Proceedings of the 2012 Winter Simulation Conference
C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, eds.

HYPERCUBE SIMULATION ANALYSIS FOR A LARGE-SCALE AMBULANCE SERVICE SYSTEM

Hozumi Morohosi

National Graduate Institute for Policy Studies
7-22-1 Roppongi, Minato-ku
Tokyo, 106-8677, JAPAN

Takehiro Furuta

Nara University of Education
Takabatake-cho, Nara-shi
Nara, 630-8528, JAPAN

ABSTRACT

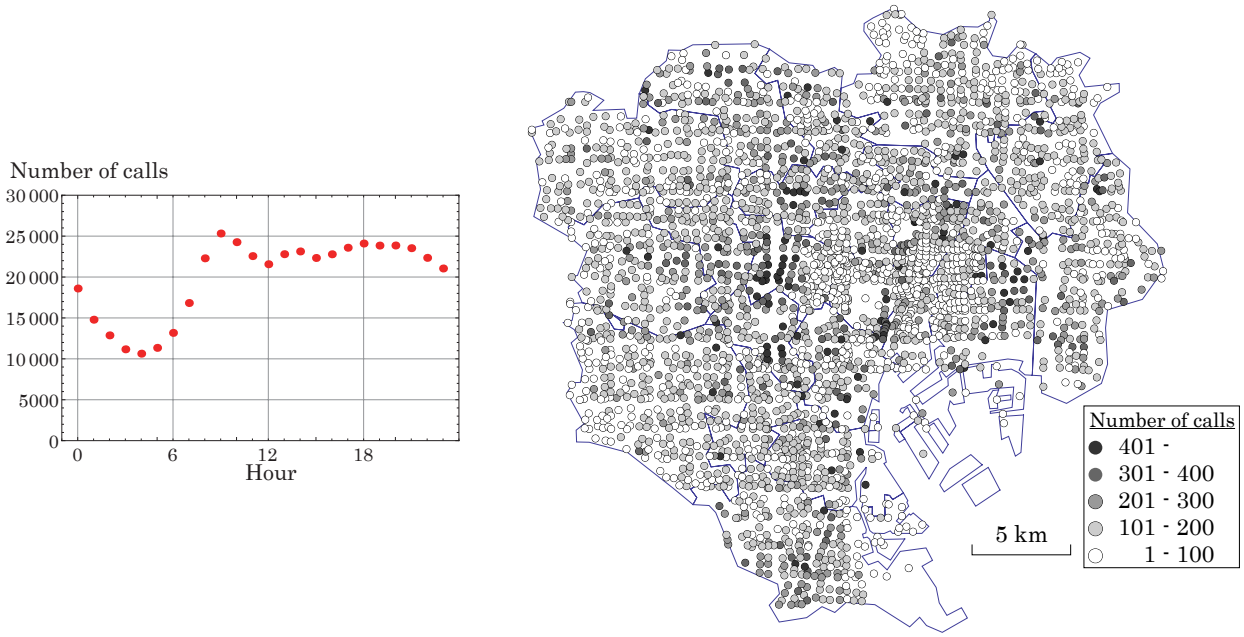
A simple yet powerful simulation model for an ambulance service system is devised and applied to a large-scale ambulance system in the Tokyo metropolis. Our simulation can provide useful measures for the location analysis of ambulance stations, which are rarely incorporated in traditional optimal location models, although they are seemingly very important for designing a reliable and efficient emergency system. Comparing simulation output and actual data enables us to investigate the present ambulance system as well as check the validity and limitations of our model. Then we use the model to evaluate solutions to the ambulance station location problem.

