

RESEARCH

Open Access



# Designing a clustering algorithm for optimizing health station locations

Pasi Fränti<sup>1\*</sup>, Sami Sieranoja<sup>1</sup> and Tiina Laatikainen<sup>2,3</sup>

## Abstract

In this paper, we define the optimization of health station locations as a clustering problem. We design a robust algorithm for the problem using a pre-calculated overhead graph for fast distance calculations and apply a robust clustering algorithm called random swap to provide accurate optimization results. We study the effect of three cost functions (Euclidean distance, squared Euclidean distance, travel cost) using real patient locations in North Karelia, Finland. We compare the optimization results with the existing health station locations. We found that the algorithm optimized the locations beyond administrative borders and strongly utilized the transport network. The results can provide additional insight for the decision-makers.

**Keywords** Facility location, Health care optimization, Clustering, Maximum coverage, Random swap

## Introduction

At present, there is substantial pressure to reduce health-care service costs by enhancing the optimization of healthcare services, potentially resulting in a reduction of the existing healthcare station network and the services they offer. Yet, a high quality of health care services would lead to healthier citizens and, in this way, reduce the overall demand for these services.

Accessibility of the service is one factor in this optimization. First, it can improve early diagnoses and treatments and support the better provision of preventive health care [1]. Accessibility is a major determinant of participation in the service, according to Gu et al. [2]. Geographic distance to services has been identified as a significant barrier to regular checkup visits and chronic

care visits, especially in rural areas, whereas acute care visits seem to be less sensitive to distance [3]. In Finland, among patients with mental health problems, distance was negatively associated both with in-person visits to health stations as well as in-home visits [4].

Second, good accessibility can save lives in case of emergency situations and prevent long-term consequences caused by delayed treatment. Patients living closer to a percutaneous coronary intervention (PCI) capable cardiac unit have a higher chance of survival than those who live far away, according to Di Domenicantonio et al. [5].

Third, improving accessibility can reduce the travel costs of the patient, both direct costs and time loss associated with indirect costs, relieving the financial burden for the patients and lowering the threshold to seek preventive care. In addition to costs for patients, long distances create societal costs, for example, in different forms of reimbursements and transport costs [6]. Accessibility is also a question of equity and fairness of service provision that should be considered [7].

Healthcare accessibility has been modeled as a *maximum coverage location* problem by maximizing the number of patients reached given some distance threshold [8]. The results in [9], however, suggested that maximizing

\*Correspondence:

Pasi Fränti

[pasi.franti@ueff.fi](mailto:pasi.franti@ueff.fi)

<sup>1</sup> Machine Learning Group, School of Computing, University of Eastern Finland, P.O. Box 111, 80101 Joensuu, Finland

<sup>2</sup> Institute of Public Health and Clinical Nutrition, University of Eastern Finland, P.O. Box 1627, 70211 Kuopio, Finland

<sup>3</sup> Joint Municipal Authority for North Karelia Social and Health Services (Siun sote), Tikkamäentie 16, 80210 Joensuu, Finland



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

coverage (minimizing the patients at risk) leads to a significant increase in the average travel time of the patients. Burkey et al. [10] reported similar results, optimization for the coverage reduced the patients at risk from 6.9 to 2.7%, on average, in the case of health care services in three US states (North Carolina, South Carolina, and Virginia), but it increased the average travel time by 13%.

Wang & Tang [11] proposed to minimize the variance of the distance or travel time for equal accessibility. Accessibility to tertiary and secondary facilities was studied in [12] using data from Shenzhen, China. The travel time was estimated by the Baidu Map application programming interface (API), and optimization was performed using the particle swarm optimization (PSO) algorithm [13].

Burkey et al. [10] reported that the existing locations already provide near-optimal geographic access to healthcare services in three US states (North Carolina, South Carolina, and Virginia). The average travel time of the existing facilities could be further reduced by only about 5% by the *p*-median clustering algorithm. The only exception was Tennessee, where the reduction was 15%. In the case of myocardial infarction patients in Finland, better optimization could decrease the average travel time to the hospital by only 3.7% [9]. A more remarkable saving was reported by Gu et al. [2], who managed to increase the accessibility of breast cancer screening services by 14% (from 0.35 to 0.40). They used Google Maps API to estimate the travel distance and time. An even more significant reduction in distance (33%) was reported by Fo & Silva Mota [14] with data from the Sao Paulo metropolitan area.

A completely different approach to measuring accessibility is the *two-step floating catchment area* (2SFCA) method by Shen et al. [15], and its hierarchical variants [16, 17]. They measure the ratio of primary-care physicians to their surrounding population (within a given travel time) and then sums up the ratios around the demand locations. This leads to a different optimization problem, which would close to resemble maximum coverage but favor bigger units. Accordingly, Tao et al. [17] observed that the healthcare facilities in Shenzhen are unevenly distributed due to the concentrated distribution of tertiary hospitals.

The use of a clustering algorithm can find theoretically better facility locations. However, existing research suggests that optimizing for one criterion, like coverage, can lead to inferior optimization for another criterion, like average travel time. The clustering process itself also includes several design parameters like the choice of the algorithm and the estimation of the travel time, which both may have a significant effect on the optimization

result. It is an open question how these design choices affect the clustering results.

To address these questions, we perform an experimental study with a sample of diabetes-related healthcare visits from SiunSote in North Karelia, Finland, between 2011 and 2014. Instead of classical *k*-means or its variant *p*-median, we use a more robust clustering algorithm to reduce the effect of algorithm artifacts. We then show the optimization results with different criteria, including Euclidean distance, squared Euclidean, travel time, and travel cost. We also consider the efficiency of the optimization process.

We also analyze the optimization results. While the results are based on a selected sub-sample of patient data and cannot be used to guide the healthcare organization, they reveal several interesting facts and trends in the area. For example, the location of existing healthcare stations and allocation of patients to these follow municipal borders, whereas the algorithm does not have such a border as it has no restrictions to allocate patients to nearby health stations in the neighboring municipality.

The results also have relevance to other healthcare services. For example, the nurse districting problem in [18] clusters the patients by simple *k*-means and tabu search based on their locations. The home care scheduling problem has been considered a two-objective optimization problem, which aims at minimizing operating costs while maximizing the quality of service at the same time [19]. Other closely related problems include ambulance location and relocation problems.

## Clustering

Next, we describe the clustering component, what components it includes, and explain the choices behind each of them. The clustering is integrated in a Web-tool described in [20].

## Distance calculation

*K*-means is the most common clustering algorithm. It minimizes squared Euclidean distances between the data (patient locations) and their nearest centroid. Squared distance is widely used even if it does not inherently correspond to real-world geographic phenomena. Some attempts have been made to make the connection, though. For example, Zhou and Li [21] modeled the cost of disaster losses as a quadratic function of the distance from the emergency facility to the disaster location.

