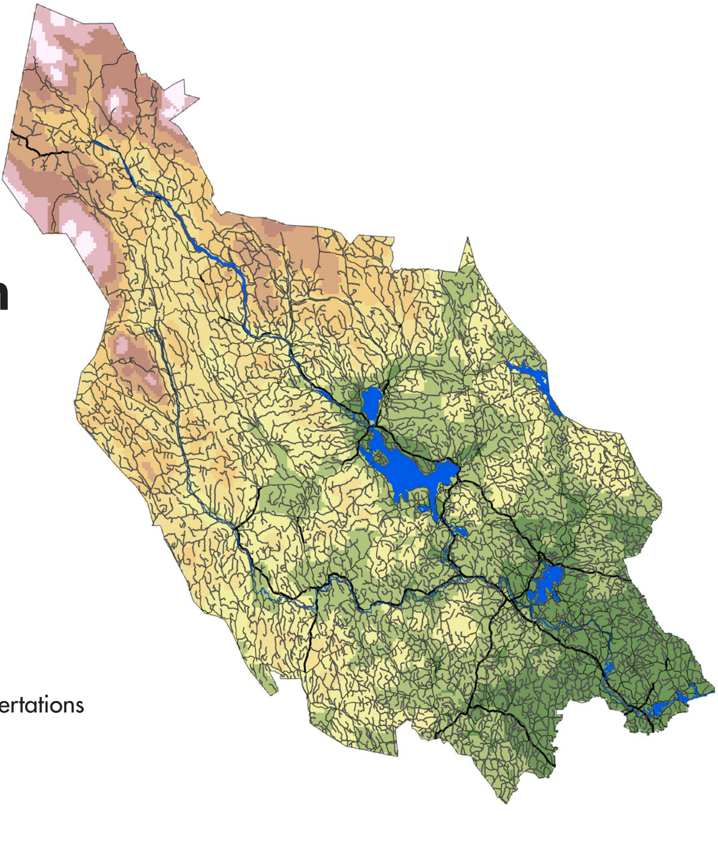
DALARNA
UNIVERSITY

Heuristic Optimization of the *p*-median Problem and Population Re-distribution



Mengjie Han

Serie: Dalarna Doctoral Dissertations
Serial number: 1
ISBN: 978-91-89020-89-4

Heuristic Optimization of the p-median Problem and Population Re-distribution

Mengjie Han

DALARNA
UNIVERSITY

Micro-data Analysis
School of Technology and Business Studies

Dalarna University, Sweden
October 2013

ISBN: 978-91-89020-89-4

Abstract

This thesis contributes to the heuristic optimization of the p-median problem and Swedish population redistribution.

The p-median model is the most representative model in the location analysis. When facilities are located to a population geographically distributed in Q demand points, the p-median model systematically considers all the demand points such that each demand point will have an effect on the decision of the location. However, a series of questions arise. How do we measure the distances? Does the number of facilities to be located have a strong impact on the result? What scale of the network is suitable? How good is our solution? We have scrutinized a lot of issues like those. The reason why we are interested in those questions is that there are a lot of uncertainties in the solutions. We cannot guarantee our solution is good enough for making decisions. The technique of heuristic optimization is formulated in the thesis.

Swedish population redistribution is examined by a spatio-temporal covariance model. A descriptive analysis is not always enough to describe the moving effects from the neighbouring population. A correlation or a covariance analysis is more explicit to show the tendencies. Similarly, the optimization technique of the parameter estimation is required and is executed in the frame of statistical modeling.

Keywords: optimization; heuristic; p-median; spatio-temporal covariance

Acknowledgements

First and foremost I, hereby, would like to show my deepest gratitude to my supervisor, the statistics professor, Kenneth Carling. Kenneth Carling is one of the most excellent, responsible and instructive supervisors I have ever seen. With his enlightening and distinctive guidance, I have finished writing six PhD dissertation theses and two additional papers within only 38 months. About three years ago, I knew nothing about my PhD life. It was Kenneth who unselfishly lead me to this enjoyable research career. I truly appreciate his dedication on my thesis. Without his profound ideas and patient correction, it was impossible for me to pass the licentiate defence. Kenneth has developed in me many useful thinking styles which can be seen from my both published and unpublished papers. I have benefited a lot from his own thinking style. I also would like to thank him for bringing me to the LILLA SS and other smaller skiing races and letting me enjoy the competition and the speed. Last but not the least, I would like to thank Kenneth for giving me a lot of training on running and the medicine ball. I really developed my endurance and strength much.

I send my gratitude to my secondary supervisor, Johan Håkansson. Johan Håkansson is a human geographer who has a different style of supervising. He not only always gives geographical insights but also dynamic ways of thinking in my papers. To be the co-author of Johan, I always have something new to think about and to do. His contribution to my papers is also highly appreciated even when we had a hard time when we responded to the comments from the journals. I really thank him for his valuable comments on my several presentations. It was he who makes my presentations better and better. I also would like to show my gratitude for his tutorials on the down hill skiing and nice arrangements of my dissertation opponent.

I would like to thank Pascal Rebreyend for his good discussion on many practical algorithms, optimization methods and C tutorial. Pascal has set up many novel ideas for improving my papers. His efforts fundamentally solved big data problems, which greatly affect the quality of the papers. I also appreciate Pascal's elegant and fast driving back and forth to the airport at midnight.

Much gratitude must be given to my dear parents, our universe sage Sakyamuni Buddha, Xiaoli Du and all my buddhists colleagues. They are my most precious gift in my life. My parents gave me an opportunity of the reincarnation. Sakyamuni Buddha subverted my understanding of the universe and PhD. My buddhists colleagues have supported me a lot when I was confused about the world. I am sincerely grateful to them for their love.

I would like to thank staff or master student in the statistics department in Dalarna University. Lars Rönnegård who is one of my co-authors and gave me a lot help on likelihood theory and statistical modeling. I would also like to thank Bo Zhu. Her initial contribution to my first paper is highly appreciated. Md.Moudud Alam gave me many active GLM lectures. He also organized many good seminars such that I can have access to many interesting talks. I also thank Xiangli Meng, Xiaoyun Zhao and Yujiao Li for their kind sharing of our office and quite a few nice discussions. Much gratitude is also sent to Xia Shen, Majbritt Felleki, Daniel Wiström, Ola Nääs, Kristin Svenson and Richard Stridbeck.

Last but not the least, much gratitude is given to Anders Forsman, Sune Karlsson, Catia Cialani, Changcheng Yao, Xier Li, Xiu Xie, Weigang Qie, Hao Wang, Ba Zhang, Dao Li and a number of persons who are not mentioned here.

Contents

1 **Introduction** 2

2 **Applications and Models** 2

2.1 p-median model 2

2.2 spatio-temporal covariance model 3

3 **Heuristics** 3

4 **Paper list** 4

5 **Concluding Remarks** 7

1 Introduction

Optimization is a process of reaching the maximum or minimum values of the objective functions. The analytical result and conclusion are always drawn according to optimization outcomes. Optimization is a necessary procedure to evaluate a model. If the optimal values are extremely difficult to obtain, we need to consider adjusting our optimization methods, the models or the restrictions.

Two main specific applications are related to optimization in the thesis. One is on the operational location problem. For location problems, we usually minimize the average or the maximum transportation cost, for example, the distance, the fuel consumption or the traveling time between the demand points and the facilities. Badly located facilities can greatly increase the average or the maximum cost. The complexity of location problem can increase fairly fast due to the increment of the number of facilities and the possible candidate locations. Since this astronomical combinatorial property can lead to suboptimal solutions, the corresponding optimization operations should be considered.

The other application is on statistical modeling. Statistical models are characterized by random variables and uncertainties. There always exists a gap between the theoretical (or optimal) parameterized curves and empirical (or estimated) curves. The gap comes from the randomness. For some parameterized models, the moment estimator or the likelihood estimator cannot be easily obtained due to the implicit form of the likelihood functions. The iterative optimization method is always employed to improve the estimation and narrow the gap.

2 Applications and Models

Our focus is mainly on location models and the spatio-temporal covariance model. The practical applications are regarding locating hospitals in a region of Sweden and Swedish population redistribution, respectively. Both of them make use of the optimization method to evaluate empirical results.

2.1 p-median model

Location models assist in the location problem by suggesting optimal locations of facilities according to an objective function. For the location models, we represent the problem by the p-median model. The idea is to optimally locate a number of facilities for a population geographically distributed in Q demand points such that the population's average distance is minimized. Hakimi (1964) offers an original and clear structure of this issue including definitions of several key concepts. Several reviews of p-median problem have been made (Farahani *et al.*, 2012; Francis *et al.*, 2009; Reese, J., 2006; and Mirchandani, 1990). Hakimi (1965) showed that the optimal solution of p-median problem can always be found on the nodes. Due to his argument, the p-median problem is always identified as discrete problem (Daskin, 1995). Thus, the definition of the linear integer programming (Rosing, *et al.*, 1979) is:

$$\text{Minimize:} \quad \sum_i \sum_j h_i d_{ij} Y_{ij} \tag{1}$$

subject to:

$$\sum_j Y_{ij} = 1 \quad \forall i \tag{2}$$

$$\sum_j X_j = P \tag{3}$$

$$Y_{ij} - X_j \leq 0 \quad \forall i, j \tag{4}$$

$$X_j = 0, 1 \quad \forall j \tag{5}$$

$$Y_{i,j} = 0, 1 \quad \forall i, j. \tag{6}$$

In (1) h_i is the weight on each demand point and d_{ij} is the cost of the edge. Y_{ij} is the decision variable indicating if a trip between node i and j is made or not. Constraint (2) ensures that every demand point must be assigned to one facility. In constraint (3) X_j is the decision variable and it ensures that the number of facilities to be located is P. Constraint (4) indicates that no demand point i is assigned to j unless there is a facility. In constraint (5) and (6) the value 1 means that the locating (X) or travelling (Y) decision is made. 0 means that the decision is not made.

The p-median model is NP-hard (Kariv and Hakimi, 1979). For the p-median model, many issues can affect the optimal locations or solutions. Thus, we have examined the effect from the distance measure, the impact of the variation on the network density, the impact of the variation on the number of facilities, the assumption that the demand or the customer gravitates to a facility because of the distance to it and the attractiveness of it, and the impact of the step size parameter in the subgradient method on the quality of the p-median optimal solutions. In these empirical studies, different methods and adaptive algorithms are considered for executing or evaluating the optimal solutions.

2.2 spatio-temporal covariance model

On the other hand, the spatio-temporal covariance model is applied to the analysis of population redistribution in Sweden. According to Håkansson (2000), the Swedish population has different redistribution tendencies at a local level and at a regional level. The spatio-temporal covariance model acts as an complementary causality analysis based on the Local Moran's I index analysis.

3 Heuristics

For a optimization problem, a heuristic is designed for solving a problem more quickly when classic methods are too slow, or for finding an approximate solution when classic methods fail to find any exact solution. It is a useful technique for NP-hard problem. Two definitions have shown the essence of the heuristic:

> “A heuristic is a rule of thumb, strategy, trick, simplification, or any other kind of device which drastically limits search for solutions in large problem space. Heuristics do not guarantee optimal solutions; in fact, they do not guarantee any solution at all; all that can be said for a useful heuristic is that it offers solutions which are good enough most of the time.” (Feigenbaum and Feldman, 1963, p.6)

> “Heuristic are criteria, methods, or principles for deciding which among several alternative courses of action promises to be the most effective in order to achive some goal.” (Pearl, 1984, p.3)

Both the location problem and statistical modeling have their limitations when the objective function or the parameterized model is optimized. For the location model, the limitation arises when the problem complexity increases. For the statistical model, both the complexity and the multidimensional non-linear objective function (e.g. likelihood function) can cause limitations. A problem is that the “best” solution is not explicit. Thus, heuristic methods or algorithms are employed.

The evolutionary heuristic method is a suitable operation for the optimization problem, because the solution in the next step or iteration always inherits the “direction” property of the current step or iteration. If this “direction” leads to the optimal solution, a fast and good solution can be obtained. Since it is difficult to identify the right direction, it is usually handled by a heuristic method. This can be seen in the thesis.

4 Paper list

Dissertation thesis:

I Carling, K., **Han, M**. and Håkansson, J., 2012. Does Euclidean distance work well when p-median model is applied in rural areas? *Annals of Operation Research* 201(1), 83–97.

The p-median model is used to locate P centers to serve a geographically distributed population. A cornerstone of such a model is the measure of distance between a service center and demand points, i.e. the location of the population (customers, pupils, patients, and so on). Evidence supports the current practice of using Euclidean distance. However, we find that the location of multiple hospitals in a rural region of Sweden with a non-symmetrically distributed population is quite sensitive to distance measure, and somewhat sensitive to spatial aggregation of demand points.

In this paper, three restrictions are put up to reduce the problem complexity such that the optimal objective function is easily evaluated by the Monte Carlo simulation.

II Carling, K., **Han, M.**, Håkansson, J. and Rebreyend, P., 2012. Distance measure and the p-median problem in rural areas, *Working paper in transport, tourism, information technology and microdata analysis*, ISSN 1650-5581; 07. Submitted.

In this paper we extend the work of Paper 1 by using of a refined network and study systematically the case when P is of varying size (2-100 facilities). We find that the network distance gives as good a solution as the travel-time network. The Euclidean distance gives solutions some 2-7 per cent worse than the network distances, the solutions deteriorate with increasing P. Our conclusions extend to intra-urban location problems.

Since problem complexity is increased, we select the heuristic simulated annealing algorithm as our operating method. The empirical parameters are decided after it was tested on a smaller scale problem.

III **Han, M.**, Håkansson, J. and Rebreyend, P., 2013. How do different densities in a network affect the optimal location of service centers? *Working paper in transport, tourism, information technology and microdata analysis*, ISSN 1650-5581; 15. Submitted to *European Journal of Operational Research.*

The optimal locations are sensitive to geographical context such as road network and demand points especially when they are asymmetrically distributed in the plane. Most studies focus on evaluating performances of the p-median model when p and n vary. To our knowledge this is not a very well-studied problem when the road network is alternated especially when it is applied in a real world context. The aim in this study is to analyze how the optimal location solutions vary, using the p-median model, when the density in the road network is alternated. To locate 5 to 50 service centers we use the national transport administrations official road network (NVDB). The road network consists of 1.5 million nodes. To find the optimal location we start with 500 candidate nodes in the network and increase the number of candidate nodes in steps up to 67,000. To find the optimal solution we use a simulated annealing algorithm with adaptive tuning of the temperature. The results show that there is a limited improvement in the optimal solutions when nodes in the road network increase and p is low. When p is high the improvements are larger. The results also show that choice of the best network depends on p. The larger p the larger density of the network is needed.

The optimal solution is examined in another perspective in this paper. The network density is varied. Our contribution provides a framework of premises before optimization.

IV Carling, K., **Han, M**. and Håkansson, J., 2012. An empirical test of the gravity p-median model. *Working paper in transport, tourism, information technology and microdata analysis*, ISSN 1650-5581; 2012:15. Submitted.

A customer is presumed to gravitate to a facility because of the distance to it and the attractiveness of it. However regarding the location of the facility, the presumption is that the customer opts for the shortest route to the nearest facility. This paradox was recently solved by the introduction of the gravity p-median model. The model is yet to be implemented and tested empirically. We implemented the model in an empirical problem of locating locksmiths, vehicle inspections, and retail stores of vehicle spare-parts, and we compared the solutions with those of the p-median model. We found the gravity p-median model to be of limited use for the problem of locating facilities as it either gives solutions similar to the p-median model, or it gives unstable solutions due to a non-concave objective function.

In this paper we continued using simulated annealing on all optimization problems. To have a idea on how good the solution is, we also examined the bounded optimal value with 99% probability.

V **Han, M.**, Håkansson, J. and Rönnegård, L., 2012. How do neighbouring populations affect local population growth over time? Submitted to *Population, Space and Place*

This study covers a period when society changed from a pre-industrial agricultural society to a post-industrial service-producing society. Parallel with this social transformation, major population changes took place. One problem with geographical population studies over long time periods is accessing data that has unchanged spatial divisions. In this study, we analyse how local population changes are affected by neighbouring populations. To do so we use the last 200 years of population redistribution in Sweden, and literature to identify several different processes and spatial dependencies. The analysis is based on a unique unchanged historical parish division, and the methods used are an index of local spatial correlation. To control inherent time dependencies, we introduce a non-separable spatio-temporal correlation model into the analysis of population redistribution. Several different spatial dependencies can be observed simultaneously over time. The main conclusions are that while local population changes have been highly dependent on the neighbouring populations, this spatial dependence has already become insignificant already when two parishes are separated by 5 kilometres.

Regarding the optimization of spatio-temporal covariance parameters, we need to find a theoretical curve that minimizes the difference to the observed curves. The corresponding iterative method is also applied on the estimation.

VI **Han, M.**, 2013. Computational study of the step size parameter of the subgradient optimization method. Manuscript.

The subgradient optimization method is a simple and flexible linear programming iterative algorithm. It is much simpler than Newton's method and can be applied to a wider variety of problems. It also converges when the objective function is non-differentiable. Since an efficient algorithm will not only produce a good solution but also take less computing time, we always prefer a simpler algorithm with high quality. In this study a series of step size parameters in the subgradient equation are studied. The performance is compared for a general piecewise function and a specific p-median problem. We examine how the quality of solution changes by setting five forms of step size parameter.

A bounded optimal solution is evaluated in this paper, which gives an idea of how good the solution is. Our contribution suggested identifying a set of parameters that produces the minimum error to improve the solution.

Additional papers not included in the thesis:

I Carling, K., **Han, M.**, Håkansson, J., Meng, X. and Rudholm, N., 2013. CO_2-emissions induced by online and brick-and-mortar retailing. Working paper.

II Rebreyend, P., **Han, M**. and Håkansson, J., 2013. How does different algorithm work when applied on the different road networks when optimal location of facilities is searched for in rural areas? Proceeding paper in *The 14th international conference on web system engineering*, Nanjing, China.

5 Concluding Remarks

We usually want to obtain the "best" solution or the almost " best" solution when we are operating optimization on a model. This makes the analytical evaluation of the computational solution very important, because we cannot guarantee whether our solution is the "best" or almost the " best". Thus, either the improvement on the algorithm or the adjustment of the model settings can be made for drawing a satisfied conclusion.

In our thesis, the evaluation of the optimized result varies through the application on the location problem and the spatio-temporal covariance model. The resulting analytical conclusion is reliable and meaningful. We can also make corresponding transportation policies and plans that can greatly improve transportation efficiency and reduce operational costs.

References

[1] Daskin, M., 1995. *Network and discrete location*, Wiley, New York.

[2] Farahani, R.Z., Asgari, N., Heidari, N., Hosseininia, M. and Goh, M., 2012. Covering problems in facility location: A review. *Computers and Industrial Engineering* 62(1), 368–407.

[3] Feigenbaum, E.A. and Feldman, J., 1963. *Computers and thought.* McGraw-Hill Inc., New York.

[4] Francis, R., Lowe, T., Rayco, M. and Tamir, A., 2009. Aggregation error for location models: survey and analysis. *Annals of Operations Research* 167, 171–208.

[5] Hakimi, S.L., 1964. Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research*, 12(3), 450–459.

[6] Hakimi, S.L., 1965. Optimum distribution of switching centers in a communications network and some related graph theoretic problems. *Operations Research* 13, 462–475.

[7] Håkansson, J., 2000. Changing population distribution in Sweden — long term contemporary tendencies, Umeå Universitet, *GERUM Kulturgeografi*, 2000:1.

[8] Kariv, O. and Hakimi, S.L., 1979. An algorithmic approach to network location problems. part 2: The p-median. *SIAM J. Appl Math* 37, 539–560.

[9] Mirchandani, P.B., 1990. “The p-median problem and generalizations”, Discrete location theory, *John Wiley & Sons, Inc.*, New York, pp 55-117.

[10] Peral, J., 1984. Heuristics: intelligent search strategies for computer problem solving. Addison-Wesley Publ. Co., London.

[11] Reese, J., 2006. Solution methods for the p-median problem: An annotated bibliography. *Network*s, 48(3), 125–142.

[12] Rosing, K.E., Revelle, C.S. and Rosing-Vogelaar, H., 1979. The p-Median and its Linear Programming Relaxation: An Approach to Large Problems. *The Journal of the Operational Research Society*, 30(9), 815–823.

PAPER I

Ann Oper Res (2012) 201:83–97
DOI 10.1007/s10479-012-1214-2

Does Euclidean distance work well when the *p*-median model is applied in rural areas?

Kenneth Carling · Mengjie Han · Johan Håkansson

Published online: 1 September 2012

Abstract The p-median model is used to locate P centers to serve a geographically distributed population. A cornerstone of such a model is the measure of distance between a service center and demand points, i.e. the location of the population (customers, pupils, patients, and so on). Evidence supports the current practice of using Euclidean distance. However, we find that the location of multiple hospitals in a rural region of Sweden with a non-symmetrically distributed population is quite sensitive to distance measure, and somewhat sensitive to spatial aggregation of demand points.

Keywords Optimal location · Euclidean distance · Network distance · Travel time · Spatial aggregation · Location model

1 Introduction

This work originated from a desire to investigate whether the current locations of two emergency hospitals in the rural region of Dalecarlia in mid-Sweden are accessible to the region's population. These hospitals serve a geographically dispersed and non-symmetrically distributed population, and the recent closure of three emergency hospitals in the region prompted this investigation. Moreover, the regional administrative division of Sweden is currently under revision, and one potential outcome is a reconfiguration of the current 21 regions into significantly fewer regions. Since the regions in Sweden are responsible for providing emergency care and are entitled to collect taxes for this purpose, it is expected that an alteration in regional division would prompt a substantial relocation of emergency hospitals.

To find the optimal locations of hospitals and to compare them with the current situation, location models such as the p-median model are useful. However, such models require a distance measure between hospitals and the population, and data on the population's locations.

K. Carling · M. Han · J. Håkansson (✉)
School of Technology and Business Studies, Dalarna University, 791 88 Falun, Sweden
e-mail: jhk@du.se

The recent location literature involving computational experimentation, primarily in an urban setting, has found that Euclidean distance works well as a distance measure. However, geographical theory suggests that Euclidean distance would work poorly in a rural setting in which the population is typically non-symmetrically distributed in the plane and the road network heterogeneous.

The aim of this paper is to examine whether Euclidean distance works well in rural areas when location models are used. This paper is the first to empirically investigate the consequences of distance measures for the optimal location of multiple service centers in rural areas. The investigation was conducted by means of a case study and several computer experiments using the p-median model. In the experiments, we use, in addition to Euclidean distance, network distance and travel time as measures of distance. Moreover, we consider the optimal allocation of service centers with two to eight hospitals for the case when the number of served residents is just above the required number for efficient emergency care.

Spatial aggregation is known to produce errors, and consequently, a study on distance measures must also consider this issue. Since the publication of Hillsman and Rhoda (1978), spatial aggregation of the population's location has attracted much interest in location literature. Spatial aggregation related to the methodological discussion of allocation of service centers appears in a large number of articles, many of which are reviewed by Love et al. (1988), Rushton (1989), Rogers et al. (1991), Hale and Moberg (2003) and Francis et al. (2009). In our experiments, we also consider a low and a high level of spatial aggregation of the population.

The paper is organized as follows: Section two presents the p-median model, and by drawing on geographical theory and location literature, it provides a critical discussion on the choice of distance measure. Section three presents the data and its sources, defines the distance measures, and provides descriptive statistics of key variables. Furthermore, maps of the Dalecarlia region put the model into an empirical context. The fourth section describes the experimental design leading to a 'what-if' analysis as well as an outline of the optimization method. Results are presented in section five, and section six presents the conclusion.

2 Location models and distance measures

Consider the problem (known as the p-median problem) of allocating P service centers to a population geographically distributed in Q demand points such that the population's average or total distance to its nearest service center is minimized (e.g. Hakimi 1964; Handler and Mirchandani 1979; Kariv and Hakimi 1979; Mirchandani 1990; Daskin 1995). Upon access to extremely detailed data, each individual in the population makes up a demand point. In many applications, the demand point would ideally be the individual's residence.

Location models assist in the location problem by suggesting an optimal location of service centers according to an objective function. For the widely used p-median model, the objective function is taken to be the minimized total (or average) distance between the demand points and their closest service centers. This is particularly the case if the service is under central control as is often the case in publicly provided services such as kindergartens and schools, museums, hospitals, courthouses, and so on. The rationale for the objective function follows from the presumption that the service is tax-funded and that access should be maximized for the population (see Church 2003 and references therein).

Arguments leading to other objective functions can be found elsewhere see e.g. Berman and Krass (1998). For instance, a heterogeneous population raises the issue of whether attributes such as the number of residents, average income, educational level, and so on should be considered. For a tax-funded service, we know of no compelling arguments for considering such factors. Therefore to maintain focus, we adhere to the objective function mentioned above.

A crucial measure and input into the objective function is the distance between the demand point and the nearest service center. Hakimi (1964) offers an original and clear structure of this issue including definitions of several key concepts. In his seminal paper, Bach (1981) conducts a thorough investigation of how to measure distance. A number of competing alternatives are the Euclidean (shortest distance in the plane), the rectilinear (or Manhattan distance), the network distance (shortest distance along an existing road or public transport network), and shortest travel time (or cost) along an existing network.

Intuitively, travel time (or cost) seems to be the most accurate measure for most settings, yet it is infrequently employed. One explanation is the difficulty and cost associated with collecting data on travel time. Another is the complication which arises in modeling the inherent variation in travel time. The second best measure would presumably be network distance while Euclidean, and rectilinear are the easiest to collect. Remarkably, Bach (1981) found that the correlation was close to one for network and Euclidean distances when he conducted an empirical examination of two densely populated German cities. Hence, his results, although difficult to generalize to other contexts where location models are applied, indicate that it does not matter whether the network or the Euclidean distance is used in location models. This viewpoint is also found in Love et al. (1988). They reason:

> Road travel between a pair of cities is seldom along a completely straight path. However, a good approximation of the average total distance between several pairs of cities in a region can often be made by using a weighted straight-line distance function. (Love et al. 1988, pp. 5–6)

This statement is further strengthened from a literature review of location models and distance estimations conducted by Rushton (1989).

Nowadays, the Euclidean distance is widely used in location literature as an adequate distance measure as shown in the survey of Francis et al. (2009). In their survey, they summarize some 40 published articles of which about half are executed on real data. In these articles, the predominant distance measure is the Euclidean (the second most common measure is the rectilinear distance which is not considered in this study as it is most naturally applied in urban areas). And as an aside, Francis and Lowe (1992) presented a case in which contractors bidding for motor vehicle inspection stations in Florida were free to choose a distance measure in their bids. All opted for the Euclidean.

However, one problem is that the road transport cost per unit distance is not constant. In many areas, particularly rural areas, this unit transport cost varies significantly and this will give rise to heterogeneous networks serving non-symmetrically distributed populations. In this setting, there may be both a difference in length between the Euclidean and the network distance, and a possible lack of correlation between them. Therefore it may be inappropriate to use Euclidean distance in location models in rural areas.

3 Data and descriptive statistics

Figures 1a–c shows the Dalecarlia region in central Sweden, about 300 km northwest of Stockholm. The size of the region is approximately 31,000 km^2. Figure 1a shows major natural structures and barriers such as topography, rivers, and major lakes in the region. The altitude of the region varies substantially; for instance in the western areas, the altitude

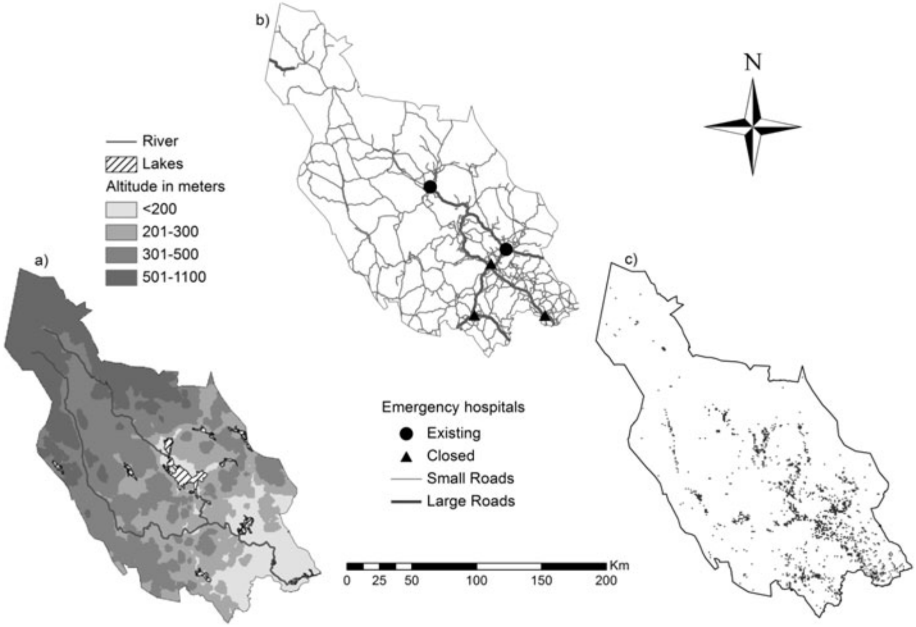


Fig. 1 Map of the Dalecarlia region showing (**a**) natural barriers, (**b**) important infrastructure and (**c**) one-by-one kilometer cells where the population exceeds 5 inhabitants

exceeds 1,000 meters above sea level, whereas the altitude is less than 100 meters in the southeast corner. Altitude variations, the rivers' extensions, and the locations of the lakes provide many natural barriers to where people could settle, and how a road network could be constructed in the region.

Figure 1b shows important infrastructure in the region. The road network is divided into small and large roads. Large roads are shown as solid black lines and small roads are indicated with thin lines. Figure 1b illustrates that the road network becomes denser and more homogeneous in areas with lower altitudes in the region's southeast corner. In the southeast and in the center of the region, a sparse network of larger roads supplements the smaller roads.

In addition, Fig. 1b indicates the two current emergency hospitals. The first hospital is located in Falun (south) and the other is located in Mora (north). Also indicated are the locations of the three closed emergency hospitals. These five hospital locations are situated in the region's largest towns.

For the population of Dalecarlia, there are adjacent hospitals east, south and west (in Norway) of the region. However, healthcare in Sweden is funded by regional taxes, and the availability to healthcare outside the region of residency is restricted. Highly specialized healthcare is an exception: in 1981, the national government appointed seven hospitals to handle this type of healthcare, and therefore the location of highly specialized healthcare is beyond the region's decision making power. As a consequence, we do not consider interactions between hospitals when conducting the experiments in Sects. 4–5.

As of December 2010, the Dalecarlia population numbers 277,000 residents. About 65 % of the population lives in towns and villages with between 1,000 and 40,000 residents. Figure 1c shows the distribution of the residents in the region by squares of 1 km by 1 km. It

Table 1 The distribution of the Euclidean distance (in kilometers) between the population and the nearest hospital

	Percentile					Mean	St. Dev.
	5	25	50	75	95		
2 current hospitals	2	14	28	54	90	32	24
All 5 hospitals	2	5	14	36	66	25	20

further indicates that the population is non-symmetrically distributed. The majority of residents live in the southeast corner, while the remaining residents are primarily located along the two rivers and around Lake Siljan in the middle of the region. Overall, the region is not only non-symmetrical, but it is also sparsely populated with an average of nine residents per square kilometer (the average for Sweden overall is 21).

The population data used in this study comes from Statistics Sweden, and is from 2002 (www.scb.se). The residents are registered at points 250 meters apart in four directions (north, west, south, and east). There are 15,729 points that contain at least one resident in the region.

The Euclidean distance between the demand points and the nearest service center, i.e. a hospital, can now be calculated. But first some notation is required. The coordinate for the qth resident is (a_q, b_q) $(q = 1, \dots, Q)$ and the number of residents at demand point q is denoted by N_q, where $N_q = 1$ since we have coordinates for each resident in Dalecarlia. The coordinate (x_p, y_p) refers to the location of the pth service point (where $p = 1, \dots, P$). The distance between the demand point and any arbitrary service point is denoted by $d(p, q)$, which equals $\sqrt{(a_q - x_p)^2 + (b_q - y_p)^2}$ for the Euclidean distance. The distance for the qth demand point to the nearest service point is

$$\overline{d}(q) = \min\bigl[d(1, q), \dots, d(P, q)\bigr].$$

The objective is to find a location of the P service points such that the sum of the shortest distances of all demand points is at its minimum. We wish to minimize

$$f_E(\bar{p}) = \sum_{q=1}^{Q} N_q \overline{d}(q).$$

The $\bar{p}$ in the right-hand side of the equation refers to the location of the P service points to be identified, whereas the subscript E refers to the distance measure, namely the Euclidean. In the following, we will drop the argument $\bar{p}$ in the function whenever it is obvious that we have evaluated the function for an optimal location.