1 INTRODUCTION

This work is an attempt to simulate a large-scale ambulance service system. Methodologically, it is a simple and straightforward application of the hypercube model introduced in Larson (1974), which is well-known in ambulance location-decision problems. The hypercube model is a spatial queueing model by which we can obtain an equilibrium equation to calculate the steady-state probability of an ambulance system, see, e.g., Morabito, Chiyoshi, and Galvao (2008), Takeda, Widmer, and Morabito (2007), Mendonça and Morabito (2001) for recent applications. Several features of a system can be obtained by using the steady-state probability. Nevertheless, if the system is very large, which is the case we are interested in, the hypercube model is difficult to solve by deterministic, i.e., numerical linear algebraic methods. Instead, we construct a random-walk simulation on the hypercube, which gives us flexible modeling of the system and can deal with the nonhomogeneous process in space and time. A typical feature of ambulance call data in large-scale cities, such as Tokyo, is its space and time heterogeneity, see Figure 1. Figure 1(a) shows the number of yearly calls for every hour of the day. We know that ambulance calls vary considerably from hour to hour. At the same time, Figure 1(b) the right panel depicts the number of calls for each town block in Tokyo metropolis by disks colored according to the number of calls and centered at the representative point of each town block. It shows high concentration in an specific district. We incorporated all these data into our simulation and built a hypercube model for an ambulance service system.

Our main concern in this study is locational analysis of ambulance stations. Since ambulances are a scarce resource, their station location has to be determined with careful consideration of efficiency and reliability. Operations researchers have worked on this problem for a long time; in fact, the ambulance location problem has extensive literature, especially in facility-location theory, see, e.g., Brotcorne, Laporte, and Semet (2003), and Li, Zhao, Zhu, and Wyatt (2011). These studies often incorporate the stochastic properties of an ambulance system into the model, such as call arrival, ambulance traveling time, and so on. Most models utilize a stochastic programming framework, and calculate, for instance, loss probability, average waiting time, etc, in view of traditional queueing theory, to define objective function and constraints (Berman and Krass 2002; Marianov and Serra 2002), whereas, curiously enough, simulating an ambulance service system has seemingly received still less attention. Since a simulation works flexibly enough to

978-1-4673-4781-5/12/$31.00 ©2012 IEEE



(a) (b)

Figure 1: Ambulance calls in Tokyo metropolis.

compute various kinds of system features, it enables us to investigate not only those traditional measures of a queueing system, but also microstructures of the system, for instance, the count of dispatches of each ambulance to each demand point. Once such simulation output is available, we can understand the service quality of the ambulance system in detail. For instance, the higher the response rate of the nearest ambulance to the call demand is, the better the ambulance system is. Our simulation can be used to analyze the details of the system and improve it, hopefully.

This work concentrates on a comparative case study of actual ambulance location and an optimal solution to the traditional location problem, which we computed in previous work (Morohosi 2008; Furuta and Morohosi 2011). Our aim is twofold: first, we analyze the dispatch pattern of a real ambulance system and simulation output by comparison. Based on those observations, second, we try to estimate the improvement from an optimal solution location.

In Section 2 we describe our simulation model and briefly introduce a location model which provides an alternative ambulance-location plan. We apply our simulation to Tokyo metropolis data in Section 3 and report findings obtained by comparison. Section 4 includes some remarks and refers to a possible extension of our work.

2 METHODOLOGY

2.1 Hypercube Model Simulation

We developed a simulation incorporating time nonhomogeneity into a hypercube model, which is often used in ambulance system analysis to calculate the stationary distribution of system variables. While it provides stationary distribution by formulating and solving the equilibrium state equation in its original formulation, since our target real-world system is too huge to be solved by simultaneous equation formulation and is largely time-dependent, we analyze the system by running a computer simulation directly. The details of the simulation components are as follows.

Let I be the set of ambulances and J be the set of demand points. Assume the sizes of I and J are $|I| = n$ and $|J| = m$, respectively. Each demand point $j \in J$ has call rate $\lambda_j(t)$, which is the number of

calls occurring at j per unit time. Looking at the detailed records of ambulance calls we investigated, we realize that they clearly depend on the time of day. Attached to each point j also is the priority list $(i_{j1}, i_{j2}, \ldots, i_{jn})$ of ambulances for it, where $i_{jl} \in I$ is the l-th highest priority ambulance to demand point j. The last parameter of the system is the average service time of ambulance $1/\mu$, which is the time interval from leaving the station to returning. Unlike the call rate, service time does not show time dependence clearly, and it is assumed to be constant for a whole day in our simulation.