Another common approach is to maximize the number of locations that are within a given distance from its nearest facility. Church [22] defined this as a *maximal covering location problem*. It is also referred to as *threshold distance* [23], where the distance cost is 0 if the distance to the centroid is less than a predefined threshold value,

otherwise, it is 1. Fränti et al. [24] minimized the number of myocardial infarction patients at risk by defining at-risk individuals based on their proximity to the nearest percutaneous coronary intervention (PCI)-capable hospital, specifically considering whether a patient resides beyond a predetermined time limit.

Absolute (non-squared) distance is the most used and natural choice in the case of location-based applications. It has a slight difference from its squared variant in the Euclidean distance case. The optimal centroid location of a cluster is its geometric center in the case of squared distance but its median in the case of absolute distance. The average (or median) can be calculated for each coordinate independently. The median is also known as the spatial median [25], and the corresponding k-means variant is known as the *p-median* [26, 27].

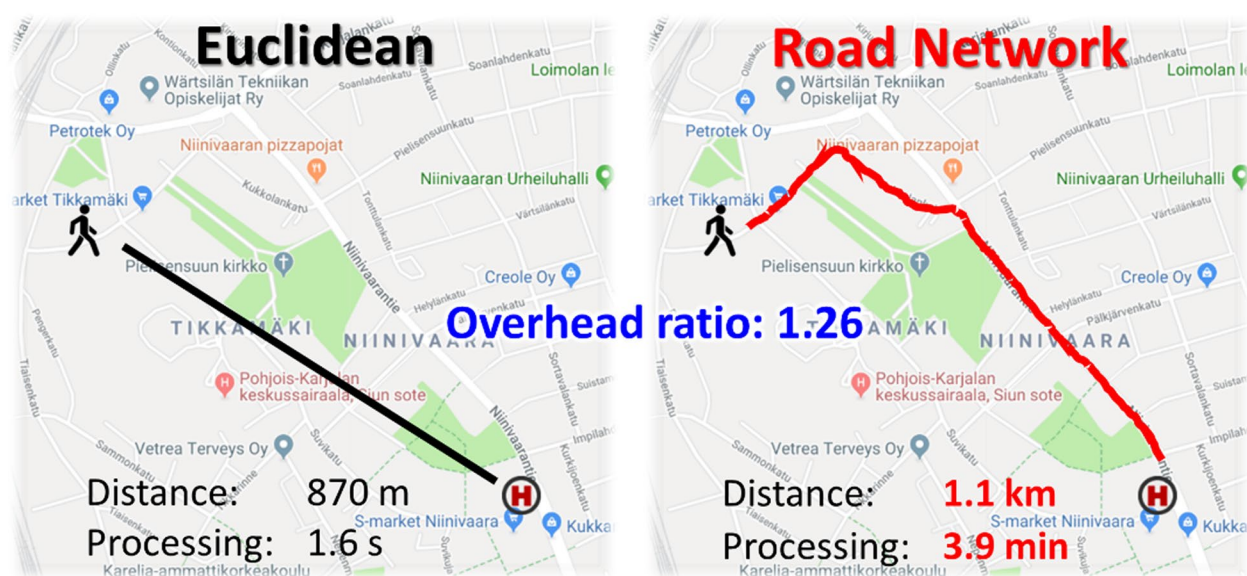
Simple Euclidean distance, however, was shown to cause bias in the facility optimization [24]. For this reason, travel distance (or time) is recommended. Calculating the shortest distance via road network is straightforward but requires lots of computation. A single shortest path calculation is fast, but facility optimization requires millions of such calculations. The locations of the stations are also dynamically changing during the optimization process, which prevents the use of a pre-calculated distance matrix. The consequence is that the optimization processing may take even days.

Boscoe [28] showed that Euclidean distance is an adequate approximation for travel distance in the United States when multiplied by a constant factor of 1.4. They

call this factor *the detour index*. It virtually equals the diagonal of a unit square corresponding to the Manhattan distance. However, our data is mostly in areas where lakes and rivers make the road network more complex, and a simple Manhattan distance would not be accurate. In city areas, the factor can be smaller than 1.4 (see Fig. 1), but in areas containing rivers and lakes, the factor can be much bigger. In general, there are large differences.

Mariescu-Istodor and Fränti [29] adapted this detour index locally by constructing a so-called *overhead graph*. The graph was built as pre-processing using patient locations in the data and the road network to identify traffic points. These points constitute the nodes of the graph. *The overhead ratio* was then calculated for all pairs of nodes in the graph as the ratio of their true travel distance and the corresponding Euclidean distance. The larger the overhead value, the longer the detour. The values are stored in the graph (represented by a matrix) and used in the optimization process.

During the optimization, when we need to estimate the travel distance between a patient's location and a given health station, we first calculate their Euclidean distance. We then find the nearest graph nodes of these two locations and obtain the stored overhead ratio between these two locations. The Euclidean distance is then multiplied by this constant to obtain an estimation of the travel distance. Travel times are derived directly from the travel distances using the average speed of each road segment.



**Fig. 1** Example of the detour index between Euclidean and road network distances from the patient location (Huvilakatu) to the nearest health station (Suvikatu station)

Dynamics like rush hours are not considered, and optimization is made merely to minimize the average.

This process is extremely efficient, requiring only a single lookup table and one multiplication operation (overhead ratio × Euclidean distance). With our data, this reduces the processing time from 2 weeks to about 15 min when using 10,000 iterations of the clustering algorithm. The huge speed-up is achieved at the cost of minor inaccuracy in the distance estimation, 2%, according to [29], and with additional memory of 512 × 512 = 0.25 MB for storing the lookup table. The process with two sample graphs is shown in Fig. 2.

**Optimization function**

K-means is by far the most common clustering algorithm and would be the most obvious choice for use here as well. However, it minimizes the sum of squared Euclidean distances, and it is unclear whether geographical distances should be squared in this application. A more common choice is to calculate the sum of Euclidean distances as such (without squaring). However, neither of these satisfied our needs as we are also interested in the travel costs of the patients. We, therefore, consider also the sum of the travel costs as the optimization function.

The three optimization functions can be written mathematically as follows:

$$\text{Euclidean} : \sum_{p \in \text{patients}} d_{L2} \left( p, \operatorname{argmin}_{h \in \text{HS}} d_{L2}(h, p) \right)$$

$$\text{Squared Euclidean} : \sum_{p \in \text{patients}} d_{L2} \left( p, \operatorname{argmin}_{h \in \text{HS}} d_{L2}(h, p) \right)^2$$

