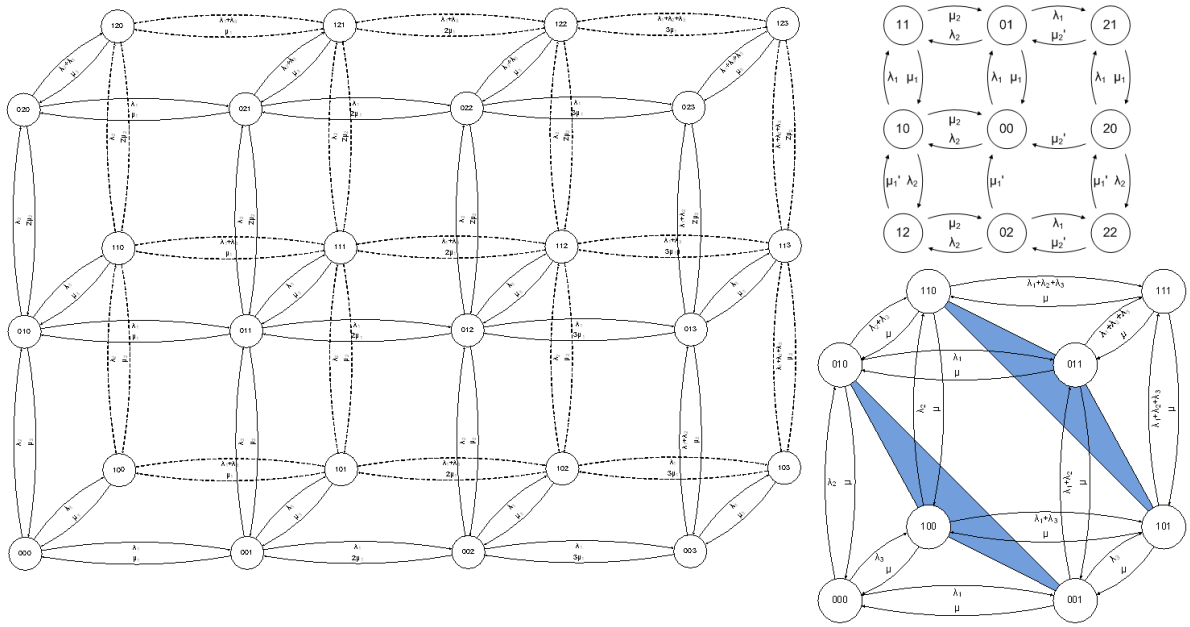
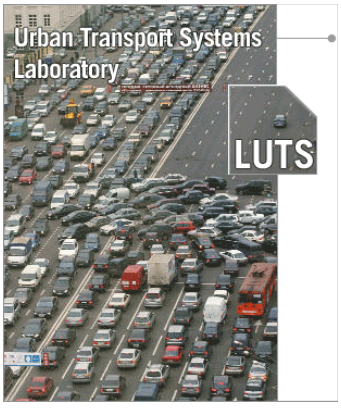


Facility Location Problem for Emergency and On-Demand Transportation Systems

Burak Boyacı

Nikolas Geroliminis

STRC 2012 **January 2012**



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

STRC 2012

Facility Location Problem for Emergency and On-Demand Transportation Systems

Burak Boyacı
Urban Transport Systems Laboratory
Ecole Polytechnique Fédérale de Lausanne (EPFL)
GC C2 406, Station 18, 1015 Lausanne, Switzerland
phone: +41-21-69-32486
fax: +41-21-69-35060
burak.boyaci@epfl.ch

Nikolas Geroliminis
Urban Transport Systems Laboratory
Ecole Polytechnique Fédérale de Lausanne (EPFL)
GC C2 389, Station 18, 1015 Lausanne, Switzerland
phone: +41-21-69-32481
fax: +41-21-69-35060
nikolas.geroliminis@epfl.ch

January 2012

Abstract

Although they have different objectives, emergency response systems and on-demand transportation systems are two similar systems in the sense that both deal with stochastic demand and service time which create congestions for moderate level of demand.

Emergency response system location problems are one of the early problems immensely dealt in the literature. These problems are modeled by either set covering or transportation models which do not give much attention to the stochastic nature of the problem. On-demand transportation is a newly developing type of transportation system and literature is not broad enough but has similarities with emergency response systems.

In this research, our aim is to solve facility location problem with stochastic demand and service time. Specifically we are dealing with temporal and spatial stochasticity which emerge because of the uncertainty in demand and service time. Recently we have developed a mixed aggregate hypercube model which are extensions to Larson (1974) and Boyacı and Geroliminis (2012). Results are promising and applicable to real life instances.