A hypercube model is a Markov chain simulation on a high dimensional unit cube. Its state space consists of n-dimensional 0-1 vector $\mathbf{S} = (S_1, \ldots, S_n)$, where S_i represents the state of ambulance: and $S_i = 1$ when ambulance i has been dispatched and busy, whereas $S_i = 0$ when ambulance i is on stand-by at the station. A state $\mathbf{S}_k = (S_{k1}, \ldots, S_{kn})$ at time step k, is identified as one vertex of an n-dimensional unit cube, hence the name. Given the current state $\mathbf{S}_t$, there are two possible kinds of state transitions:

(i) A dispatched ambulance i finishes taking a patient to the hospital and returns to the home station. The consequent state transition is given by $S_{k+1,i} = 0$, and $S_{k+1,i'} = S_{k,i'}$, $i' \neq i$.

(ii) A new call arrives and the ambulance i nearest to the patient is dispatched. The state transition is $S_{k+1,i} = 1$, and $S_{k+1,i'} = S_{k,i'}$, $i' \neq i$.

Each probability of transition is given in terms of parameters μ and $\lambda_j(t)$. Let $\Lambda(t) = \sum_j \lambda_j(t)$ and

$$p = \mu \sum_i S_{ki} / (\mu \sum_i S_{ki} + \Lambda(t)), \tag{1}$$

which is the probability that one of the busy ambulances finishes transportation and returns to its home station.

Transition (i) occurs with probability p, and every ambulance can be chosen with equal probability, i.e., with probability $1/(\sum_i S_{ki})$. Or, if transition (ii) occurs with probability $1-p$, and a new call demand point j is chosen with probability $\lambda_j(t)/\Lambda(t)$, then the nearest stand-by ambulance i is dispatched to point j.

Another important ingredient of the simulation is timekeeping to know the current time of the simulation model. We advance the simulation time when an ambulance call happens. Since we assume the time intervals D between consecutive call arrivals are independent, exponentially distributed with parameter $\Lambda(t)$, we use a nonhomogeneous Poisson point process technique for generating them, cf. Fishman (2001). In summary, the algorithm of the simulation is described as follows.

Ambulance simulation algorithm

Input: I: ambulance stations, J: demand points, λ_j: call arrival rate at $j \in J$, μ: service rate, $(i_{j1}, \ldots, i_{jn})$: priority list of ambulances for j. T: total time of simulation.

Output: Some features $Y(\cdot)$ of system calculated from $\{\mathbf{S}_k\}$.

Variables: $\mathbf{S}_k$: state vector. t: simulation clock time. D: interval time of consecutive call arrivals.

Method:

Set $t := 0$. $k := 0$

Give initial state $\mathbf{S}_0$.

Repeat:

Calculate p by using Eq. (1).

if $U \leq p$ then [return case]

Choose one ambulance i from busy ones $\{i : S_{ki} = 1\}$ with equal probability, set $S_{k+1,i} := 0$.

else [dispatch case]

Choose one demand point j with probability $\lambda_j(t)/\Lambda(t)$, then choose the nearest stand-by ambulance i_{jl} to j such that $S_{k,i_{j1}} = 1, \ldots, S_{k,i_{j,l-1}} = 1, S_{k,i_{jl}} = 0$. If such i_{jl} is not found, continue to next repeat [call loss].

Generate exponential random variable D with parameter $\Lambda(t)$, then advance the time as $t := t + D$.

Set $S_{k+1,i_{jl}} := 1$.
Set $k := k+1$.
Until $t \geq T$

One key feature of our simulation is introducing a priority list of ambulances for each demand point, by which the simulation works very quickly. The choice of a priority list greatly affects the simulation result. We try a few possible choices of list in the numerical studies in Section 3.

2.2 Location Model

There are diverse studies on the ambulance station location problem, and several optimal location models have been proposed. Among them, we focus on a queueing-theory-based one: the maximum expected covering location problem (MEXCLP) (Daskin 1983), which takes account of ambulance congestion and the probability of their absence. The objective of MEXCLP is to maximize the total expectation of covered ambulance demand. A fundamental assumption of MEXCLP is that ambulances operate independently from one another, and busy probability is identical for all ambulances. This assumption is a necessary simplification for building an integer programming model for ambulance location decisions, while the hypercube model simulation considers interactions among ambulances and is seemingly closer to reality. In another study, we applied MEXCLP to Tokyo metropolis data, with constant arrival rate, to find optimal locations of ambulance stations (Furuta and Morohosi 2011).