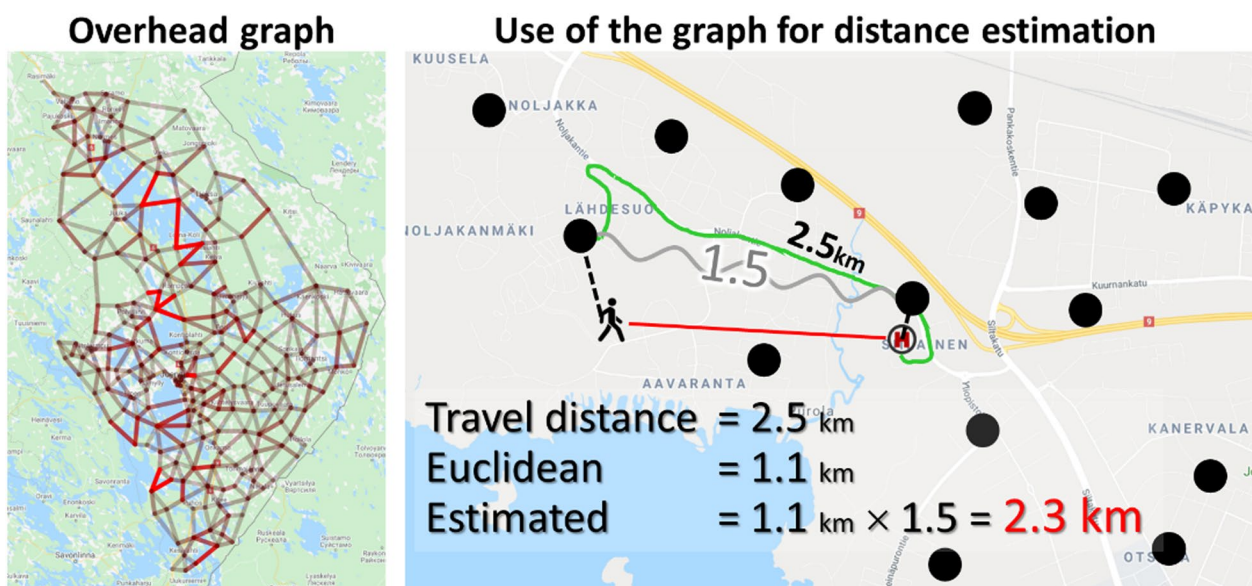
$$\text{Travel cost} : \sum_{p \in \text{patients}} d_{\text{travel}} \left( p, \operatorname{argmin}_{h \in \text{HS}} d_{\text{travel}}(h, p) \right)$$

where *HS* refers to the health stations.

For estimating the travel costs, we adopt the cost model presented by [30] tailored for the local region. The model assumes that patients use the bus when the distance to the nearest bus stop is less than 200 m, otherwise, own car is used. Exceptions are patients living within 1 km from the hospital who are expected to walk to the health station with 0 € cost. People 80 years or older are assumed to use taxis.


The model’s key elements are summarized in Table 1. Our model is slightly simplified; instead of considering different zones for bus fares, we apply a flat rate of 5.1€ per bus trip. Leminen et al. model also considers the cost of the patient’s time during travel based on the average hourly gross wage in the respective zip code area. We have omitted this component to streamline the optimization process.

The ideal goal is to maximize accessibility, but it is not clear how to measure it. We use distance, travel time, and travel costs. Travel time is derived directly from travel distance based on the average speeds of the road segments used and is the most obvious measure. However, people also tend to minimize costs in case of non-urgent



**Fig. 2** Overhead graph with 256 nodes constructed for the North Karelia region and an example of its use

**Table 1** Travel cost model as presented in Leminen et al. [30]

Movement	Dist. to Health Center	Dist. to Bus Stop	Age	Cost
	< 1 km		< 80	0
	> 1 km	> 200 m	< 80	0.45 €/km
	> 1 km	≤ 200 m	< 80	Zone A, B, C : 3.3 € Zone AB or BC : 4.5 € Zone ABC : 7.5 €
			≥ 80	Start: 5.9 € + 1.55 €/km

visits, so the travel cost is also a relevant indicator of accessibility.

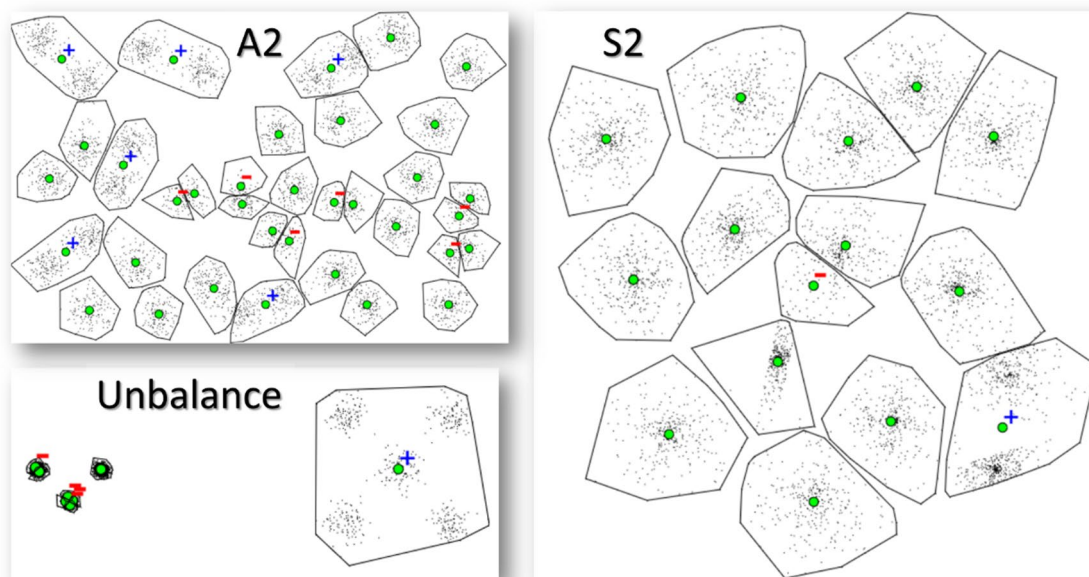
### Clustering algorithm

K-means and its variants like p-median would be the most obvious choices for the clustering algorithm, but they can be inaccurate, see Fig. 3. It was shown in [31] that k-means works worse when the number of clusters is high, and the clusters are well separated. Part of these problems can be compensated by repeating the algorithm multiple times [32] at the cost of increased processing time or by better initialization using methods like Maxmin [33] or its variant called k-means++ [34]. Despite k-means working well for most data, neither of

these alternatives was able to cluster all benchmark datasets correctly [32].

P-median [26, 27] suffers from the same problems as k-means. P-median uses a median for the cluster centroid instead of a mean. It is potentially more robust on noise, but recent results have shown Medoid performing poorly in the context of averaging GPS trajectories [35]. Our data is from patients' real home locations, which is much less noisy than the result of some medical measurement processes.