Keywords

emergency response, spatial queues, hypercube queueing models, location models

1 Introduction

Location-allocation of *emergency response systems* is one of the oldest problems in the operations research literature. Locating ambulances, fire brigades and police-beats were the pioneer problems mathematically modeled and solved. Although there are quite a few number of works on the subject, most of them does not take the probabilistic nature of the problem and solves it with deterministic assumptions. However, this is one of the most important property of the problem that differs it from other types of facility location-allocation problems. This randomness creates congestions and causes unexpected losses.

On-demand transportation (also known as demand responsive transport, dial-a-ride transit) is an advanced, user-oriented form of public transport with flexible routing and scheduling of vehicles operating in shared-ride mode between pick-up and drop-off locations according to passengers needs. These systems provide service in areas with low passenger demand where regular bus service is not economically feasible or applicable. Shuttle bus services, paratransit, shared taxis and taxicabs are some types of on-demand transportation systems. Although they have different purposes and priorities, the environment that these systems work has similarities with the emergency response systems. They have probabilistic demand and the main aims are rapid and reliable response and adequate coverage.

2 Literature Survey

The early models dealing with the location of emergency response systems assume deterministic demand. They ignored stochastic nature of the problem and dealt on coverage and median models. Hakimi (1964) proposed p-median problem that locates p facilities on a finite set of candidate locations in such a way that minimizes total transportation cost of n customers. A recent survey of the literature about this subject has conducted by Mladenović *et al.* (2007).

Coverage models are used to locate limited number of facilities (i.e. emergency response systems) in such a way to maximize total coverage. Toregas *et al.* (1971), Church and ReVelle (1974), Marianov and ReVelle (1996), Daskin and Stern (1981) and Gendreau *et al.* (1997) are some of the different versions of coverage models.

Although the literature mainly covers static and deterministic location models, in recent models uncertainty is also taken into account. This uncertainty can be either related to planning of future periods (dynamic models) or input model parameters (probabilistic models). For urban problems, *probabilistic models* are the most suitable ones. For location and allocation of emergency response systems and other service on-demand vehicles (e.g. taxis), it is more convenient to model both the demands and the duration of the time the facility serving these demands with

probabilistic models. In these models, with some probability, it is always possible to have demand which cannot be intervened by any facility, because of stochasticity in both demand and service times. Manne (1961), Daskin (1983), ReVelle and Hogan (1989) and, Marianov and ReVelle (1996) are some of the important articles written in this literature.

Larson (1974) proposed a *hypercube queueing model* (HQM) which is the first model that embeds the *queueing theory* in facility location allocation problems. This model analyzes systems such as emergency services, door-to-door pickup and delivery services , neighborhood service centers and transportation services which has response district design and service-to-customer mode (Larson and Odoni, 1981). The solution of this model provides state probabilities and associated system performance measures (e.g. workload, average service rate, loss rate) for given server locations. "The HQM is not an optimization model; it is only a descriptive model that permits the analysis of scenarios" (Galvão and Morabito, 2008). HQM models the current state as a continuous-time Markov process but does not determine the optimal configuration.

The first model proposed by Larson (1974) assumes that the service time is independent of the locations of the calls for service and the dispatched unit. This argument was supported by the idea that time spend on the road is negligible compared to service time. This can be a fact for fire brigades but not for the ambulances and on-demand vehicles. However even with this simplification, as number of servers (n) is increased, number of states (2^n) grows exponentially. That's why Larson (1975) also proposed a heuristic method because of this exponential behavior. As an extension, Atkinson *et al.* (2008) assume different service rates for each server in the system with equal interdistrict or intradistrict responses which increases number of states (3^n) significantly.

Iannoni and Morabito (2007); Iannoni *et al.* (2008) embedded hypercube in a genetic algorithm framework to locate emergency vehicles along a highway. They extend the problem to enable multiple dispatch (e.g. more than one server can intervene for the same incident).Geroliminis *et al.* (2009, 2011) integrate the location and distracting decisions in the same optimization and solve the problem by using steepest descent (Geroliminis *et al.*, 2009) and genetic algorithms (Geroliminis *et al.*, 2011). Recently Boyacı and Geroliminis (2012) proposed two new aggregate models which group servers into bins of servers. These two approaches dramatically decrease number of states and make HQM applicable for medium sized problems.