The problem description is as follows. Let K be the total number of ambulances to locate and L be the set of potential stations. Total call demand at $j \in J$ is denoted by a_j. The busy probability of an ambulance is obtained as $q = \frac{1}{\mu K} \int \Lambda(t)\,\mathrm{d}t$. We introduce two kinds of decision variable, x_i, $i \in L$, which is equal to the number of ambulances to locate at i, and y_{jk}, $j \in J$, $1 \leq k \leq K$, which is equal to 1 if the point j is covered by ambulances of more than or equal to k, and 0 otherwise. Denote the set of potential stations covering demand point j by $N_j = \{i \in L, d_{ij} \leq R\}$, where d_{ij} is the distance between i and j, and R is the distance to be determined as a coverage standard. Setting $R = 1500$ in our computation, we solve following MEXCLP:

$$
\begin{aligned}
\text{max.} \quad & C = \sum_{j \in J} \sum_{k=1}^{K} a_j (1-q) q^{k-1} y_{jk} \\
\text{s.t.} \quad & \sum_{i \in N_j} x_i \geq \sum_{k=1}^{K} y_{jk}, \quad j \in J, \\
& \sum_{i \in L} x_i = K, \\
& x_i : \text{ integer}, \quad y_{jk} \in \{0,1\}.
\end{aligned}
\tag{2}
$$

We are interested in knowing whether its solution shows good performance and in trying to check it by simulation.

3 NUMERICAL STUDIES

Using the hypercube simulation model, we studied two kinds of location analysis on ambulance stations in Tokyo. The Tokyo fire department operates 162 ambulance teams to meet the calls of more than four million people from 3110 town blocks, which are the demand points in our context, i.e., $|I| = 162$, $|J| = 3110$ in our notation. As shown in Figure 1, ambulance call arrivals vary from hour to hour during a day, so we set four time segments; 0:00–5:59, 6:00–11:59, 12:00–17:59, 18:00–23:59, and calculated the average arrival rate $\lambda_j(t)$ at demand point j for those time segments.

All the dispatch records, which contain the correspondence of an ambulance team to demand points with time and geographical data, are kept. Those data are the basis of our simulation. The priority list of

each demand point j is made from the actual count of dispatches so that the ambulance of first priority is the one with the most frequent dispatches to that demand point, and second priority is given to the second-most frequently dispatched ambulance, and so on.

There may be another possibility for defining the priority list by a road network distance d_{ij} between station i and point j, which is calculated using a geographical information system. In the distance-based priority list, the ambulance nearest to the demand point j in terms of the distance d_{ij} is chosen for the first priority, and the second nearest one for second priority, and so on. Nevertheless, we find the priority list based on distance does not give simulation output with good agreement to actual dispatch data. Our result seems interesting for understanding how the actual system is working, and the reason this disagreement is caused is probably a lack of tacit knowledge of detailed road traffic information, such as congestion, road width, and other conditions. Call center operators seem to decide which ambulance to dispatch after careful consideration based on their experience. Since the dispatch data are influenced by those actual situations and fit for our simulation use, the following computation uses the empirical priority list.

All the estimates are made using the algorithm in Section 2.1 with the batch-means method after a warm-up interval (Fishman 2001). First, we check our simulation model by comparing the dispatch data and simulation output. Second, we investigate the optimal solution in MEXCLP. The main measure of our comparison study is the response rate of the nearest ambulance to each demand point. We count the dispatch of each ambulance to demand points in the simulation and compare it with the number in the real data.

3.1 Comparison of Simulation Result and Actual Data

We compare the actual number of dispatches per year s_i for each ambulance $i \in I$, with simulation count $\bar{s}_i$. Given the fact that there are some stations equipped two ambulances in Tokyo area, since those ambulances have the same priority to any demand points and they are essentially indistinguishable, we calculate the sum of those two ambulance dispatches and depict it in Figure 2. Figure 2(a) shows the plot $(s_i, \bar{s}_i)$ of actual