The accuracy of k-means is usually sufficient when applied as a data processing component within a more complex pattern recognition system. Kinnunen et al. [36] reported that the choice of the algorithm was negligible



**Fig. 3** Incorrectly detected clusters (+ sign signifies too many centroids,—sign too few) happen in k-means when there are many clusters and some of them are well separated. K-means may fail even with seemingly easy datasets (named A2, S2, Unbalance) to find all clusters correctly because the algorithm is incapable of moving the centroids across empty areas (deserts). Repeating the algorithm compensates for this only partly but relies too much on luck

## Random Swap (RS)

```

Random Swap( $X$ )  $\rightarrow$   $C, P$ 

 $C \leftarrow$  Select random representatives( $X$ );
 $P \leftarrow$  Optimal partition( $X, C$ );
REPEAT  $T$  times
  ( $C^{new}, j$ )  $\leftarrow$  Random swap( $X, C$ );
   $P^{new} \leftarrow$  Local repartition( $X, C^{new}, P, j$ );
   $C^{new}, P^{new} \leftarrow$  Kmeans( $X, C^{new}, P^{new}$ );
  IF  $f(C^{new}, P^{new}) < f(C, P)$  THEN
    ( $C, P$ )  $\leftarrow$   $C^{new}, P^{new}$ ;
RETURN ( $C, P$ );

```

P. Fränti, "Efficiency of random swap clustering", *Journal of Big Data*, 2018.

## Genetic Algorithm (GA)

```

GeneticAlgorithm( $X$ )  $\rightarrow$  ( $C, P$ )
FOR  $i \leftarrow 1$  TO  $Z$  DO
   $C^1 \leftarrow$  RandomCodebook( $X$ );
   $P^1 \leftarrow$  OptimalPartition( $X, C^1$ );
SortSolutions( $C, P$ );
REPEAT
  ( $C, P$ )  $\leftarrow$  CreateNewSolutions( $\{C, P\}$ );
  SortSolutions( $C, P$ );
UNTIL no improvement;
CreateNewSolutions( $\{C, P\}$ )  $\rightarrow$  ( $C^{new}, P^{new}$ )
 $C^{new1}, P^{new1} \leftarrow C^1, P^1$ ;
FOR  $i \leftarrow 2$  TO  $Z$  DO
  ( $a, b$ )  $\leftarrow$  SelectNextPair;
   $C^{newi}, P^{newi} \leftarrow$  Cross( $C^a, P^a, C^b, P^b$ );
  IterateK-Means( $C^{newi}, P^{newi}$ );
Cross( $C^1, P^1, C^2, P^2$ )  $\rightarrow$  ( $C^{new}, P^{new}$ )
 $C^{new} \leftarrow$  CombineCentroids( $C^1, C^2$ );
 $P^{new} \leftarrow$  CombinePartitions( $P^1, P^2$ );
 $C^{new} \leftarrow$  UpdateCentroids( $C^{new}, P^{new}$ );
RemoveEmptyClusters( $C^{new}, P^{new}$ );
IS( $C^{new}, P^{new}$ );
CombineCentroids( $C^1, C^2$ )  $\rightarrow$   $C^{new}$ 
 $C^{new} \leftarrow C^1 \cup C^2$ 
CombinePartitions( $C^{new}, P^1, P^2$ )  $\rightarrow$   $P^{new}$ 
FOR  $i \leftarrow 1$  TO  $N$  DO
  IF  $\|x_i - c_{p_i}\|^2 \leq \|x_i - c_{p_j}\|^2$  THEN
     $p_i^{new} \leftarrow p_j^1$ 
  ELSE
     $p_i^{new} \leftarrow p_i^2$ 
END-FOR
UpdateCentroids( $C^1, C^2$ )  $\rightarrow$   $C^{new}$ 
FOR  $j \leftarrow 1$  TO  $|C^{new}|$  DO
   $c_j^{new} \leftarrow$  CalculateCentroid( $P^{new}, j$ );

```

P. Fränti, "Genetic algorithm with deterministic crossover for vector quantization", *Pattern Recognition Letters*, 2000.

**Fig. 4** Pseudo codes of two good clustering algorithms. We selected RandoSwap due to its simplicity

on the overall speaker recognition performance as long as a reasonably good algorithm was chosen (including repeated k-means). However, clustering is the core component of our analysis, and high accuracy is required to avoid any bias caused by the algorithm. For this reason, we have selected a more robust algorithm.

Many potentially good clustering algorithms exist including Ward's agglomerative clustering method [37], its enhanced variant called iterative shrinking [38], splitting algorithm [39], global k-means [40], and evolutionary algorithms of which the genetic algorithm (GA) [41] and the self-adaptive genetic algorithm (SAGA) [42] have shown to be the most accurate in terms of minimizing the clustering objective function.

Among the many good choices, we select random swap [43]. It performs virtually as well as the more complex genetic algorithms while having the benefit of straightforward implementation, see the pseudo-codes of random swap and the genetic algorithm in Fig. 4. Its simplicity is important because it allows easier adaptation of the algorithm to work with different distance functions such as travel cost. Several implementations with different programming languages<sup>1</sup> are also publicly available, including a version supporting parallel processing [44].

The random swap algorithm works as follows. It starts with random initial locations for the stations and then uses two k-means iterations to reallocate the patients to their nearest station and then re-optimize the stations' locations. Random swap is a wrapper around the

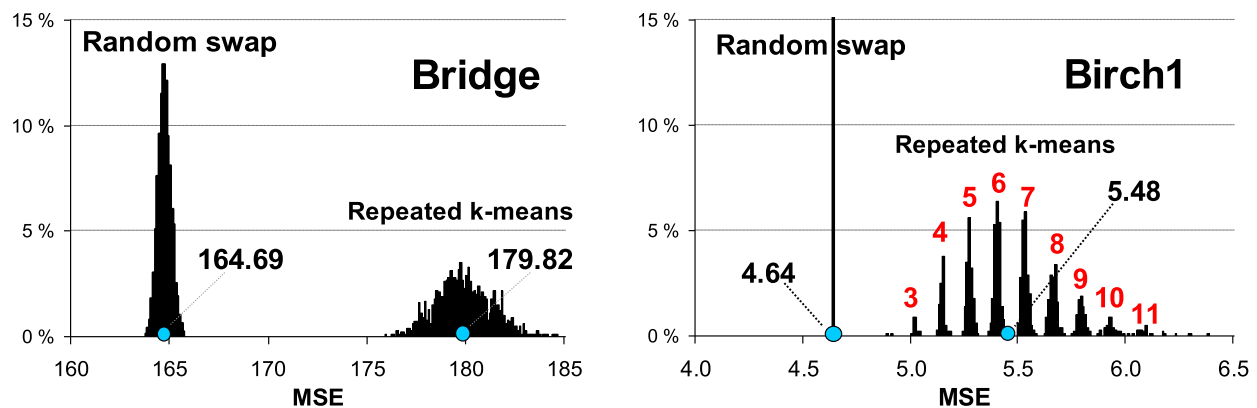
k-means. It selects a random station, relocates it to a new (random) location, and then iterates k-means twice. If the new solution improves the cost function, it is kept; otherwise, the previous solution is restored. While seemingly naïve, this simple trial-and-error approach is effective as the trial swaps can be implemented efficiently.

Here, we use  $T=5000$  trial iterations, which is the original recommendation. The algorithm is not particularly sensitive to this parameter, and the exact value is not even important in offline optimization as we can easily run the algorithm as long as we want, e.g., let the algorithm run 100,000 iterations overnight just to be sure. A theory about how to set this value more accurately can be found in [43]. Visual animation of the algorithm can be found here <https://cs.uef.fi/ml/software/> and tutorial here <https://www.youtube.com/@pasifranti541>.