By dividing the value of the objective function f_E for a given solution of service points with the size of the population, one obtains the average Euclidean distance to the nearest hospital. Table 1 presents statistics on the average Euclidean distance for the population in Dalecarlia to the two current hospitals, and to all of the five hospitals.

A more refined understanding of the distance for the residents can be obtained by examining the distribution of $\overline{d}_E(q)$. Table 1 also shows the distribution by means of some percentiles. The percentiles show the population proportion having a certain Euclidean distance or shorter to its nearest hospital. For instance, with the two current hospitals, 75 % of the population must travel 54 km or less. In comparing current hospital locations to previous locations, one observation is that 25 % of the residents with the shortest distance to the hospitals experienced a 180 % increase in the Euclidean distance while this increase was merely

50 % for the 25 % living furthest away from the hospital. However, the actual increase in distance is not more than 9 (14 − 5) kilometers for the 25th percentile of the residents, but it is as much as 18 (54 − 36) kilometers for the 75th percentile.

The Swedish road system is divided into national roads, local streets and private roads. The local streets are managed by the municipalities. The national roads are public, funded by a state tax, and administered by a government agency called the Swedish Transport Administration. The national roads are of varying quality, and are, in practice, distinguished by a speed limit. Parts of the road network in the cities are local streets usually with low and uniform speed limits.

Figure 1b shows the national roads in Dalecarlia. The data for the road network comes from Sweden's Mapping, Cadastral and Land Registration Authority (www.lantmateriet.se). The road network data describes the situation as of 2001.

The national road system in the region totals 5,437 kilometers; and the computer model there off divided them into 1,977 digitally stored road segments. The road segments vary in length and range from a few meters (typically at intersections) to 52 kilometers, although the typical road segment is a couple of kilometers.

There are many possible routes to travel between any two points. However, we assume that residents opt for the shortest route. We identified 778 nodes as being all the intersections or the ends of road segments in the region's road network. We then created a distance matrix with the dimension of 778 by 778 to represent the shortest network distance between all node-pairs. The creation of the distance-matrix was conducted according to the Dijkstra algorithm (Dijkstra 1959), and the naïve version of the algorithm was implemented by us in the program-package R (see www.r-project.org). The naïve implementation of the Dijkstra algorithm works in this case since the complexity is modest, and it is easy to implement. However, Zhan and Noon (1998) recommend other implementations of the Dijkstra algorithm or other shortest-path algorithms for more complex problems.

We did not have digital access to private roads and local streets. We assumed that residents can travel to the nearest node on a road network with a length equal to the Euclidean distance, and that the network distance between a resident and a node is the Euclidean distance to the nearest node and the shortest network distance between the nearest node and the node of interest. Potentially this might induce a bias in the network distance measure. To ascertain the magnitude of this error, we examined a random sample of 100 residents and retrieved their network distances to a node by using a route-finder program (www.eniro.se). The differences in distances between our own calculations and those made by the route-finder were insignificant and almost always less than one percent.

In line with the notation for Euclidean distance, we denote by $\overline{d}_N(q)$ the qth individuals shortest network distance to the nearest of the service points, and the objective function by f_N. We also represent heterogeneity in the road network by assuming the travel speed to be 65 km per hour on the small roads, and 90 km per hour on the large roads (see Fig. 1b). The travel speeds of 65 and 90 km/h are, of course, a rough approximation to the actual travel speed of residents which varies with road conditions.

Small roads are predominant and constitute 85.2 % of the network while large roads constitute the remaining 14.8 %. We use subscript T to denote the travel time measure (in minutes). To obtain a matrix of distances consisting of the shortest travel time between all node-pairs, we followed the same procedure used for the network distance matrix after converting the lengths of the road segments into travel times depending on the segment being a large or a small road.

Table 2 presents a number of statistics for the network distance for the Dalecarlia population to the nearest hospital, i.e. $\overline{d}_N(q)$. It shows the statistics for the existing two hospitals,

Table 2 The distribution of the network distance (in kilometers) between the population and nearest hospital

	Percentile					Mean	St. Dev.
	5	25	50	75	95		
2 current hospitals	3	18	36	64	116	40	30
All 5 hospitals	3	7	20	45	89	33	25

and, prior to the closure of three hospitals, for the five hospitals that were in existence at that time. By comparing the statistics for the Euclidean distance and the network distance in Tables 1 and 2, an observation can be made: the network distance is on average about 30 % longer than the Euclidean distance.

Table 2 shows that the median resident currently travels about 36 kilometers to the nearest hospital, whereas if all five hospitals were operational, the distance would be only 20 kilometers. Yet after the closure of three hospitals, the mean distance indicates an increase in distance to the nearest hospital by 7 kilometers. By comparing the current and previous distances to the nearest hospital for the 25th and 75th percentiles, an increase of the distance by some 150 % and some 50 % respectively can be noted. The reduction of the number of hospitals in the region from five to two reduced accessibility to the hospitals for the residents of densely populated areas whereas the population in remote areas suffered comparably less, measured as the relative difference in travel distance.

4 Experiments and optimization

In the experiment, we vary three factors. The first is the three distance measures of $\overline{d}_E(q)$, $\overline{d}_N(q)$ and $\overline{d}_T(q)$. The second is the number of service points (hospitals) which is varied from two to eight. We have conducted experiments with nine and more hospitals as well, but we found that the number of residents in the hospitals' service areas was below the number (about 20,000 residents) needed to efficiently run an emergency hospital (see Phelps 2003).

The third factor is the level of spatial aggregation of the demand points. The demand points are registered 250 meters apart from each other in four directions, which is a low level of aggregation. We spatially aggregate the population by joining demand points into aggregated demand points in which there is 5,000 meters in four directions between them. Note that this means that this is an aggregation by 400 times, implying substantial aggregation.

Before presenting the results, details about the optimization technique are required. Explicitly stated, the objective function f_E, in the case of P hospitals, is

$$f_E(p^*) = \sum_{q=1}^{Q} \min\left[\sqrt{(a_q - x_1)^2 + (b_q - y_1)^2}, \ldots, \sqrt{(a_q - x_P)^2 + (b_q - y_P)^2}\right].$$

It is infeasible to find a tractable mathematical solution to a problem involving multiple hospitals. Instead, we compute the objective function for all possible configurations of P hospitals under a set of restrictions. The configuration that yields the smallest value of the objective function is regarded as optimal. The optimum for the network and the travel time is found in the same way thereby replacing the Euclidean distance between a resident and a hospital by the shortest network distance and the shortest travel time.

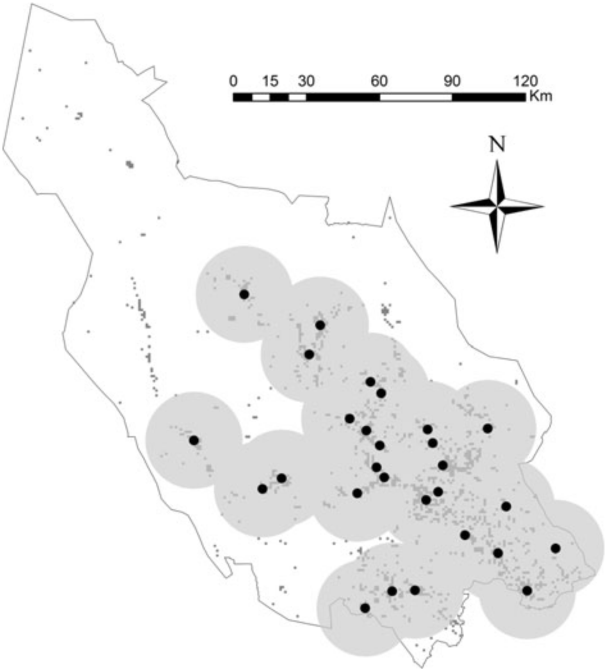


Fig. 2 The permissible area for locating hospitals in the experiments (*grey shaded*). The *black dots* show the position of the region's 28 towns and villages with 1,000 or more residents. The *grey shaded circles* illustrate their surrounding area. The *grey dots* illustrate the residents' location in Dalecarlia

The first restriction is that the hospitals must be located at one of the 778 nodes in the network. From an applied perspective, this is reasonable since a hospital's function is contingent on a road infrastructure. Furthermore, this restriction fixes the potential locations such that the three distance measures are comparable. However, it is difficult to evaluate all potential configurations since they amount to $\binom{778}{P}$. The computational approach is therefore divided in two steps, first into a global search and secondly into a local search.

The second restriction is that the hospitals must be located in a town or village with at least 1,000 residents (the global search). There are 28 towns and villages of this size in the region. Figure 2 shows their positions on the map. To do the local search for a configuration of hospitals, we further allow for location in the surroundings of the towns or villages. The surrounding is defined by a circle with a radius of 20 km of the town's (or village's) center. Figure 2 illustrates the surroundings by grey shaded circles. The non-shaded area illustrates the impermissible area for the location of hospitals in the experiments. There are 156 nodes in the impermissible area which implies that the potential number of configurations is reduced to $\binom{622}{P}$. Looking at Figs. 1 and 2, one notes that the restriction essentially implies that no hospital will be located in the unpopulated mountain area. Less than 4 % of the population lives in the impermissible area, and a location in the impermissible area would be impractical due to both a lack of labor and a lack of other inputs for the operation of a hospital.

The third restriction is that at the most one hospital may be located in a town or a village. The restriction is sensible since the largest town in the region has less than 40,000 residents and the minimum number of residents for efficient operation of one hospital is, as pointed out, above 20,000 residents. However, Fig. 2 shows that the surroundings of many towns and villages overlap which means that two hospitals may be very closely located.

A positive effect of imposing the restrictions is that the number of possible configurations might be reduced such that it is feasible to evaluate all possible configurations of hospitals thereby avoiding a heuristic solution to the optimization problem. In the first step, global search, we evaluate the objective function for all $\binom{28}{P}$ configurations of towns and villages at their center. The P towns and villages with the smallest value on the objective function are selected. In the second step, local search, we evaluate for these towns and villages the objective function for all possible $(1 \times P)$ vectors containing one node from each of the P towns' and villages' surroundings. For example, for $P = 2$ with surrounding nodes a_1 and b_1 of the first town and nodes a_2 and b_2 of the second town we would also try location pairs (a_1, a_2), (b_1, a_2), (a_1, b_2), and (b_1, b_2). We regard the vector with P nodes giving the smallest value of the objective function as the optimal configuration of hospitals.

Our approach of finding optimal configurations of hospitals is not generally feasible. Assume that the 622 nodes are evenly distributed in the 28 towns and villages such that each one has 22 nodes. The number of possible vectors would then be 22^P, which equals $\approx 5.5 \times 10^{10}$ for $P = 8$. The computational burden might be much worse than this since the nodes are concentrated in the surroundings of the bigger towns were one might expect the hospitals to be located. In our case, some smaller remote towns with relatively few nodes reduced the computations. For $P = 5$ we ended up evaluating $((\binom{28}{5}) + (23 \cdot 38 \cdot 31 \cdot 36 \cdot 11)) \approx 1.1 \times 10^7$ combinations for the travel time distance and about twice as many for the other two distance measures.

Some practical remarks are: the hospitals are usually located on the central node in the towns and villages, and in most towns and villages, there are nearby competing nodes giving almost the same value on the objective function, i.e. the objective function does not always have a distinct minimum.

Given the computational burden for $P \geq 6$ and the fact that the locations tended to be at or close to the central node of the towns and villages, we decided to shrink the surroundings to five km in the local search phase. As one of the worst computations, i.e. the case with the locations of eight hospitals and travel time distance as objective function, we ended up evaluating little more than 1.4×10^7 combinations $((\binom{28}{8}) + (15 \cdot 4 \cdot 12 \cdot 8 \cdot 9 \cdot 4 \cdot 8 \cdot 7))$. Furthermore, we confirmed that the optimal configuration was not at the border of the surroundings, in which case we once again set the surrounding of the town to a radius of 20 km.

While the second and third restrictions are justified in this case, one might be interested in how these restrictions affect the configuration. We dropped the restrictions and searched for a configuration based on Euclidean distance and travel time distance for $P = 5, 7, 8$. Without the restrictions, we resorted to heuristics (classical heuristics as the p-median problem was of small magnitude, cf. Mladenovic et al. 2007). We implemented the heuristic given in Daskin (1995) on pages 208–221. The heuristic algorithm gave the same configuration without restrictions 2–3 as our approach with the restrictions for $P = 5, 7$. For $P = 8$, the heuristic algorithm located a hospital in the impermissible area in the far northwest. The service area for the hospital was less than 5,000 residents, and the value of the objective function was the same as for our approach. Hence, restrictions 2–3 seem to be inconsequential in this case.

5 Results

The Euclidean and the travel time distances give different optimal configurations of the hospitals in the computer experiments. The most pronounced difference was found in our experiments with five and eight hospitals using the objective functions f_E (Figs. 3a and 4a) and f_T (Figs. 3b and 4b). These figures outline the configurations of the hospitals. They

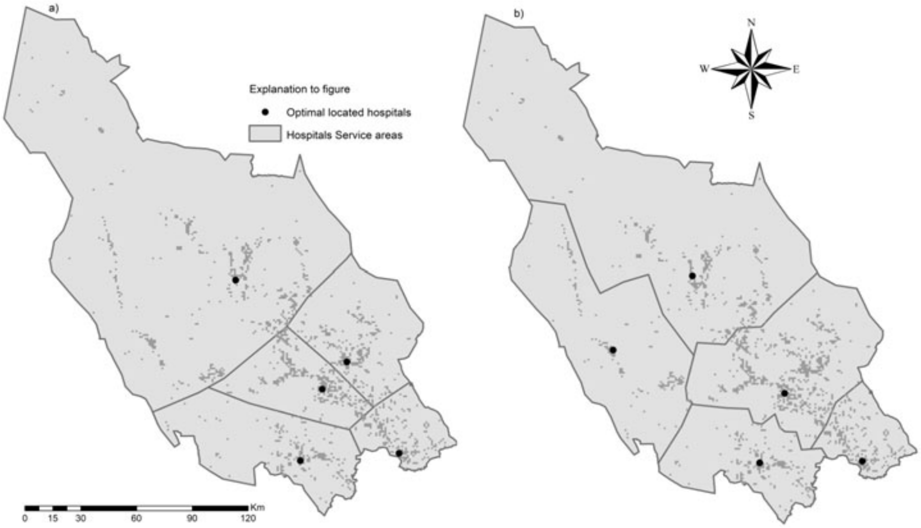


Fig. 3 Differing locations of the configurations of 5 optimally located hospitals and their service areas. In (**a**) the objective function is f_E, and in (**b**) the function is f_T. The distribution of residents is indicated by *dark grey dots*

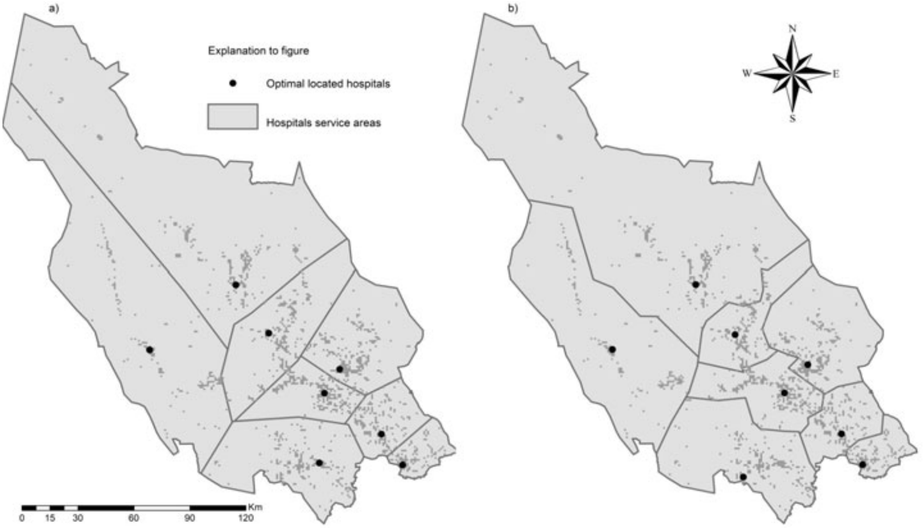


Fig. 4 Differing service regions for the configurations of 8 optimally located hospitals and their service areas. In (**a**) the objective function is f_E, and in (**b**) the function is f_T. The distribution of residents is indicated by *dark grey dots*

also indicate the service area to each hospital, i.e. the area at which the hospital is the nearest service point. The geographical distribution of residents is also shown in the figures.

In our experiments involving five hospitals, f_T locates a hospital in the western part of the region (Fig. 3b). The residents in this area are hindered by natural barriers and rely on

Table 3 The population's Euclidean and network (in parenthesis) distances in kilometers to the nearest hospitals when they are optimally located

No. of hospitals	Percentile					Mean	St. D
	5	25	50	75	95		
2	3 (4)	13 (16)	28 (36)	41 (50)	59 (78)	29 (36)	21 (26)
3	2 (3)	9 (9)	17 (24)	35 (45)	58 (78)	23 (30)	20 (26)
4	2 (3)	5 (6)	14 (19)	26 (34)	58 (76)	19 (25)	19 (24)
5	1 (2)	3 (5)	9 (13)	23 (30)	58 (76)	16 (22)	20 (25)
6	1 (2)	3 (6)	8 (11)	22 (29)	37 (52)	14 (19)	17 (21)
7	1 (2)	3 (5)	9 (11)	19 (23)	37 (49)	13 (17)	16 (19)
8	1 (2)	3 (6)	8 (11)	16 (22)	37 (50)	12 (16)	16 (19)

a sparse network of small roads in their travels eastwards. These hindrances increase travel time which the Euclidean distance fails to account for.

The choice of objective function also affects the service areas. In our 8 hospital experiment, the configuration of the hospitals is similar for f_E and f_T (see Fig. 4). However, the hospitals' service areas differ. The reason for this differentiation is that a resident's nearest hospital depends on whether it is identified by the Euclidean distance or by travel time. In the case of 5 hospitals, the service areas have between 36,000 and 80,000 residents when Euclidean distance is used in the objective function, while the service areas have between 18,000 and 136,000 residents when travel time is used. In the case of 8 hospitals, the service areas have between 17,000 and 59,000 residents when Euclidean distance is used. When travel time is used it has between 17,000 and 64,000 residents.

The fact that the service areas differ, despite a similar configuration of the hospitals in the 8 hospital experiment, suggests residents are not equally affected by the objective functions. This will be examined further below.

Table 3 illustrates the distribution of $\overline{d}_E(q)$ and $\overline{d}_N(q)$ for experiments with a varying number of hospitals and Table 4 illustrates the distribution of $\overline{d}_T(q)$.

The following explanation focuses on the Euclidean distance. The experiment with two hospitals is of particular interest as it can be contrasted to the current locations of two hospitals (cf. Table 1). The improvement in average distance is about 10 %, since the mean is decreased from 32 to 29 km. In fact, the median distance is unaltered at 28 km. However, residents living far away from the two current hospitals would have greatly benefited, e.g. the 95th percentile would be reduced from 90 km to 59 km.

The minimal objective function value f_E decreases, at a decreasing rate, as the number of hospitals is increased in the experiment as manifested by the decreasing mean distance. However, there is no simple link between the mean distance and the distribution as described by the percentiles. Take as an example five hospitals as opposed to eight, the average distance is decreased by about 25 %, yet the 25 % living closest to a hospital would experience no shortening of distance whereas the 25 % living furthest to a hospital would experience a reduction of 30 % in distance. This example and the results presented in Table 4 further highlight the need to examine not only the mean but also the distribution when locating service points.

To summarize the outcome of the experiments regarding the location of hospitals, we compute the time (in minutes) to travel between adjacent optimally located hospitals. Consider as an example the first row in Table 5 and the experiment with two hospitals, the positions of the two hospitals were obtained by the objective function f_E as well as by f_N,

Table 4 The population's travel time (in minutes) to the nearest hospitals when they are optimally located

No. of hospitals	Percentile					Mean	St. D
	5	25	50	75	95		
2	3.4	11.0	24.6	37.4	60.0	26.4	19.6
3	1.7	6.8	17.0	30.6	58.6	21.2	18.7
4	1.7	4.2	13.6	23.0	57.0	17.8	18.7
5	1.7	5.1	13.6	22.1	39.1	16.2	15.3
6	1.7	4.2	8.5	19.6	36.6	14.4	15.3
7	1.7	4.2	7.6	17.0	33.2	12.8	14.4
8	1.7	4.2	7.6	16.2	34.0	11.9	14.4

Table 5 Average travel time distance (in minutes) between closest optimally located hospitals across experiments

Measure used in objective function	No. of hospitals						
	2	3	4	5	6	7	8
Distance measures							
Euclidean vs. network	1.7	5.1	1.8	2.8	2.9	5.3	2.2
Euclidean vs. travel time	1.7	5.7	2.7	6.2	3.0	6.3	3.9
network vs. travel time	0	0.5	0.9	4.8	2.1	2.0	2.5

and the closest pairs of hospitals from the two experiments are compared. In this case, the position of the northern hospital differed by 3.4 minutes in the travel time-network, whereas the southern hospitals were at the same location giving 0 minutes travel time between. The average difference is 1.7 minutes.

Consider as another example the experiment with 5 hospitals, and the Euclidean distance versus travel time where the positions of the hospitals differ by 6.2 minutes on average. 6.2 minutes is a significant difference considering that the average travel time to the nearest hospital for the population was found to be 16.2 minutes (cf. Table 4). Clearly, hospital location is quite sensitive to the distance measure (see also Figs. 3a–b).

The fact that the distance measures in some experiments give substantially different locations of hospitals does not imply that the population would be greatly affected by the choice of measure. The population was also not greatly affected by the amount of aggregation. We complement Table 5 by computing how the travel time distance to the nearest hospital for the population is affected by the location obtained from the different objective functions.

Table 6 shows the mean travel time for residents along the travel time-network to the nearest hospital. The configuration of hospitals is obtained using the three different distance measures as well as low (the demand points 250 meters apart) and high (the demand points 5000 meters apart) levels of spatial aggregation. Table 5 showed that the locations of the two hospitals were insensitive to the choice of distance measure. Not surprisingly, the residents' travel time to their nearest hospital (the mean time increases by 1.1 % if f_E is compared with f_T) is similar in Table 6. The use of f_E causes an increase in travel time by about 4 % with a range of 1.1 % to 7.0 %, depending on the number of hospitals. Spatial aggregation of the population was inconsequential, and upon seeing the results, we ignored the experiments for 4, 6 and 7 hospitals. This matches similar conclusions regarding data aggregation and p-median location problems found in an extensive study by Murray and Gottsegen (1997).

Table 6 Mean travel time (in minutes) for the population to the nearest hospital, where the hospital's location is obtained under three different objective functions and two levels of spatial aggregation

Measure used in objective function	No. of hospitals						
	2	3	4	5	6	7	8
Low spatial aggregation							
Euclidean	26.7	22.0	18.6	17.2	15.0	13.7	12.4
network	26.4	21.2	18.1	16.4	14.5	12.9	11.9
travel time	26.4	21.2	17.8	16.2	14.4	12.8	11.9
High spatial aggregation							
Euclidean	27.1	22.5		16.3			11.9
network	26.5	21.6		16.3			11.9
travel time	26.5	21.1		16.2			11.9

Table 7 The Spearman rank correlation of the residents' distances in the travel time-network to nearest hospital where the configuration of hospitals is obtained under different objective functions and levels of spatial aggregation. Correlations below 0.9 are marked with bold text

Measure used in objective function	No. of hospitals						
	2	3	4	5	6	7	8
Distance measures							
Euclidean vs. network	0.99	0.95	0.99	0.98	0.95	0.91	0.95
Euclidean vs. travel time	0.99	0.94	0.98	**0.84**	0.96	**0.86**	**0.87**
Network vs. travel time	1.00	0.99	0.99	**0.83**	0.96	0.95	0.91
Spatial aggregation							
Euclidean vs. Euclidean, aggr.	0.99	0.99		0.97			**0.89**
Network vs. network, aggr.	0.99	0.99		0.99			0.91
Travel time vs. travel time, aggr.	0.99	0.99		0.99			**0.83**

Figures 3–4 illustrate that the configuration of hospitals and their service areas differ by distance measure. Consequently, the distance measure might affect the resident's distance to a hospital. Table 7 shows the Spearman rank correlation (see Wackerly et al. 2002) between the travel time-network distances to the nearest hospital when the location of the nearest hospital is obtained under different objective functions. If all residents were equally affected by the choice of objective function, then the correlation would be one.

Again, we note the experiments with locating two hospitals. Tables 5–6 show that the configuration, and consequently, the residents' distance to a nearest hospital is insensitive to the objective function. The correlations in the experiments with two hospitals are next to one in Table 7. This implies that a resident would travel the same route on the network irrespective of the objective function applied. However, this is not the case for all experiments. The choice of objective function between travel time and network has the most extreme consequence on a resident's traveling distance and traveling route with a correlation of only 0.83 for the 5 hospitals experiment and 0.91 for the 8 hospitals experiment. Clearly, the choice of objective function affects different parts of the population differently.

6 Conclusion

Does the Euclidean distance work well when the p-median model is applied in non-homogeneous rural areas? This case study answers no. We find that the Euclidean distance leads to sub-optimally located hospitals with two main consequences. The first is that the residents' travel time to a nearest hospital is increased. The second is that the Euclidean distance obscures the hospitals' service areas, which may cause planning errors for healthcare managers and infrastructure authorities.

If at all, we expected the discrepancy between using the Euclidean distance and the network distance measures to rise with an increasing number of service points. However, the computer experiments did not support this expectation. Hence, it seems unforeseeable in any given application as to whether the outcome of the location models using the Euclidean distance or a network distance will differ. Of course, we stopped the computer experiments at $P = 8$. In many location problems, this is a small number of service points.

In our computer experiments, we considered two levels of spatial aggregation. Firstly, the spatial aggregation did not alter the conclusion regarding the working of the Euclidean distance for the p-median model. Secondly, the effect of spatial aggregation on the location problem considered here was inconsequential compared with the choice of distance measure. However, spatial aggregation did mask the residents' travel path to their nearest hospital, thereby potentially causing planning errors.

In essence, the objective function used in this study is to minimize the population's average distance to its nearest hospital. The computer experiments showed that the change in the mean distance for different configurations of service points is a very crude description of how the population is affected. We suggest that the mean is complemented by percentiles to obtain a better understanding of how sub-populations are affected by competing configurations of service points.

We were originally interested in the location of emergency hospitals in a Swedish region called Dalecarlia. We primarily found that an increased number of hospitals would greatly improve accessibility to hospitals for the residents, but also that the current configuration is sub-optimal from a traveling point of view.

This study leaves various topics to be investigated further, among which is the choice of a region for experimentation. Given that this study leads to a conclusion in contrast to the current consensus in the computational location literature, it is important that more case studies are conducted in rural areas. Hopefully, such studies could consider the location of many service points, consider a more refined and heterogeneous network, and elaborate with competing objective functions. An alternative objective function known as the p-center model is to minimize the maximum distance for the population to the nearest service point. For emergency care, the p-center objective function is not far-fetched. Such an objective function is very sensitive to the skewness of the distribution. Consequently, if we were to conduct further experiments using the p-center model then we would expect to find a substantial impact of the choice of distance measure on the location of hospitals.

Acknowledgements We are grateful to Bo Zhu who contributed in preparing an early draft of this work, and to Siril Yella who carefully read and commented on a previous version. We also acknowledge many useful comments by an anonymous reviewer which greatly improved the work. Financial support from the Swedish Retail and Wholesale Development Council is gratefully acknowledged.

References

Bach, L. (1981). The problem of aggregation and distance for analyses of accessibility and access opportunity in location-allocation models. *Environment & Planning A*, *13*, 955–978.

Berman, O., & Krass, D. (1998). Flow intercepting spatial interaction model: a new approach to optimal location of competitive facilities. *Location Science*, *6*, 41–65.

Church, R. L. (2003). COBRA: A new formulation of the classic p-median location problem. *Annals of Operations Research*, *122*, 103–120.

Daskin, M. S. (1995). *Network and discrete location: models, algorithms, and applications*. New York: Wiley.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, *1*, 269–271.

Francis, R. L., & Lowe, T. J. (1992). On worst-case aggregation analysis for network location problems. *Annals of Operations Research*, *40*, 229–246.

Francis, R. L., Lowe, T. J., Rayco, M. B., & Tamir, A. (2009). Aggregation error for location models: survey and analysis. *Annals of Operations Research*, *167*, 171–208.

Hakimi, S. L. (1964). Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research*, *12*(3), 450–459.

Hale, T. S., & Moberg, C. R. (2003). Location science research: a review. *Annals of Operations Research*, *32*, 21–35.

Handler, G. Y., & Mirchandani, P. B. (1979). *Location on net works: theorem and algorithms*. Cambridge: MIT Press.

Hillsman, E. L., & Rhoda, R. (1978). Errors in measuring distances from population to service centers. *The Annals of Regional Science*, *12*, 74–88.

Kariv, O., & Hakimi, S. L. (1979). An algorithmic approach to network location problems. Part 2: The p-median. *SIAM Journal on Applied Mathematics*, *37*, 539–560.

Mirchandani, P. B. (1990). The p-median problem and generalizations. In *Discrete location theory* (pp. 55–117). New York: Wiley.

Love, R. F. Morris, J. G., & Wesolowsky, G. O. (1988). *Facilities location—models & methods*. New York: North-Holland.

Mladenovic, N., Brimberg, J., Hansen, P., & Moreno Pérez, J. A. (2007). The p-median problem: a survey of metaheuristic approaches. *European Journal of Operational Research*, *179*(3), 927–939.

Murray, A. T., & Gottsegen, J. M. (1997). The influence of data aggregation on the stability of p-median location model solutions. *Geographical Analysis*, *29*, 200–213.

Phelps, C. (2003). *Health economics* (3rd ed.). Boston: Addison Wesley.

Rogers, D. F., Plante, R. D., Wong, R. T., & Evans, J. R. (1991). Aggregation and disaggregation techniques and methodology in optimization. *Operations Research*, *18*, 25–42.

Rushton, G. (1989). Applications of location models. *Annals of Operations Research*, *39*, 553–582.

Zhan, F. B., & Noon, C. E. (1998). Shortest path algorithms: an evaluation using real road networks. *Transportation Science*, *32*(1), 65–73.

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. (2002). *Mathematical statistics with applications* (6th ed.). Belmont: Wadsworth.

PAPER II

Distance measure and the *p*-median problem in rural areas

Authors[♦]: Kenneth Carling, Mengjie Han, Johan Håkansson[♣], and Pascal Rebreyend

Abstract: The *p*-median model is used to locate *P* facilities to serve a geographically distributed population. Conventionally, it is assumed that the population patronize the nearest facility and that the distance between the resident and the facility may be measured by the Euclidean distance. Carling, Han, and Håkansson (2012) compared two network distances with the Euclidean in a rural region with a sparse, heterogeneous network and a non-symmetric distribution of the population. For a coarse network and *P* small, they found, in contrast to the literature, the Euclidean distance to be problematic. In this paper we extend their work by use of a refined network and study systematically the case when *P* is of varying size (2-100 facilities). We find that the network distance give as good a solution as the travel-time network. The Euclidean distance gives solutions some 2-7 per cent worse than the network distances, and the solutions deteriorate with increasing *P*. Our conclusions extend to intra-urban location problems.

Key words: dense network, location model, optimal location, simulated annealing, travel time, urban areas

[♦] Kenneth Carling is a professor in Statistics, Mengjie Han is a PhD-student in Micro-data analysis, Johan Håkansson is a professor in Human Geography, and Pascal Rebreyend is a professor in Computer Science at the School of Technology and Business Studies, Dalarna university, SE-791 88 Falun, Sweden.

[♣] Corresponding author. E-mail: jhk@du.se. Phone: +46-23-778573.

1. Distance measures in the *p*-median model

Consider the problem of allocating *P* facilities to a population geographically distributed in *Q* demand points such that the population's average or total distance to its nearest service center is minimized. Hakimi (1964) considered the task of locating telephone switching centers and showed that, in a network, the optimal solution of the *p*-median model existed at the nodes of the network. Thereafter, the *p*-median model has come to use in a remarkable variety of location problems (see Hale and Moberg, 2003).

However, there are three, main challenges with applying the *p*-median model on a specific location problem. The first is computational due to the combinatorial feature of the problem. Enumeration of all possible locations, in search of the optimal one, is a formidable task even for *P* and *Q* small. Hence, much research has been devoted to efficient (heuristic) algorithms to solve the *p*-median model (see Handler and Mirchandani 1979, Daskin 1995, and Murray and Church 1996 as examples).

The second challenge is the aggregation error arising from the common practice of aggregating demand points. Hillsman and Rhoda (1978) analysed the errors that may arise in measuring the distance between the population to be served and the facilities. One source of error comes from the aggregation of the population in an area to a single point, where the point shall represent the position of all members of the population in the area. Their research spurred an on-going investigation of this error and techniques to reduce the error (see Francis, Lowe, Rayco, and Tamir 2009 and references therein).