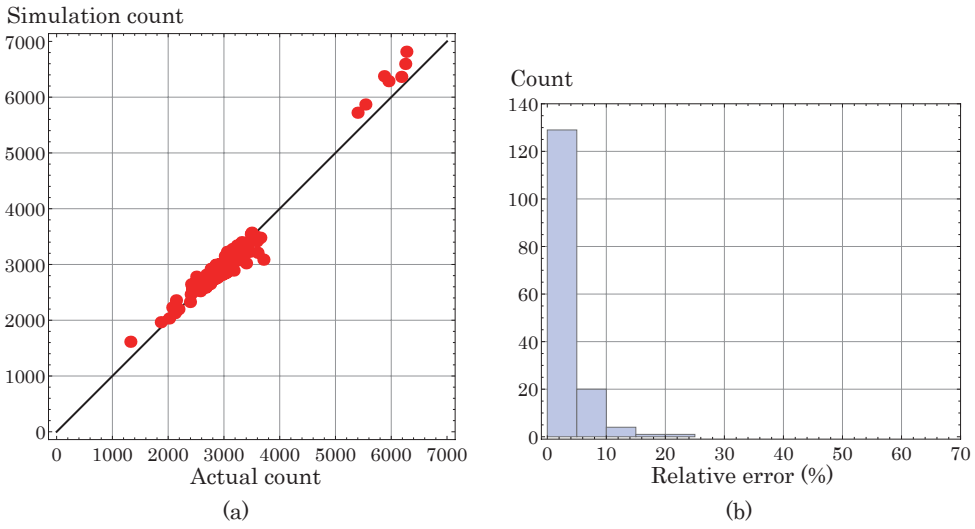


Figure 2: Distribution of relative error in simulation output relative to actual data.

and simulation counts for all the ambulances. Applying regression to the data, we have a good fit between them and $R^2 = 0.97$. Figure 2(b) shows the distribution of relative errors $|\bar{s}_i - s_i|/s_i$. The average of relative errors is 3.1% and the standard deviation is 3.0%. Among 162 ambulances there are 131 ambulances which have smaller than 5% relative error. Some outliers are found. To look at them in detail we plot the stations on the map in Figure 3 with their relative error by color. Most outliers with positive error are located

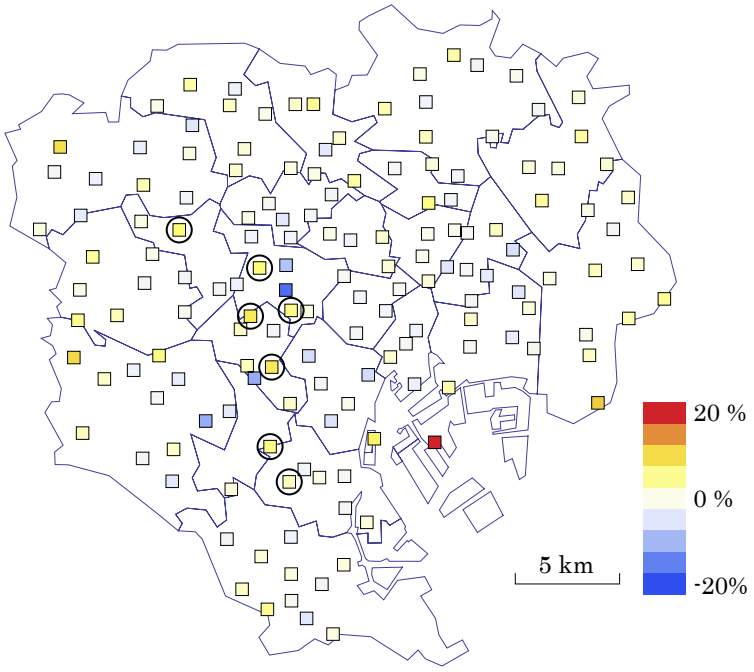


Figure 3: Ambulance stations with relative error of simulation output on Tokyo map. Stations with two ambulances are marked with circle.

at a marginal area of a region, are on islands, or they are stations with two ambulances. Geographical conditions seem to affect the simulation output, as can be readily understood. In fact, other ambulances are deployed outside of the region and can be dispatched to a call from this region, but those ambulances cannot be taken into consideration in our simulation; we cannot include such a marginal effect. On the other hand, the reason stations with two ambulances have more dispatch counts in simulation than in the actual data is not obvious. We keep in mind this bias in the following analysis.