Figure 5 shows the benefit of random swap over k-means. The average squared distance (MSE) is 10% smaller for data with lots of clusters (Bridge). Repeating k-means 100 times can improve accuracy, but it cannot reach the same accuracy level. For data with 100 separate clusters (Birch1), the difference is 15%. K-means locates 7 clusters wrongly. If repeated 2500 times, the number will be reduced to 3. Even with a small number, it is too much when accuracy is important.

For the centroid, we use arithmetic means of all the patient locations in the cluster. This is not the optimal choice for minimizing the travel cost, but it is the best we can think of, and it can be calculated fast. Finding a better location could theoretically be obtained by a local search around the current location, but re-calculating all distances would be time-consuming. Instead, we consider fine-tuning the centroid location to the nearest existing

<sup>1</sup> <https://cs.uef.fi/ml/software/> and <https://github.com/uef-machine-learning>



**Fig. 5** Effect of the optimization by random swap algorithm versus k-means

**Table 2** Summary of data

Data	Values
Time range	2011–2014
Cohort	Type 2 Diabetes patients
Number of patients	9333
Male	47% (4387)
Female	53% (4946)
Mean age (in 2011)	67 years
Number of visits	175,039
Geographical location	North-Karelia, Finland

building. While new buildings can (and probably should) be constructed, this at least restricts the centroids from being located on lakes and inaccessible places without any infrastructure nearby.

### Case study: SiunSote healthcare data

As a case study, we used data from type 2 diabetes patients in North Karelia, Finland. The reason for selecting this cohort is that it includes the exact locations of the patients that we need for optimization. For the full data, we only have the postal code accuracy. It is accurate enough for optimizing the locations of PCI units throughout Finland [24], but it is not necessarily for optimizing the locations of local health stations.

The data is extracted from Siun Sote's electronic patient records, which organizes the regional health services in North Karelia. The data includes all type 2 diabetes patients ( $n = 9333$ ) diagnosed by the end of 2012 and having visits to primary health care between 2011 and 2014 (175,039 visits in total, averaging 4.7 visits annually per person), see Table 2. Type 2 diabetes patients need frequent follow-up and often have co-morbidities resulting in heavy use of health services. The study restricted the

use of primary health care services, and thus, information on specialized care visits was excluded.

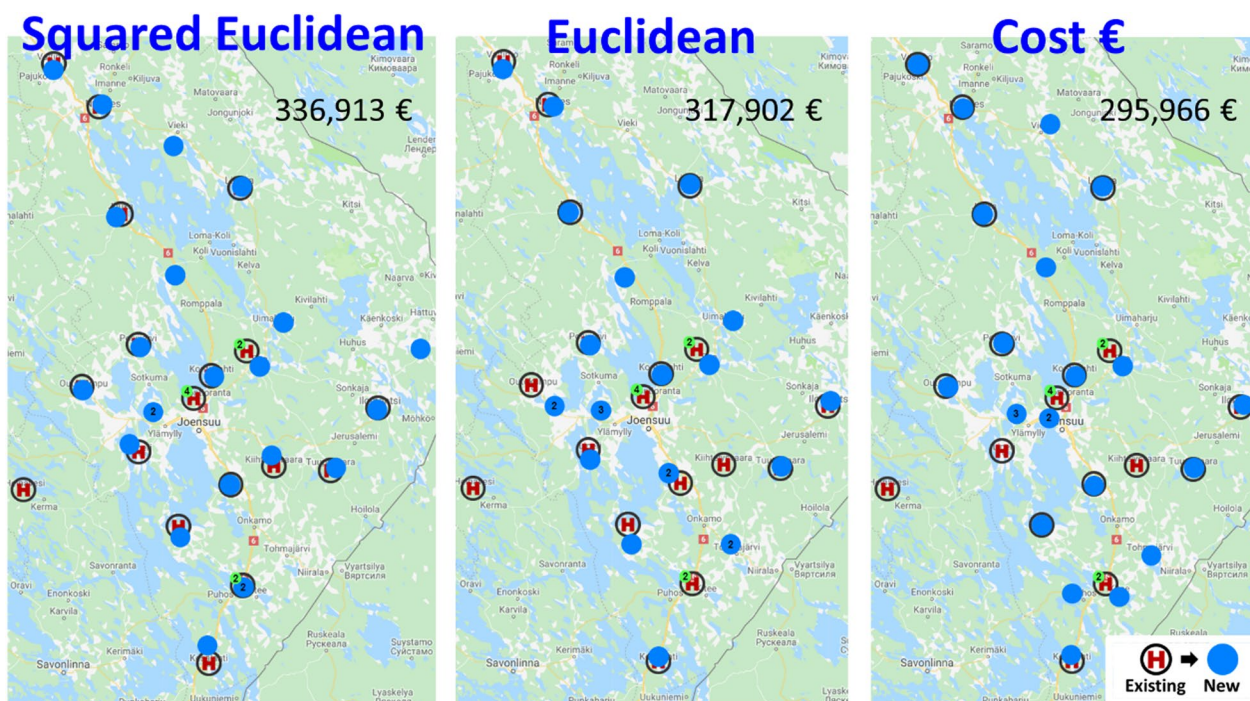
The exact locations were extracted by geocoding the patient home addresses using ArcGIS 10.3 [45]. The information on primary health care visits includes information on the health station where the visit occurred. Their locations were also geocoded based on the address.

The distances to the existing health stations and nearest bus stops were calculated using road network distance information from the OpenStreetMap (OSM) server. Distances to the optimized health station locations were estimated using the overhead graph, as explained in Sect. "Distance calculation". The travel costs were estimated using the model by [30], as discussed in Sect. "Optimization function". It uses factors like patients' age, distance to the health station, and the possibility of using public transport; the most likely travel mode was decided.

### Results

Figure 6 shows the overall optimization results using three alternative distance functions. First, we can see that in all cases, most of the optimized locations are roughly at the same locations as the current health stations, with minor variations. First, squared Euclidean distance is used in standard k-means packages, and it penalizes long distances by using a quadratic function. This creates stations more easily in sparsely inhabited areas. Euclidean distance is more conservative in this regard. The effect of travel cost is more difficult to predict since it utilizes both the road network, the location of bus stops, and the age of the patients (in the case of those older than 80 years).

The total travel cost is the lowest (295 k€/year, 6.76€/visit) when optimized directly for the travel costs. Using Euclidean distance is somewhat worse (317 k€/year,



**Fig. 6** The optimization results are presented for three different distance functions, with the optimized locations indicated by blue points. In some cases, multiple locations are clustered and represented by a single green dot, along with a numerical value denoting the count of stations within the cluster. Additionally, the number above corresponds to the total annual travel cost for all patients, measured in euros

7.26€/visit) but still slightly better than using squared Euclidean distance (336 k€/year, 7.70€/visit).