3 Hypercube Queueing Models

The conventional HQM models (Larson, 1974) include *hypercube* in the name since the transition graph of the continuous time Markov chain used to model this structure has a hypercube structure when the number of servers is more than three. State variables contain n binary vari-

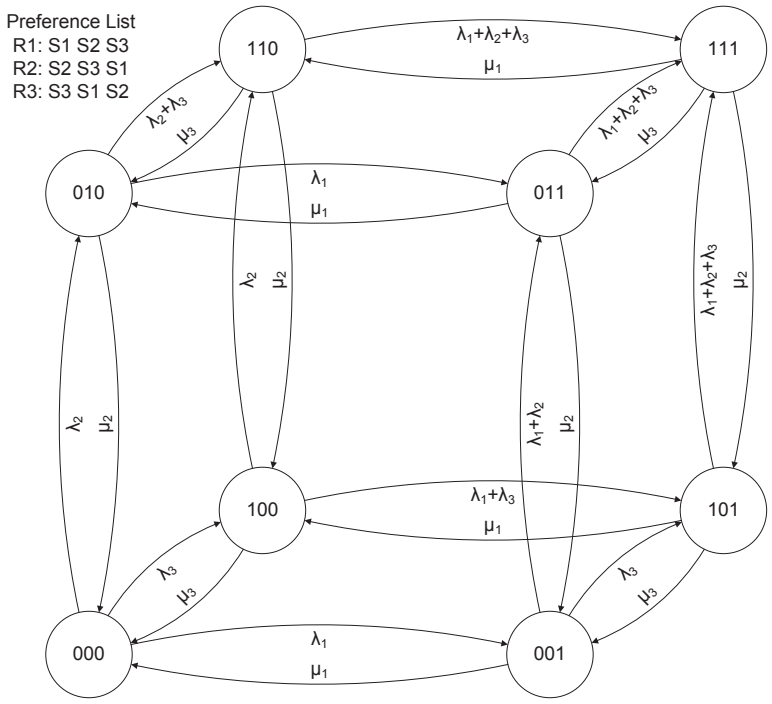


Figure 1: 2^n HQM for three servers

ables which shows if server i is available (0) or busy (1). In other words, each state is a number in base 2 and each digit shows the state of the corresponding server. For each region which are called *atoms* (j) there exist a priority list of servers. Incidents in each region are served by the available server with the highest priority for this atom. If there does not exist any available servers, either the call is lost (i.e. call for ambulance is dispatched by a backup) or joins a queue to be served (i.e. taxi customers are asked to be waited until there is one available), depending on the problem assumptions. Service requests arrive from each atom according to an independent Poisson process with parameter λ_j.

As it is mentioned before, Boyacı and Geroliminis (2012) have proposed two new aggregate models. In 2^n (Larson, 1974) and 3^n (Atkinson *et al.*, 2008) models number of states is exponential function of the number of servers, n (2^n and 3^n respectively) whereas in the aggregated models number of states is exponential function of the number of bins. Number of states in 2^n and 3^n aggregate models equal to $((n_1+1)(n_2+1)...(n_I+1))$ and $\prod_i \binom{n_i+2}{2}$ respectively where n_i is defined as number of servers in bin i for $i = 1, ..., I$. As an example, a system of 3 bins with 9, 6 and 5 servers in each for 3^n aggregate model has 32340 whereas this number would be 420 if we use 2^n aggregate model which assumes equal service rates for inter and intradistrict responses ($\mu_i = \mu_i'$ for $\forall i$). For the same total number of servers, the conventional two models need around million and more than three billion states respectively. For more information, interested readers can apply to Boyacı and Geroliminis (2012)