The third challenge is to measure the distance between the demand point and the nearest service center. In his seminal paper, Bach (1981) conducted a thorough investigation of how to measure distance. A number of competing alternatives are the Euclidean (shortest distance in the plane), the rectilinear (or Manhattan distance), the network distance (shortest distance along an existing road or public transport network), and shortest travel time (or cost) along an existing network. Remarkably, Bach (1981) found that the correlation was close to one for network and Euclidean distances when he conducted an empirical examination of two densely populated German cities. Hence, his results indicate that it does not matter whether the network or the Euclidean distance is used as distance measure. After the publication of Bach (1981), there is little research on the choice of distance measure.

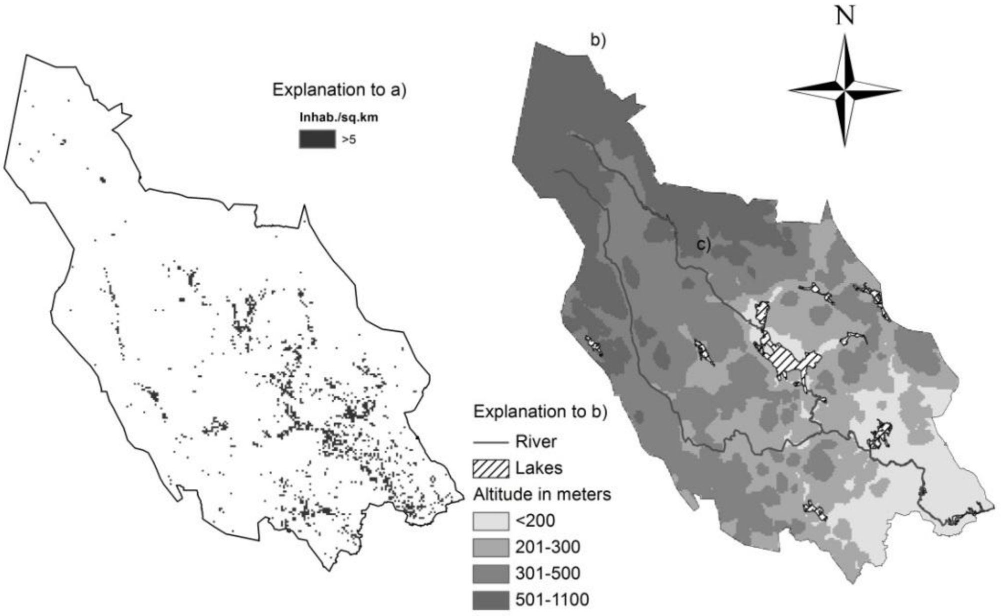
Carling, Han, and Håkansson (2012) compared the Euclidean distance with a coarse road network distance, and travel-time in a two-speed network. They compared the outcome of the *p*-median model for the three distance measures for a problem where *P* was varied from 2 to 8 facilities (*Q* was large and the population spatially disaggregated). They concluded that the Euclidean distance was problematic as it led to suboptimal location of facilities and a distorted understanding of the facilities service area. Spatial aggregation was however found to be inconsequential.

Carling et al (2012) was limited in scope with regard to the *p*-median model as it studied the choice of distance measure for *P* small in a rural setting with a coarse representation of the network. The aim of this paper is to test whether their conclusion for the *p*-median model is of more generality. We do this by systematically vary *P* from small to medium in size (2-100 facilities). The experiment is conducted on a refined network in Dalecarlia in Sweden with more than 1,500,000 nodes in which the speed limit for a road segment varies between 30 km/h to 110 km/h. Moreover, there are more than 15,000 demand points representing the population with an error of at the most 175 meters.

The paper is organized as follows: Section two presents the empirical setting and the distance measures. Section three gives the computational approach. Section four presents the results. And the fifth section concludes.

2. The empirical setting: Geography and Network

Figure 1 shows the Dalecarlia region in central Sweden, about 300 km northwest of Stockholm. The size of the region is approximately 31,000 km^2. Figure 1a gives the geographical distribution of the region[1]. As of December 2010, the Dalecarlia population numbers 277,000 residents. About 65 % of the population lives in 30 towns and villages with between 1,000 and 40,000 residents, whereas the remaining third of the population resides in small, scattered settlements. The figure shows the distribution of the residents in the region by squares of 1 km by 1 km. It indicates that the population is non-symmetrically distributed, and also sparsely populated with an average of nine residents per square kilometer (the average for Sweden overall is 21).



[1] The population data used in this study comes from Statistics Sweden, and is from 2002 (www.scb.se). The residents are registered at points 250 meters apart in four directions (north, west, south, and east). There are 15,729 points that contain at least one resident
in the region.

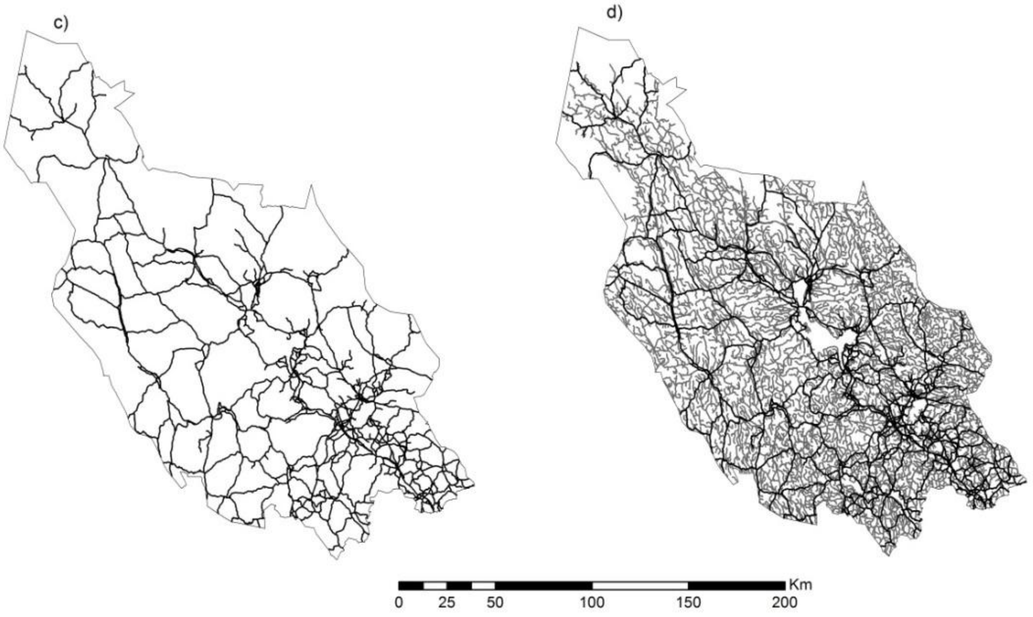


Figure 1: Map of the Dalecarlia region showing (a) one-by-one kilometer cells where the population exceeds 5 inhabitants, (b) landscape, (c) national road system, and (d) national road system with local streets and subsidized private roads.

Figure 1b shows the landscape and gives a perception of the geographical distribution of the population. The altitude of the region varies substantially; for instance in the western areas, the altitude exceeds 1,000 meters above sea level, whereas the altitude is less than 100 meters in the southeast corner. Altitude variations, the rivers' extensions, and the locations of the lakes provide many natural barriers to where people could settle, and how a road network could be constructed in the region. The majority of residents live in the southeast corner, while the remaining residents are primarily located along the two rivers and around Lake Siljan in the middle of the region.

Figure 1c shows the national road network in the region. The Swedish road system is divided into national roads and local streets that are public, and subsidized and non-subsidized private roads and in Dalecarlia the total length of the road system is 39,452 km.[2] The non-subsidized private roads is the most extensive network amounting

[2] The road networks are provided by the NVDB (The National Road Data Base). NVDB was formed in 1996 on behalf of the government and now operated by Swedish Transport Agency. NVDB is divided into national roads, local road

to more than 50 per cent of the country's roads and it is primarily built and maintained by companies, and in Dalecarlia for the purpose of transporting timber. The national road system in Dalecarlia totals 5,437 km with roads of varying quality that are, in practice, distinguished by a speed limit.

Table 1: The distribution of speed limits (km/h) in the public road network of Dalecarlia.

	Speed limit							
	-30	40	50	60	70	80	90	100-
Proportion (%)	9	3	31	2	24	19	10	2

Figure 1d adds the local streets and subsidized private roads to the national road network with an additional extension of 14,803 km. This network is very dense compared with the national roads alone. The reason to also depict the subsidized private roads is that they provide an opportunity for the residents to reach the public roads.

The speed limit varies between 30 to 110 km/h in the region's road network. Table 1 gives the proportion of road-kilometers by speed limit for the public road network. The speed limit of 70 km/h is default and the national roads usually have a speed limit of 70 km/h or more. The road network in the towns consists mostly of local streets with low and uniform speed limits (30-50 km/h). Han, Håkansson, and Rebreyend (2012) used the *p*-median model on this road network, and they noted that it is imperative to include local streets unless *P* is small.

3. The *p*-median model and computational aspects

The problem is to allocate *P* facilities to the population geographically distributed in *Q* demand points such that the population's average or total distance to its nearest facility is minimized. The *p*-median objective function[3] is $\sum_{q \in N} w_q \min_{p \in P}\{d_{qp}\}$, where *N* is the

and streets. The national roads are owned by the national public authorities, and the construction of them funded by a state tax. The local roads or streets are built and owned by private persons or companies or by the municipalities. Data was extracted spring 2011 and represents the network of the winter of 2011. The computer model is built up by about 1.5 million nodes and 1,964,801 road segments.

[3] Arguments leading to other objective functions can be found elsewhere see e.g. Berman and Krass (1998) and Drezner and Drezner (2007). For instance, a heterogeneous population raises the issue of whether attributes such as the

number of nodes, q and p indexes the demand and the facility nodes respectively, w_q the demand at node q, and d_{qp} the shortest distance between the nodes q and p.[4]

The shortest Euclidean distance, d_{qp}^E say, is simply the distance in the plane between the nodes q and p. To find the shortest network distance and shortest travel-time distance, d_{qp}^N and d_{qp}^T say, between the nodes q and p is trickier since there may be many possible routes between the nodes in a refined network. We implemented the Dijkstra algorithm (Dijkstra 1959) and retrieve the shortest distance from the center to the residents in each evaluation of the objective function. To obtain the travel-time we assumed that the attained velocity corresponded to the speed limit in the road network.

The p-median problem is NP-hard (Kariv and Hakimi, 1979). Han et al (2012) discussed and examined exact solutions to the problem as well as heuristic solutions. They advocated the simulated annealing algorithm for the problem at hand and we comply. This randomized algorithm is chosen due to its easiness to implement and the quality of results in case of complex problems. Most important, in our case, the cost of evaluating a solution is high and therefore we prefer an algorithm which keeps the number of evaluated solutions low. This excludes for example algorithms such as Genetic Algorithm and some extended Branch and Bound. Moreover, we may have good starting points obtained from pre-computed trials. Therefore a good candidate is Simulated Annealing (Kirkpatrick, Gelatt, and Vecchi, 1983).

The simulated annealing (SA) is a simple and well described meta-heuristic. Al-khedhairi (2008) gives the general SA heuristic procedures. SA starts with a random initial solution s and the initial temperature T_0 and the temperature counter $t = 0$. The next step is to improve the initial solution. The counter $n = 0$ is set and the operation is repeated until $n = L$. A neighbourhood solution s' is evaluated by randomly exchanging one facility

number of residents, average income, educational level, and so on should be considered. To maintain focus, we adhere to the objective function mentioned above.

[4] Facilities are always located at a node in line with the result of Hakimi (1964). Residents are assumed to start the travel at their nearest node, and reaching it by a travel of the Euclidean distance. This assumption is of no importance in this dense road network.

in the current solution to the one not in the current solution. The difference, Δ, of the two values of the objective function is evaluated. We replace s by s' if $\Delta < 0$, otherwise a random variable $X \sim U(0,1)$ is generated. If $X < e^{(\Delta/_T)}$, we still replace s by s'. The counter $n = n + 1$ is set whenever the replacement does not occur. Once n reaches L, $t = t + 1$ is set and T is a decreasing function of t. The procedure is stopped when the stopping condition for T is reached.

The main drawback of the SA is the algorithms sensitivity to the parameter settings. To overcome the difficulty of setting efficient values for parameters like temperature, an adaptive mechanism is used to detect frozen states and if warranted re-heat the system. In all experiments, the initial temperature was set at 400 and the algorithm stops after 2000 iterations. Each experiment was computed three times with different, random starting points to reduce the risk of local solutions. Among the three trials, we selected the solution with the lowest value of the objective function. The three solutions for each experiment varied slightly, but in an identical manner across the experiments. Hence, for the comparison of distance measure this choice is inconsequential.

Our adaptive scheme to dynamically adjust temperature work as follow: after 10 iteration with no improvement the temperature is increased according to $newtemp = temp * 3^{\beta}$, where β starts at 0.5 and is increased by 0.5 each time the system is reheated. As a result, the SA will never be in a frozen state for long. The temperature is decreased each iteration with a factor of 0.95. The settings above are a result of substantial, preliminary testing on this data and problem. In fact, some of the solutions were compared to those obtained by alternative heuristics.

The number of facilities is varied in the experiments. We consider locating small to medium number of facilities, $P \in (2,100)$. The location problem differs as a consequence, not only because P is varied. Figure 2a shows the solution for $P = 5$. The facilities lay far apart in the region and interurban travelling on the national road network is required for a large proportion of the population. Hence, in this case the rural

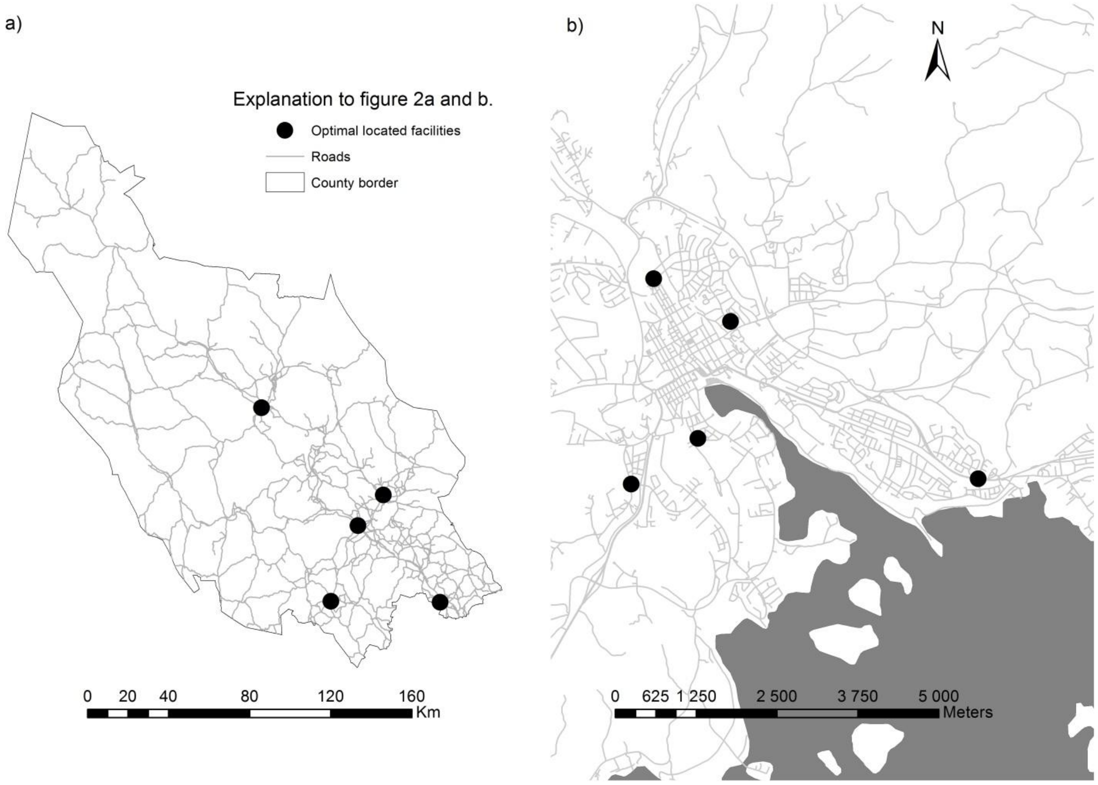

a)
b)
Explanation to figure 2a and b.
● Optimal located facilities
Roads
County border
0 20 40 80 120 160 Km
0 625 1 250 2 500 3 750 5 000 Meters
N

Figure 2: Solution of the *p*-median model in Dalecarlia for 5 and 100 facilities. (a) the solution of five facilities and the national road network. (b) the solution of 100 facilities and the road network with both national roads and local streets, focusing on the downtown area of the city of Falun.

landscape with its natural barriers and so forth affects the solution indirectly since it has affected the infrastructural setting of national roads and the location of settlements. Consider on the other hand the experiment with $P = 100$.

Figure 2b shows the solution in the downtown area of the largest city in the region – Falun. There are five facilities located in this area and the population travels to the nearest facility primarily on the local streets in the city. In conclusion, the experiments for which P are small characterizes a *p*-median problem on a rural region with a non-symmetrical distribution of the population and a highly heterogeneous road network. For the experiments with a larger P, the setting resembles a problem in an urban area. Consequently, the results of the experiments may have some external validity outside this region which is under study.

4. Results

In this section, we take, as the benchmark, the solution to the *p*-median model when the travel-time is used as distance measure. Table 2 shows the average travel-time in seconds for the residents to their nearest facility in the experiments with P varying. For $P = 2$ the average trip is about 25 minutes, a value that decreases to slightly more than 3 minutes for $P = 100$. The solutions based on the network distance are virtually identical to those of the travel-time distance as can been seen in Table 2 by comparing the average travel-time for the two measures. To complement the experimental results given in seconds, the travel distance in km on the road network for the residents to the nearest facility is shown on the last row of the table. To sum up, the finding is that the network distance, not accounting for the quality in the road network, produces the same solution to the *p*-median model as an elaborated distance measure that accounts for those aspects.

Table 2: The residents' average travel time in seconds to the nearest facility. The travel-time is evaluated for the solutions of the *p*-median model for the travel-time and the network measures. Last row gives the average network distance to the nearest facility.

	P													
Measure	2	5	8	10	15	20	25	30	35	40	45	50	75	100
Travel-time	1546	973	704	617	505	444	387	348	323	301	290	273	224	198
Network	1540	988	704	618	505	444	387	348	325	301	296	272	224	198
Network (km)	33.7	20.2	13.7	12.1	9.2	7.4	6.6	6.0	5.4	5.1	4.7	4.5	3.6	3.2

Solutions for the *p*-median model was also obtained based on the Euclidean measure, and the travel-time between the residents and their nearest facility computed. Generally, these solutions increased the residents' travel-time. Figure 3 shows a relative comparison between the Euclidean solution and the travel-time solution. As an instance, for $P = 2$, the average travel-time was found to be 1,630 seconds for the Euclidean solution and 1,546 seconds for the travel-time solution, giving a relative difference of 5.4 per cent. The relative difference was 3.6 per cent on average ranging from 0.0 per cent to 7.0 per cent.

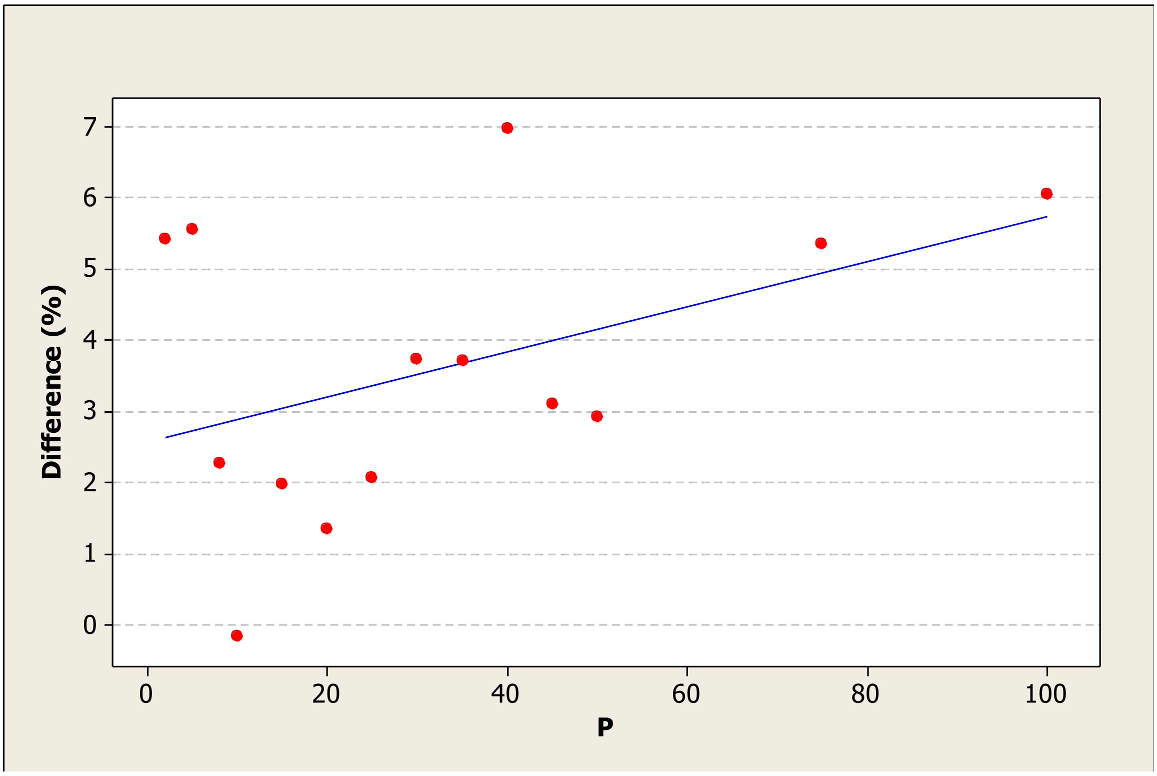


Figure 3: The relative difference between the solutions of the *p*-median model based on the Euclidean and the travel-time measure of distance.

In Figure 3, a regression line is imposed as a function of P. The significant estimate of the intercept is 2.6 and the estimate of the regression coefficient is 0.03, where the regression coefficient is borderline significant with a p-value of 0.06. Taken at face-value, the regression coefficient implies a one percentage point worsening of the Euclidean solution for each increment of P of 30 facilities. To conclude, the Euclidean measure is potentially problematic since it may provide solutions to the p-median problem that leads to excessive travel times and distances for the population.

5. Conclusions

In this study we have examined whether or not the distance measure is of importance when the *p*-median model is used to locate facilities. To do this, we have systematically varied P from small ($P = 2$) to medium size ($P = 100$) in a very dense network with attributed speed limits.

Two main conclusions can be drawn from this investigation. The first is that the Euclidean distance provides solutions to the *p*-median model that lead to excessive travel-time for the residents of as much as 7 per cent. The excess seems to increase with the number of facilities to locate.

The second conclusion is that the network distance provided equally good solutions to the *p*-median problem as an elaborated network. In spite of the fact that the elaborated network accounted for heterogeneity in the network due to variation in speed limits and the implied variation in road quality. This finding is startling as the elaborated network showed substantial heterogeneity in terms of speed limits and implied road quality. It should be noted however that the network studied here is very refined and that the findings may not extend to a sparse network.

As a final remark, note that the variation in *P* has some implications for interpreting the findings for a rural setting. For *P* small, the setting is a problem of locating facilities in inter-urban environment where a large fraction of the population travels between towns to patronize the nearest facility. For the larger values of *P*, it is a setting where multiple facilities are located within the towns and the residents travel primarily on local streets within the towns. Hence, we assert that the findings bear some relevance for location problems in urban settings, in addition to rural ones.

Acknowledgements

Financial support from the Swedish Retail and Wholesale Development Council is gratefully acknowledged.

References

Al-khedhairi. A., (2008), Simulated annealing metaheuristic for solving p-median problem. *International Journal of Contemporary Mathematical Sciences*, 3:28, 1357-1365, 2008.

Bach, L. (1981). The problem of aggregation and distance for analyses of accessibility and access opportunity in location-allocation models. *Environment & Planning A*, 13, 955–978.

Berman, O., and Krass, D. (1998). Flow intercepting spatial interaction model: a new approach to optimal location of competitive facilities. *Location Science*, 6, 41–65.

Carling K., Han M., and Håkansson J, (2012). Does Euclidean distance work well when the *p*-median model is applied in rural areas?, *Annals of Operation Research* 201(1), 83-97.

Daskin, M.S., (1995). Network and discrete location: models, algorithms, and applications. New York: Wiley.

Dijkstra, E.W., (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1, 269–271.

Drezner, T., and Drezner, Z, (2007). The gravity *p*-median model, *European Journal of Operational Research*, 179, 1239-1251.

Francis, R. L., Lowe, T. J., Rayco, M. B., & Tamir, A. (2009). Aggregation error for location models: survey and analysis. *Annals of Operations Research*, 167, 171–208.

Hakimi, S.L., (1964). Optimum locations of switching centers and the absolute centers and medians of a graph, *Operations Research*, 12:3, 450-459.

Hale, T.S., and Moberg, C.R. (2003). Location science research: a review. *Annals of Operations Research*, 32, 21–35.

Han, M., Håkansson, J., and Rebreyend, P., (2012). How does the use of different road networks effect the optimal location of facilities in rural areas?, Working papers in transport, tourism, information technology and microdata analysis, ISSN 1650-5581.

Handler, G.Y., and Mirchandani, P.B., (1979). Location on networks: *Theorem and algorithms*, MIT Press, Cambridge, MA.

Kariv, O., and Hakimi, S.L., (1979), An algorithmic approach to network location problems. part 2: The p-median. *SIAM Journal of Applied Mathematics*, 37, 539-560.

Kirkpatrick, S., Gelatt, C., and Vecchi, M., (1983), Optimization by simulated annealing. *Science*, 220:4598, 671-680.

Murray, A.T., and Church, R.L., (1996). Applying simulated annealing to location-planning models, *Journal of Heuristics*, 2, 31-53.

PAPER III

How do different densities in a network affect the optimal location of service centers?

Authors[1]: Mengjie Han[2], Johan Håkansson, and Pascal Rebreyend

Abstract: The *p*-median problem is often used to locate *p* service centers by minimizing their distances to a geographically distributed demand (*n*). The optimal locations are sensitive to geographical context such as road network and demand points especially when they are asymmetrically distributed in the plane. Most studies focus on evaluating performances of the *p*-median model when *p* and *n* vary. To our knowledge this is not a very well-studied problem when the road network is alternated especially when it is applied in a real world context. The aim in this study is to analyze how the optimal location solutions vary, using the *p*-median model, when the density in the road network is alternated. The investigation is conducted by the means of a case study in a region in Sweden with an asymmetrically distributed population (15,000 weighted demand points), Dalecarlia. To locate 5 to 50 service centers we use the national transport administrations official road network (NVDB). The road network consists of 1.5 million nodes. To find the optimal location we start with 500 candidate nodes in the network and increase the number of candidate nodes in steps up to 67,000. To find the optimal solution we use a simulated annealing algorithm with adaptive tuning of the temperature. The results show that there is a limited improvement in the optimal solutions when nodes in the road network increase and *p* is low. When *p* is high the improvements are larger. The results also show that choice of the best network depends on *p*. The larger *p* the larger density of the network is needed.

Key words: location-allocation problem, inter-urban location, intra-urban location, *p*-median model, network distance, simulated annealing heuristics.

[1] Mengjie Han is a PhD-student in Micro-data analysis, Johan Håkansson is an assistant professor in Human Geography, and Pascal Rebreyend is an assistant professor in Computer Science at the School of Technology and Business Studies, Dalarna University, SE-791 88 Falun, Sweden.

[2] Corresponding author. E-mail: mea@du.se. Phone: +46-23-778000.

1. Introduction

To have an as accurate representation of a road network as possible is important for many researchers and planners using the network for transportation and planning optimization. The focus is mainly on how the roads are used and maintained. Less focus is given to using the road network to locate facilities in order to minimize transportation. In this study we focus on the inter-urban and the intra-urban location allocation problem in relation to the density of the road network. To do so we turn to the *p*-median problem.

The p-median location problem is well-studied (Farahani *et al*., 2012). However, most studies are not based on real road distances. Francis *et al*. (2009) made an explicit review of the p-median location problem. Among the 40 published articles, about half of them are studies based on real data. From that survey it is also obvious that almost all of the distance measures are Euclidean distance and rectilinear distance. In a recent study by Carling *et al.* (2012) the performance of the *p*-median model was evaluated when the distance measure was alternated between Euclidian, network and travel time. It was shown that for region with an asymmetrical distributed population and road network due to natural barriers the choice of distance measure has affected the optimal locations, and that the use of Euclidian distance leads to sub optimal solutions.

The work in this study follows the work of Carling *et al*. (2012). In Carling *et al*. (2012) the road network was limited to 1579 nodes and there was no analysis done of the effects on the suggested solutions by varying the number of nodes in the road network. However, differences in accuracy of the road networks could also influence the optimal location of service centers.

In a discrete location allocation problem complexity varies due to the number of demand points, number of service centers to locate and/or number of nodes in a network. However the *p*-median model is NP-hard (Kariv and Hakimi, 1979) and so aggregation has often been used to reduce the size of the problem. In our study we use a real world road network which consists of about 1.5 million nodes. To our knowledge there is no study which has used such a large real world network density applied on a discrete *p*-median problem. Based on that, the aim of this paper is to analyze how the optimal location solutions vary, using the *p*-median model, when both the number of service centers and the density of the road network are alternated. The investigation is conducted by the means of a case study in a region of Sweden, Dalecarlia. The population is distributed at 15,000 weighted demand points. The road network we elaborate is from the Swedish digital road system: NVDB (The National Road Database) and it is administrated by the Swedish Traffic administration. We start with 500 candidate nodes to locate on and increase them in steps up to 67,000.

To evaluate the effects of different road networks on the optimal location solutions in different situations we compare the results from the experiments in which we have alternated both the density in the road network and the number of service centers that are located within Dalecarlia. In this study we simulate an inter-urban location problem and both inter-urban and intra-urban location problems. The location of emergency hospitals and courts is typical inter-urban location problems in a region like Dalecarlia. To locate for instance high schools and post offices could be seen as typical inter-urban as well as intra-urban location problems. In this study we therefore systematically alternate *P* between 5 and 50.

To do this, several computer experiments using the *p*-median model were implemented. Since the exact optimal solution is difficult to obtain, the experiments are conducted by use of a simulated annealing algorithm.

The remaining parts of this paper are organized as follows. In section 2 we discuss some relevant literature. In section 3 we present the data used. In the fourth section we present the simulated annealing methods used. In section five we present and comment results and in section six we have a concluding discussion.

2. Literature Review

The discrete *p*-median model was first introduced by Hakimi (1964). The goal with the model is to find *p* service centers which minimize the summed distances between demands and their nearest centers. This problem can be formulated as follows.

Minimize $f = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i d_{ij} x_{ij}$, subject to $\sum_{j=1}^{n} x_{ij} = 1$ and $\sum_{j=1}^{n} x_{jj} = p$, where f is the value of objective function. n is the number demand locations. w_i is the weight of each demand location. d_{ij} is the distance from demand location i to the center j. x_{ij} is a dummy variable: taking 1 if location i is allocated to center j.

Since we model our problem as a *p*-median problem, our objective function will be to minimize the value f which is the sum of all network distances between a person and the closest service center. (d_{ij} is one for the closest location in our case). By dividing this value by the total population, we obtain the average distance between a person and its closest service center.

To find the optimal location for p-service centers in relation to the demand using the *p*-median model is NP-hard, Kariv and Hakimi (1979). The complexity depends both on the number of service centers to be located, the number of demand points, as well as on how distance is measured.

Although Euclidean distance is most widely used, the network distance is in most cases more accurate in measuring the travel distance between two points (e.g. Carling *et al.* 2012). Further, a refined network should give the possibility to more accurate distance measures between two points compared to a sparser network. There are a few studies which evaluate network effects on optimal locations. Peeters and Thomas (1995) examined the *p*-median problem for different types of networks by changing the nature of the links. They found that there was a difference in optimal solutions when the links were changed but they registered no differences in computational effort.

Morris (1978) tested the linear programming algorithm for 600 random generated data sets. He generated a benchmark to simulate the effect of a road network by adding a random noise to the Euclidian distance. His conclusion was that regardless whether he was using the pure Euclidian distance or the simulated networks he was able to solve the problem, implying that the choice of distance measure is not significant. However, the data set were very small and it was only a simulated network with values close to the Euclidian distance. Further he did not really evaluate the effect of the choice of the distance measure to the quality of the solutions.