Next, we will provide a more detailed discussion of the implications of the optimization results. Out of the 23 stations, 17 have remained near their original locations. These locations benefit from a road network and public transportation system already optimized for their accessibility. However, six health stations were located differently by the algorithm. We will briefly summarize these changes and their underlying reasons as follows:

- Heinävesi removed: lack of patient data.
- Viinijärvi and Ylämylly added: ignoring municipality borders.
- Joensuu Center merged with downtown+Lehmo was removed.
- Uimaharju removed and Kopravaara added: serves more people.
- Mätäsvaara and Multala added: in the middle of nowhere.
- Kiihtelysvaara removed + Rääkkylä relocated logistically.

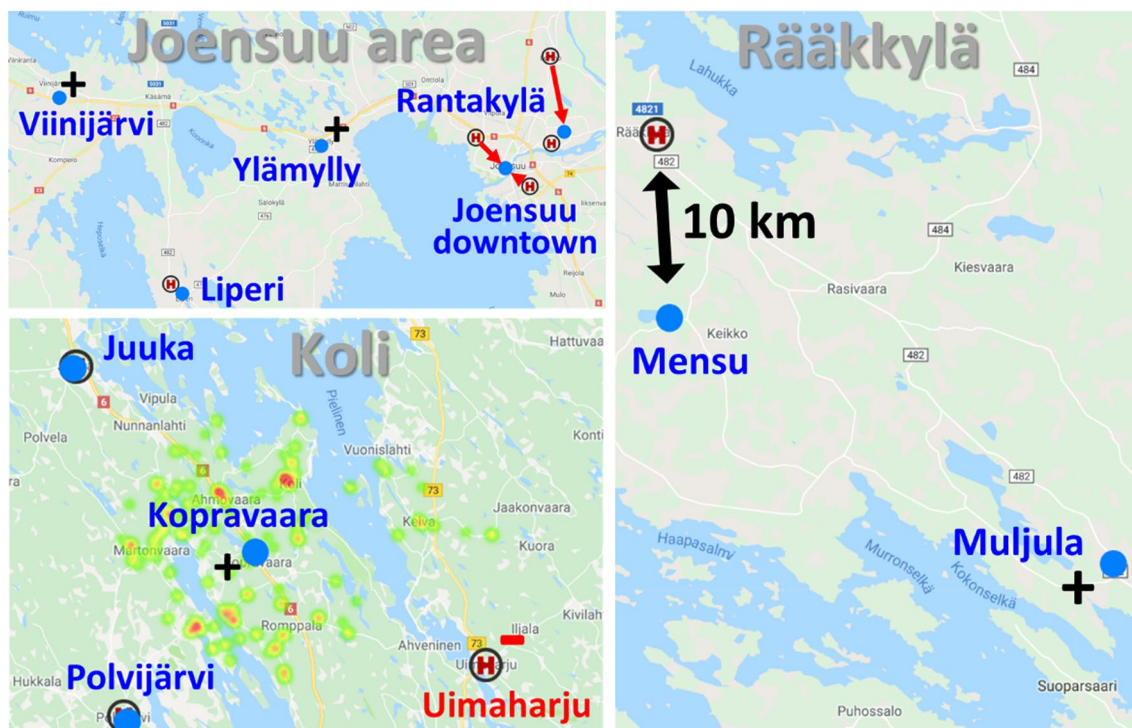
The first change is merely a data artifact. At the time of data collection, Heinävesi health station belonged to Siun-Sote (wellbeing services) county but administratively did

not belong to the North Karelia province (joined later). As a result, the data had Heinävesi health station but not the patients. For this reason, the algorithm naturally relocated the unused center elsewhere.

The second change is surprising, although logical. The current locations adhere primarily to municipal boundaries. This results in three centers (Siilainen, Niinivaara, Rantakylä) within the Joensuu urban area (~59,000 inhabitants), see Fig. 7. In contrast, the neighboring municipality Liperi (12,104 inhabitants) has only one. Liperi itself is an interesting case consisting of three distinct hubs: Liperi center (1,401 inhabitants), Viinijärvi village (746 inhabitants), and Ylämylly suburban center (~6,000 inhabitants). Ylämylly is well connected to Joensuu City through a fast motorway and is essentially considered a suburban extension of Joensuu.

The algorithm disregards municipal boundaries and allocates one station to each of the three hubs in Liperi. Placing a station in Ylämylly is logical, as it serves not only Ylämylly but also the adjacent westernmost Joensuu suburban area, Marjala, with a population of 2,328. Relocating one station to Viinijärvi is less obvious but is explained by the significant savings in travel costs. The total cumulative costs of the three new stations are (90 k€) consisting of Liperi (38 k€), Viinijärvi (27 k€), and Ylämylly (25 k€), whereas the cost to the current station in Liperi is 160 k€. This results in cost savings of 70 k€.





**Fig. 7** Places with significant changes. The heat map shows the home locations of the patients to whom the new Kopravaara is the nearest health station

The costs of the removed stations in the city of Joensuu are Siilainen (125 k€) and Niinivaara (44 k€), and their replacement at Joensuu downtown is only 139 k€. In other words, the two stations in Joensuu can be simply placed in a better location at the heart of the city downtown, still achieving a 30 k€ reduction in travel costs for the patients. The unused station can then be relocated elsewhere in Ylämylly.

The fourth change is the removal of Uimaharju station and the creation of a new one in Kopravaara, near Koli. Uimaharju is a small town (1,300 inhabitants), best known for its pulp mill and sawmill, but with a steadily decreasing population. Although the removal of the Uimaharju<sup>2</sup> health station resulted in an increase in travel costs for the Eno health station from 42 to 80 k€, the more strategic placement in the Koli area yielded greater reductions in the overall travel costs for Juuka, Polvijärvi and Kontiolahti. The importance of the Koli area has been increasing due to its famous national park, including the national view of Finland, which has increased both the number of visitors and the population in the area.

One new addition is Mätäsvaara (917 inhabitants), located roughly halfway between Nurmes and Lieksa, see Fig. 8. Another one is Muljula (382 inhabitants), somewhere between Rääkkylä and Kitee. Neither of them serves a major population, but their effect on travel costs is apparently so high that the model allocated a station there. For example, adding Mätäsvaara reduced the calculative costs of Nurmes from 128 to 88 k€ and Lieksa from 118 to 93 k€.

The removal of Kiihtelysvaara station (3622 inhabitants) did not cause any other changes in the area. The patients are simply treated cost efficiently either in Hammaslahti or Tuupovaara health stations. They all belong to Joensuu municipality.

Rääkkylä is one of the rare tiny municipalities that have kept their independence despite having only about 3,000 inhabitants. It is relatively large (530.85 km<sup>2</sup>) and sparsely populated (4.7 inhabitants per square kilometer), even on a Finland scale. The number of inhabitants per square meter is 9.4 in North Karelia and 17.6 in the entire Finland, on average. Due to the sparsely populated area, the health station is optimized for a logistically better place in Mensu, seemingly in the middle of nowhere, 10 km to the south of the current location in the Rääkkylä town.