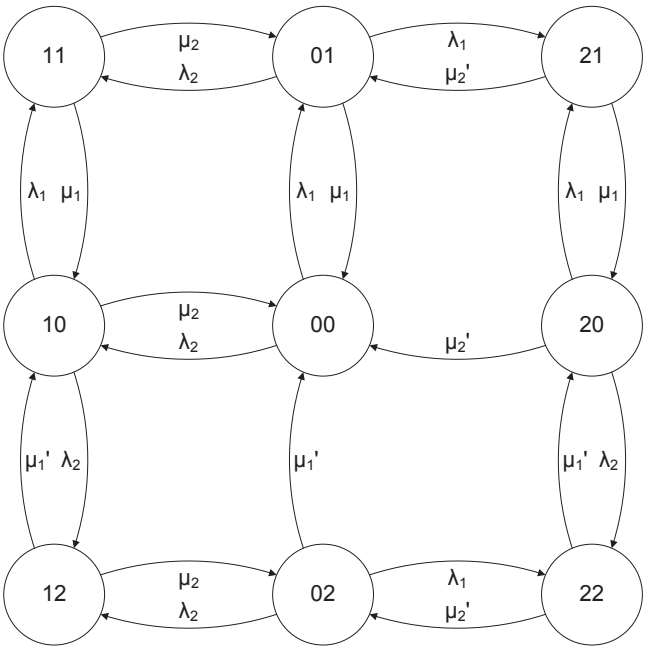


Figure 2: 3^n HQM for two servers with different inter (μ_i) and intradistrict (μ_i') service rates

4 Mix Aggregate Hypercube Algorithm

As it is stated before, although aggregate models decrease number of states in the HQM, the number is still exponential function of the number of bins in the problem. Especially 3^n aggregate model is more representative since it assumes different service rates for different bin-subregion pairs but it has also more states for the same problem than 2^n aggregate models. Because of these reasons, we are proposing an approximation algorithm that takes a mix structure of these two models.

The approximation algorithm basically contains three steps. In the first step, the whole problem is clustered into regions by taking subregion demands and location of the servers into consideration. This clustering should be binary in each step. In other words, clustering algorithm starts dividing the whole area into two and divides each part into two if needed in every iteration. This approach is important since we will use the same pattern in the opposite direction when each region is merged. The second step contains solving 3^n models in each divided region. In order to keep algorithm efficient, there should be at most 6 servers in each region which makes problems of size 729 states. In the last step, we start merging each region as they are clustered by using a mix aggregate hypercube model that has bins of 2^n and 3^n aggregate models. There are two reasons for that: First of all, we want to keep our algorithm efficient, so that is the reason a pure 3^n aggregate model is not used. Secondly, although we cluster the whole area into regions, there are still servers that can serve the requests of the other regions quite often which are the servers that are close to the borders between two merged regions. So the algorithm uses three bins, the one that is composed of servers close to the border has two different