Schilling *et al*. (2000) examined the Euclidean distance, network distance and a randomly generated network distance. Their conclusion is that it is much easier for the Euclidean and network to obtain the optimal solution and with less computational effort. However, the problem is small scale and they did not provide the effect of network with different numbers of nodes in the networks. In our study we are dealing with large networks and we systematically alternate the number of nodes in it to evaluate the quality in the optimal solutions. None of the previous studies provided any analysis of network aggregation.

In a recent study Avella *et al*. (2012) tested a large size *p*-median problem using a new heuristic based on Lagrangean relaxation. The number of nodes varies from 3,038 to 89,600. They compared their computational results to the results found by Hansen *et al*. (2009) under 4 instance sets (from Birch and TSP library). The largest data set is Birch 1. The Birch data set are synthetically generated, designed to test clustering algorithms. Birch 1 and 3 differ in two significant ways. Birch 1 is the largest data set used (89,600 nodes) and it consists of symmetrical distributed demand points and nodes in the network which are also organized in tight clusters. Birch 3 consists of up to 20,000 nodes and the demand points and the nodes in the network are more asymmetrically distributed and the clusters also vary more in their characteristics. They found that the new heuristic is fast and efficient. They also showed that the quality of the optimal solutions was quite different when Birch 1 was used compared to when Birch 3 was used. Instances of type Birch 3 also took longer computing time to be solved. Larger instances exhibit worse results. However, they did not consider a real world network, when the number of nodes in the network is alternated systematically.

3. Data

3.1 Demand Points and Service Centers

The demand points represent the distribution of the population's residence in Dalecarlia. In this study we use the population in 2002. The figures are public produced and controlled data from Statistics Sweden (www.scb.se). The populations' residents are registered on 250 meter by 250 meter squares. We generalize each square is to its central point. Each point is then weighted by the number of people living in each square. The populations' residence location is represented by 15,729 weighted points. In total 277,725 lived in Dalecarlia during the study year. The distribution of the residents is shown in Figure 1a. The figure illustrate that the population in the region is asymmetrically distributed. The majority of residents live in the southeast corner, while the remaining residents are primarily located along the county's major rivers and lakes. Overall, the region is not only non-symmetrical distributed, but it is also sparsely populated with an average of nine residents per square kilometer (the average for Sweden overall is 21).

Figure 1b shows some important features of the natural landscape in Dalecarlia. Firstly it is shown that the altitude in Dalecarlia vary a lot. From the south east corner with altitude below 200 the altitude increase in general towards north east. Secondly it is shown that a major river (Dalecarlia River) and some large lakes also act as natural barriers. Clearly, when comparing the distribution of the population (Figure 1a) with the natural barriers (Figure 1b) there is a correlation.

Concerning the service centers, in this study, we search for optimal locations for *p* equal 5, 10, 15, 20, 25 30, 35, 40, 45 and 50.

3.2 The Road Network

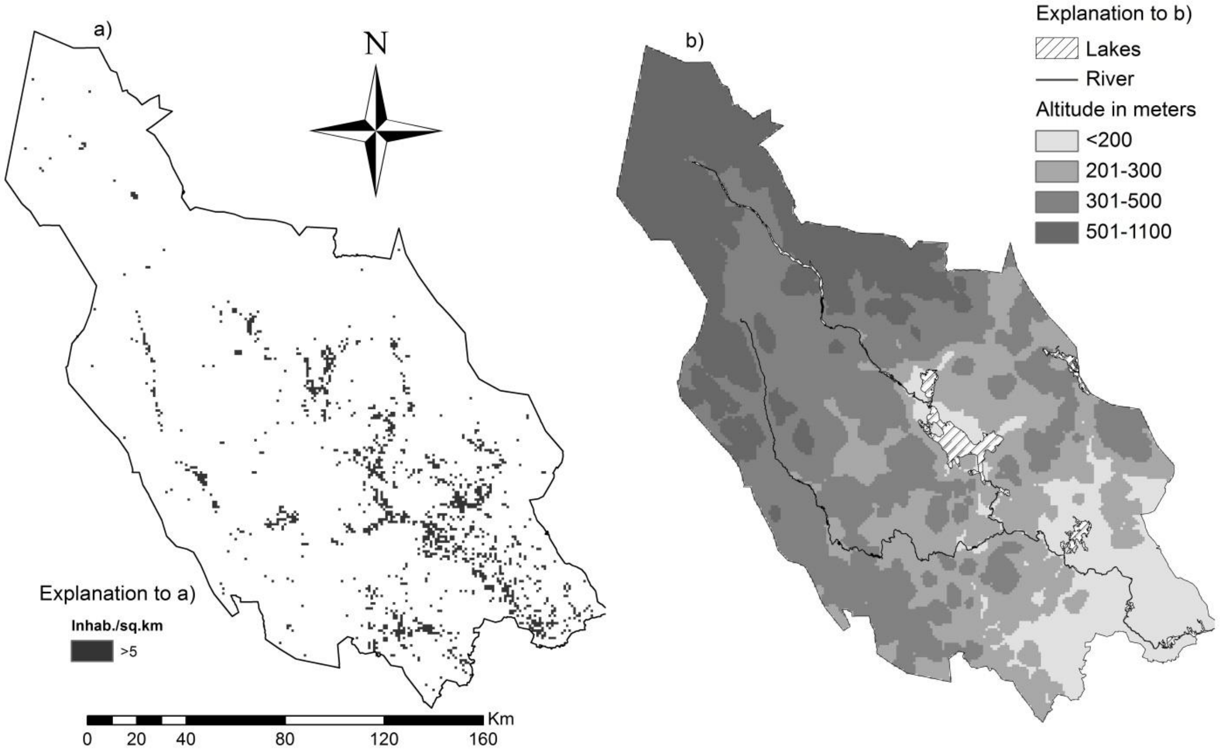


Figure 1. The distribution of the population on 1 by 1 km squares (a) and natural landscape (b) in Dalecarlia.

The road network used is the 2011 national road database (NVDB) for Dalecarlia. NVDB was formed in 1996 on behalf of the government. It is organized and updated by the National Transport Administration (Trafikverket) in Sweden. In total the road network for Dalecarlia contains about 1.5 million nodes and 1,964,801 segments. The total length is 39,452 kilometers. The average distance between the nodes in NVDB is about 40 meters. The minority of the nodes is nodes in intersections or at points where roads starts or begin. Most nodes describe the geographical shape of the road and by that they give a precise description of the length of the road. We use this network to calculate the distance between the demand points and the closest service center. To do so we use the Euclidian distance to identify the closest node on the road network. Then we add the shortest network distance. To find the shortest network distance the Dijkstra algorithm has been used (Dijkstra 1959).

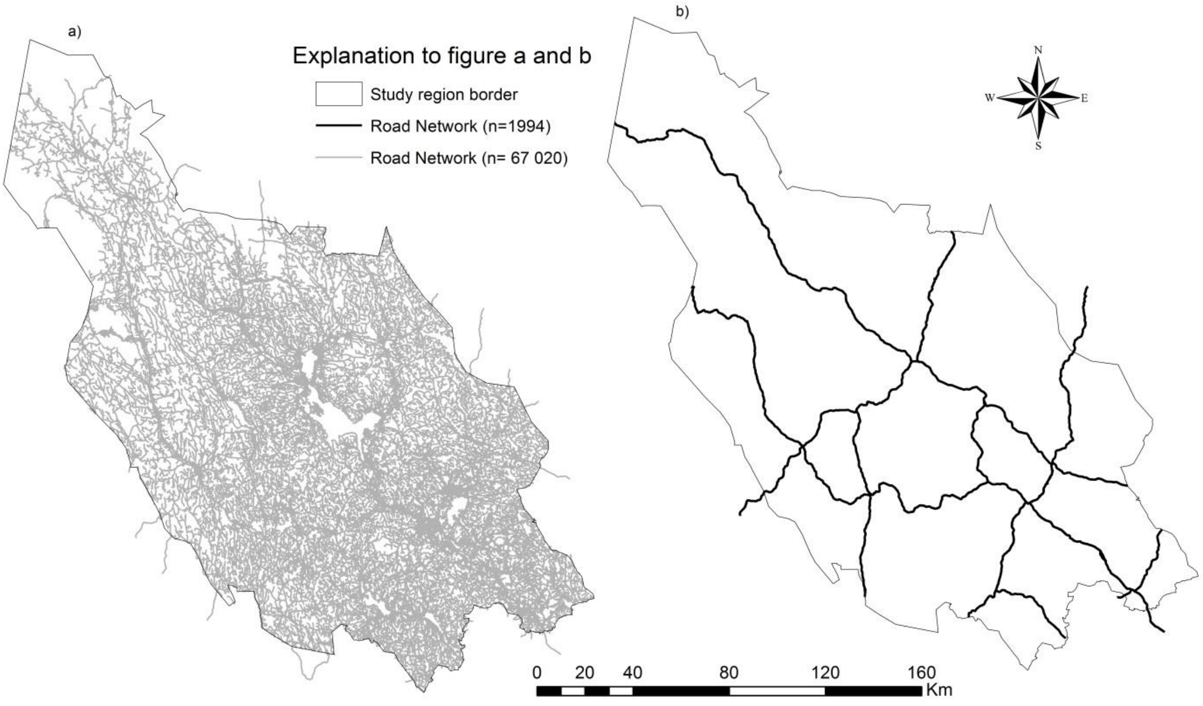


Figure 2. All roads in a dense network with 67,020 candidate nodes (a) and all roads in a sparse network with 1,994 candidate nodes (b) in Dalecarlia

To identify the candidate nodes to locate on we select one node in each 500 by 500 meter square in which the roads pass through. By reducing the number of nodes within a square an in-built location error occurs. However by selecting the center of the square as the representative node the maximum location error due to this could be 354 meters in Euclidian metric. Finally we used at the most 67,020 nodes in the road network as candidate nodes to locate on. (see Figure 2a).

NVDB is divided into 10 different categories according to the quality of the roads (see Table 1). To alternate the density in the road network we used those road classes. In Dalecarlia there is just one road (class 0) which is a European highway. For this reason, class 0 roads are merged into class 1 in this study. By just taking into account the largest roads (class 0 and 1) the set of candidates to find an optimal location of a service center are as many as about 2000 nodes distributed in a rather sparse network (see Figure 2b). This is still quite large; so to decrease the density in the road network further we add two new classes which consist of 500 and 1000 candidates to locate service centers in. We select these candidates randomly from candidates in class 0 and 1. From Table 1 we can see that average distance between the candidate nodes varies rather little when the road classes 0 to 9 are concerned. However, the average distances between candidate nodes become significant longer when the density in the road network is decreased further.

Table 1. Number of candidate nodes, road length and average road distance between candidate nodes with different road classes on the road network in Dalecarlia.

Road classes	Number of nodes	Length (km)	Meters between Candidate nodes
0 to 9	67020	39454	588
0 to 8	45336	23086	509
0 to 7	20718	10964	529
0 to 6	12552	5631	449
0 to 5	12417	5479	441
0 to 4	6735	2923	434
0 to 3	3926	1725	439
0 to 2	2909	1299	446
0 to 1	1994	883	443
0 to 1 (randomized)	1000	883	883
0 to 1 (randomized)	500	883	1766

Figures 2a and 2b illustrate that the road network becomes denser and more homogenous in areas in the region’s southeast corner. In the southeast and in the center of the region, a sparse network of larger roads supplements the smaller roads. From Figure 2a it is obvious that the smaller local roads and streets are oriented to the larger roads. It is also evident that the smaller roads make the road network more homogenous when it comes to its distribution in the region.

4. Simulated Annealing

4.1 Algorithm

Since the *p*-median problem is NP-hard, for large number problems, the exact optimal solution is difficult to obtain. That is why there are only a few studies examining the exact solutions (Hakimi, 1965; Marsten, 1972; Galvão, 1980; Christofides and Beasley, 1982). Instead most studies regarding *p*-median problem use heuristics and meta-heuristics (e.g. Kuehn and Hamburger, 1963; Maranzana, 1964; Rahman and Smith, 1991; Rolland *et al.,* 1996 Crainic, 2003; and Ashayeri, 2005). In our case, the cost of evaluating a solution is rather high therefore we focus on an algorithm which tries to keep the needed number of evaluated solutions low. This excludes, for example, algorithms such as the Genetic and to some extent Branch and Bound algorithms.

Another sub-class in meta-heuristics is simulating the annealing method, which we will use in this paper (e.g. Kirkpatrick 1983, Chiyoshi and Galvão, 2000; Al-khedhairi, 2008; and Murray and Church, 1996). This randomized algorithm has been chosen due to its flexibility, its ease of implementation and the quality of results in the case of complex problems. Al-khedhairi (2008) gave the general SA heuristic procedures.

SA starts with a random initial solution s, a choice of a control parameter named the initial temperature T_0, and the corresponding temperature counter $t = 0$. The next step is to improve the initial solution. The counter of the number of iterations is initially set as $n = 0$ and the procedure is repeated until $n = L$, where L is the pre-specified number of iterations of the algorithm. A neighborhood solution s' is evaluated by randomly exchanging one facility in the current solution to the one not in the current solution. The difference, Δ, of the two values of the objective function is evaluated. We replace s by s' if $\Delta < 0$, otherwise a random variable $X \sim U(0,1)$ is generated. If $X < e^{\left(\frac{\Delta}{T}\right)}$, s still replaces s'. The counter is updated as $n = n + 1$ whenever the replacement does not occur. Once n reaches *L*, the temperature counter is updated as

$t = t + 1$ and *T* is a decreasing function of *t*. The procedure stops when the stopping condition for *t* is reached.

Given *p* we start the simulated annealing by randomly selecting points to locate the service centers. We then randomly select one of the suggested service center location sites and define a neighborhood around it. As the neighborhood we apply a square of 25 km centered on the selected site. If we have less than 50 candidates for a service center location we increase the neighborhood by steps of 2.5 km until this criterion is satisfied. This was necessary in just a few cases.

4.2 Adaptive Tuning and Parameters

The parameters used here have been tuned after prior testing. In our study we start with the initial temperature T_0 of 400. We multiply the temperature by 0.95 at each new iteration. To avoid having our algorithm blocked in a local minimum, we have an adaptive scheme to reheat the system. If 10 times in a row we refuse a solution, we increase the temperature multiplying the temperature by 3^{β}. A suitable value of β is 0.5. Therefore, the initial value of β is 0.5 and if no solution is accepted between two updates of the temperature we increase beta β by 0.5. β will be reset to 0.5 as soon as we accept a solution. Experiments done with 2000 and 20,000 iterations have shown that for our cases 20,000 leads to significantly better results. The number of iterations has been fixed at 20,000. Our experiments have been conducted on an Intel Core2 duo E8200 cpu working at 2.66 GHz. The operating system used is Linux and programming has been done in C and compiled with gcc. It took us about 24 hours to compute 20,000 iterations.

5. Results

Table 2 shows some results from the computer experiments when different density in the Dalecarlia road network for the location of a different number of service centers is alternated. The table gives information on the mean travel distance in the road network from their residence to the closest service center for the inhabitants in Dalecarlia.

Highlighted figures in the table indicate the best solution found for a given number of service centers (*p*). When *p* is set to 15 the solutions computed continue to be better until road class 3 is added to road classes 0, 1 and 2. The best solution gives an average travel distance in the complete road network from the inhabitants' homes to the closest service center of 8.53 kilometers.

The main result which can be drawn from Table 2 is that a more complex location problem can take advantage of a more complex network. This is shown by the fact that when the number of service centers is below 20 the best solutions are found already with the density given by the road classes up to two while when the number of service centers is above 20 the best solutions are found with a higher density of road network.

Table 2. The mean network distance in kilometers to the closest service center given different *p* and densities of the road network to locate on.

	Road classes in the road network										
p	500pt	1000pt	0-1	0-2	0-3	0-4	0-5	0-6	0-7	0-8	0-9
5	22.71	20.34	19.74	**19.73**	20.20	20.17	20.25	19.99	20.07	20.33	20.52
10	11.20	11.23	11.18	**11.17**	11.20	11.22	11.26	11.38	11.44	11.74	11.80
15	8.63	8.64	8.63	**8.53**	8.58	8.58	8.66	8.63	8.75	8.92	9.21
20	7.67	7.61	7.11	7.11	7.03	7.08	7.18	**6.99**	7.24	7.54	7.82
25	7.15	7.31	7.19	6.19	6.24	**6.12**	6.16	6.15	6.30	6.61	6.94
30	6.90	6.93	6.94	5.78	5.56	**5.34**	5.58	5.55	5.68	5.88	6.05
35	6.72	6.67	6.67	5.33	5.10	**4.96**	5.15	5.13	5.16	5.29	5.43
40	6.52	6.47	6.54	5.09	4.71	4.70	4.71	4.72	**4.69**	4.89	5.29
45	6.34	6.31	6.33	4.90	4.45	**4.29**	4.40	4.42	4.49	4.69	4.84
50	6.27	6.24	6.25	4.69	4.24	4.14	**4.09**	4.12	4.24	4.40	4.46

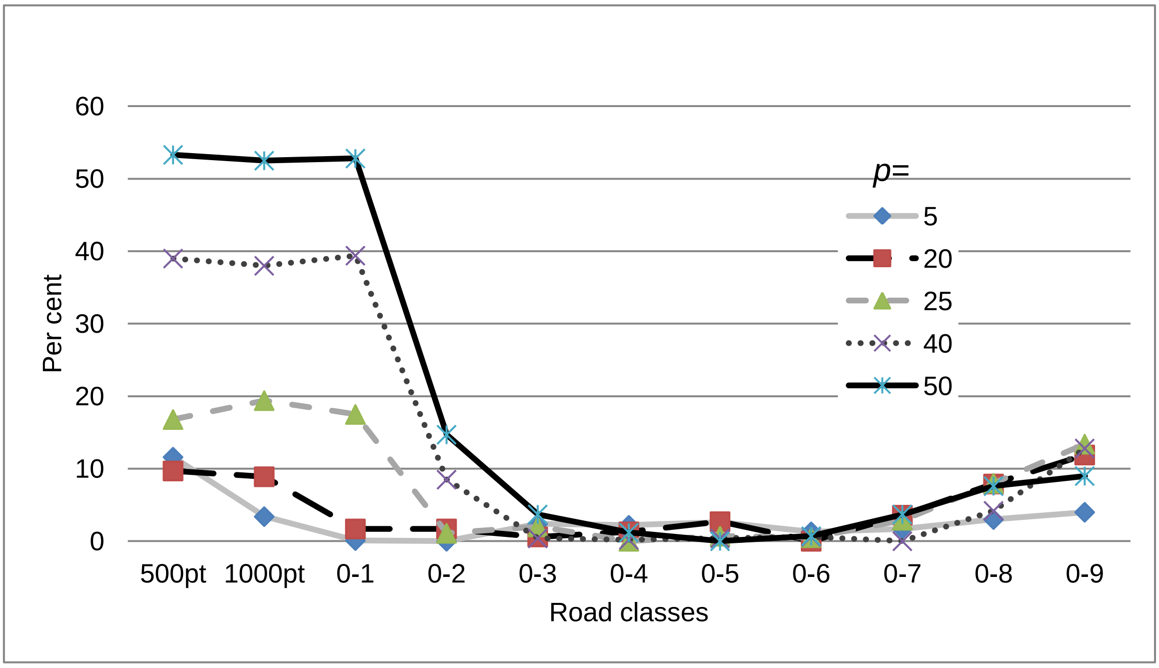


Figure 2. Variations in excess distances (in per cent) compared to the best solutions when different density in the network has been used to find an optimal location on.

Figure 2 illustrates how much worse (in per cent) solutions are in relation to the best solution for different densities in the network. In the figure this is illustrated with a selection of different *p*. The conclusion is that there is more to gain in choosing the right density level on the network when *p* is higher. This is clearly shown since when the number of service centers is 20 or less the worse solution found is not less than 12 per cent longer than the best one. On the other hand for location problems with more than 25 service centers the worst solution is at least 30 per cent longer than the best one.

6. Conclusions and Discussions

The paper aims to examine the effect of alternating the density in a road network when service center location problem is studied. To do so, we use a large scale real world road network with 1.5 million nodes in the region of Dalecarlia in Sweden and we alternate the density of the road network used to locate on from 500 to 67,000 candidate nodes. As demand points we use the population in the region registered on squares of 250 by 250 meters. The population and the network are asymmetrical distributed in the region due to natural barriers. To scrutinize the problem we also alternate the number of *p* between 5 and 50. In doing so, we cover inter-urban location problems as well as intra-urban location problems. We use the *p*-median model and meta-heuristics to find the optimal solutions.

It has earlier been shown that it is important to use the network distances when optimal locations are sought. In this study we add the result that an increased density of the road network is only necessary up to a certain level. We also show that when the number of service centers increases the density needed in the network tends to be higher. This implies that for inter-urban location problems (like for instance locating emergency hospitals or courts) with lower *p* in a region of the size used here it is sufficient to use fairly simple networks, while dealing with inter-urban as well as intra-urban location problems (like for instance locating high schools or post offices) simultaneously with higher *p* the need for a more refined network is larger.

The road network used here was not constructed for the purpose of service centers location. The structure of it is probably suitable for a lot of issues related to what happens on the road. However, in organizing this network to be suitable for the purpose of being used in location problem we turn out to have between 500 candidate nodes up to 67,000 candidate nodes which are the extremes in our case. It turns out that these two extreme densities of the road network were not suitable for solving the location problem here. One possible future research question could be how the road network should be arranged to be suitable for location allocation problems.

In this study we use simulated annealing. It has obvious drawbacks. It would however be interesting to evaluate how other algorithms would perform in this kind of setting.

Further, the case here is quite a small geographical rural area, Dalecarlia. As such more case studies are needed. In addition, the important roads are first and foremost designed to be efficient in a national transportation system. Further, many public activities but also private businesses are taken conducted at a national level. There is a need to better evaluate the efficiency in present situations of where these activities are carried out. One suggestion for future research is therefore to scale up the present case study to national level. Advanced methods (e.g. more aggressive heuristics, distributed computing) will be needed to keep the computing time acceptable and still reach excellent solutions.

Acknowledgements

Financial support from the Swedish Retail and Wholesale Development Council is gratefully acknowledged. The funder has exercised no influence on this research all views expressed are solely the responsibility of the authors. The authors would also like to thank David Glémarec and Kevin Jaouen for their help in writing, testing and tuning the Simulated Annealing algorithm.

References

Al-khedhairi, A., 2008. Simulated annealing metaheuristic for solving *p*-median problem. *Int. J. Contemp. Math. Sciences* 3(28), 1357-1365.

Ashayeri, J., Heuts, R. and Tammel, B., 2005. A modified simple heuristic for the *p*-median problem, with facilities design applications. *Robotics Computer-Integrated Manufacturing* 21, 451-464.

Avella, P., Boccia, M., Salerno, S. and Vasilyev, I., 2012. An aggregation heuristic for large scale *p*-median problem. *Computers and Operations Research* 39, 1625-1632.

Carling, K., Han, M. and Håkansson, J., 2012. Does Euclidian distance work well when the *p*-median model is applied in rural areas? *Annals of Operation Research* 201(1), 83-97.

Chiyoshi, F. and Galvão, R.G., 2000. A statistical analysis of simulated annealing applied to the *p*-median problem. *Annals of Operations research* 96, 61-74.

Christofides, N. and Beasley, J., 1982. A tree search algorithm for the *p*-median problem. *European Journal of Operational Research* 10(2), 196-204.

Crainic, T., Gendreau, M., Hansen, P. and Mladenović, N., 2003. Parallel variable neighborhood search for the *p*-median. *Les Cahiers du GERAD G* 4.

Dijkstra, E. W., 1959. A note on two problems in connection with graphs, *Numerische Mathematik* 1, 269-271.

Farahani, R.Z., Asgari, N., Heidari, N., Hosseininia, M. and Goh, M., 2012. Covering problems in facility location: A review. *Computers and Industrial Engineering* 62(1), 368-407.

Francis, R., Lowe, T., Rayco, M. and Tamir, A., 2009. Aggregation error for location models: survey and analysis. *Annals of Operations Research* 167, 171-208.

Galvão, R.D., 1980. A dual-bounded algorithm for the *p*-median problem. *Operations Research* 28(5), 1112-1121.

Hakimi, S.L., 1964. Optimum locations of switching centers and the absolute centers and medians of graph. *Operations Research* 12(3), 450-459.

Hakimi, S.L., 1965. Optimum distribution of switching centers in a communications network and some related graph theoretic problems. *Operations Research* 13, 462-475.

Hansen, P., Brimberg, J., Urošević, D. and Mladenović, N., 2009. Solving large *p*-median clustering problems by primal-dual variable neighborhood search. *Data Min Knowl Disc* 19, 351-375.

Kariv, O. and Hakimi, S.L., 1979. An algorithmic approach to network location problems. part 2: The *p*-median. *SIAM J. Appl Math* 37, 539-560.

Kirkpatrick, S., Gelatt, C. and Vecchi, M., 1983. Optimization by simulated annealing. *Science* 220(4598), 671-680.

Kuehn, A. and Hamburger, M., 1963. A heuristic program for locating warehouses. *Manage Sci* 9, 643-666.

Maranzana, F., 1964. On the location of supply points to minimize transport costs. *Operations Research* 15, 261-270.

Marsten, R., 1972. An algorithmic for finding almost all the medians of a network. Technical report 23, Center for Math Studies in Economics and Management Science, Northwestern University.

Morris, J., 1978. On the extent to which certain fixed-charge depot location problems can be solved by LP. Journal of the Operational Research Society 29, 71-76.

Murray, T.A. and Church, R.L., 1996. Applying simulated annealing to location-planning models, *Journal of Heuristics* 2, 31-53.

Peeters, D. and Thomas, I., 1995. The effect of spatial structure on *p*-median results. *Transportation Science* 29, 366-373.

Rahman, S. and Smith, D., 1991. A comparison of two heuristic methods for the *p*-median problem with and without maximum distance constraints. *Int J Open Product Manage* 11, 76-84.

Rolland, E., Schilling, D., and Current, J., 1996. An efficient tabu search procedure for the *p*-median problem. *Eur J Oper Res* 96, 329-342.

Schilling, D., Rosing, K. and Revelle, C., 2000. Network distance characteristics that affect computational effort in *p*-median location problems. *European Journal of Operational Research* 127(3), 525-536.

PAPER IV

An empirical test of the gravity *p*-median model

Authors[♦]: Kenneth Carling, Mengjie Han, Johan Håkansson[♣], and Pascal Rebreyend

Abstract: A customer is presumed to gravitate to a facility by the distance to it and the attractiveness of it. However regarding the location of the facility, the presumption is that the customer opts for the shortest route to the nearest facility. This paradox was recently solved by the introduction of the gravity *p*-median model. The model is yet to be implemented and tested empirically. We implemented the model in an empirical problem of locating locksmiths, vehicle inspections, and retail stores of vehicle spare-parts, and we compared the solutions with those of the *p*-median model. We found the gravity *p*-median model to be of limited use for the problem of locating facilities as it either gives solutions similar to the *p*-median model, or it gives unstable solutions due to a non-concave objective function.

Key words: distance decay, market share, network, retail, simulated annealing, travel time

[♦] Kenneth Carling is a professor in Statistics, Mengjie Han is a PhD-student in Micro-data analysis, Johan Håkansson is a professor in Human Geography, and Pascal Rebreyend is a professor in Computer Science at the School of Technology and Business Studies, Dalarna university, SE-791 88 Falun, Sweden.

[♣] Corresponding author. E-mail: jhk@du.se. Phone: +46-23-778573.

1. The background to the gravity *p*-median model

Consider a market area with already existing facilities (or service points) competing for customers. Conventionally, a model for estimating market shares is based on the gravity model presented by Huff (1964, 1966). He proposed the probability that a customer patronizes a certain facility to be a function of the distance to and attractiveness of the facility. The model defines for each customer a probability distribution of patronage for each facility in a market area. Thereby, the market share of a facility can be evaluated by aggregating all the customers and corresponding probabilities in the area of interest.

The same model may be used for investigating the effect of adding or removing a single facility in the market area contingent to a specific location of that facility (see Lea and Menger, 1990). Moreover, an optimal location with regard to some outcomes can be identified (Holmberg and Jornsten, 1996).

However, the general problem of allocating *P* facilities to a population geographically distributed in *Q* demand points is usually executed in a different manner. Hakimi considered the task of locating telephone switching centers and formalized what is now known as the *p*-median model. The *p*-median model addresses the problem of allocating *P* facilities to a population geographically distributed in *Q* demand points such that the population's average or total distance to its nearest facility is minimized (e.g. Hakimi 1964, Handler and Mirchandani 1979, and Mirchandani 1990). The *p*-median objective function is $\sum_{q \in N} w_q \min_{p \in P} \{d_{qp}\}$, where *N* is the number of nodes, *q* and *p* indexes the demand and the facility nodes respectively, w_q the demand at node *q*, and d_{qp} the shortest distance between the nodes *q* and *p*. Hakimi (1964) showed that the optimal solution of the *p*-median model existed at the network's nodes. After Hakimi's work, the *p*-median model has been used in a remarkable variety of location problems (see Hale and Moberg, 2003).

However, it has been argued that the *p*-median model is inappropriate for locating facilities in a competitive environment because of the assumption that customers opt for the nearest facility (see e.g. Hodgson, 1978 and Berman and Krass, 1998). Recently, Drezner and Drezner (2007) presented the gravity *p*-median model that integrates the gravity rule with the *p*-median model. In their paper, they restate arguments for the gravity rule that can be found elsewhere: 1) the population is often spatially aggregated and approximately represented by the center of the demand point, 2) customers might act on incomplete information regarding the distance to each of the facilities, and 3) facilities vary in attractiveness to customers. There is also a fourth argument namely that the choice of facility may depend on other purposes for a trip (Carling and Håkansson, 2013).

Up to now, the computational aspects of the gravity *p*-median model have been studied with the intention of finding good solutions to the NP-hard problem (Drezner and Drezner, 2007 and Iyiguna and Ben-Israel, 2010). The same holds for Drezner and Drezner's (2011) extension of the model to a multiple server location problem.

The aim of this paper is to put the gravity *p*-median model to an empirical test. We consider the problem of locating 7 locksmiths, 11 vehicle inspections, and 14 retail stores of vehicle spare-parts in a Swedish region where we have detailed network data and precise geo-coding of customers. The *p*-median model ought to be appropriate in the vehicle inspection problem, whereas the gravity p-median model is presumably more suitable for the retail store problem. The problem of locating locksmiths may be regarded both a *p*-median problem and a gravity *p*-median problem.

This paper is organized as follows: section two presents the empirical setting and discusses the implementation of the gravity *p*-median model. Section three presents the results. And the fourth section concludes this paper.

2. Implementing the gravity *p*-median model

2.1 Geography

Figure 1 shows the Dalecarlia region in central Sweden, about 300 km northwest of Stockholm. The size of the region is approximately 31,000 km^2. Figure 1a depicts the location of customers in the region[1]. As of December 2010, the Dalecarlia population numbers 277,000 residents. About 65 % of the population lives in 30 towns and villages of between 1,000 and 40,000 residents, whereas the remaining third of the population resides in small, scattered settlements.

Figure 1b shows the landscape and it gives a perception of the geographical distribution of the population. The altitude of the region varies substantially; for instance in the western areas, the altitude exceeds 1,000 meters above sea level whereas the altitude is less than 100 meters in the southeast corner. Altitude variations, the rivers' extensions, and the locations of the lakes provide many natural barriers to where people could settle and how a road network could be constructed in the region. The majority of residents live in the southeast corner while the remaining residents are located primarily along the two rivers and around Lake Siljan in the middle of the region. The region constitutes a secluded market area as it is surrounded by extensive forest and mountain areas which are very sparsely populated. Hence, in the following we ignore potential influence of customers and facilities outside the region.

2.2 Distance measure

Carling, Han, and Håkansson (2012) found the Euclidian distance measure to perform poorly for the *p*-median problem, leading to suboptimal locations and biased market shares in this rural area. In the empirical analysis we have tested the Euclidean measure but because of its shortcomings we focus on what follows from the travel-time distance. To obtain the travel-time, we assumed that the attained velocity corresponded to the speed limit on the road network.

[1] The population data used in this study comes from Statistics Sweden, and is from 2002 (www.scb.se). The residents are registered at points 250 meters apart in four directions (north, west, south, and east) implying a maximum error of 175 meters in the geo-coding of the customers. There are 15,729 points that contain at least one resident in the region.