<sup>2</sup> Uimaharju health station was stopped in 2019. No new station was founded in Koli, though.



**Fig. 8** Other places that lost or gained their health station. Heinävesi patient data were missing, so this removal is merely an artifact of the data. Kiihtelysvaara previously operated as a separate municipal station, maintaining its own health station despite serving a relatively small patient population. The algorithm optimized a new station in Mätäsvaara instead

**Table 3** Effect of the optimization on the average travel distance and travel time

	Optimized for	Travel time	Travel distance	Travel cost
Original	–	10.1 min	8.4 km	367 399 €
Optimized	Squared Euclidean	9.1 min	6.8 km	336 913 €
Optimized	Euclidean	8.4 min	6.2 km	317 902 €
Optimized	Travel cost	8.1 min	6.2 km	295 966 €
Saving		20%	26%	19%

To summarize the findings, we can see that the algorithm ignores administrative and historical boundaries and can suggest better locations. Employing three distinct cost functions for optimization, we observed that while squared Euclidean distance (commonly used in standard k-means) enhanced the original locations, the non-squared alternative yielded superior results. The best results came directly from using the travel cost of the patients as an optimization function.

The optimized locations would reduce the total travel costs by 19%, from 367 to 296 k€ (Table 3). This means better accessibility and reducing the average travel time of a single visit from 10.1 to 8.1 min and travel distance from 8.4 to 6.2 km. This would be a remarkable achievement if it could be implemented in real life.

## Conclusions

We have designed a clustering algorithm to optimize the locations of health stations. The advantage of the proposed approach is its simplicity. It should also generalize to entire Finland and other countries if data is available. The travel cost needs to be tuned to the region applied,

but the distance and travel time are generic. The clustering approach itself can be tuned relatively easily to be used with other optimization objectives.

The results with SiunSote data in North Karelia, Finland, showed that the algorithm would ignore municipal borders and emphasize the distances and travel costs instead. It allocated a new station in Ylämylly and removed one from Lehmo as there is another station close to Rantakylä, even in a different municipality. The algorithm also removed Uimaharju station, which actually happened in practice after the data was collected. No new station has been founded in the Koli area, though, despite what the algorithm has suggested.

The choice of the cost function has a significant effect on the result. For example, optimizing for squared Euclidean distance (as in standard k-means) would penalize bigger distances more and, therefore, allocate more resources to remote places.

Optimizing travel costs exploits the existing transportation routes. Evidence of this is that the algorithm allocated one station in downtown Joensuu, which is basically next to the start/end points of most bus routes. The travel cost model does not emphasize the distance as much as squared Euclidean, but it also allocates one station halfway between Lieksa and Nurmes, an area having less public transportation and higher travel costs due to the more frequent need for car transportation.

The algorithm can provide better optimization of the resources and would be applicable to data from other areas and different patient cohorts. However, the results cannot be easily applied to real life. Optimizing the overall healthcare service is more complex, and we should also consider factors such as volume, specializations, quality of care, and effectiveness.

The optimization exercise can still provide added value, especially when there is a need to cut down or increase the number of health stations or hospitals that provide different services. Such situations easily occur when health services are reorganized, as happened in Finland at the beginning of 2023 when a large national social and health care reform occurred. The algorithm could also be used for forecasted data of forecasted future population, which would better consider aging and internal migration.

### Limitations

We have done our best to make sure the optimization is well done, and the possibility of results caused by algorithm artifacts remains quite low, contrary to the known limitations of algorithms like k-means and also p-median. However, when dealing with real data, there are always issues that affect the results.

One known limitation in the optimization process is the use of a simple geometric center for the health station location. This may not be the optimal location when minimizing the travel cost. Theoretically, better locations could be found by tuning the locations using an iterative local search algorithm. Starting from the geometric center, the algorithm could consider its neighboring locations in a trial-and-error manner utilizing the overhead graph [29]. However, this would increase the processing time considerably, and it is uncertain if the potentially more accurate optimization would be worth the additional computing.

Another limitation was the lack of Heinävesi population data. The consequence of this is that this particular health station was removed without any cost, and additional resources were placed elsewhere. Another issue is that the volume and specialization of the stations were not considered in the optimization. All health stations were assumed to be of equal importance, having full service for all patients, which is not always the case. Many smaller stations only provide basic primary care services, whereas larger stations can also provide specialized services.

A third limitation of this case study is that only one patient cohort was used: type 2 diabetes patients. This does not, of course, represent the service use of the whole population and is biased toward the elderly population. However, patients with type 2 diabetes are high-service users using primary care services and can be regarded as a typical patient group in primary health care.

Finally, we only considered the geographic distance to measure accessibility. It would be interesting to repeat the optimization process with other objectives and more parameters. For example, the ratio of physicians to patients is an interesting measure, but the simplistic

thresholding in the method should be removed; otherwise, it would suffer the same problem as all other methods optimizing maximum coverage.

### Acknowledgements

This study is a part of the consortium project "Improving the Information Base and Optimising Service Solutions to Support Social Welfare and Healthcare Reform (IMPRO)." The research was funded by the Strategic Research Council at the Academy of Finland, funding decision numbers 336325 and 336330.

### Author contributions

P.F. and S.S. constructed the research, prepared the illustrations, and wrote the paper. T.L. provided the data. The analysis of the results was done jointly by all authors.

### Availability of data and materials

No datasets were generated or analysed during the current study.

### Declarations

#### Competing interests

The authors declare no competing interests.

Received: 27 September 2024 Accepted: 21 February 2025

Published online: 22 March 2025

### References

- Collet JP, et al. ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: the task force for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation of the European Society of Cardiology (ESC). *Eur Heart J*. 2021;42(14):1289–367.
- Gu W, Wang X, McGregor SE. Optimization of preventive health care facility locations. *Int J Health Geogr*. 2010;9:1–16.
- Arcury TA, Gesler WM, Preisser JS, Sherman J, Spencer J, Perin J. The effects of geography and spatial behavior on health care utilization among the residents of a rural region. *Health Serv Res*. 2005;40:135–55.
- Lankila T, Laatikainen T, Wikström K, Linna M, Antikainen H. Association of travel time with mental health service use in primary health care according to contact type—a register-based study in Kainuu, Finland. *BMC Health Serv Res*. 2022;22(1):1458.
- Di Domenicantonio R, Cappai G, Sciattella P, Belleudi V, Di Martino M, Agabiti N, et al. The tradeoff between travel time from home to hospital and door-to-balloon time in determining mortality among STEMI patients undergoing PCI. *PLoS ONE*. 2016;11(6):e0158336.
- Pyykönen M, Linna M, Tykkyläinen M, Delmelle E, Laatikainen T. Patient-specific and healthcare real-world costs of atrial fibrillation in individuals treated with direct oral anticoagulant agents or warfarin. *BMC Health Serv Res*. 2021;21(1):1299.
- Ahmadi-Javid A, Seyedi P, Syam SS. A survey of healthcare facility location. *Comput Oper Res*. 2017;79:223–63.
- Murray AT. Maximal coverage location problem: impacts, significance, and evolution. *Int Reg Sci Rev*. 2016;39(1):5–27.
- Fränti P, Măriescu-Istodor R, Akram A, Satokangas M, Reissell E. Can we optimize locations of hospitals by minimizing the number of patients at risk? *BMC Health Serv Res*. 2023;23(1):415.
- Burkey ML, Bhadury J, Eiselt HA. A location-based comparison of health care services in four US states with efficiency and equity. *Socioecon Plann Sci*. 2012;46(2):157–63.
- Wang F, Tang Q. Planning toward equal accessibility to services: a quadratic programming approach. *Environ Plann B Plann Des*. 2013;40(2):195–212.
- Tao Z, Wang Q, Han W. Towards health equality: optimizing hierarchical healthcare facilities towards maximal accessibility equality in Shenzhen. *China Appl Sci*. 2021;11(21):10282.