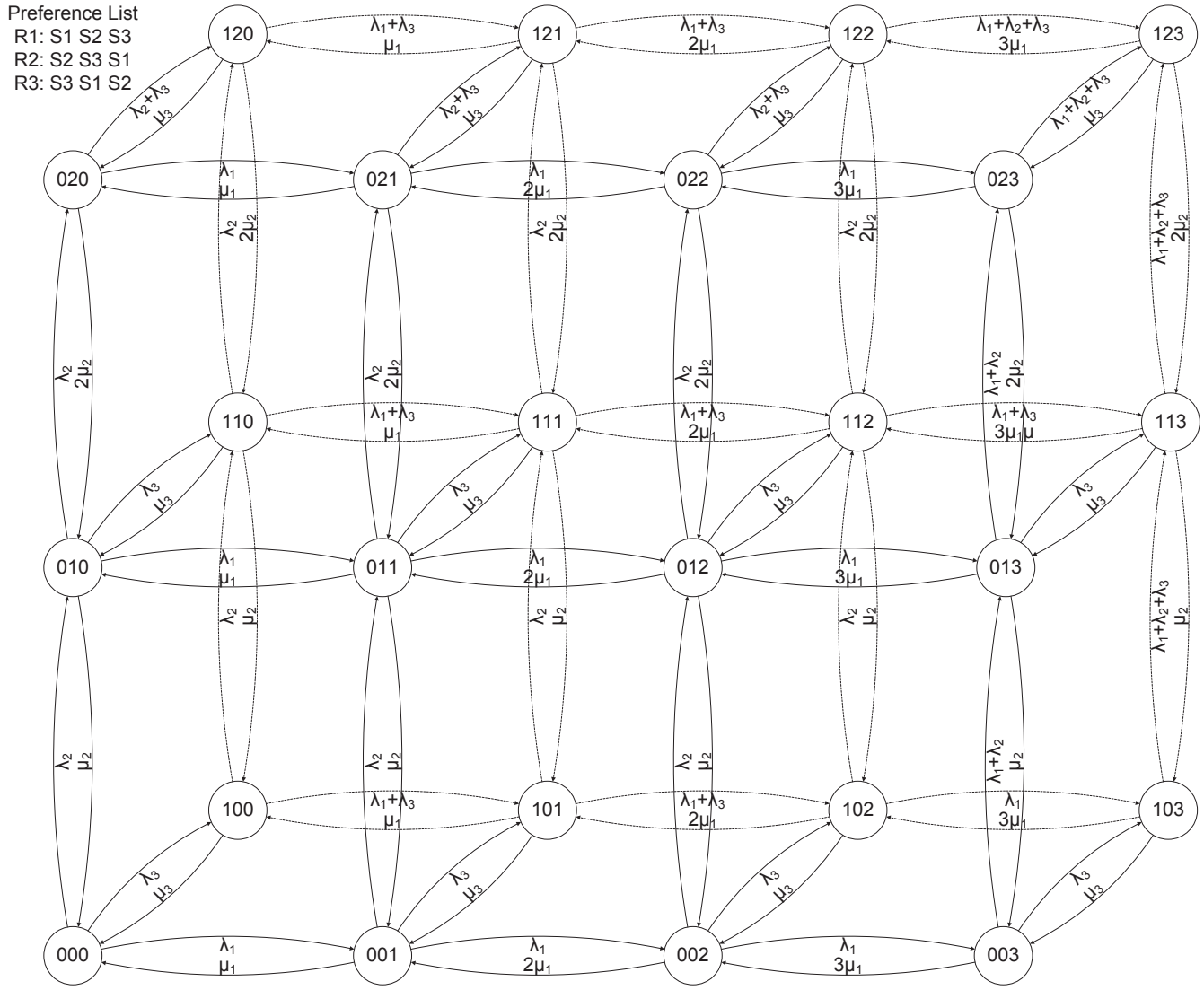


Figure 3: 2^n aggregate HQM for three bins with equal inter and intradistrict service rates (μ_i)

response rates (3^n). The two other bins that are composed of servers away from the border between two regions have one response rate (2^n). The pseudo-code of the algorithm and its simple representation can be seen in Algorithm 1 and Figure 6 respectively.

There are two points which need to be described precisely. The first thing is, as it is described above, we are calculating the service rates for each subregion and then using these values for a bigger subregion that contains two small subregions and a new subregion which is generated by taking some of the servers of these two small subregions. As a result new subregions has less servers than the ones that are calculated before. In order to update this value we are taking the additional service rate of the servers for number of servers in the common area of that subregion. With this approach, heuristic gives service rates for number of servers close to the exact values. However, the lost rate is underestimated and needs to be adjusted as well. In reality, this value is underestimated because when a bin is formed it is assumed that every demand can be served as long as there is an available server in that bin. However this is