3.2 Investigation on Location Problem Solution

We use the simulation output for the ambulance location obtained by solving MEXCLP to find the dispatch count for each ambulance. It is an *ex post facto* evaluation of the MEXCLP solution.

Let $\hat{y}_{jk}$ be the optimal solution of MEXCLP. The expected coverage probability of the demand point j can be obtained by $p_j = \sum_k (1-q)q^{k-1}\hat{y}_{jk}$ and the optimal value of MEXCLP is $C = \sum_j a_j p_j$. Running the algorithm in Section 2.1 with a small modification to count the correspondence $\bar{c}_{ij}$ of ambulance i to demand point j, we can obtain the estimate of expected coverage probability $\bar{p}_j$ of the demand j as

$$\bar{p}_j = \frac{\sum_{i \in N_j} \bar{c}_{ij}}{\sum_{i \in I} \bar{c}_{ij}}, \tag{3}$$

where N_j is the set of potential stations covering j defined in Section 2.2 with the coverage standard $R = 1500$, and the estimates of the objective function of MEXCLP by simulation is $\bar{C} = \sum_j a_j \bar{p}_j$. In our computation we have $C = 366993$, while $\bar{C} = 305029$: that is, our simulation estimates approximately 20% lower expected coverage than MEXCLP provides. After taking into account of the bias of our simulation, this gap in number is significant. The main reason for the difference is probably, as is often said, that the assumption of independent operation among ambulances is not adequate.

To look closely into the simulation output, we define the gap g_j as the difference between p_j and $\bar{p}_j$: $g_j = p_j - \bar{p}_j$, and show the distribution of g_j in Figure 4. Figure 4(a) shows the plot of gaps to the calls, i.e.,

(a_j, g_j), $j \in J$. Many demand points, especially those of large numbers of calls, have a positive gap value, which means they are less covered in the simulation than MEXCLP calculates. Geographical distribution is shown in the right panel, which maps the gaps g_j on the region. We can see that demand points with negative gaps, which means they are better covered in the simulation output, are often at marginal areas of the region. As stated in Section 3.1, ambulances at the marginal areas often have positive bias in dispatch count, which may affect this apparently better coverage. On the other hand, demand points in the central area often have positive gaps. Many of these points have a large number of calls and they are not as well covered as the MEXCLP solution predicts.