13. Tao Z, Cheng Y, Dai T, Rosenberg MW. Spatial optimization of residential care facility locations in Beijing, China: maximum equity in accessibility. *Int J Health Geogr*. 2014. <https://doi.org/10.1186/1476-072X-13-33>.
14. Fo ARAV, da Silva Mota I. Optimization models in the location of health-care facilities: a real case in Brazil. *J Appl Operat Res*. 2012;4(1):37–50.
15. Shen Q. Location characteristics of inner-city neighborhoods and employment accessibility of low-wage workers. *Environ Plann B Plann Des*. 1998;25(3):345–65.
16. Jin M, Liu L, Tong D, Gong Y, Liu Y. Evaluating the spatial accessibility and distribution balance of multi-level medical service facilities. *Int J Environ Res Public Health*. 2019;16(7):1150.
17. Tao Z, Cheng Y, Liu J. Hierarchical two-step floating catchment area (2SFCA) method: measuring the spatial accessibility to hierarchical healthcare facilities in Shenzhen, China. *Int J Equity Health*. 2020;19(164):1–16.
18. Tran TC, Dinh TB, Gascon V. Meta-heuristics to solve a districting problem of a public medical clinic. *Int Sympos Inform Commun Technol*. 2017. <https://doi.org/10.1145/3155133.3155146>.
19. Braekers K, Hartl RF, Parragh SN, Tricoire F. A bi-objective home care scheduling problem: analyzing the trade-off between costs and client inconvenience. *Eur J Oper Res*. 2016;248(2):428–43.
20. P. Fränti, R. Mariescu-Istodor and A. Akram, (2022), Web-tool for optimizing locations of health centers, *Int. Conf. on Health and Social Care Information Systems and Technologies (HCist'2022)*, Lisboa, Portugal.
21. Zhou, W., & Li, Z. (2013). The multi-covering emergency service facility location problem with considering disaster losses. *IET International Symposium on Operations Research and its Applications in Engineering, Technology and Management (ISORA)*, 1–6.
22. Church RR, ReVelle C. The maximal covering location problem. *Papers Region Sci Assoc*. 1974;32(1):101–18.
23. T. Lähderanta, L. Lovén, L. Ruha, T. Leppänen, M. Kuismin, I. Launonen, S. Piirtikangas, J. Riekkö and M.J. Sillanpää (2023). The intersection of location-allocation, partitionial clustering and model-based clustering techniques. *Manuscript*. (submitted)
24. Fränti P, Mariescu-Istodor R, Akram A, Satokangas M, Reissell E. Can we optimize locations of hospitals by minimizing the number of patients at risk? *BMC Health Serv Res*. 2023;23(415):1–12.
25. Äyrämö S, Kärkkäinen T, Majava K. Robust refinement of initial prototypes for partitioning-based clustering algorithms. Singapore: World Scientific; 2007. p. 473–82.
26. Mladenović N, Brimberg J, Hansen P, Moreno-Pérez JA. The p-median problem: a survey of metaheuristic approaches. *Eur J Oper Res*. 2007;179(3):927–39.
27. Daskin MS, Maass KL. The p-median problem. In: Laporte G, Nickel S, Saldanha Gama F, editors. *Location science*. Cham: Springer; 2015. p. 21–45.
28. Boscoe FP, Henry KA, Zdeb MS. A nationwide comparison of driving distance versus straight-line distance to hospitals. *Prof Geogr*. 2012;64(2):188–96.
29. Mariescu-Istodor R, Fränti P. Fast travel distance estimation using overhead graph. *J Locat-Based Serv*. 2021;15(4):261–79.
30. Leminen A, Tykkyläinen M, Laatikainen T. Self-monitoring induced savings on type 2 diabetes patients' travel and healthcare costs. *Int J Med Inform*. 2018;115:120–7.
31. Fränti P and Sieranoja S. K-means properties on six clustering benchmark datasets. *Applied Intelligence*. 2018;48(12):4743–59.
32. Fränti P, Sieranoja S. How much k-means can be improved by using better initialization and repeats? *Pattern Recognit*. 2019;93(95–112):2019.
33. Gonzalez TF. Clustering to minimize the maximum intercluster distance. *Theoret Comput Sci*. 1985;38:293–306.
34. Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. Stanford: Stanford; 2006.
35. Jimoh B, Mariescu-Istodor R, Fränti P. Is medoid suitable for averaging GPS trajectories? *ISPRS Int J Geo Inf*. 2022;11(2):133.
36. Kinnunen T, Sidoroff I, Tuononen M, Fränti P. Comparison of clustering methods: a case study of text-independent speaker modeling. *Pattern Recogn Lett*. 2011;32(13):1604–17.
37. Ward JH Jr. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*. 1963;58(301):236–44.
38. Fränti P, Virmajoki O. Iterative shrinking method for clustering problems. *Pattern Recogn*. 2006;39(5):761–75.
39. Fränti P, Kaukoranta T, Nevalainen O. On the splitting method for vector quantization codebook generation. *Optical Engineering*. 1997;36(11):3043–51.
40. Likas A, Vlassis N, Verbeek JJ. The global k-means clustering algorithm. *Pattern Recogn*. 2003;36(2):451–61.
41. Fränti P. Genetic algorithm with deterministic crossover for vector quantization. *Pattern Recogn Lett*. 2000;21(1):61–8.
42. Kivijärvi J, Fränti P, Nevalainen O. Self-adaptive genetic algorithm for clustering. *J Heuristics*. 2003;9(2):113–29.
43. Fränti P. Efficiency of random swap clustering. *J Big Data*. 2018;5(13):1–29.
44. Nigro L, Cicirelli F, & Fränti P (2022). Parallel Random Swap: An Efficient and Reliable Clustering Algorithm in Java. *Simulation Modelling Practice and Theory*, 102712.
45. Environmental Systems Research Institute (ESRI) (2014). ArcGIS, version 10.3. Redlands, CA: ESRI.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.