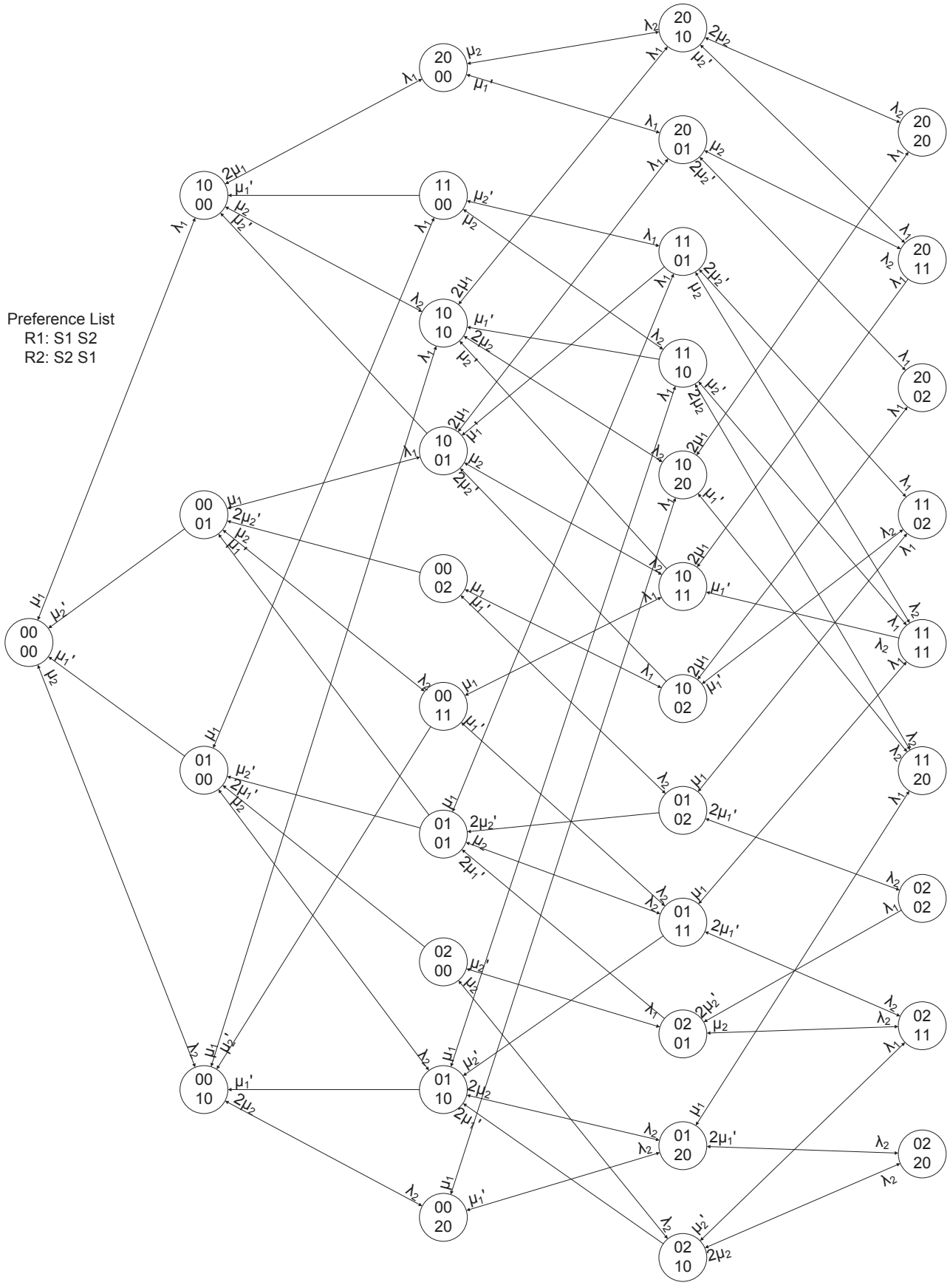


Figure 4: 3^n aggregate HQM for two bins containing two servers in each bin with different intra (μ_i) and interdistrict(μ_i') service rates