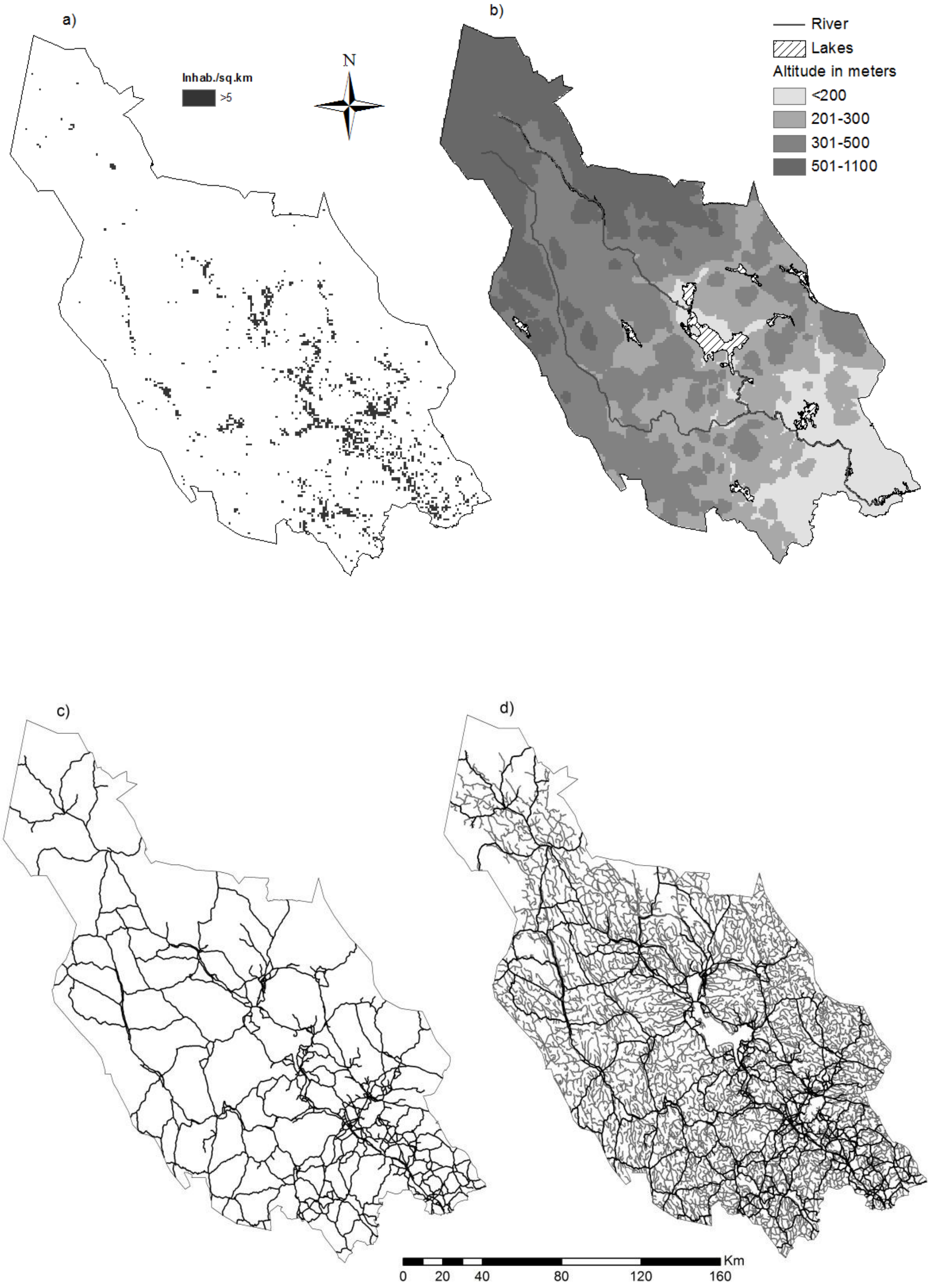


Figure 1: Map of the Dalecarlia region showing (a) one-by-one kilometer cells where the population exceeds 5 inhabitants, (b) landscape, (c) national road system, and (d) national road system with local streets and subsidized private roads.

The Swedish road system is divided into national roads and local streets which are public as well as subsidized and non-subsidized private roads. In Dalecarlia, the total length of the road system in the region is 39,452 km (see Figure 1d).[2] Han, Håkansson, and Rebreyend (2013) used the *p*-median model on this road network, and they noted that for *P* small the national road network was sufficient. Therefore, we only use the national roads in this study.

Figure 1c shows the national road network in the region. The national road system in the region totals 5,437 km with roads of varying quality which are in practice distinguished by a speed limit. The speed limit of 70 km/h is default and the national roads usually have a speed limit of 70 km/h or more.

2.3 Objective function and parameters

The objective function for the gravity *p*-median model is similar to the objective function of the *p*-median model with the addition of a term specifying the probability that a customer located at node *q* will visit a facility at node *p*. Drezner and Drezner (2007) specify the probability term as $\frac{A_p e^{-\lambda d_{qp}}}{\sum_{p \in P} A_p e^{-\lambda d_{qp}}}$, where A_p is the attractiveness of the facility and λ is the parameter of the exponential distance decay function[3]. As a consequence, the gravity *p*-median objective function is $\min_{p \in P}\left\{\sum_{q \in N}\left[w_q \frac{\sum_{p \in P} d_{qp} A_p e^{-\lambda d_{qp}}}{\sum_{p \in P} A_p e^{-\lambda d_{qp}}}\right]\right\}$.

As noted above, we use travel-time as the distance measure which means that the quickest path between *q* and *p* needs to be identified. We implemented the Dijkstra algorithm (Dijkstra 1959) and retrieved the shortest travel time from the facilities to residents in each evaluation of the objective function. We impose that facilities are located at the nodes of the network even though the Hakimi-property does not generally

[2] The road network is provided by the NVDB (The National Road Data Base). The NVDB was formed in 1996 on behalf of the government and is now operated by the Swedish Transport Agency. NVDB is divided into national roads, local roads and streets. The national roads are owned by the national public authorities, and their construction is funded by a state tax. The local roads or streets are built and owned by private persons, companies, or by the municipalities. Data was extracted in spring 2011 and represents the network of winter 2011. The computer model is built up by about 1.5 million nodes and 1,964,801 road segments.

[3] The exponential function and the inverse distance function dominate in the literature as discussed by Drezner (2006).

Table 1: Swedes self-estimated network distance for purchases of durable goods.

	Travel distance (km)						
	<2.5	2.5-5	5-25	25-50	50-125	125-250	>250
Proportion (%)	14	22	32	17	9	4	2

apply to the gravity *p*-median model (Drezner and Drezner, 2007). The reason for this choice is to enable a fair comparison with the *p*-median solutions which will be at the nodes. Moreover, all customers are assigned to the facilities which means that we abstract from the possibility of lost demand, i.e. the case when some customers seek substitutes because of the facilities being inaccessible for them (Drezner and Drezner, 2012).

The attractiveness parameter, A_p, is discussed under subsection 2.5 but it is varied for only one of the businesses.

The value of lambda[4] is decisive on how far a customer is likely to travel for patronize a facility. For λ=0, all (equally attractive) facilities are equally likely to be patronized by the customer, irrespective of the customer's distance to them. The larger the value of lambda, the more attached the customer is to the nearest facility. Drezner (2006) derived λ=0.245 for shopping malls in California whereas Huff (1964, 1966) reported, albeit using the inverse distance function, on larger values for grocery and clothing stores. We use Drezner's value converted from Euclidean distance and English miles into the corresponding value for the network distance and in kilometres. By assuming the network distance[5] to be 1.3 times the Euclidean distance we have λ=0.11.

A value of lambda specific for the applications here is λ=0.035. We obtained this value as the maximum likelihood estimate of the parameter based on grouped data from the Swedish Trade Federation (Svensk Handel). The data values are shown in Table 1. In the empirical part, we only consider goods and services requiring infrequent trips which

[4] The solutions to the location models are obtained in the travel time network. To conform to the existing literature, we discuss lambda in terms of a parameter for a road network. In the algorithm we adjust lambda to the corresponding value in the travel time network.

[5] Love and Morris (1972) found a coefficient of 1.78, however the relationship has been observed elsewhere in the literature and found relevant for this network in Carling et al (2012).

ought to be like durables.

2.4 Implementation of simulated annealing

The *p*-median problem is NP-hard (Kariv and Hakimi, 1979) and so is the gravity *p*-median problem. Han et al (2013) discussed and examined solutions to the *p*-median problem for the region's network. They advocated the simulated annealing algorithm which is used here and also used for the gravity *p*-median model.[6] This randomized algorithm is chosen due to its ease of implementation and the quality of results regarding complex problems. Most important in our case is that the cost of evaluating a solution is high and therefore we prefer an algorithm which keeps the number of evaluated solutions low. This excludes for example algorithms like Genetic Algorithm and some extended Branch and Bound. Moreover, we have good starting points obtained from pre-computed trials. Therefore a good candidate is simulated annealing (Kirkpatrick, Gelatt, and Vecchi, 1983).

The simulated annealing (SA) is a simple and well described meta-heuristic. Al-khedhairi (2008) describes the general SA heuristic procedures. SA starts with a random initial solution s, the initial temperature T_0, and the temperature counter $t = 0$. The next step is to improve the initial solution. The counter $n = 0$ is set and the operation is repeated until $n = L$. A neighbourhood solution s' is evaluated by randomly exchanging one facility in the current solution to the one not in the current solution. The difference, Δ, of the two values of the objective function is evaluated. We replace s by s' if $\Delta < 0$, otherwise a random variable $X \sim U(0,1)$ is generated. If $X < e^{(\Delta/T)}$, we still replace s by s'. The counter $n = n + 1$ is set whenever the replacement does not occur. Once n reaches L, $t = t + 1$ is set and T is a decreasing function of t. The procedure stops when the stopping condition for t is reached.

The main drawback of the SA is the algorithm's sensitivity to the parameter settings. To

[6] Drezner and Drezner (2007) discuss alternative heuristic algorithms.

Table 2: Average value of the objective function as well as the lower bound of a 99% confidence interval for the minimum of the objective function (in parenthesis).

	Location model		
Business	PM	GPM (λ=0.11)	GPM (λ=0.035)
Vehicle Insp.	611.09 (597.16)	794.06 (756.36)	1724.86 (1671.46)
Locksmiths	798.45 (778.91)	946.59 (907.23)	1756.08 (1713.88)
Spare-parts	545.80 (518.53)	745.23 (708.12)	1716.51 (1669.63)
- twofold A_p	n a	754.57 (739.23)	1716.86 (1664.78)
- fivefold A_p	n a	757.89 (718.12)	1702.54 (1669.79)

overcome the difficulty of setting efficient values for parameters such as temperature, an adaptive mechanism is used to detect frozen states and if warranted re-heat the system.[7] In all experiments, the initial temperature was set at 400 and the algorithm stopped after 10,000 iterations. Each experiment was computed twice with different random starting points to reduce the risk of local solutions. To ascertain the quality of the solution we also applied a method for computing a 99% confidence interval for the minimum, to which the obtained solution can be compared. In doing so, we follow the recommendation in Carling and Meng (2013) who studied alternative approaches to statistically estimating the minimum of an objective function for the *p*-median problem. Table 2 gives the average of the objective function obtained as a solution to its minimum as well as the lower bound of the confidence interval. The businesses under study are described in the ensuing subsection.

Typically, the solutions are some 10 to 40 seconds away from a lower bound of the minimum which we consider sufficiently precise for this type of applications.

2.5 Businesses under study

The problem of locating vehicle inspections appears frequently in the literature on the *p*-median model (see e.g. Francis and Lowe, 1992). In Sweden, vehicle inspection was a state monopoly until 2009 when the market was deregulated. A state monopoly may be clearly regarded as a central planner and we therefore expect current locations of the

[7] Our adaptive scheme to dynamically adjust temperature works as follow: after n=10 iterations with no improvement, the temperature is increased according to newtemp=temp*3^β, where β starts at 0.5 and is increased by 0.5 each time the system is reheated. As a result, the SA will never be in a frozen state for long. The temperature is decreased each iteration with a factor of 0.95. The settings above are a result of substantial preliminary testing on this data and problem. In fact, some of the solutions were compared to those obtained by alternative heuristics.

inspections to resemble the *p*-median solution.

As of October 2012 there are eleven vehicle inspections operated by two companies in Dalecarlia. The inspections perform vehicle safety checks of vehicles according to EU protocol; hence there is no reason to expect the inspections to vary in attractiveness. Furthermore, the owner of a vehicle is required to regularly have the vehicle inspected. Older vehicles are subject to annual inspections whereas newer ones, inspections are triennial. Thus, a trip to the vehicle inspection is an infrequent patronage.

There are seven locksmiths in the region. These are small business without any central control. The virtue of the business makes it far-fetched that locksmiths differ much in attractiveness. Putting these two facts together, it is difficult to decide whether to expect locksmiths to follow a *p*-median or a gravity *p*-median location pattern.

The third business is retail stores of vehicle spare-parts. There are two competitors in the region. One has 12 facilities in the region and the other has 2 facilities. However, the stores of the latter competitor are large and offer an ample selection of spare-parts as well as many complementing products. We expect these two stores to be quite more attractive. We consider two assumptions. The first is the case where the two stores are twice as attractive as the competitor’s stores. The second is the case where the two stores are assumed to be five times as attractive.

3. Results

Figure 2 shows the current location of the 11 vehicle inspections (Figure 2a) and the 7 locksmiths (Figure 2b) in the region. Imposed on the map in the figure is the solution to the *p*-median model (hereafter PM) for the two businesses. As expected, the current location of the vehicle inspections is quite near to the PM solution where ten out of eleven facilities coincide. The current locations of the seven locksmiths differ from the PM solution, but not by much.

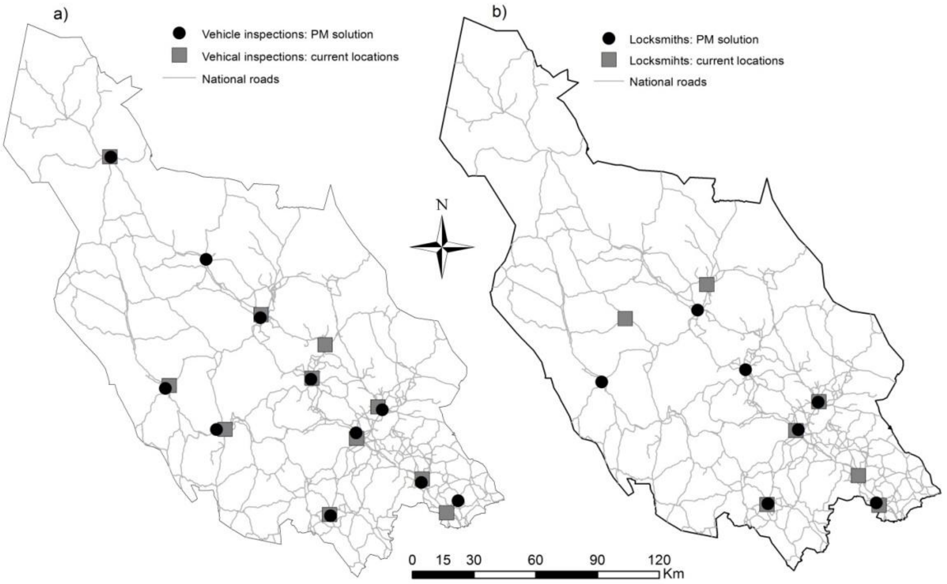

a)
Vehicle inspections: PM solution
Vehical inspections: current locations
National roads
b)
Locksmiths: PM solution
Locksmihts: current locations
National roads
N
0 15 30 60 90 120
Km

Figure 2: Map of the Dalecarlia region showing the current locations and the *p*-median (PM) solution for (a) vehicle inspections and (b) locksmiths.

We now turn to the gravity *p*-median model (hereafter referred to as GPM followed by λ used) and how it compares to PM. Figure 3 shows that the GPM(0.11) solution is similar to the PM solution; for the vehicle inspections problem, the results of the models coincides almost completely. The similarity is also apparent in the case of locksmiths.

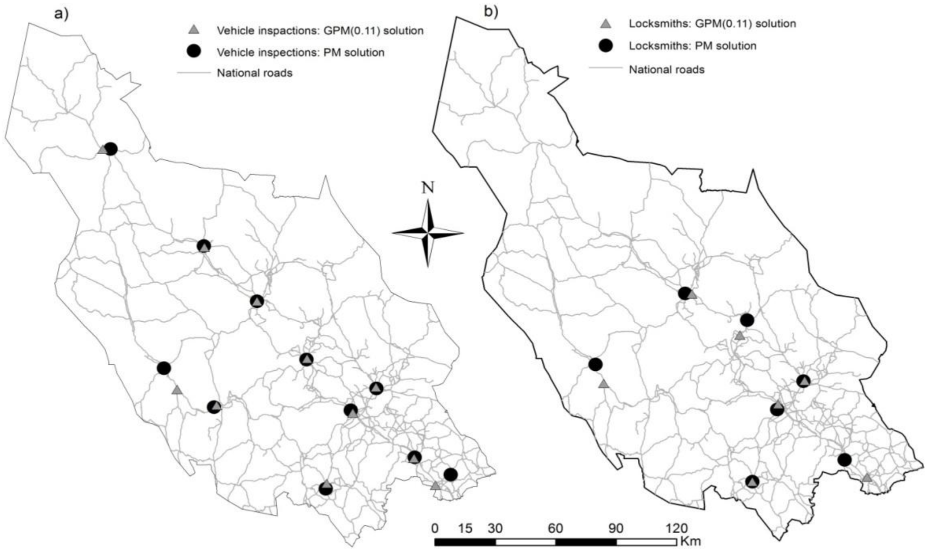

a)
Vehicle inspections: GPM(0.11) solution
Vehicle inspections: PM solution
National roads
b)
Locksmiths: GPM(0.11) solution
Locksmiths: PM solution
National roads
N
0 15 30 60 90 120
Km

Figure 3: Map of the Dalecarlia region showing the *p*-median (PM) solution and the gravity p-median (GPM) solution with $\lambda = 0.11$ for (a) vehicle inspections and (b) locksmiths.

Table 3: The customers' average travel-time (seconds) to the nearest facility for current locations and *p*-median (PM) as well as gravity *p*-median (GPM) solutions.

	Location model			
Business	Current	PM	GPM (λ=0.11)	GPM (λ=0.035)
Vehicle Insp.	612.65	611.09	629.59	863.77
Locksmiths	1014.36	798.45	815.92	1188.09
Spare-parts	789.94	545.80	551.97	808.19
- twofold A_p	n a	n a	588.29	823.73
- fivefold A_p	n a	n a	583.83	897.11

To understand the practical difference between the solutions of the PM and the GPM(0.11) models, we compute the travel-time to the nearest facility for customers in the region. Table 3 shows the average travel-time to the current locations, the PM, and the GPM solutions. The GPM(0.11) gives solutions that imply some two per cent longer travel time to the nearest vehicle inspection or locksmiths compared to the PM solutions.

Table 3 also gives the average travel-time for the GPM(0.035) solutions. Recall that this model is the best estimate of how Swedish customers patronize facilities of durable goods and services. The GPM(0.035) solutions differ substantially from the PM where the GPM(0.035) solutions imply some 50 per cent longer trips to the nearest facility on average.

Following up on the findings in Table 3, Figure 4 contrasts the GPM(0.035) solutions to the PM solution for vehicle inspections (Figure 4a) and locksmiths (Figure 4b). The models provide distinctively different geographical configuration of locations. For the GPM(0.035), facilities tend to be clustered in some towns, and we stress that it is not because the algorithm entered local minima as we have tested several starting values and the clustering pattern repeated itself.

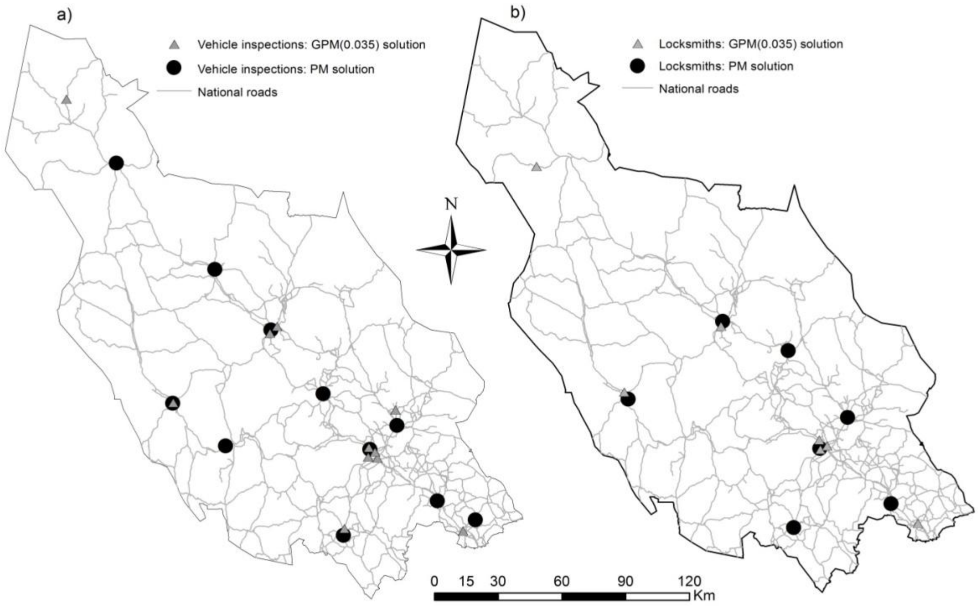

a)
Vehicle inspections: GPM(0.035) solution
Vehicle inspections: PM solution
National roads
b)
Locksmiths: GPM(0.035) solution
Locksmiths: PM solution
National roads
N
0 15 30 60 90 120
Km

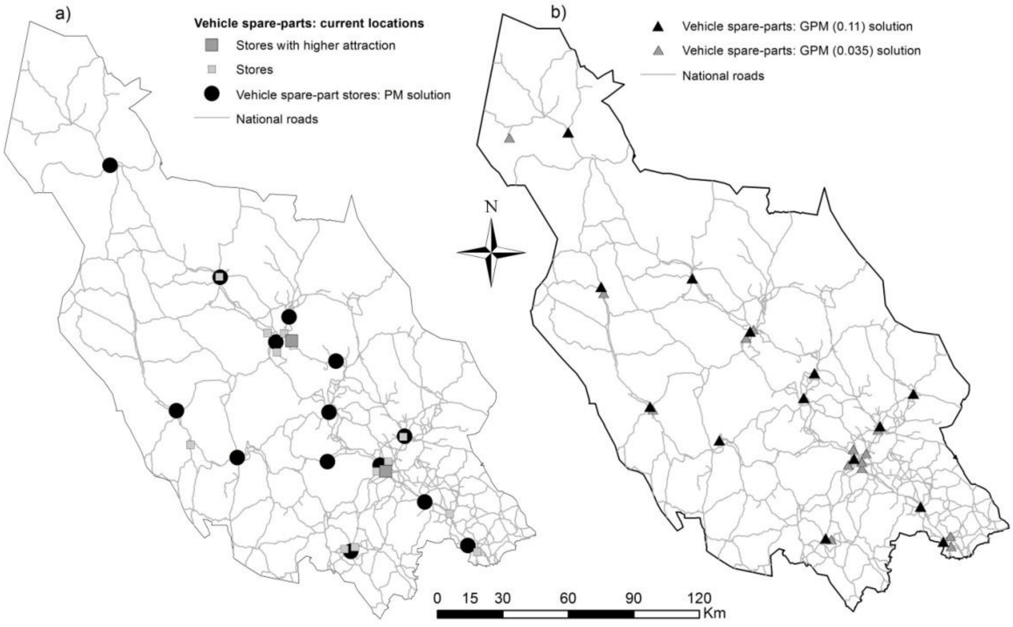
Figure 4 Map of the Dalecarlia region showing the *p*-median solution (PM) and the gravity *p*-median solution (GPM) with $\lambda = 0.035$ for (a) vehicle inspections and (b) locksmiths.

The clustering pattern indicates a difficulty to identify potential locations which give a unique market area for a facility. Consider that λ=0.035 implies that a customer's expected travel distance is about 30 kilometers, and consequently facilities cover vast market areas leaving no or only remote areas uncovered in this spatially saturated market. And in a spatially saturated market, market shares will not be found in uncovered areas but in large market areas with relatively few competing facilities; thus the clustering pattern of facilities.

Consider now the more challenging business of spare-parts for vehicles. Figure 5 shows the geographical configuration of locations for the three models and current locations. In Figure 5a the current locations of spare-parts stores is contrasted with the PM solution of 14 facilities showing a substantial difference between them. In Figure 5b configuration of GPM(0.11) and GPM(0.035) are contrasted. Again, the two values of λ lead to substantially different configurations where the clustering pattern of GPM(0.035) is pronounced. By comparing Figure 5a with 5b, there is a notable similarity between the PM and GPM(0.11) solutions on the one hand whilst on the other hand a similarity between GPM(0.035) and current location of the stores of vehicle spare-parts.

As noted above, there are two existing facilities in the region which are substantially more attractive than the competitor's twelve stores. We postulate that the difference in attractiveness is either twofold or fivefold. Figures 5c-d give the configuration of stores for the GPM solutions as well as indicate the two more attractive stores. In spite of introducing heterogeneity in attractiveness, GPM(0.11) continues to produce a solution similar to the PM. The GPM(0.035) solution gives a strong clustering with a remarkable location of facilities in the north-west of the region. This aberrant solution points at an instability of the model because of a spatially saturated market.

The GPM(0.035) has given unstable solutions in several of the problems as indicated by multiple locations at the same node and several facilities being located close to the region's border. To examine the problem of a spatially saturated market we conduct an experiment. Figure 6 gives the attained value of the objective function for the three models when locating two to twenty facilities in steps of two. It shows that the attained value of the objective function consistently decreases for the PM solutions when the


a)
Vehicle spare-parts: current locations
Stores with higher attraction
Stores
Vehicle spare-part stores: PM solution
National roads
b)
Vehicle spare-parts: GPM (0.11) solution
Vehicle spare-parts: GPM (0.035) solution
National roads
N
0 15 30 60 90 120
Km

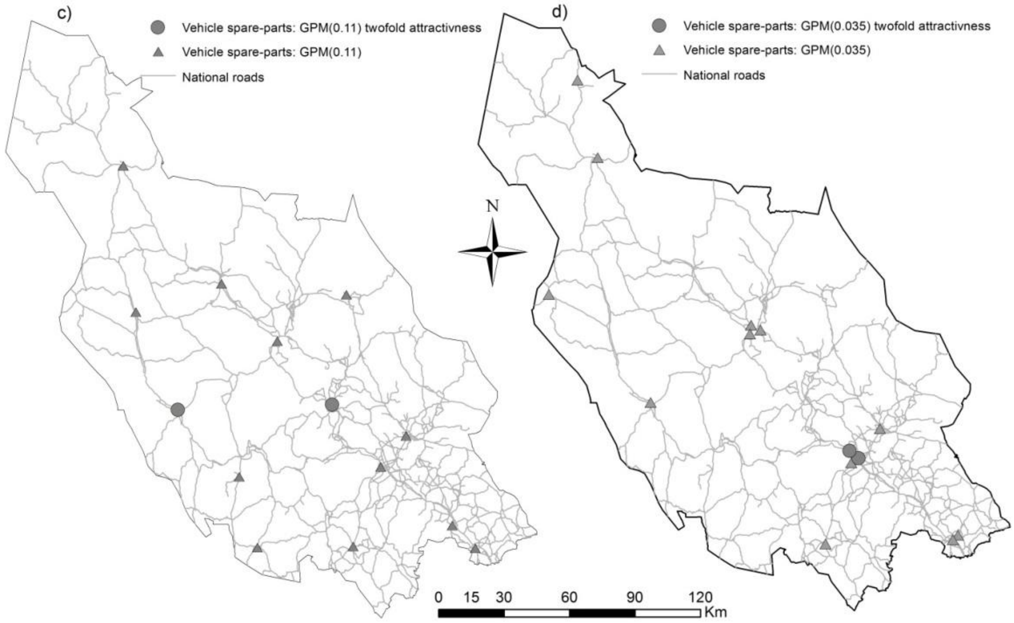

c)
Vehicle spare-parts: GPM(0.11) twofold attractivness
Vehicle spare-parts: GPM(0.11)
National roads
d)
Vehicle spare-parts: GPM(0.035) twofold attractivness
Vehicle spare-parts: GPM(0.035)
National roads
N
0 15 30 60 90 120
Km

Figure 5: Map of the Dalecarlia region showing (a) the current location and the *p*-median solution, (b) the gravity *p*-median solution with $\lambda = 0.11$ and $\lambda = 0.035$ and $A_p = 1$, (c) twofold attractiveness and $\lambda = 0.11$, and (d) twofold attractiveness and $\lambda = 0.035$ for retail stores of vehicle spare-parts.

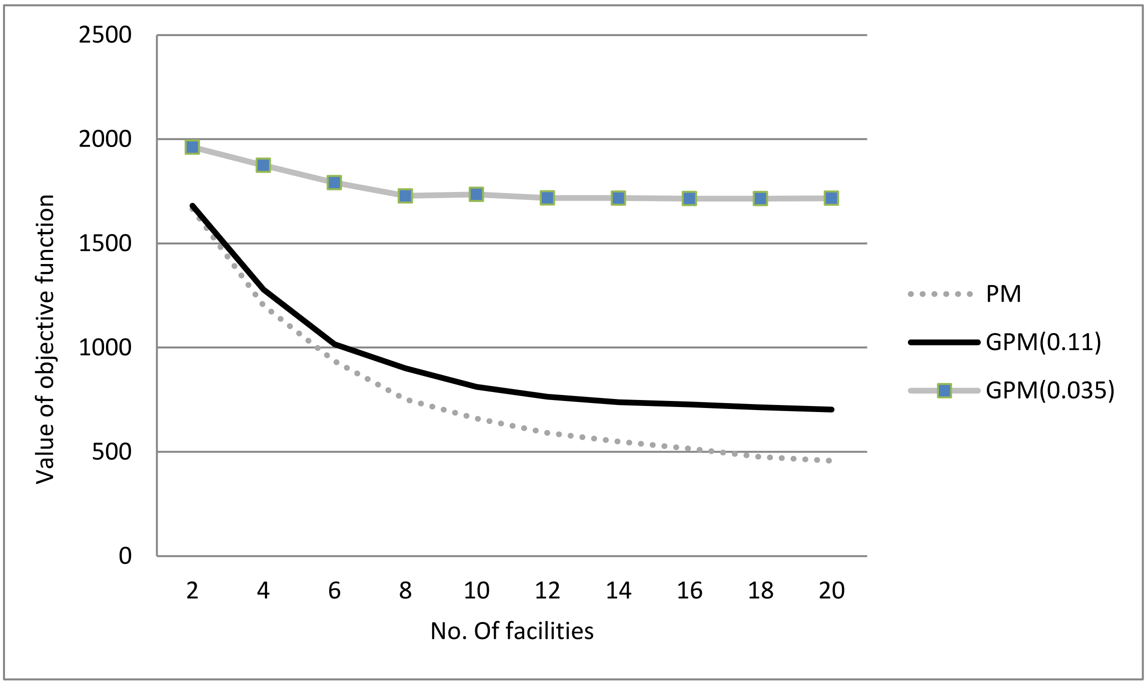


Figure 6: The attained value of the objective functions for the different location models in an experiment with locating 2 to 20 facilities in steps of 2.

Table 4: The market share for seven locksmiths in the region.

	Location model		
Facility	Current	PM	GPM (λ=0.11)
1	16.30%	12.45%	13.23%
2	14.21%	14.33%	13.96%
3	27.46%	23.85%	24.08%
4	21.76%	19.93%	19.84%
5	13.37%	13.53%	13.10%
6	=0	9.89%	9.56%
7	6.90%	6.02%	6.23%

number of facilities is increased. For GPM(0.035) the objective function decreases slowly initially and then flattens out at about 8 facilities. Hence, in the location of 8 or more facilities the objective function lacks a unique configuration of the facilities associated with the minimum because of its non-concave form. The practical interpretation of this is in a spatially saturated market there is no geographical location that will make a facility successful from offering an improved accessibility to the customers.

Before concluding that the PM and GPM(0.11) solutions are interchangeable, we need to verify that they give a similar market share and market area of the facilities. In doing so we take locksmiths as an example simply because it is easy to match PM-facilities to GPM(0.11)-facilities in this case. Table 4 gives the expected proportion of customers patronizing the seven locksmiths. In calculating the expected proportion, we stipulate that the customers patronize the facilities in accordance with the probability $\frac{e^{-0.11d_{qp}}}{\sum_{p\in P} e^{-0.11d_{qp}}}$, i.e. the gravity model with $\lambda = 0.11$. The table shows that the PM solution and GPM(0.11) solution matches. In the table the market shares for the current locksmiths is also shown, setting the market share at zero for the sixth facility as found in the PM and GPM solutions but not in reality.