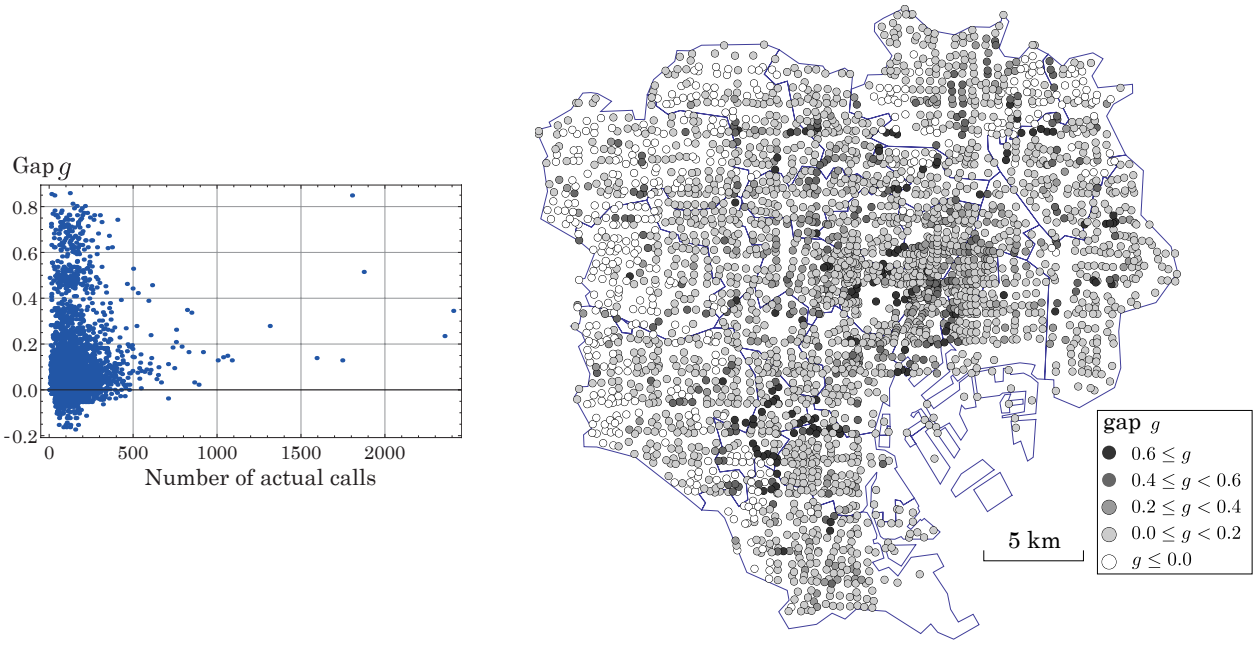


Figure 4: The difference of coverage probability for each town block. (a) plots gap against the number of calls. (b) shows gap distribution on the map.

4 CONCLUDING REMARKS

We introduced a simple hypercube simulation model for an ambulance system. The simulation output is compared to actual data and shows relatively good agreement. We used the model to evaluate the MEXCLP solution and found that the MEXCLP objective function over estimates coverage, although there are both demand points of over- and under-estimated coverage rate. Our model is simple and provides a quick simulation even for a large system like Tokyo, and our work could be a first step toward the optimization of ambulance location by simulation.

ACKNOWLEDGMENTS

This work is partially supported by Grant-in-Aid for Scientific Research 21510141, 22510169, 22710152, and 24241054.

REFERENCES

Berman, O., and D. Krass. 2002. “Faility location problems with stochastic demands and congestion”. In *Facility Locations: application and theory*, edited by Z. Drezner and H. W. Hamacher, 329–371. New York: Springer.

Brotcorne, L., G. Laporte, and F. Semet. 2003. "Ambulance location and relocation models". *European Journal of Operational Research* 147:451–463.

Daskin, M. S. 1983. "A maximum expected locatin model: formulation, properties and heuristic solutions". *Transportation Science* 17:48–70.

Fishman, G. S. 2001. *Discrete-Event Simulation*. New York: Springer.

Furuta, T., and H. Morohosi. 2011. "Applying covering models to ambulance system of megalopolitan area in Japan". *manuscript*.

Larson, R. C. 1974. "A hypercube queueing model for facility location and redistricting in urban emergency services". *Computers and Operations Research* 1:67–95.

Li, X., Z. Zhao, X. Zhu, and T. Wyatt. 2011. "Covering models and optimization techniques for emergency response facility location and planning: a review". *Mathematical Methods of Operations Research* 74:281–3109.

Marianov, V., and D. Serra. 2002. "Location problems in the public sector". In *Facility Locations: application and theory*, edited by Z. Drezner and H. W. Hamacher, 119–150. New York: Springer.

Mendonça, F. C., and R. Morabito. 2001. "Analysing emergency medical service ambulance deployment on a Brazilian highway using the hypercube model". *Journal of the Operational Research Society* 52:261–270.

Morabito, R., F. Chiyoshi, and R. D. Galvao. 2008. "Non-homogeneous servers in emergency medical systems: Practical applications using the hypercube queueing model". *Socio-Economic Planning Sciences* 42:255–270.

Morohosi, H. 2008. "A case study of optimal ambulance location problems". In *Proceedings of the 7th International Symposium on Operations Research and Its Applications*, edited by X.-S. Zhang, D.-G. Liu, and Y. Wang, 125–130. Lijiang, China: World Publishing Corporation.

Takeda, R. A., J. A. Widmer, and R. Morabito. 2007. "Analysis of ambulance decentralization in an urban emergency medical service using the hypercube queueing model". *Computers and Operations Research* 34:727–741.

AUTHOR BIOGRAPHIES

HOZUMI MOROHOSI is a professor at National Graduate Institute for Policy Studies, Tokyo, Japan. He received a Ph.D. from University of Tokyo in Engineering. His email address is morohosi@grips.ac.jp.

TAKEHIRO FURUTA is an associate professor at Nara University of Education, Nara, Japan. His email address is takef@fw.ipsj.or.jp.