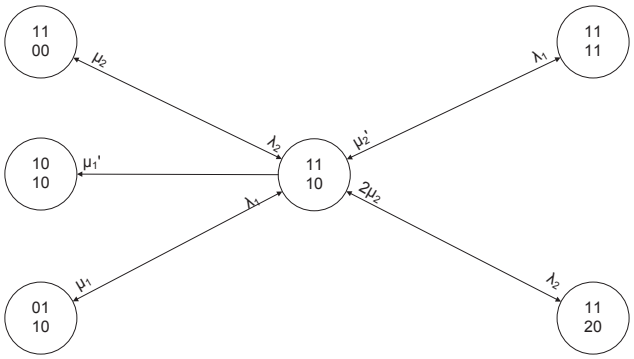


Figure 5: Single state with its connected states from 3^n aggregate HQM

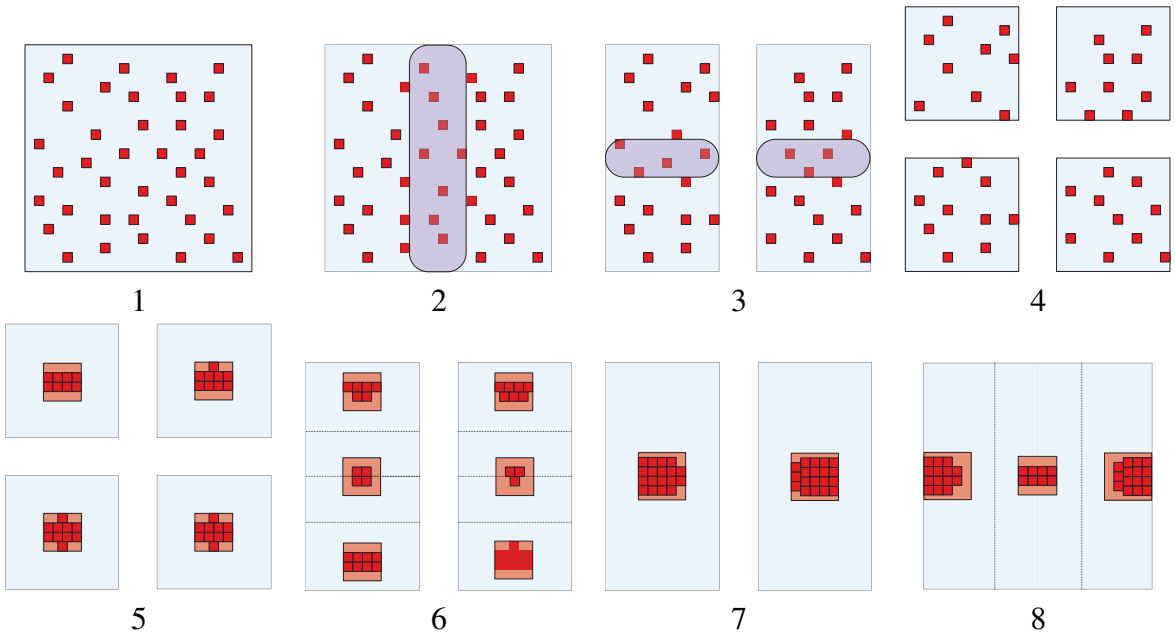


Figure 6: The steps of mix aggregate hypercube algorithm

not the case and the probability should be adjusted for this fact. One of the ways to get rid of this is not to use aggregate models and represent each server state separately. However this increases number of states extremely. What we applied here is simply calculating the probability of having an available server for that demand explicitly and assigning service and lost rates accordingly. Here we have used two approaches. In the first one we assume each server has equal probability to get busy and in the second one we assume each server has different probabilities to get busy and this value is proportional to the demand that they are serving as primary server. Applying this approach improves the lost rates and gives very close results to exact solutions. Detailed experimental results can be seen in the next section. The probability of serving a demand point can be formulated as:

Algorithm 1 Mix aggregate hypercube algorithm

1. Divide the region into subregions iteratively
   (a) Size of each problem in the leaves should be solvable
   (b) Common area servers should be decided
2. Solve the problems in the leaves by using 3^n hypercube model and calculate average service rates for number of busy servers (kind of $M/M/n$)
3. For each subregion not in the leaves
   (a) Solve 3^n hypercube model for the common area servers and calculate their average service rates for number of busy servers (intradistrict service rate)
   (b) Calculate average service rate for atoms that are not in common area servers primary service area (interdistrict service rate)
   (c) Solve $3^n + 2^n$ mix aggregate hypercube model and calculate average service rates for number of busy servers
4. Calculate the lost rate for the first subregion (which is the main region)

$$P_b = \begin{cases} 1 \text{ if } b \leq k-1 \\ 1 - \dfrac{\prod_{i=1}^{k} p_{s_i} \left(\sum_{\tilde{\mathcal{K}} \in \overline{\mathcal{K}}, |\tilde{\mathcal{K}}|=b-k, \tilde{\mathcal{K}}=\left\{\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_{\tilde{k}}\right\}} \prod_{i=1}^{\tilde{k}} p_{\tilde{s}_i} \right)}{\sum_{\tilde{\mathcal{N}} \in \mathcal{N}, |\tilde{\mathcal{N}}|=b, \tilde{\mathcal{N}}=\{\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_b\}} \prod_{i=1}^{b} p_{\tilde{q}_i}} \text{ otherwise} \end{cases} \tag{1}$$

where

$$P_b : \text{probability of serving the demand point for } b \text{ busy servers} \tag{2}$$

$$\mathcal{N} = \{q_1, q_2, ..., q_n\} : \text{the set of all servers} \tag{3}$$

$$\mathcal{K} = \{s_1, s_2, ..., s_k\} : \text{the set of servers that can serve the demand point} \tag{4}$$

$$\tilde{\mathcal{K}} = \{\tilde{s}_1, \tilde{s}_2, ..., \tilde{s}_{\tilde{k}}\} : \text{the set of servers that cannot serve the demand point} \tag{5}$$

$$p_\star : \text{the probability of being busy for server } \star . \tag{6}$$

Equation 1 is also applicable for the case where the probability of being busy is equal. For any busy probability $p > 0$, the formulation returns the same value.

5 Computational Results

The performance of the mixed aggregate approximation algorithm is evaluated by comparing the results of the method with the exact 3^n hypercube model instances. However, since 3^n