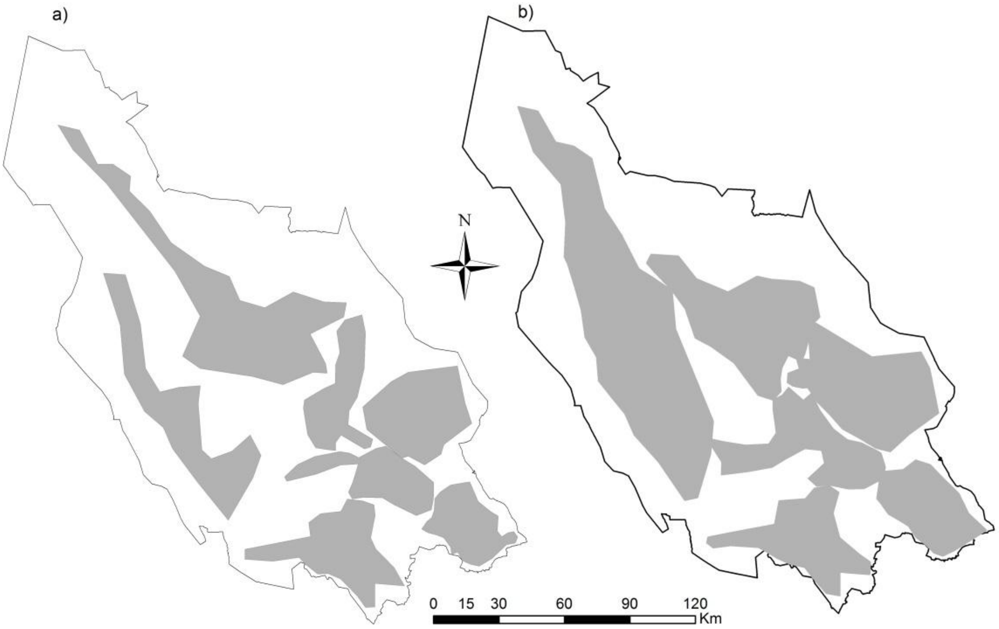


Figure 6: Map of the Dalecarlia region showing the market areas for the locksmiths; (a) areas for PM location of locksmiths, (b) areas for current location of locksmiths.

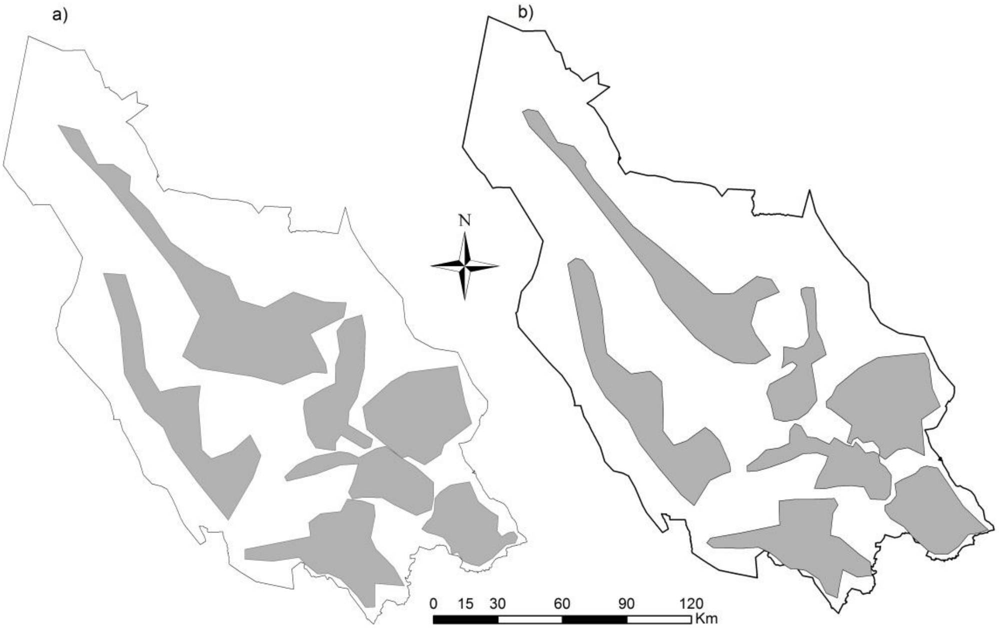


Figure 7: Map of the Dalecarlia region showing the market areas for the locksmiths; (a) areas for PM location of locksmiths, (b) areas for GPM ($\lambda = 0.11$) location of locksmiths.

The similarity in the geographical extension of the market areas for the locksmiths is illustrated in Figures 6-7. The figures show the market areas for the locksmiths including only dedicated customers i.e. those who have at least 50 per cent probability of patronizing the facility.[8] In figure 6 the current market areas is compared with market areas of the PM solution. The PM solution suggests a market area in the middle of the region which partly contributes to making the market areas quite different even though the location of facilities is similar between current and the PM solution (see Figure 2).

Figure 7 illustrates the similarity in market areas for the PM and the GPM(0.11) solutions. In summary, the PM and the GPM(0.11) solutions are found to give similar location of facilities, similar market shares, and also similar market areas. Hence, they appear interchangeable as location models.

4. Concluding discussion

The *p*-median model is used when optimal locations are sought for facilities. It is assumed that customers travel to the nearest facility along the shortest route. In a competitive environment, such as the retail sector, this is not necessarily realistic. To address the location problem more realistically, the gravity *p*-median model has recently been suggested as a tool for seeking location of multiple facilities in competitive environments. This model is not yet tested empirically. In this study we implemented the gravity *p*-median model in an empirical problem of locating locksmiths, vehicle inspections, and retail stores of vehicle spare-parts. In doing so, we contrasted the solutions of gravity *p*-median model to those of the *p*-median model.

We find that the *p*-median model gives solutions similar to the current location of vehicle inspections as expected and fairly similar to the current location of locksmiths. The current location of retail stores of vehicle spare-parts does not match the solution of the *p*-median model which indicates that the model is unrealistic in this case.

[8] Drezner, Drezner, and Kalczynski (2012) discusses and reviews several views on customers in defining market areas.

The gravity *p*-median model requires a parameter defining the reach of the facility to customers. We examined two values. The first is $\lambda = 0.11$ which is a derived value for shopping malls in California implying that the expected travel length in the road network is about 9 km (Drezner, 2006). The second value, $\lambda = 0.035,$ was obtained from a Swedish survey with an implied expected travel length in the road network of about 30 km. For $\lambda = 0.11,$ the gravity *p*-median model gives solutions that coincide with the *p*-median solutions irrespective of heterogeneity in attractiveness of the facilities. Note, however, that we introduced heterogeneity in attractiveness only in the case of stores of vehicle spare-parts where such heterogeneity was realistic.

For the most realistic value of $\lambda = 0.035$, we find the model to produce unstable solutions for at least the cases of vehicle inspections and stores of vehicle spare-parts. The instability results from a spatially saturated market in which no improvement in the objective function can be made from adding facilities. We illustrate that the market here is saturated for *P* at around 6-8 facilities. Given a small value of lambda, the competitive edge of a facility in a spatially saturated market is not given by its location, but by its attractiveness. In summary, we find the gravity *p*-median model to add little improvement over the classical *p*-median model.

Acknowledgements

Financial support from the Swedish Retail and Wholesale Development Council is gratefully acknowledged. The funder has exercised no influence on this research all views expressed are solely the responsibility of the authors.

References

Al-khedhairi. A., (2008), Simulated annealing metaheuristic for solving p-median problem. *International Journal of Contemporary Mathematical Sciences*, 3:28, 1357-1365, 2008.

Berman, O., and Krass, D., (1998). Flow intercepting spatial interaction model: a new approach to optimal location of competitive facilities, *Location Science*, 6, 41-65.

Carling K., Han M., and Håkansson J, (2012). Does Euclidean distance work well when the *p*-median model is applied in rural areas?, *Annals of Operations Research*, 201, 83-97.

Carling, K., and Håkansson, J., (2013). A compelling argument for the gravity *p*-median model, *European Journal of Operational Research*, 226:3, 658-660.

Carling, K., and Meng, X., (2013). A stopping rule while searching for optimal solution of facility-location, Working papers in transport, tourism, information technology and microdata analysis, 2013:20.

Dijkstra, E.W., (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1, 269–271.

Drezner, T., (2006). Derived attractiveness of shopping malls, *IMA Journal of Management Mathematics*, 17, 349-358.

Drezner T., and Drezner Z., (2007). The gravity *p*-median model, *European Journal of Operational Research*, 179, 1239-1251.

Drezner T., and Drezner Z., (2011). The gravity multiple server location problem, *Computers & Operations Research*, 38, 694-701.

Drezner T., and Drezner Z., (2012). Modelling lost demand in competitive facility location, *Journal of the Operational Research Society*, 63, 201-206.

Drezner T., Drezner Z., and Kalczynski, P., (2012). Strategic competitive location: improving existing and establishing new facilities, *Journal of the Operational Research Society*, 63, 1720-1730.

Francis, R. L., and Lowe, T. J., (1992). On worst-case aggregation analysis for network location problems, *Annals of Operations Research*, 40, 229–246.

Hakimi, S.L., (1964). Optimum locations of switching centers and the absolute centers and medians of a graph, *Operations Research*, 12:3, 450-459.

Hale, T.S., and Moberg, C.R. (2003). Location science research: a review. *Annals of Operations Research*, 32, 21–35.

Han, M., Håkansson, J., and Rebreyend, P., (2013). How does the use of different road networks effect the optimal location of facilities in rural areas?, Working papers in transport, tourism, information technology and microdata analysis, 2013:15.

Handler, G.Y., and Mirchandani, P.B., (1979). Location on networks: *Theorem and algorithms*, MIT Press, Cambridge, MA.

Hodgson, M.J., (1978). Toward more realistic allocation in location - allocation models: an interaction approach, *Environment and Planning* A 10:11, 1273-1285.

Holmberg, K., and Jornsten, K., (1996). Dual search procedures for the exact formulation of the simple plant location problem with spatial interaction, *Location Science*, 4, 83–100.

Huff, D.L., (1964). Defining and estimating a trade area, *Journal of Marketing*, 28, 34–38.

Huff, D.L., (1966). A programmed solution for approximating an optimum retail location, *Land Economics*, 42, 293–303.

Iyiguna, C., and Ben-Israel, A., (2010). A generalized Weiszfeld method for the multi-facility location problem, *Operations Research Letters*, 38:3, 207-214.

Kariv, O., and Hakimi, S.L., (1979), An algorithmic approach to network location problems. part 2: The p-median. *SIAM Journal of Applied Mathematics*, 37, 539-560.

Kirkpatrick, S., Gelatt, C., and Vecchi, M., (1983), *Optimization by simulated annealing. Science*, 220:4598, 671-680.

Lea, A.C., Menger, G.L., (1990). An overview of formal methods for retail site evaluation and sales forecasting: Part 2. Spatial interaction models, *The Operational Geographer* 8, 17–23.

Love, R.F., and Morris, J.G., (1972), Modelling inter-city road distances by mathematical functions. *Operational Research Quarterly*, 23:1, 61-71.

Mirchandani, P.B., (1990). “The p-median problem and generalizations”, *Discrete location theory*, John Wiley & Sons, Inc., New York, pp 55-117.

PAPER V

How do neighbouring populations affect local population growth over time?

Mengjie Han, Johan Håkansson, Lars Rönnegård

Dalarna University

Borlänge, Sweden

Abstract: This study covers a period when society changed from a pre-industrial agricultural society to a post-industrial service-producing society. Parallel with this social transformation, major population changes took place. One problem with geographical population studies over long time periods is accessing data that has unchanged spatial divisions. In this study, we analyse how local population changes are affected by neighbouring populations. To do so we use the last 200 years of population redistribution in Sweden, and literature to identify several different processes and spatial dependencies. The analysis is based on a unique unchanged historical parish division, and the methods used are an index of local spatial correlation. To control inherent time dependencies, we introduce a non-separable spatial temporal correlation model into the analysis of population redistribution. Several different spatial dependencies can be observed simultaneously over time. The main conclusions are that while local population changes have been highly dependent on the neighbouring populations, this spatial dependence have become insignificant already when two parishes is separated by 5 kilometres. Another conclusion is that the time dependency in the population change is higher when the population redistribution is weak, is it currently is and as it was during the 19th century until the start of industrial revolution.

Keywords: population redistribution, spatial dependency, Moran's I, non-separable time space correlation model, Sweden

1 Introduction

This study extends over a period when society changed from a pre-industrial agricultural society to an industrial society with mechanisation and wage labour and, from an industrial to a post-industrial service-producing society during the latter part of the period. Parallel with this social transformation, major population changes have took place. Consequently, the geographical distribution and redistribution of the population has been a constantly recurring research theme in geography and in other disciplines.

Over last decades substantial research focused on urbanization. However, the research touched upon concentration and dispersion and structured the population redistribution phenomena at different geographical levels internationally (e.g. Champion and Hugo 2004, Geyer and Kontuly 1996, Pounds 1990, Van der Woude, De Vries and Hayami 1990) and in Sweden (Eneqvist 1960, Norborg 1968, Andersson 1987, Håkansson 2000a, Nilsson 1989, Norborg 1999).

One problem with geographical population studies over long time periods is accessing data that has unchanged spatial divisions (e.g. Gregory and Ell 2005). This problem has forced much of the research to be either case studies often with relatively detailed information except for a limited geographical area, or studies with larger study areas, such as countries or even larger areas, spanning over long time periods with often relative low spatial resolutions.

The long term redistribution in Sweden was recently studied with a high spatial resolution (Håkansson 2000a). It was shown that the distribution of a population on a regional level at 75% was the same in 1990 as it was in 1810. It was also shown that on a local level the distribution at 50% was the same in 1990 compared to 1810. Hence, it was concluded that the redistribution of a population has mainly been a local redistribution. The reasons for this are that most migration covered a short distance and that migration was a selective process. The implication is that there should be a measurable statistical dependency between population growths in neighbouring areas. The nature of this relationship depends on what redistribution process is at work at the time. Therefore, our aim is to analyse how and to what extent neighbouring populations affect local population growth.

In this study, we adopt a national perspective on the local population growth in Sweden between 1810 and 2000. To do so, we use a unique data set with population figures in parishes for every 10th year. The parish division change over time. However, an unchanged geographical division over time has been constructed. The unchanged parish division consists of 1840 parishes. To our knowledge this spatial division is the lowest possible geographical level that is feasible to use for population studies of this kind and with this time perspective in Sweden. Even in an international perspective we are not aware of studies with this fine spatial division covering such a long time period and large geographical area. Based on the population figure, each parish's population share of the total population in Sweden and its change is calculated. To conduct the spatial statistical analysis, we first use an index of spatial autocorrelation, local Moran's I (Anselin 1995). Furthermore to control for temporal correlation, we also develop a non-separable statistical spatial-temporal correlation model to analyse how the population changes over time and space (see Cressie and Wikle 2011, Gneiting 2002). To our knowledge it is the first time such a model is used to analyse population redistribution over long time periods.

This paper is organized as follows: section two presents a short literature review over the main processes that have redistributed the population in Sweden since the beginning of the 19th century. In section three, the data and the empirical setting are presented and discussed. Section four presents the methods used in the spatial analysis. Section five gives the results. Section six concludes the paper.

2 Literature review

Several processes that have redistributed the population in Sweden have been described in the literature. Many of them, especially those dealing with the redistribution during the 19th and early 20th century, are conducted as case studies with a relatively limited geographical area as the study area.

Combining them together gives a picture of the redistribution in Sweden and that several different processes can be at work at the same time. For instance, it is obvious that the colonization of the interior parts of northern Sweden occurred at the same time as the emigration to the US.

In this section we shortly review the major processes described in the literature, and we discuss how they affect the population redistribution between neighbouring parishes.

(1) Colonisation: several studies show how the frontier of colonisation has been moved inland in Sweden's northern regions (Norrland) throughout the 19th century (e.g. Enequist 1937, Hoppe 1945, Bylund 1956 & 1968, Rudberg 1957, Egerbladh 1987). It appears that a large part of this colonisation took place through the population already living in northern Sweden starting new settlements constantly further inland from the coast. High fertility levels are an important explanation as to why a pool of colonizers evolved. However, migration from southern Sweden also took place. Colonisation in Norrland continued for a couple of decades into the 20th century. Since colonisation is a means in which new settlements evolve close to each other we expect a spatial dependency in which a parish with a population growth is surrounded by other parishes also expiring population growth.

A number of settlement history studies also show a course of colonisation at the micro level in southern and central Sweden during the early and mid-part of the 19th century due to population pressure and to enclosure revision (e.g. Dahl 1941, Arpi 1951, Eriksson 1955, Hoppe and Langton 1994). These reveal that the division of villages led to a colonisation of the thinly populated outlying lands. We expect this process to find a spatial dependency in which a parish with population decrease is surrounded by other parishes, population increase due to colonisation.

(2) Emigration: There was a great drain of population to North America (e.g. Sundbärg 1910, Atlas över Sverige 1960, Tedebrand 1972, Norman 1974, De Geer 1977, Norman & Rundblom 1980). From emigration studies it is clear that the emigration during the latter part of the 19th and early part of the 20th centuries was relatively greater from urban areas than from rural areas (Norman 1974, De Geer 1977, Norman & Rundblom 1980). However at first,

emigration was mainly from southern Sweden. Later when emigration from Norrland occurred, fewer people were involved. The loss of around million individuals undeniably had spatial consequences, as did the later addition of return migrants from North America (Tedebrand 1972, Lindblad 1995). We expect emigration to be a process in which a parish with a population decline is surrounded by other parishes undergoing the same development.

(3) Depopulation of rural areas – Countryside urbanisation: at the micro level the depopulation process began relatively early in southern Sweden (e.g. Nordström 1952, Edestam 1955, Eriksson 1974). When urbanisation started, it first took place in the countryside where relatively many smaller towns developed. Several economic historical studies about sawmill and industrial communities show how the population moved in from the immediate surroundings (e.g. Godlund 1954, Hjulström & Arpi & Lövgren 1955). Others also demonstrated the connection between the building of railways and the growth of new towns along their routes (e.g. Heckscher 1907, Elander, and Jonasson 1949).

Urbanisation: A larger number of studies show that the process of urbanisation changed after the end of the Second World War. People did not move merely from the countryside to towns. Instead, major towns experienced a powerful growth in population, while migration to southern Sweden increased, above all from Norrland (e.g. Bylund & Norling 1966). A number of studies focused on explaining the migration patterns in this stage of the urbanisation process (e.g. Godlund 1964, Wärneryd 1968, Jakobsson 1969, SOU 1970, Falk 1976). Selective migration during urbanization changed the age structure so much that the regional differences in mortality and fertility patterns have changed to such an extent that the natural population changes currently concentrate the population (Håkansson 2000b). Due to urbanisation we expect to find parishes with population growth surrounded by other parishes with population decline.

(4) Immigration: In the 1930s Sweden became a net-immigration country. It is clear from the literature that immigration is one of the major contributions to population distribution during the post-war period. During the 1960s there was a boom of labour immigration. This went mainly to the major metropolitan areas and to industrial towns in southern Sweden

(Hammar 1975, Andersson 1993, Borgegård & Håkansson 1997). In the 1970s and 1980s reasons for immigration changed. Immigration due to war and persecution became the most common reason. To an extent larger than before new immigrant groups settled in the three metropolitan regions in Sweden, Stockholm, Gothenburg and Malmö, even though there was a policy at work during the 1980s that first dispersed the immigrants. Immigration can be assumed to concentrate the population towards the largest urban areas. Since the immigration population is growing in the larger cities, we expect they are going to live in more and more parishes. Therefore, we expect a similar spatial dependency as for colonisation on local level implying that a parish with population growth is surrounded by others with population growth.

(5) Suburbanisation: During the 1960s and 1970s a growing number of dwellings began to be constructed in the fringe areas of towns. At the same time, suburban areas were built up outside the towns (e.g. Lewan 1967, Bodström, Lindström & Lundén 1979, Nyström 1990). Many smaller settlements on the fringes of towns, were also incorporated within the expanding towns (Johansson 1974). Explanations for this spread of built-up residential areas within the urban landscapes have been analysed in a number of studies (e.g. Lewan 1967, Holmgren, Listérus, Köstner & Nordström 1979, Lövgren 1986). The fringe areas and suburbs became places of residence for an increasing part of the population. We therefore expect to see a spatial dependency pattern in which a parish with population decline is surrounded by other parishes with population increase.

(6) Counterurbanisation: During the 1970s the patterns of migration were changed as the larger towns experienced outmigration while the smaller towns and the countryside experienced inmigration (Ahnström 1980, Forsström & Olsson 1982, Nyström 1990, Forsberg 1994, Borgegård, Håkansson & Malmberg 1995, Amcoff 2001). A few studies have pointed out the reasons for the stagnation in the big cities (e.g. Ahnström 1980, 1986). Several studies deal with the expansion and condition of middle-sized towns (e.g. Andersson & Strömgren 1988, Eriksson 1989, Kåpe 1999). Some studies demonstrate the importance of the demographic components for population development in a number of different types of

Table 1 Concentration and Dispersion of the population in Sweden at local and regional level.

Geographical levels	**1810-1840**	**1840-1880**	**1880-1960**	**1960-2000**
Local	*dispersion*	concentration	concentration	*dispersion*
Regional	*dispersion*	*dispersion*	concentration	concentration

municipality (Borgegård & Håkansson 1997, Håkansson 2000b). Other studies also point to the continued expansion of the major cities' commuter districts and to the continued spread of settlements that are not tied to the suburban areas (Forsström & Olsson 1982, Nyström 1990, Forsberg & Carlbrand 1994, Amcoff 2001, Lindgren 2003). Based on the literature, we expect to see the same spatial dependency pattern as for suburbanization.

Most of the work about the redistribution of population referred to above is highly limited time-wise. However in some studies, population redistribution is dealt with over long periods and therefore partially bridges the temporal limitations (e.g. Eneqvist 1960, Lewan 1967, Norborg 1968 and 1974, Hofsten & Lundström 1976, Guteland, Holmberg, Hägerstrand, Karlqvist & Rundblad 1975, Andersson 1987, Söderberg & Lundgren 1982, Hägerstrand 1988, Nilsson 1989, Borgegård, Håkansson & Malmberg 1995, Norborg 1999, Bäcklund 1999, Håkansson 2000a). From these studies and the ones reviewed above, it is relevant to divide the redistribution during the last 200 years into different time periods. The time periods and the dominating direction in the population redistribution are shown in Table 1. In Figure 1 the distribution of the populations in 1810 and 1990 are given. It illustrates that the effects of 200 years of population redistribution have led to a distribution where there are large differences in population densities between nearby located parishes. This is a pattern of a highly urbanised population. Beside that, the similarity in how the populations' are distributed in 1810 and 2000 is striking. From the figure it is also obvious that the large numbers of parishes that have undergone a population decline are located in the southern part of Sweden. Their distribution across southern Sweden is intermingled with the parishes with population increase. This together lends support to the idea that the redistribution in Sweden has mainly been a process in which nearby parishes are dependent on each other.

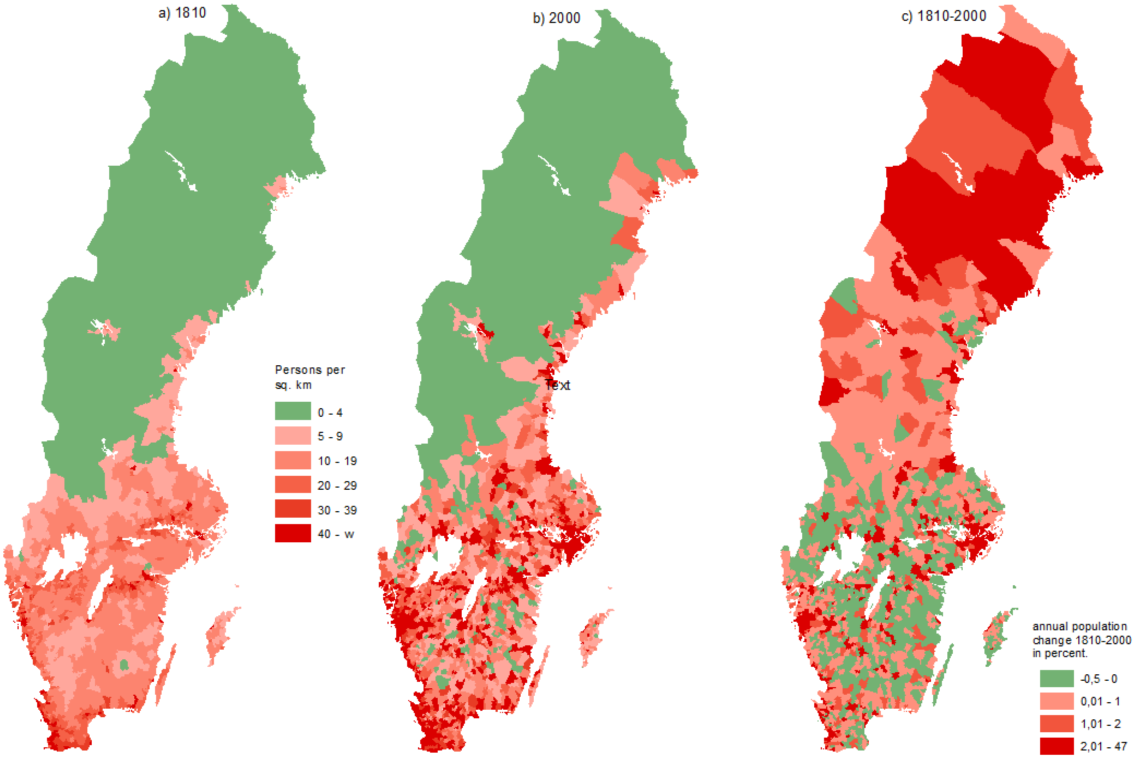


Figure 1 Population density in Swedish parishes in 1810 (a) and 1990 (b) as well as the annual population change between 1810 and 2000 (c).

Based on this literature review, we can identify four different expected local spatial dependencies that work under different population redistribution processes (Table 2). As the population has grown significantly in Sweden since the beginning of the 19th century, we look at these spatial dependencies as changes in the share of the total population. All of these four different forms of spatial dependencies can be measured. However a fifth form, the non-spatial dependency, could exist. The non-spatial dependency can have different meanings depending on how and when it occurs. One obvious reason as to why a non-spatial dependency occurs is that the spatial dependencies defined here are wrong. Another reason could be that the spatial structure used in this study is too crude and does not capture the spatial dimension. Another explanation is that the processes that are evaluated here are too weak as population redistribution processes, and they just have minor impact. These could be described as social processes with a spatial dimension.

Table 2 Expected local spatial dependencies between a parish's population change and the population change in its surrounding parishes

Population change in a parish's surrounding	A parish's population change	
	Increase (H)	Decrease (L)
Increase (H)	- Colonisation in Northern Sweden - Immigration	- Colonisation on southern Sweden - Suburbanisation - Counter urbanisation
Decrease (L)	- Urbanisation	- Emigration

3 The data and an unchanged parish division

This study is based on a material based on population numbers for administrative parish units and certain parish registrations from Tabellverket and SCB. The population returns are for every tenth year between 1810 and 2000. Altogether the material contains 2,615 geographical units. Regarding the reliability of the information, these population figures are impaired by all of the flaws accompanying the employed sources (e.g. Nilsson 1989).

The parish division changes over time. Different methods on how to create a consistent spatial division over time is discussed by Gregory and Ell (2005). Their aim is also to find automatic methods for doing this. This is not necessary in this paper. To create a spatial division over time, we started with assigning the population in each parish to a church co-ordinate from 1972 (SCB 1972). The church coordinate is chosen because the church in most parishes is located with relatively high centrality in relation to the parish inhabitants. Information about boundary changes merges and divisions that involve transfers of people (see Sveriges församlingar genom tiderna 1989) have been used to organize a spatial parish division over time. This is achived by merging parishes whenever a parish change through merging, division or boundary change has occurred. In a last step, these churches are given the parish boundaries that existed in 1990. Every parish without a church coordinate is identified and merged to a parish had been merged to or divided from. The spatial division is then further adjusted to the 2000 spatial parish division. After merging, the unchanged spatial division consists of 1840 historical parishes (see Figure 2a).

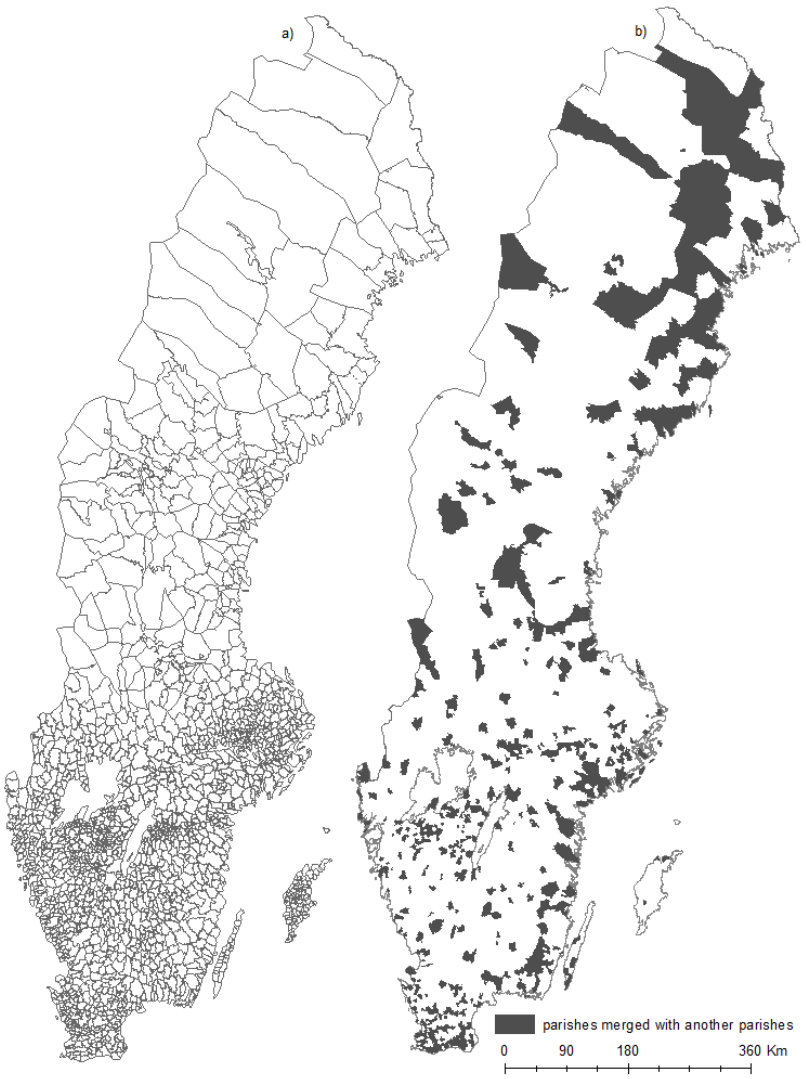


Figure 2 A historical unchanged parish division (a) and present time parish areas which have been merged with other parishes due to changes in the parish division 1810-2000 (b) in Sweden.

The reduction of parishes varies regionally (Figure 2b). The losses are largest in the sparsely populated areas in the interior parts of the northern Sweden and in the cities. Further, the regional differences in the number of analysis units naturally influence the analysis of the population changes. In principle, however, the effect of parish changes is eliminated and if one wants to conduct a study about population distribution in Sweden with this long time perspective, this division is the lowest possible level of observation that can be attained.

A more exhaustive description of the data and the unchanged parish division is given in Håkansson 2000.

4 The spatial correlation index

As shown by the literature review; we can expect that several different processes that redistribute the population are at work simultaneously, and that they result in different spatial dependencies. To obtain an understanding of how neighbouring populations affect the local population growth in a parish, we first analysed the spatial autocorrelation in the population redistribution from each parish separately. We therefore implemented Anselin's Local Moran's I in a GIS. The index (*I*) here is given from the parishes weighted by their population change rate. The method then identifies parishes whose percentage populations change rates correlates. To do this, we calculate a Local Moran's I index (I_i) and a Z score as well as the type of spatial correlation that are at work for each parish. Local Moran's I value is formulated as:

$$I_i = \frac{x_i - \bar{X}}{S_i^2} \sum_{j=1, j \neq i}^{n} \omega_{i,j}(x_i - \bar{X})$$

where

$$S_i^2 = \frac{\sum_{j=1, j \neq i}^{n} \omega_{i,j}}{n-1} - \bar{X}^2$$

x_i is an attribute feature, $\bar{X}$ is the mean of the attribute, and $\omega_{i,j}$ is the spatial weight between location i and j. The Z score is the normalized value of I_i, which indicates if I_i is significant or not. A positive value for I indicates that a parish is surrounded by other parishes with similar percentage populations change rates. Such a correlations is part of two types of spatial clusters (HH and LL in table 3) if they are statistical significant at a (0.05 level). A negative value for I indicates that a parish with a certain percentage population change rate is surrounded by other parishes with different percentage population change rates. Such a parish is considered as an outlier in a cluster if the correlation is statistically significant (0.05

Table 3 Spatial correlations in the population redistribution measured with Local Moran's I

Spatial correlations	The spatial relationship between parishes in the population redistribution
HighHighValues (HH)	Parishes with relative high population increase surrounded by other parishes with high relative population increase
HighLowValues (HL)	Parishes with relative high population increase surrounded by other parishes with fast relative population decrease.
LowHighValues (LH)	Parishes with fast relative population decrease, surrounded by other parishes with high relative population increase
LowLowValues (LL)	Parishes with fast relative population decrease, with similar developments in surrounding parishes
No significant relationship	Parishes relative population change is taken place randomly in space

level) and this gives two other types of spatial clusters (HL and LH in Table 3). The different types considered in this study are summarized in Table 3. In the table, a fifth category of a non-statistical significant relationship between parishes' percentage population change rates is added. These defined spatial dependencies correspond with those expected as identified in the literature review (see Table 2).

