COMPARISON OF COVER RATE

Cover rate %	1	2	3	4	5	average	variance
Original Network	82.72	81.14	75.02	81.04	86.10	81.20	12.91
K-means Clustering	81.90	93.30	94.28	85.32	94.56	89.87	27.49
TDA methods	88.96	92.00	90.62	95.22	93.84	92.13	4.96

Table 4 shows that the average cover rate of TDA methods was 92.13% and the variance was 4.96. Compared to other two average (Original Network 81.20%, K-means Clustering 89.87%) and variance (Original Network 12.91, K-means Clustering 27.49) of the cover rate, TDA had the most stable performance. Compared to the original base station network, the base station network through TDA have a higher cover rate.

This experiment theoretically verifies that the Topological Data Analysis method has a more significant effect on this type of problem than Clustering. The drawback is that only this specific type of problem is discussed. If optimal base station network problems need to be solved more extensively, the background must be reconsidered and refreshed, similar to the life of the hardware in base stations. In the future, when the background of this problems gets more abundant, the heuristic algorithm will be compared with the above two methods more deeply to prove the role of Topological Data Analysis.

VI. CONCLUSION

This article mainly introduced the application of Topological Data Analysis on the problem of the removal of the base stations due to the cost. When removing the base station with the lowest cost and negative efficiency, TDA methods can help decision maker make it clearer to decide which points should be removed and maximize the covered area of base station network. From the experiment above, it is obvious that TDA methods have a better efficiency on optimal base station network. To think it reversely, it also can be used when it is unclear which position to choose. By drawing the barcode of the position, we can satisfy our needs by choosing how many the longest bars are. However, there are still some shortages of using barcodes to identify the base station. For example, it is hard when the number of base stations is too large to count the longest and shortest ones. Furthermore, it has great potential and can be improved to be a completer and more proven tool.\

Through the research above, it is proved that topological structure and its properties can be closely connected to such kind of optimal network problem by Topological Data Analysis. With TDA, we may discover more properties that are not easy to find and use them to solve the problems better.

ACKNOWLEDGMENT

I would like to thank my parents for their support during the special period. I also want to thank the teaching assistant of the program I have took for his suggestions of modelling, with whose help I have made improvement in programming.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

The author finished this paper by himself.

REFERENCES

[1] F. Chazal, "High-dimensional topological data analysis," *Handbook of Discrete and Computational Geometry*, chapter 27. CRC Press. 2017.

[2] S. S. Dhillon and K. Chkrabarty, "Sensor placement for effective coverage and surveillance in distributed sensor networks," in *Proc. IEEE Wireless Communications & Networking Conference (WCNC)*, pp. 1609–1614, 2004.

[3] B. Liu and D. Towsley, "A study of the coverage of large-scale sensor networks," in *Proc. IEEE International Conference on Mobile Ad-hoc and Sensor Systems*, 2004.

[4] H. Koskinen, "On the coverage of a random sensor network in a bounded domain," in *Proc. 16th ITC Specialist Seminar*, pp. 11-18, 2004.

[5] Y. T. Hou, Y. Shi, H. D. Sherali, and S. F. Midfkiff, "Prolonging sensor network lifetime with energy provisioning and relay node placement," in *Proc. IEEE International Conference on Sensor and Ad Hoc Communications and Networks (SECON)*, pp. 295–304, 2005.

[6] Y. Zou and K. Chakrabarty, "Sensor deployment and target localization based on virtual forces," in *Proc. IEEE Annual Joint Conference of the IEEE Computer and Communications Societies (Infocom)*, pp. 1293–1303, 2003.

[7] G. Carlsson, A. Zomorodian, A. Collins, and L. Guibas, "Persistence barcodes for shapes," *Intl. J. Shape Modeling*, vol. 11, pp. 149-187, 2005.

[8] F. Chazal, M. Glisse, C. Labruère, and B. Michel, "Convergence rates for persistence diagram estimation in topological data analysis," *To Appear in Journal of Machine Learning Research*, 2014.

[9] N. Atienza, R. Gonzalez-Diaz, and M. Rucco, "Persistent entropy for separating topological features from noise in vietoris-rips complexes," *J. Intell. Inf. Syst*, vol. 52, no. 3, pp. 637–655, 2019.

Minhao Lyu now is majoring in MSc business analytics at Warwick Business School, UK. He gained knowledge related to optimization problems and ways to deal with different categories of data. He graduated from Hefei University of Technology, China and majored in information and computing science and got his bachelor's degree in 2020. During the undergraduate period, he is interested in data science and statistics. In 2019, he took part in 2019 ICM competition and won the Meritorious winner, From July 2019 to September 2019, he was an data analyst intern in China Mobile company.