In the analysis, the parishes influence on a spatial cluster is weighted depending on their distance to the evaluated parish. We used an inverse distance decay function to weight the surrounding parishes. Further, we chose to limit the area taken into account in the Local Moran's I analysis around the parishes to 86 km. The distance limit is chosen so that every parish that is evaluated has a least one other parish to be evaluated against. However, we tested to set the outer limit for the area of interest to 20, 50, 60 and 70 km. The results of the analysis remain more or less the same.

5 The spatial-temporal correlation model

In the analysis of local spatial autocorrelations so far, we only consider observations between parishes located in spatial proximity to each other. However, it is important to include a temporal dimension because observations of population change are time dependent and there are tendencies for each spatial unit to inherit features from the previous time period.

To study correlation both in space and time, we now consider a spatial-temporal correlation model.

To do so we first need to define the spatial temporal process. Let $p_{i,t}$ be the proportion of the entire population at time t living in parish i , and let the observed change z be defined as $z_{i,t} = p_{i,t} - p_{i,t-1}$. For these observations we have a spatial-temporal Gaussian process for $Z(\boldsymbol{s};t)$
where Z$(\boldsymbol{s};t)$ is a random variable of the population change in space $\boldsymbol{s}$ and time t. In the spatial-temporal analysis, the covariance C describes the relationship between nearby observation in time and space $cov\big(Z(\boldsymbol{s};t),\ Z(\boldsymbol{s}+\boldsymbol{h};t+u)\big) = C(\boldsymbol{h};u)$. Here $\boldsymbol{s}+\boldsymbol{h}$ is the increase in spatial distance and $t+u$ is the increase in time, and all elements in C are assumed to be non-negative.

For a covariance model assumed to be separable (as it often is), $C(\boldsymbol{h};u)$ can be written as $C(\boldsymbol{h}) \cdot C(u)$. However, these kinds of separable covariance models often produce erroneous results when applied to real world data (e.g. Cressie 2011). Due to this, we turn to a non-separable covariance model which not only considers a product of spatial and temporal covariances, but also the interaction between them (e.g. Cressie and Huang 1999, Gneiting 2002, Stein 2005). We use a non-separable covariance model suggested by Gneiting (2002):

$$C(\boldsymbol{h};u) = \frac{\sigma^2}{(|u|^{2\alpha}+1)^{\beta}} exp\left(\frac{-c||\boldsymbol{h}||^{2\omega}}{(|u|^{2\alpha}+1)^{\omega\beta}}\right)$$

and the parameters to be estimated are $\sigma^2, \alpha, \beta, \omega$ and c . The parameter β describes the interactions between space and time and can take values from 0 to 1. For $\beta = 0$, the covariance function is separable, and for large β there is a strong dependency. The parameters were estimated by minimizing the difference between observed and fitted variograms (see Appendix).

6 Results

6.1 Local population change and spatial dependencies with the neighbouring populations

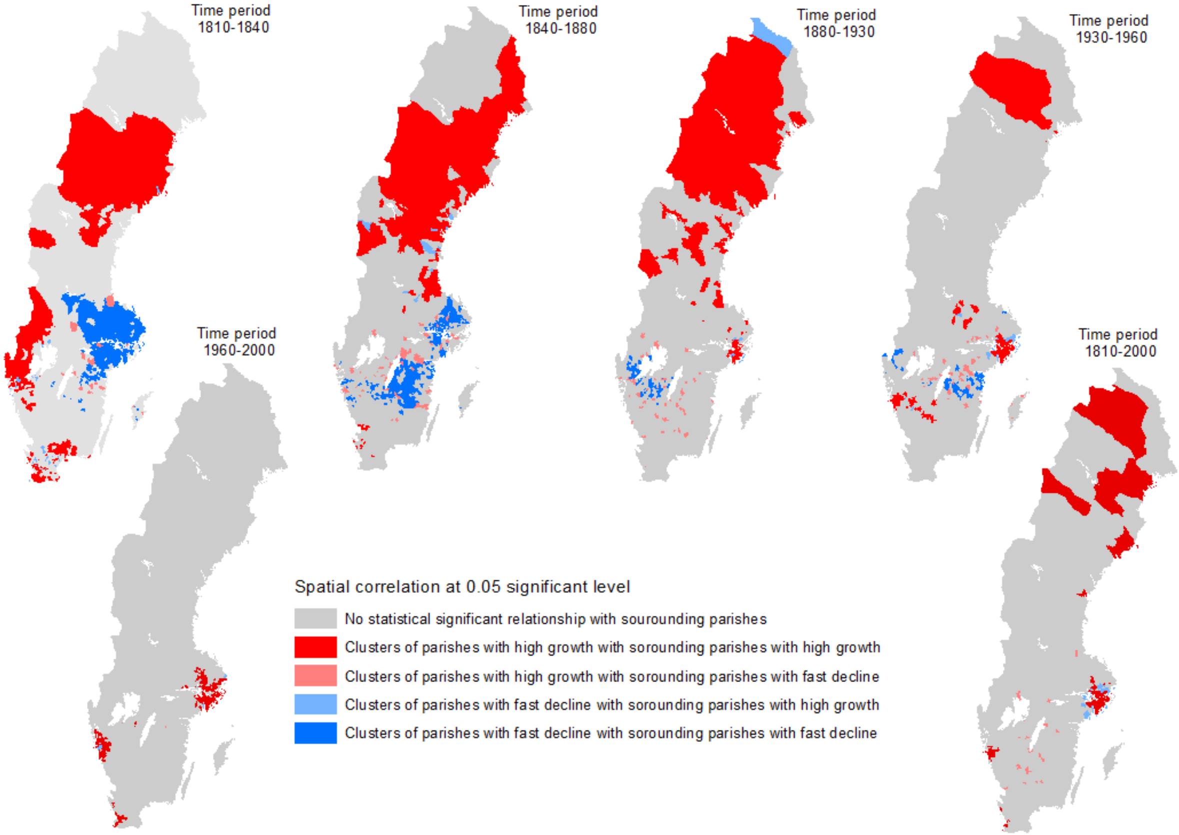

Time period 1810-1840
Time period 1840-1880
Time period 1880-1930
Time period 1930-1960
Time period 1960-2000
Time period 1810-2000
Spatial correlation at 0.05 significant level
No statistical significant relationship with sourounding parishes
Clusters of parishes with high growth with sorounding parishes with high growth
Clusters of parishes with high growth with sorounding parishes with fast decline
Clusters of parishes with fast decline with sorounding parishes with high growth
Clusters of parishes with fast decline with sorounding parishes with fast decline

Figure 3 Spatial correlations between proximity located parishes in the population redistribution in Sweden 1810 to 2000 during different sub periods.

To analyse the question of how local population change is affected by neighbouring populations, we first turn to the question regarding the extent of which different local spatial dependencies existed in the population redistribution. To answering this question, we used Moran's I to search for clusters of different spatial dependencies defined in Table 3. Figure 3 shows the clusters of spatial dependencies from that analysis with the time divided into the 5 sub periods as discussed above as well as for the entire 200 hundred year study period.

It is obvious from Figure 3 that different clusters of spatial dependencies affecting the population redistribution co-exist at the same time. It is also obvious that the spatial dependencies in the population growth change over time. In addition, note that all of the four different the spatial clusters identified and defined in this study (HH, HL, LH and LL see Table 2 and 3) have existed in the redistribution of the Swedish population.

The most wide spread and long lived form of cluster of spatial dependency in Swedish population redistribution is the one with a parish that has a fast population increase and which is surrounded by neighbouring parishes experiencing fast population increase (HH-clusters). This type of spatial autocorrelation is at work early in the study period and is common in the northern parts of Sweden as well as in and around the three metropolitan areas. Expected spatial dependencies of colonisation in northern Sweden and from immigration could therefore be observed.

Resulting from urbanisation the expected spatial dependency, with a single growing parish surrounded by parishes with population decrease (HL-clusters), can also be seen in the figure. However, HL-clusters in the population redistribution are mainly at work in the southern part of Sweden. Even though HL-clusters can be noted in the early 19th century, it is most common during the 20th century. Within the southern part of Sweden it is also notable that the frequency by which HL-clusters can be observed alternate over time between the different parts.

Clusters with parishes with a population decrease and surrounded by other parishes with a similar population change (LL-clusters), is, for a long time, co-existing in the southern part of Sweden with mainly HL-clusters. At the beginning of the study period, the LL-clusters is at work first in almost every parish in a areas around the capital city of Stockholm stretching throughout the district of Bergslagen. Later, the centre of gravitation for this type of cluster moved further south to some of the agricultural heartlands in Sweden. The population redistribution behind this type of cluster corresponds well with the overall migration out of these areas, first to colonize the northern part of Sweden and second to North America. In the 20th century, the LL-clusters become increasingly mixed up with HL-clusters. It also alternate in a similar way with its centre of gravitation between different parts of southern Sweden. This happened over time during the urbanisation, and it leads us to interpretat that this is part of the urbanisation that does not involves just a movement of people from the closest surrounding countryside, but also from a countryside at a longer distance to the growing cities.

The last defined spatial dependency, a parish with decreasing population surrounded by parishes with increasing populations (LH-clusters), can also be found, even though it is not that common either in time or space. However in early 19th century, LH-clusters could be seen in the northern parts of Scania in the most southern part of Sweden. Here it corresponds to enclosure revision, the colonisation of locally marginalised, and unproductive agricultural land. Beside that, LH-clusters could be observed around the metropolitan city of Stockholm during the urbanisation process.

Even though these clusters of spatial dependencies can be observed to be at work, perhaps the most striking feature in the population redistribution throughout the study period is seemingly the lack of spatial dependency in population change between parishes. This is given by the fact that for a majority of the parishes there is no measurable significant spatial autocorrelation between neighbouring parishes. The development over time in this respect is also clear. The share of the parishes where population change does not correlate with surrounding neighbouring parishes increased from 63 per cent in the first sub period to above 93 per cent during the last period at the end of the 20th century. This means that the spatial relations in the redistribution of the population, as described in the literature, was at work in Sweden during the last 200 years either are counter acting each other, or are at work on an even lower geographical level which is impossible to measure here.

6.2 Spatial-tempral dependency in the local population change.

We first turn to the analysis of the temporal dependencies in the Swedish population redistribution for when spatial dependencies are controlled. Table 4 shows the time parameter and the interaction parameter from equation 1. As shown in the table, the fitted covariance function was far from 0, and therefore, since the estimated interaction parameter β varies between 0.33 and 0.95 (Table 4), it shows the importance of including time-space interaction in a non-separable covariance model when analysing population redistribution. In table 4 the correlation in time is also shown. The strongest temporal is in the 19th century with a peak for the period of 1840-1880 with a temporal correlation of 0.54. Therefore, this shows that during this period, the population growth in each parish was heavily dependent

Table 4 Estimated parameters for temporal changes within parishes.

Time period	α	β	Correlation between annual changes
1810-1840	0.89	0.33	0.25
1840-1880	0.02	0.84	0.54
1880-1930	0.64	0.92	0.06
1930-1960	0.84	0.95	0.02
1960-2000	0.62	0.76	0.11

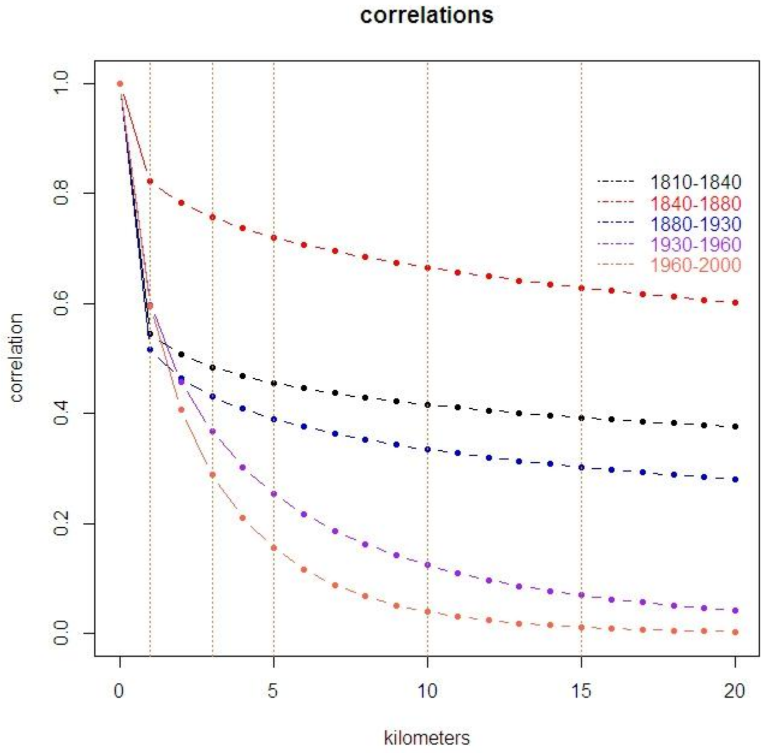


Figure 4 Correlations, when inherent time dependency is controlled for, on different distances between parishes' population changes in Sweden 1810-2000 divided into 5 sub periods

on the previous years' growth. From being very low, the correlation once again increases at the end of the study period. Therefore according to this study, it seems as if time dependency in the local population change is low when the population and the distribution change substantially as it is after the industrial revolution the 1880s and approximately at a time when became a common good in the 1960s.

We now turn to the analysis of the spatial dependencies in the Swedish population redistribution for when inherent time dependencies are controlled. Figure 4 give the correlation in population change between an average parish and other parishes lying with a increasing distance (between 0 and 20 km) from it for a time lag of 0 years (ie $u = 0$) during the different sub periods. Unsuprisingly this shows, for instance, that the correlation with it own population change is 1. However it also shows that the spatial correlation decrease with increasing distance. Also, it is when the correlation curves for different time periods are compared that the spatial dependency in the population redistribution have underwent changes over the 200 year study period. For instance, compare the curves stretching over the 19th century with the ones stretching over the 20th century. In the 19th century the correlation between parishes population change was still as high as about 60 perscent when they were as far away from each others as 20 km. This changed significantly, and at the end of the 20th centrury the distance decay function had become much steaper. The spatial correlation in population change between parishes is on average already non-existing when the distance between them is 5 km. To conclude, spatial dependency on how local population change is affected by neighbouring populations have gone from a situation in which there was a strong dependency even with parishes located far away from each other to a situation where there is a very limited covariation between parishes as close to each other as 5 km.

6 Concluding discussion

In this study, we analyse how local population change is affected by neighbouring populations. To do so we use the last 200 years of population redistribution in Sweden. From the literature several different processes and spatial dependencies can be identified. The analysis is based on a unique unchanged historical parish division, and the methods used are an index of local spatial correlation (Anselin Local Moran's I). To control for inherent time dependencies we introduce a non-separable spatial temporal correlation model into the analysis of population redistribution.

We found that the correlation between neighbouring parishes' population change have diminished over time. From a situation in the 19th century when there was a strong spatial dependency even between parishes as far apart as 20 kilometres, it has change so that, nowadays, the correlation is already marginal when the distances between parishes is 5 kilometres. The conclusions that can be drawn from this are: firstly that the local population changes have been rather dependent on the neighbouring populations and secondly spatial dependency in this respect is nowadays very low.

Another finding is that the temporal dependency in the local population change increases when the geographical distribution of population becomes more stable.

We also found several different spatial dependencies at work influencing the redistribution of population. For instance, all local spatial dependencies defined by Local Moran's I can be observed. In fact it is shown that for most of the time, two or more local spatial dependencies are at work in redistributing the population at the same time. However, which of the four spatial dependencies analysed here that are at work at the same time change over time. Also note that the 4 spatial dependencies defined by Moran's I (see table 3) do not capture all the spatial combinations that are at work simultaneously in the redistribution. A mixture of different spatial dependencies at work simultaneously in the same area lends us to add interpretations which combine the defined spatial dependencies. Lastly, the only significant spatial dependencies in the population redistribution in Sweden over the last 40 years can be observed around the three metropolitan areas. The conclusion drawn from this pattern is that the redistribution in Sweden is related to immigration and high fertility rates.

It is sometimes argued that population redistribution is a complex process. To make it understandable, the spatial patterns are often summarized and simplified to a single spatial measure, or to the rural urban dimension, or urban hierarchy, or to a very high geographical level. Further, the inherent time dependency in the redistribution is seldom controlled. The long population redistribution in Sweden is certainly a result of different processes at work creating complex patterns of spatial dependencies. Applied to Sweden, we suggest some methodologies that on a low geographical level are able to both visualize the complexity in the population redistribution and to summarize this when the inherent time dependency is controlled for.

Acknowledgments

We are grateful to Christian Swärd, Lund, who has made the compilation of data set and the Demographic database in at Umeå University who have given us access to it. We also acknowledge useful comments by an anonymous reviewer.

REFERENCES

Ahnström, L. (1980) Turnaround-trenden och de nordiska huvudstadsregionernas utveckling efter 1950 (The Turnaround trend and development in the nordic capital regions after 1950.), *NordREFO* 1980, no. 3-4.

Ahnström, L. (1986) The turnaround trend and the economically active population of seven capital regions in western Europe, *Norsk Geografisk Tidskrift* 40: pp. 50-64.

Amcoff, J. (2001) *Samtida bosättning på svensk landsbygd* (Contemporary settlements on the Swedish countryside.), Uppsala University, Geografiska regionstudier, ISSN 0431-2023; 41

Andersson, R. (1987) Den svenska urbaniseringen. Kontextualisering av begrepp och processer (The Swedish urbanisation. Interpretation of concepts and processes). Uppsala universitet, Geografiska Regionstudier, no. 18, Kulturgeografiska institutionen,.

Andersson, R. (1993) Immigration policy, and the geography of ethnic integration in Sweden, *Nordisk Samhällsgeografisk Tidskrift* 16: pp 3-29.

Andersson, Å. E. and Strömquist, U. (1988*):* K-samhällets framtid (The future for the 'knowledge society'), Prisma, Värnamo.

Anselin, L., (1995) Local Indicators of Spatial Association – LISA*, Geographical Analysis*, 27(2): 93-115.

Arpi, G. (1951) *Den svenska järnhanteringens träkolsförsörjning 1830 – 1950* (The supply of charcoal in the Swedish iron industry 1830-1950), Stockholm, Jernkontorets Bergshistoriska Skriftserie, 14.

Atlas över Sverige (1960) Emigrationen (The Emigration), kartblad 57-57, Stockholm, Esselte.

Bodström, K., Lindström, B.& Lundén, T. (1979) *Bostad – arbete - samhälle –Möjligheter och restriktioner vid samlokalisering* (The residence - the work place - the society –Possibilities and restrictions when co-located), Byggforskningen, Stockholm, R77:1979.

Borgegård, L. E., Håkansson, J. & Malmberg, G. (1995) Population Redistribution in Sweden – Long Term Trends and Contemporary Tendencies, *Gegrafiska Annaler* 77B: pp 31-45.

Borgegård, L-E. & Håkansson, J (1997) Population concentration and dispersion in Sweden since the 1970s, in L-E. Borgegård, A. M. Findlay, E. Sondell (eds) *Population, Planning and Policies*, Umeå, Cerum Report no. 5, pp 9-30.

Bylund, E. (1956) *Koloniseringen av Pite lappmark t o m år 1867* (The colonisation of interior parts of Pite lappland). Uppsala, Geographica no. 30.

Bylund, E. (1968) Generationsvågor och bebyggelsespridning (Generation waves and dispersion of settlements) , *Ymer*, pp 64-71.

Bylund, E. & Norling, G. (1966) Bebyggelsen (The settlement structure), in Ahlman, A., Arpi, G., Hoppe, G. & Man-Nerfelt, C. (eds), *Sverige land och folk*. Del 1, Natur och Kultur, Stockholm.

Bäcklund, D. (1999) *Befolkningen och regionerna – Ett fågelperspektiv på regional ekonomisk utveckling i Sverige från 1820 och framåt* (The population and the regions – an aerial view on regional economic development in Sweden from 1820), Östersund, Swedish institute for Regional Research, Rapport 100.

Champion, T., and Hugo, G., (2004) New Forms of Urbanization, Beyond the Urban-Rural Dichotomy, Ashgate, 420p.

Cressie, N (1993) Statistics for spatial data, *John Wiley & Sons, Inc.*

Cressie, N. and Wikle, C. (2011) Statistics for spatio-temporal data, *John Wiley & Sons, Inc.*

Cressie, N. and Huang, H.C. (1999) Classes of non-separable, patio-temporal stationary covariance functions, *Journal of the American Statistical Association*, 94 1330-1340.

Dahl, S. (1941) *Storskiftes och enskiftes genomförandet i Skåne* (The implementation of the enclosure reforms in Scania) , Lund, Scandia.

De Geer, E. (1977) *Migration och influensfält. Studier av emigration och intern migration i Finland och Sverige 1816-1972* (Migration and influence fields. Studies of emigration and internal migration in Finland and Sweden 1816-1972), Uppsala, Studia Historica Upsaliensia no. 97.

Edestam, A. (1955) Dalslands folkmängd år 1880 och år 1950 (The population in Dalsland in 1880 and 1950), *SGÅ,* no. 31.

Egerbladh, I. (1987) *Agrara bebyggelseprocesser. Utvecklingen i Norrbottens kustland fram till 1900-talet* Agricultural Settlement processes. The development in Norrbottens coastal area until the 20th century) Umeå, Geografiska institutionen, GERUM, no. 7.

Elander, N. & Jonasson, O. (1949) Trafikförvaltningarna Göteborg–Stockholm-Gävle och Göteborg-Dalarna-Gävle. Befolkningen och näringslivet i Mellansverige 1865-1940 (The transportadministrations Gothenburg-Stockholm-Gävle and Gothenburg-Dalarna-Gävle. The population and the comersial and indistrial sector in mid-Sweden between 1865 and 1940), Göteborg, Trafikförvaltningen.

Enequist, G. (1937) *Nedre Luledalens byar* (The villages in the lower parts of the Lule river valley), Uppsala, Geographica, no. 4.

Enequist, G. (1960) Advance and retreat of rural settlement in north-western Sweden. *Geografiska Annaler*, vol 42: pp 211-220.

Eriksson, G.A. (1955) Bruksdöden i Bergslagen efter år 1850 (The decline of iron mills in Bergslagen after 1850)., Geografiska institutionen, Uppsala universitet, medelanden ser A. no. 100.

Eriksson, P-G. (1974) *Kolonisation och ödeläggelse på Gotland. Studier av den agrara bebyggelseutvecklingen från tidig medeltid till 1600-talet* (Colonisation and retreat on the iland of Gotland. Studies of the development of the agricultural settlement from the early Medival times until the 17th century), Stockholm University, Kulturgeografiska institutionen, Meddelanden no. B 27.

Eriksson, O. (1989) *Bortom Storstadsidéerna. En regional framtid för Sverige och Norden på 2010-talet* (Beyond the metropolitans. A regional future for Sweden and the Nordic countries at the 2010th decade), Carlssons, Stockholm.

Falk, T. (1976) *Urban Sweden. Changes in the Distribution of Population –the 1960s in Focus*, Stockholm, The Economic Research Institute at Stockholm School of Economics.

Forsberg, G. & Carlbrand, E (1994) *Mälarbygden - en kreativ region? En studie av Mälardalens landsbygd i förändring* (Is the valley of Mälaren a creative region? A study of a countryside in change), Uppsala University, Forskningsrapport no. 107.

Forsström, Å. & Olsson, R. (1982): *Boende i glesbygd.* (People in the countryside), Gothenburg, BFR, R100:1982.

Geyer, H.S., and Kontuly, T.M., (1996), Differential Urbanization, Integrating Spatial Models, Arnold, 343p.

Gneiting, T. (2002): Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association,* Vol 97, No 458, pp590-600.

Godlund, S. (1954): *Busstrafikens framväxt och funktion i de urbana influensfälten* (The expansion of the bus traffic in the urban fields of influence), Lund, Gleerup.

Godlund, S. (1964) *Den Svenska urbaniseringen* (The Swedish urbanisation), Medelande från Göteborgs universitets geografiska institution, no. 76.

Gregory, I.N. and Ell, P.S., (2005), Breaking the boundaries: geographical approaches to integrating 200 years of the census, Journal of Statistical Society A, 168, part 2, pp. 419-437.

Guteland, G., Holmberg, I., Hägerstrand, T., Karlqvist, A. & Rundblad, B. (1975): *Ett folks biografi. Befolkning och samhälle i Sverige från historia till framtid* (A biography of a people. The population and the society in the history and in the future), Stockholm, Publica, LiberFörlag.

Hammar, T. & Ko-Chih Tung, R. (1975) *Invandrare och politik: finländare i Södertälje: En intervjuunderökning våren 1973 om samhällsinformation, myndigheter och politik* (Immigrants and politics: Finnish immigrants in Södertälje: An interview study about information, public authorities and politics in spring 1973), Informationskontoret, Södertälje kommun.

Heckscher, E. F. (1907) Till belysning af järnvägarnas betydelse för Sveriges ekonomiska utveckling (The importance of the rail way for the economic development in Sweden), Stockholm.

Hjulström, F., Arpi, G. & Lövgren, E. (1955) Sundsvallsdistriktet 1850-1950 (The district of Sundsvall 1850-1950), Uppsala, Geographica, no. 26.

Holmgren, B., Listérus, J., Köstner, E. & Nordström, L. (1979) Skatteavdrag för arbetsresor. Avdragens betydelse för bebyggelsestruktur och inkomstfördelning. En förstudie (Tax reduction and commuting to work. The impact of tax reduction on settlement structure and income distribution. A pilot study.). Kulturgeografiska och nationalekonomiska institutionerna, Göteborgs universitet, Göteborg.

Hoppe, G. (1945): Vägarna inom Norrbottens län. Studier av den trafikgeografiska utvecklingen från 1500-talet till våra dagar (The roads in the county of Norrbotten. Studies of the development of the road infrastructure from the 16th century until present time), Uppsala, Geographica no. 16.

Hoppe, G. & Langton, J. (1994) Peasentry to capitalism. Western Östergötland in the nineteenth century, Cambridge Studies in Historical Geography, Cambridge, Cambridge University Press.

Hofsten, E. & Lundström, H. (1976) Swedish Population History, Main trends from 1750 to 1970, Stockholm, Urval no. 8, Skriftserie utgiven av statistiska centralbyrån.

Håkansson, J. (2000a) Changing Population Distribution in Sweden –Long Term trends and Contemporary Tendencies, Umeå universitet, *GERUM Kulturgeografi*, 2000:1.

Håkansson, J. (2000b) Impact of Migration, Natural Population Change and Age Composition on the Redistribution of the Population in Sweden 1970 - 1996, *Cypergeo*, 123.

Hägerstrand, T. (1988) Krafter som format det svenska kulturlandskapet (Forces that have transformed the landscape affected by human activities in Swedish), in *Mark och Vatten 2010*, Stockholm, Bostadsdepartementet, pp16-55.

Jakobsson, A. (1969) Omflyttningen i Sverige 1950-1960. Komparativa studier av migrationsfält, flyttningsavstånd och mobilitet (Migration in Sweden 1950-1960. Comparative studies of migration fields, migration distances and mobility), Meddelande Från Lunds Universitet Geografiska Institutionen, Avhandlingar LIX.

Johansson, I. (1974) Den stadslösa storstaden. Förortsbildning och bebyggelseomvandling kring Stockholm 1870 – 1970 (The non-urban metropolitan. The development of suburbs around Stockholm 1870-1970), Statens råd för byggforskning, R1974:26

Kåpe, L. (1999) Medelstora svenska städer. En studie av befolkning, näringsliv och ortssystem (The middle sized cities. A study of population, business sector and the city system). Karlstad University Studies, 1999: 2.

Lewan, N. (1967) Landsbebyggelse i förvandling (The changing settlement structure on the countryside)., Geografiska institutionen, Lunds Universitet, Medelande, L.

Lindblad, H. & Henricsson, I. (1995): Tur och retur Amerika : utvandrare som förändrade Sverige (To and from America: Emigrants that changed Sweden), Stockholm, Fischer.

Lindgren, U., (2003): Who is the Counter-Urban Mover? Evidence from the Swedish Urban System, *International Journal of population Geography*, 9, pp.399–418.

Lövgren, S. (1986) Så växer tätorten: diskussion om en modell för urban utbredning (The growth of urban areas : a discussion on a model for urban expansion). Kulturgeografiska institutionen, Umeå universitet, GERUM Kulturgeografi, No. 5.

Nilsson (1989): *Den urbana transitionen. Tätorterna i svensk samhällsomvandling 1800 – 1980* (The Urban transition. The urban settlements in the swedish socal and economic transition 1800-1980), Stockholm, Stadshistoriska institutet, Studier I stads- och kommunhistroria 5.

Norborg, K. (1968) Jordbruksbefolkningen i Sverige. Regional struktur och förändring under 1990-talet (The Swedish agricultural population. Regional structure and change during the 20th century)., Geografiska institutionen, Lunds Universitet, Medelande, LVI. Lund.

Norborg, K. (1974) Befolkningens fördelning och flyttningar i Sverige (The distribution And migration of the Swedish population)., LiberLäromedel, Lund.

Norborg, L-A. (1999) Sveriges historia under 1800- och 1900-talen. Svenska samhällsutveckling 1809-1998 (The Swedish history during the 19th and 20th centuries.The swedish social and economic development 1809 to 1998.)., Almqvist & Wiksell, Lund.

Nordström, O. (1952): Befolkningsutveckling och arbetskraftsproblem i östra Småland 1800-1850 (Population development and lbour supply problems in eastern Småland 1800-1850), Geografiska institutionen, Lunds universitet, medelande XXXIII..

Norman, H. (1974): Från Bergslagen till Nordamerika : studier i migrationsmönster, social rörlighet och demografisk struktur med utgångspunkt från Örebro län 1851-1915 (From Bergslagen to North America: studies on migration patterns, social mobility and demographic structure with the county of Örebro as startingpoint 1851-1915), Uppsala, Studia historica Upsaliensia.

Norman, H. & Rundblom, H. (1980) *Amerikaemigrationen* (The emigration to America). Cikada, Helsingborg.

Nyström, J. (1990): Stockholms Stadsland – om förändringsprocesser i en storstads ytterområden (Stockholm urban landscape –a study about change in the outer parts of the metropolitan area), Stockholm, Kulturgeografiska Institutionen, B 72.

Pounds, N.J.G., (1990), An Historical Geography of Europe, Cambridge University Press.

Rudberg, S. (1957): Ödemarkerna och den perifera bebyggelsen i inre Nordsverige (The wastelands and the peripheral settlements in the interior parts of Norrland)., Uppsala university, Geogarphica, no. 33.

SCB (1972): Koordinatregister över Sveriges församlingar (Coordinates of the Swedish parishes), Stockholm, Promemorior från SCB, no. 4,

Sherman, M (2011): Spatial statistics and spatio-temporal data, *John Wiley & Sons, Inc.*

SOU (1970): Urbaniseringen i Sverige – en geografisk samhällsanalys The Swedish urbanisation –a geographical analysis)., SOU 1970: 14, Bilaga 1, Stockholm, Allmänna förlaget.

Stein, M.L. (2005): Space-time covariance functions, *Journal of the American Statistical Association*, 100, 310-321

Sundbärg, G. (1910): Emigrationsutredningen The emigration investigation., Bilaga VIII, Bygdeundersökningar, Stockholm.

Sveriges församlingar genom tiderna (1989): Skatteförvaltningen, Riksskatteverket, Stockholm, Allmänna förlaget.

Söderberg, J and Lundgren, N. G (1982): *Ekonomisk och geografisk koncentration 1850 – 1980* Economic and spatial concentration 1850-1980)., Stockholm, LiberFörlag.

Tedebrand, L-G. (1972): Västernorrland och Nordamerika 1875 – 1913. Utvandring och återvandring (Västernorrland and North America 1975-1913. Emigration and return migration)., Uppsala, Studia Historica Upsaliensia 42.

Van der Woude, A., De Vries, J., and Hayami, A., (1990), Urbanization in History. A Process of Dynamic Interactions, Claredon Press Oxford, 371p.

Wärneryd, O. (1968): Interdependence in Urban Systems, geografiska institutionen, Gothenburg University, Medelande B:1.

Appendix

To estimate the parameters in the spatio-temporal covariance model we use the variogram function (see Sherman 2011). The variogram is related to the covariance model as $\gamma(\boldsymbol{h}; u) = C(\boldsymbol{0}; 0) - C(\boldsymbol{h}; u)$ and simplifies parameter estimation in (1). The variogram can be reformulated as:

$$var\big(Z(\boldsymbol{s} + \boldsymbol{h}; t + u) - Z(\boldsymbol{s}; t)\big) = 2\gamma(\boldsymbol{h}; u).$$

and consequently a moment estimate of the observed $\gamma(\boldsymbol{h}; u)$ is:

$$\hat{\gamma}(h; u) = \frac{1}{2|N(h; u)|} \sum_{N(h;u)} \{[Z(s_i; t_i) - Z(s_j; t_j)]^2\}$$

where $N(h; u)$ is the number of pairs of observations, truncated at 200 km and ± 20 years with no correlation assumed beyond these limits.

The fitting algorithm was implemented in R (www.rproject.org). First we need to find a model variogram γ curve that minimizes the difference to the observed variogram $\hat{\gamma}$ as $\hat{\gamma} = \{\hat{\gamma}(h_1; u_1),\ \hat{\gamma}(h_2; u_2),\ \ldots,\ \hat{\gamma}(h_m; u_m)\}$ (see Cressie 1993 and Sherman 2011). To do so we assume that the real parameter of covariance is θ. For each θ, $\gamma(\theta)$ is defined as the value of the m-dimensional variogram. Therefore, the minimization criterion is $Q(\theta) = \hat{\gamma} - \gamma(\theta)$. The model variogram is then fitted using weighted least square (WLS) (see Sherman, 2011) such that we only need to minimize:

$$Q(\theta)^T W Q(\theta) = \sum_{i=1}^{m} w_i^2 Q_i^2(\theta)$$

where W is a diagonal matrix, with elements w_i^2 on the diagonal. The choice of weight is $w_i^2 = N(h_i; u_i)/2\gamma^2(h_i; u_i)$, because $2\gamma^2(h_i; u_i)/N(h_i; u_i)$ characterize the variance of

$\hat{\gamma}(h_i; u_i)$. Plugging in w_i^2 the final expression to be minimized is $\sum_{i=1}^{m} N(h_i; u_i)\left[\frac{\hat{\gamma}(h_i;u_i)}{\gamma_i(\theta)} - 1\right]^2$. To find the "best" θ, a set of initial combination values between 0 and 1 was given and we implement Nelder–Mead simplex algorithm to calculate the θ. To avoid local minimum, we ended up with 84 thousand different initial value combinations for each period and selected the solution with the minimal function value. The running time for fitting the five correlation curves is approximate 2.5 hours.

PAPER VI

Computational study of the step size parameter of the subgradient optimization method

Mengjie Han[1]

Abstract

The subgradient optimization method is a simple and flexible linear programming iterative algorithm. It is much simpler than Newton's method and can be applied to a wider variety of problems. It also converges when the objective function is non-differentiable. Since an efficient algorithm will not only produce a good solution but also take less computing time, we always prefer a simpler algorithm with high quality. In this study a series of step size parameters in the subgradient equation are studied. The performance is compared for a general piecewise function and a specific p-median problem. We examine how the quality of solution changes by setting five forms of step size parameter α.

Keywords: subgradient method; optimization; convex function; p-median

1 Introduction

The *subgradient optimization* method is suggested by Kiwiel (1985) and Shor (1985) for solving non-differentiable functions, such as constrained linear programming. As to the ordinary gradient method, the subgradient method is extended to the non-differentiable functions. The application of the subgradient method is more straightforward than other iterative methods, for example, the interior point method and the Newton method. The memory requirement is much lower due to its simplicity. This property reduces the computing burden when big data is handled.

However, the efficiency or the convergence speed of the subgradient method is likely to be affected by pre-defined parameter settings. One always likes to apply the most efficient empirical parameter settings on the specific data set. For example, the efficiency or the convergence speed is related to the step size (a scalar on the subgradient direction) in the iteration. In this paper, the impact of the step size parameter in the subgradient equation on the convex function is studied. Specifically, an application of the subgradient method is conducted with a p-median problem using Lagrangian relaxation. In this specific application, we study the impact of the step size parameter on the quality of the solutions.

Methods for solving the p-median problem have been widely studied (see Reese, 2006; Mladenović, 2007; Daskin,1995). Reese (2006) summarized the literature on solution methods by surveying eight types of methods and listing 132 papers or books. Linear programming (LP) relaxation accounts for 17.4% among the 132 papers or books. Mladenović (2007) examined the metaheuristics framework for solving a p-median problem. Metaheuristics has led to substantial improvement in solution quality when the problem scale is large. The Lagrangian heuristic is a specific representation of LP and metaheuristics. Daskin (1995) also showed that the Lagrangian method always gives good solutions compared to constructive methods.

[1] PhD student in the School of Technology and Business Studies, Dalarna Unversity, Sweden. E-mail: mea@du.se

Solving p-median problems by Lagrangian heuristics is often suggested (Beasley, 1993; Daskin, 1995; Beltran, 2004; Avella, 2012; Carrizosa, 2012). The corresponding subgradient optimization algorithm has also been suggested. A good solution can always be found by narrowing the gap between the lower bound (LB) and the best upper bound (BUB). This property provides an understanding of how good the solution is. The solution can be improved by increasing the best lower bound (BLB) and decreasing the BUB. This procedure could stop either when the critical percentage difference between LB and BUB is reached or when the parameter controlling the LB's increment becomes trivial. However, the previous studies did not examine how the LB's increment affects the quality of the solution. The LB's increment is decided by the step size parameter of the subgradient substitution. Given this open question, the aim of this paper is to examine how the step size parameter in the subgradient equation affects the performance of a convex function through a general piecewise example and several specific p-median problems.

The remaining parts of this paper are sectionally organized into subgradient method and the impact of step size, p-median problem, computational results and conclusions.

2 Subgradient method and the impact of step size

The subgradient method provides a framework for minimizing a convex function $f : \boldsymbol{R}^n \rightarrow \boldsymbol{R}$ by using the iterative equation:

$$x^{(k+1)} = x^{(k)} - \alpha(k)g^{(k)}. \tag{2.1}$$

In (2.1) $x^{(k)}$ is the kth iteration of the argument x of the function. $g^{(k)}$ is an arbitrary subgradient of f at $x^{(k)}$. $\alpha(k)$ is the step size. The convergence of (2.1) has been proved by Shor (1985).

2.1 step size forms

Five typical rules of step size are listed in Stephen and Almir (2008). They can be summarized as:

- constant size: $\alpha(k) = \xi$
- constant step length: $\alpha(k) = \xi/\|g^{(k)}\|_2$
- square summable but not summable: $\alpha(k) = \xi/(b+k)$
- nonsummable diminishing: $\alpha(k) = \xi/\sqrt{k}$
- nomsummable diminishing step length: $\alpha(k) = \xi(k)/\|g^{(k)}\|_2$

The form of the step size is pre-set and will not change. The top two forms, $\alpha(k) = \xi$ and $\alpha(k) = \xi/\|g^{(k)}\|_2$, are not examined since they are constant size or step length which are lack in variation for p-median problem. On the other hand, the bottom three forms are studied. We restrict $\xi(k)$ such that $\xi(k)/\|g^{(k)}\|_2$ can be represented by an exponential decreasing function of $\alpha(k)$. Thus, we study $\alpha(k) = \xi/k$, $\alpha(k) = \xi/\sqrt{k}$, $\alpha(k) = \xi/1.05^k$, $\alpha(k) = \xi/2^k$ and $\alpha(k) = \xi/\exp(k)$ in this paper. We first examine the step size impact on a general piecewise convex function and then on the p-median problem.

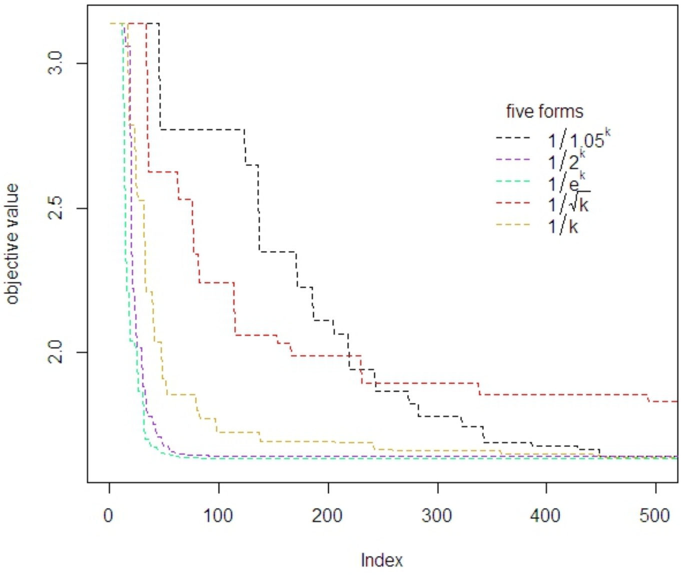


Figure 1: Objective values of a picewise convex function when five forms of step sizes are compared

2.2 general impact on convex function

We consider the minimization of the function:

$$f(x) = \max_i(\boldsymbol{a}_i^T x + b_i)$$

where $x \in \boldsymbol{R}^n$ and a subgradient $\boldsymbol{g}$ can be taken as $\boldsymbol{g} = \nabla f(x) = \boldsymbol{a}_j$ ($\boldsymbol{a}_j^T x + b_j$ maximizes $\boldsymbol{a}_i^T x + b_i$, $i = 1, \ldots, m$). In our experiment, we take $m = 100$ and the dimension of x being 10. Both $\boldsymbol{a} \sim MVN(\boldsymbol{0}, \boldsymbol{I})$ and $b \sim N(0, 1)$. The initial value of constant ξ is 1. The initial value of $\boldsymbol{x}$ is $\boldsymbol{0}$. We run the subgradient iteration 1000 times. Figure 1 shows the non-increased objective values of the function against the number of iterations. The objective value is taken when there is a improvement in the objective function. Otherwise, it is taken as the minimum value in the previous iterations.

In Figure 1, $\alpha(k) = 1/2^k$ and $\alpha(k) = 1/\exp(k)$ have similar converging patterns and quickly approach the "optimal" bottom. The convergence speed of $\alpha(k) = 1/\sqrt{k}$ is a bit slower, but it has a steep slope before 100 iterations as well. $\alpha(k) = 1/\sqrt{k}$ does not have a fast improvement after 200 iterations, while $\alpha(k) = 1/1.05^k$ has an approximately uniform convergence speed. However, $\alpha(k) = 1/\sqrt{k}$ is still far away from the "optimal" bottom. In short, $\alpha(k) = 1/2^k$ and $\alpha(k) = 1/\exp(k)$ provide uniformly good solutions which would be more efficient when dealing with big data.

3 *p*-median problem

An important application of the subgradient method is solving p-median problems. Here, the p-median problem is formulated by integer linear programming. It is defined as follows.

$$\text{Minimize:} \quad \sum_i \sum_j h_i d_{ij} Y_{ij} \tag{3.1}$$

subject to:

$$\sum_j Y_{ij} = 1 \quad \forall i \tag{3.2}$$

$$\sum_j X_j = P \tag{3.3}$$

$$Y_{ij} - X_j \leqslant 0 \quad \forall i, j \tag{3.4}$$

$$X_j = 0, 1 \quad \forall j \tag{3.5}$$

$$Y_{i,j} = 0, 1 \quad \forall i, j \tag{3.6}$$

In (3.1), h_i is the weight on each demand point and d_{ij} is the cost of the edge. Y_{ij} is the decision variable indicating whether if a trip between node i and j is made or not. Constraint (3.2) ensures that every demand point must be assigned to one facility. In (3.3) X_j is a decision variable and it ensures that the number of facilities to be located is P. Constraint (3.4) indicates that no demand point i is assigned to j unless there is a facility. In constraint (3.5) and (3.6) the value 1 means that the locating (X) or travelling (Y) decision is made. 0 means that the decision is not made.

To solve this problem using the sugradient method, the Lagrangian relaxation must be made. Since the number of facilities, P, is fixed, we cannot relax the locating decision variable X_j. Consequently, the relaxation is necessarily put on the travelling decision variable Y_{ij}. It could be made either on (3.2) or on (3.4). In this paper, we only consider the case for (3.2), because the same procedure would be applied on (3.4). We do not repeat this for (3.4). What we need to do is to relax this problem for fixed values of the Lagrange multipliers, find primal feasible solutions from the relaxed solution and improve the Lagrange multipliers (Daskin, 1995). Consider relaxing constraint (3.2), we have

$$\begin{aligned} \text{Minimize:} \quad & \sum_i \sum_j h_i d_{ij} Y_{ij} + \sum_i \lambda_i (1 - \sum_j Y_{ij}) \\ = & \sum_i \sum_j (h_i d_{ij} - \lambda_i) Y_{ij} + \sum_i \lambda_i \end{aligned} \tag{3.7}$$

with constraints (3.3)–(3.6) unchanged. In order to minimize the objective function for fixed values of λ_i, we set $Y_{ij} = 1$ when $h_i d_{ij} - \lambda_i < 0$ and $Y_{ij} = 0$ otherwise. The corresponding value of X_j is 1. A set of initial values of λ_is are given by the mean weighted distance between each node and the demand points.

Lagrange multipliers are updated in each iteration. The step size value in the kth iteration for the multipliers $T^{(k)}$ is :

$$T^{(k)} = \frac{\alpha^{(k)}(BUB - \mathcal{L}^{(k)})}{\sum_i \{\sum_j Y_{ij}^{(k)} - 1\}^2}, \tag{3.8}$$

where $T^{(k)}$ is the kth step size value; BUB is the minimum upper bound of the objective function until the kth iteration; $\mathcal{L}$ is the value evaluated by (3.7) at the kth iteration; $\sum_j Y_{ij}^{(k)}$ is the current optimal value of the decision variable. The Lagrangian multipliers, λ_i, are updated by:

$$\lambda_i^{(k+1)} = \max\{0, \lambda_i^{(k)} - T^{(k)}(\sum_j Y_{ij}^{(k)} - 1)\}. \tag{3.9}$$

A general working scheme is:

- step 1 Plug the initial values or updated values of λ_i into (3.7) and identify the p medians according to $h_i d_{ij} - \lambda_i$;
- step 2 According to p medians in step 1, evaluate the subgradient $g^{(k)} = 1 - \sum_j Y_{ij}$, BUB and $\mathcal{L}^{(k)}$. If the stopping criteria is met, stop. Otherwise, go to step 3;
- step 3 Update $T^{(k)}$ using (3.8);
- step 4 Update Lagrangian multipliers $\lambda_i^{(k)}$s using (3.9). Then go to step 1 with new λ_is.

The lower bound (LB) in each iteration is decided by the value of λ_is. The step size $T^{(k)}$ will affect the update speed of λ_is. It goes to 0 when the number of iterations tends to infinity. When it goes slowly, the increment of LB would be fast but unstable. This leads to inaccurate estimates of the LB. On the other hand, when the update speed goes too fast, the update of LB is slow. The non-update would easily happen such that the difference between BUB and BLB remains even though more iterations are made. The danger will arise if the inappropriate step size is computed. Thus, a good choice of executed parameter controlling the update speed would make the algorithm more efficient.

4 Computational results

In this section, we study the parameter, α, controlling the step size. Daskin (1995) suggested an initial value of 2 and a halved decreasing factor after 5 failures of changing; Avella (2012) suggested an initial value of 1.5 and a 1.01 decreasing factor after one failure of changing. We could also consider other alternative initial values instead of those in the previous studies. However, that is only a minor issue and not related to the step size function. Thus, we skip the analysis of the initial values.

The complexity in our study is different from Daskin (1995) and Avella (2012). We take medium sized problems from the OR-library (Beasley, 1990). The OR-library consists of 40 test p-median problems. The optimal solutions are given. We pick eight cases. N varies from 100 to 900 and P varies from 5 to 80. A subset is picked in our study by only selecting two cases for each $N = 100, 200, 400, 800$. The parameter α take five forms. Following

Table 1: Lagrangian settings testing a subset of OR-library

$\alpha(k)$	form 1: ξ/k form 2: $\xi/\sqrt{k}$ form 3: $\xi/1.05^k$ form 4: $\xi/2^k$ form 5: $\xi/\exp(k)$
n (number of failures before changing α)	5
restart the counter when α changed	Yes
critical difference	0.01
initial values of λ_is	$\sum_j h_i d_{ij} Y_{ij} / \sum j$
maximum iterations after no improvement on BUB	$m = 1000$ and $m = 100$

Stephen and Almir (2008), we take the forms of $\alpha(k) = \xi/k$, $\alpha(k) = \xi/\sqrt{k}$, $\alpha(k) = \xi/1.05^k$, $\alpha(k) = \xi/2^k$ and $\alpha(k) = \xi/\exp(k)$.The procedure settings are shown in Table 1.

In Table 1, $\alpha(k)$ is the step size function of. We take $\xi = 1$ as we did for the piecewise function $f(x)$. n is a counter recording the number of 5 consecutive failures. As suggested by Daskin (1995), we do not further elaborate the impact of the counter. The critical difference takes the value of 1% of the optimal solution. This is only a criterion for known optimal values and it can be largely affected by the type of the problem. Considering that, the algorithm also stops if no improvement of BUB is found after preset number of iterations. Here we compare 100 and 1,000. Given the settings, the results are shown in Table 2 and Table 3.

In Table 2 and Table 3, optimal solution values are given for two stopping criteria. The problem complexity varies. We compare different forms of $\alpha(k)$. BLBs (best lower bound), BUBs (best upper bound), deviations ($\frac{\text{BUB}-\text{Optimal}}{\text{Optimal}} \times 100\%$), U/L ($\frac{\text{BUB}}{\text{BLB}}$) and the number of iterations. The optimal BUB and U/L are marked in bold.

Table 2 shows the solutions for $m = 100$. For pmed 1 and pmed 6 of the OR-library, the exact optimal solutions are obtained. For pemd 35, an almost exact solution is also obtained. On the other hand, for pmed 4, pmed 9, pmed 18 and pmed 37, the BLB is much closer to the optimal. For most of the cases, the step size function with the minimum U/L ratio gives the lowest BUB. It is an indication of the good quality of the algorithm even though $1/1.05^k$ performs very badly in pmed 18 and pmed 37. It is no surprise that more exact solutions appear when the number of iterations is increased, for example, $1/1.05^k$ in pmed 1 and pmed 6; $1/k$ in pmed 4 and pmed 35 in Table 3. Similarly, we also improve the quality of BLBs. The worst deviation is 17.70 for $m = 1000$ instead of 44.14 for $m = 100$.

There are several overall tendencies we can draw from Table 2 and Table 3. Firstly, $1/2^k$ and $1/\exp(k)$ are relatively stable which is also in accordance with piecewise function we studied before. This can be seen not only for less complicated problem but also for the complicated case. However, there is no obvious tendency of which one will dominate. Secondly, it is difficult for $1/\sqrt{k}$ to perform better that the rest of 4 forms to have an optimal BUB, which is

Table 2: Comparison of optimal solutions for different step size decreasing speed ($m = 100$)

File No.	$f_n(\alpha)$	BLB	BUB	Optimal	Deviation (%)	U/L	Iterations
pmed 1 ($N = 100$ $P = 5$)	$1/k$	5803	5821	5819	0.03	1.003	65
	$1/\sqrt{k}$	5811	5821	5819	0.03	**1.002**	98
	$1/1.05^k$	5521	6455	5819	10.93	1.169	103
	$1/2^k$	5796	**5819**	5819	0.00	1.005	46
	$1/exp(k)$	5796	5821	5819	0.03	1.005	53
pmed 4 ($N = 100$ $P = 20$)	$1/k$	3032	3265	3034	7.61	1.077	300
	$1/\sqrt{k}$	3030	3297	3034	8.67	1.088	435
	$1/1.05^k$	3034	**3182**	3034	4.88	**1.049**	535
	$1/2^k$	3034	**3182**	3034	4.88	**1.049**	249
	$1/exp(k)$	3034	**3182**	3034	4.88	**1.049**	164
pmed 6 ($N = 200$ $P = 5$)	$1/k$	7770	8238	7824	5.29	1.060	143
	$1/\sqrt{k}$	7760	8195	7824	4.74	1.056	202
	$1/1.05^k$	7459	8948	7824	14.37	1.200	153
	$1/2^k$	7753	**7824**	7824	0.00	**1.009**	145
	$1/exp(k)$	7751	**7824**	7824	0.00	**1.009**	56
pmed 9 ($N = 200$ $P = 40$)	$1/k$	2732	**3051**	2734	11.59	**1.117**	471
	$1/\sqrt{k}$	2719	3264	2734	20.04	1.200	386
	$1/1.05^k$	2725	3239	2734	18.47	1.189	451
	$1/2^k$	2732	3069	2734	12.25	1.123	282
	$1/exp(k)$	2732	3127	2734	14.37	1.145	297
pmed 16 ($N = 400$ $P = 5$)	$1/k$	8090	8253	8162	1.411	1.020	231
	$1/\sqrt{k}$	8086	8240	8162	0.96	1.019	261
	$1/1.05^k$	8092	**8185**	8162	0.28	**1.011**	534
	$1/2^k$	8088	8239	8162	0.94	1.019	210
	$1/exp(k)$	8080	8206	8162	0.54	1.016	156
pmed 18 ($N = 400$ $P = 40$)	$1/k$	4807	5021	4809	4.22	1.043	256
	$1/\sqrt{k}$	4801	5150	4809	7.09	1.073	516
	$1/1.05^k$	3848	6913	4809	43.75	1.797	101
	$1/2^k$	4805	**4865**	4809	1.16	**1.012**	216
	$1/exp(k)$	4803	4902	4809	1.93	1.021	269
pmed 35 ($N = 800$ $P = 5$)	$1/k$	10288	10504	10400	0.01	1.021	124
	$1/\sqrt{k}$	10296	**10401**	10400	0.01	**1.010**	254
	$1/1.05^k$	10183	10710	10400	2.98	1.052	144
	$1/2^k$	10286	**10401**	10400	0.01	1.011	201
	$1/exp(k)$	10282	**10401**	10400	0.01	1.012	239
pmed 37 ($N = 800$ $P = 80$)	$1/k$	5056	5248	5057	3.78	1.038	306
	$1/\sqrt{k}$	5033	5577	5057	10.28	1.108	342
	$1/1.05^k$	3820	7289	5057	44.14	1.908	101
	$1/2^k$	5055	5137	5057	1.58	1.016	314
	$1/exp(k)$	5051	**5100**	5057	0.85	**1.010**	161

Table 3: Comparison of optimal solutions for different step size decreasing speed ($m = 1000$)

File No.	$\alpha(k)$	BLB	BUB	Optimal	Deviation (%)	U/L	Iterations
pmed 1 ($N = 100$ $P = 5$)	$1/k$	5804	5821	5819	0.03	1.003	65
	$1/\sqrt{k}$	5811	5821	5819	0.03	1.002	98
	$1/1.05^k$	5815	**5819**	5819	0.00	**1.001**	239
	$1/2^k$	5796	**5819**	5819	0.00	1.004	46
	$1/exp(k)$	5796	5821	5819	0.03	1.004	53
pmed 4 ($N = 100$ $P = 20$)	$1/k$	3034	**3182**	3034	4.88	**1.049**	1975
	$1/\sqrt{k}$	3031	3259	3034	7.42	1.075	1580
	$1/1.05^k$	3034	**3182**	3034	4.88	**1.049**	1435
	$1/2^k$	3034	**3182**	3034	4.88	**1.049**	1162
	$1/exp(k)$	3034	**3182**	3034	4.88	**1.049**	1064
pmed 6 ($N = 200$ $P = 5$)	$1/k$	7782	8086	7824	3.35	1.039	1417
	$1/\sqrt{k}$	7783	7867	7824	0.66	1.011	1853
	$1/1.05^k$	7783	**7824**	7824	0.00	**1.005**	698
	$1/2^k$	7753	**7824**	7824	0.00	1.009	145
	$1/exp(k)$	7751	**7824**	7824	0.00	1.009	56
pmed 9 ($N = 200$ $P = 40$)	$1/k$	2733	**3051**	2734	11.59	**1.116**	1371
	$1/\sqrt{k}$	2720	3217	2734	17.70	1.183	1400
	$1/1.05^k$	2734	3098	2734	13.31	1.133	1674
	$1/2^k$	2732	3069	2734	12.25	1.123	1182
	$1/exp(k)$	2732	3073	2734	12.40	1.125	1359
pmed 16 ($N = 400$ $P = 5$)	$1/k$	8091	8219	8162	0.70	1.016	1685
	$1/\sqrt{k}$	8088	8240	8162	0.96	1.019	1161
	$1/1.05^k$	8092	**8162**	8162	0.00	**1.009**	859
	$1/2^k$	8088	8183	8162	0.26	1.012	1433
	$1/exp(k)$	8080	8206	8162	0.54	1.016	1056
pmed 18 ($N = 400$ $P = 40$)	$1/k$	4808	4943	4809	2.79	1.028	1499
	$1/\sqrt{k}$	4807	4957	4809	3.08	1.031	3453
	$1/1.05^k$	4809	4894	4809	1.77	1.018	2707
	$1/2^k$	4805	**4841**	4809	0.67	**1.007**	314
	$1/exp(k)$	4803	4877	4809	1.41	1.015	1726
pmed 35 ($N = 800$ $P = 5$)	$1/k$	10297	**10401**	10400	0.01	**1.010**	1453
	$1/\sqrt{k}$	10297	**10401**	10400	0.01	**1.010**	348
	$\alpha/1.05^k$	10302	10481	10400	0.78	1.017	1696
	$\alpha/2^k$	10286	**10401**	10400	0.01	1.011	1011
	$\alpha/exp(k)$	10282	**10401**	10400	0.01	1.012	1139
pmed 37 ($N = 800$ $P = 80$)	$1/k$	5057	5124	5057	1.32	1.013	1779
	$1/\sqrt{k}$	5056	5201	5057	2.85	1.029	3281
	$1/1.05^k$	5057	5140	5057	1.64	1.016	2159
	$1/2^k$	5055	5123	5057	1.31	1.013	2009
	$1/exp(k)$	5051	**5100**	5057	0.85	**1.010**	161

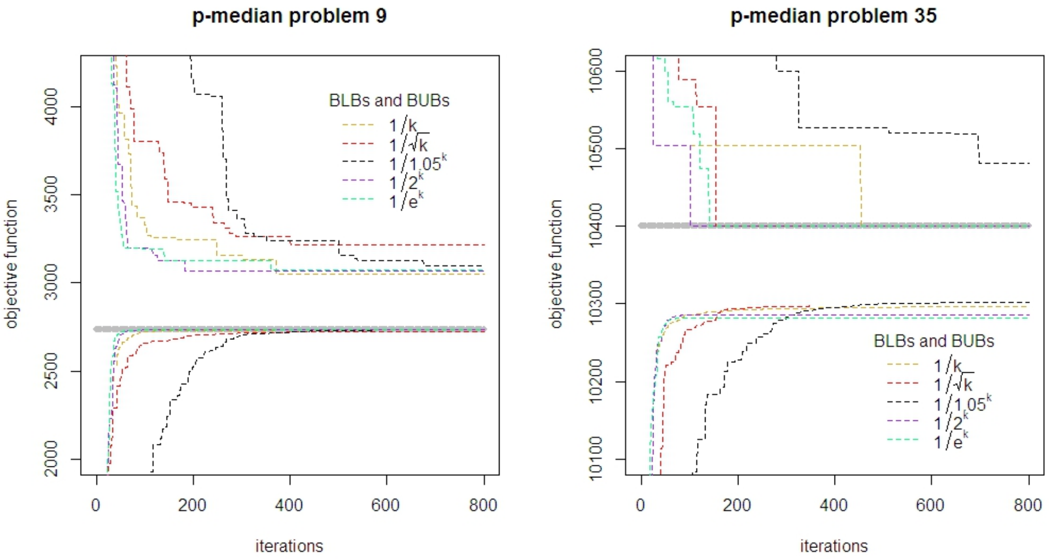


Figure 2: Changes for BLB and BUB for file No.9 and No.35

in accordance with the piecewise function. One reason is that when the number of iterations is large, a slightly short step size is required. Too large steps can bring infeasible solutions, which to some extent enlarge the gaps between BLBs and BUBs. Thirdly, $1/k$ and $1/1.05^k$ are too sensitive to the stopping criterion, which is not seen in the general piece wise function. The decision to stop the algorithm should be very carefully made. One suggested way is to visualize the convergence curve and to terminate the iteration when the curve becomes flat.

Generally speaking, the BLB and the BUB tend to complement each other. In other words, one can always make an inference that either the BLB or the BUB would be the benchmark when there is a gap between BLB and BUB. In Figure 2, for example, two extreme cases are shown. The grey line represents the optimal value. The left panel shows the first 800 objective values for five forms of step size functions in problem 9 (pmed 9). The right one shows the values in problem 35 (pmed 35). For pmed 9, the BLBs quickly converge to the optimal. However, only sub-optimal BUBs are obtained. On the contrary, pmed 35 has good BUBs and bad BLBs. Thus, either the BLB or BUB is likely to reach the sub-optimal. When this happens, a complement algorithm could be involved to improve the solution.

5 Conclusion

In this paper, we studied how the decreasing speed of step size in the subgradient optimization method affects the performance of the convergence. The subgradient optimization method is simpler in solving linear programming. However, the choice of the step function in the subgradient equation can bring uncertainties to the solution. Thus, we conduct our study by examining how the step size function parameter α affects the performance. Both a general piecewise function and a specific p-median problem are studied. The p-median problem is represented by linear programming and the corresponding Lagragian relaxation is added.

We examined five forms of the step size parameters α. One is the square summable but not summable form $\alpha(k) = \xi/(b+k)$. One is the nonsummable diminishing form $\alpha(k) = \xi/\sqrt{k}$. Three are nonsummable diminishing step length forms $\alpha(k) = \xi/1.05^k$, $\alpha(k) = \xi/2^k$ and $\alpha(k) = \xi/\exp(k)$. We evaluated the best upper bound, best lower bound, and the required iterations to reach our stopping criteria. We have the following conclusions.

Firstly, the nonsummable diminishing step size function $\alpha(k) = \xi/\sqrt{k}$ has its limitation when the number of iterations are large. For both the general piecewise function and the p-median problem, the nonsummable diminishing step size function performs badly and easily goes into the suboptimal solution. Two nonsummable diminishing step length functions $\alpha(k) = \xi/2^k$ and $\alpha(k) = \xi/\exp(k)$ have similar behaviors and stable solutions. As long as the problem is not likely to lead to the suboptimal solutions, step size functions $\alpha(k) = \xi/2^k$ and $\alpha(k) = \xi/\exp(k)$ always give fast convergence for both BLB and BUB. This is found both in general piecewise function and p-median problems. The square summable but not summable form $\alpha(k) = \xi/(b+k)$ as well as nonsummable diminishing form $\alpha(k) = \xi/1.05^k$ are unstable. They are also sensitive to the number of iterations.

Secondly, from our empirical result, the quality of the solution will be largely affected by the specific type of the problem. The problem characteristic may have influence on the difficulties of avoiding suboptimal solutions. If it is easy to avoid suboptimal solutions for a specific step size function α, one can make a good inference. On the other hand, if the subgradient method can always produce the suboptimal solution, a complement algorithm can be considered to get out from the suboptimal.

Thirdly, the problem complexity has little impact. We cannot assert that good solutions can be found for a less complex problem and bad solutions for a complex solution for a subgradient method.

References

[1] Avella, P., Boccia, M., Salerno, S. and Vasilyev, I., 2012. An aggregation heuristic for large p-median problem, *Computers and Operations Research*, 32, 1625–1632.

[2] Beasley, J.E., 1990. OR Library: distributing test problems by electronic mail, *Journal of the Operational Research*, 41(11), 1069-1072.

[3] Beasley, J.E., 1993. Lagrangian heuristics for location problems, *European Journal of Operational Research*, 65, 383–399.

[4] Beltran, C., Tadonki, C. and Vial, J.-Ph., 2004, Solving the p-median problem with a semi-lagrangian relaxation, *Logilab* Report, HEC, University of Geneva, Switzerland.

[5] Carrizosa, E., Ushakov, A. and Vasilyev, I., 2012. A computational study of a nonlinear minsum facility location problem, *Computers and Operations Research*, 32, 2625–2633.

[6] Daskin, M., 1995, *Network and discrete location*, Wiley, New York.

[7] Kiwiel, K.C., 1985, Methods of decent for nondifferentiable optimization. Springer Verlag, Berlin.

[8] Mladenović, N., Brimberg, J., Hansen, P. and Moreno-Pérez, JA., 2007. The p-median problem: a survey of mataheuristic approaches, *European Journal of Operational Research*, 179(3), 927–939.

[9] Reese, J., 2006. Solution methods for the p-median problem: an annotated bibliography, *Networks*, 48(3), 125–142.

[10] Shor, N.Z., 1985, Minimization methods for non-differentiable functions. Springer Verlag, New York.

[11] Stephen, B. and Almir, M., 2008, Notes for EE364b, Stanford University, Winter 2006–2007.