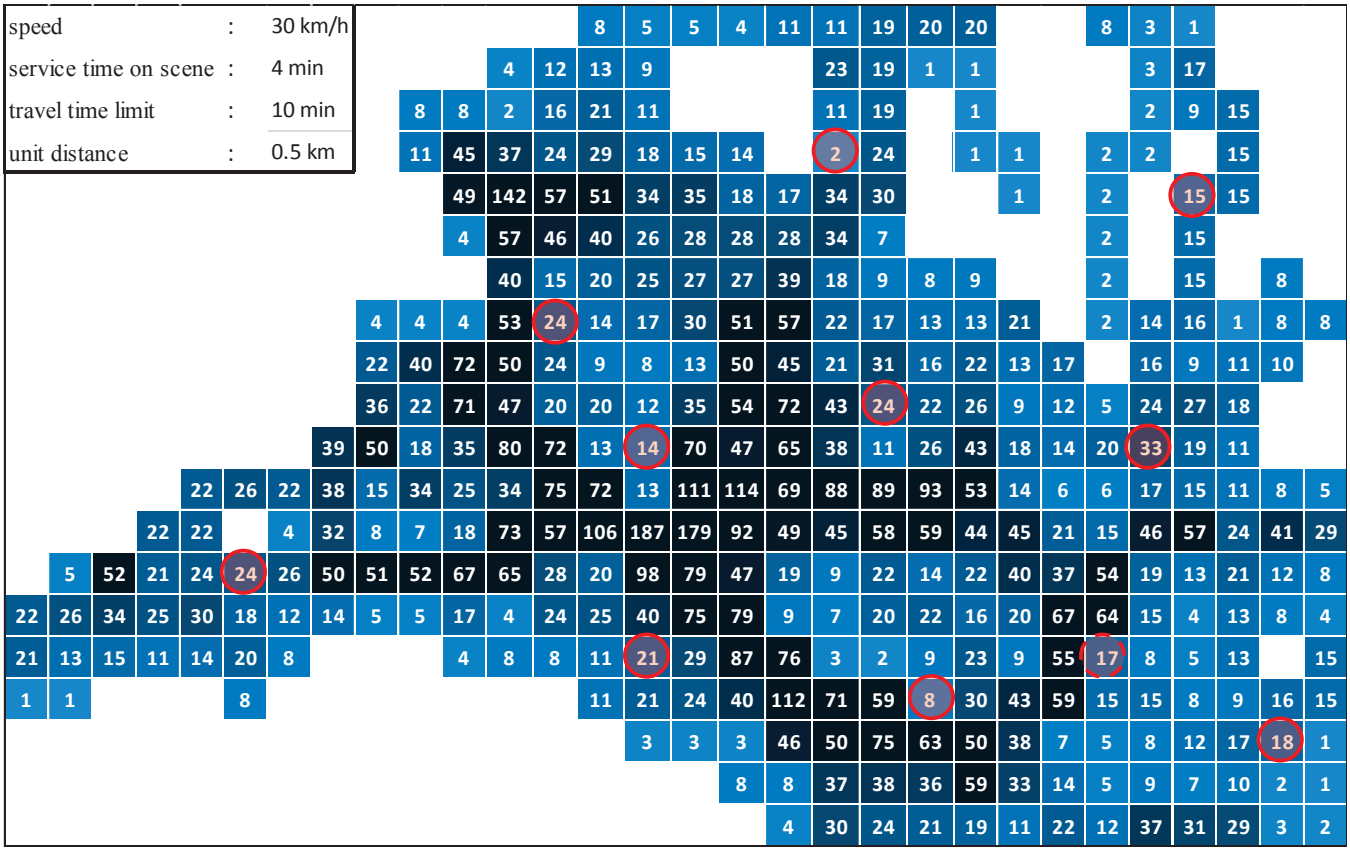


Figure 7: Demand and potential locations for central Athens network

hypercube model has exponential number of states, the compared instances have at most 11 servers. The lost rate of the system and service rates per number of busy servers are compared and represented in the Figure 8 and Figure 9.

We demonstrate the model for locating repair and tow-away vehicles for public transport in Athens (Greece) surface transportation network. This network contains around 3000 buses of different size. This system is used by 1.7 million passengers. Although the whole area is about 650 km^2 we deal with the 150 km^2 area of the highly populated part which contains more than 85% of the demand. In Athens, the buses are handled by city's bus company (ETHEL) whereas the Athens Public Transportation Organization (OASA) is responsible for planning and managing the bus system. In Figure 7 you can see incident percentages that are derived from 10-year historical data and normalized to 10000 per cells that are squares with 0.5 km in each side. In this example, 10 and 11 candidate locations (the additional one to the first 10 candidate locations is shown with dotted line) for transit mobile repair units (TMRUs) are selected (pointed out with red circles) and number of TMRUs needed in each candidate location is calculated for given demand intensity. The reader can refer to Karlaftis *et al.* (2004) for more information about the data.

As it can be seen in Figure 8, service rates of exact method and approximate algorithms are

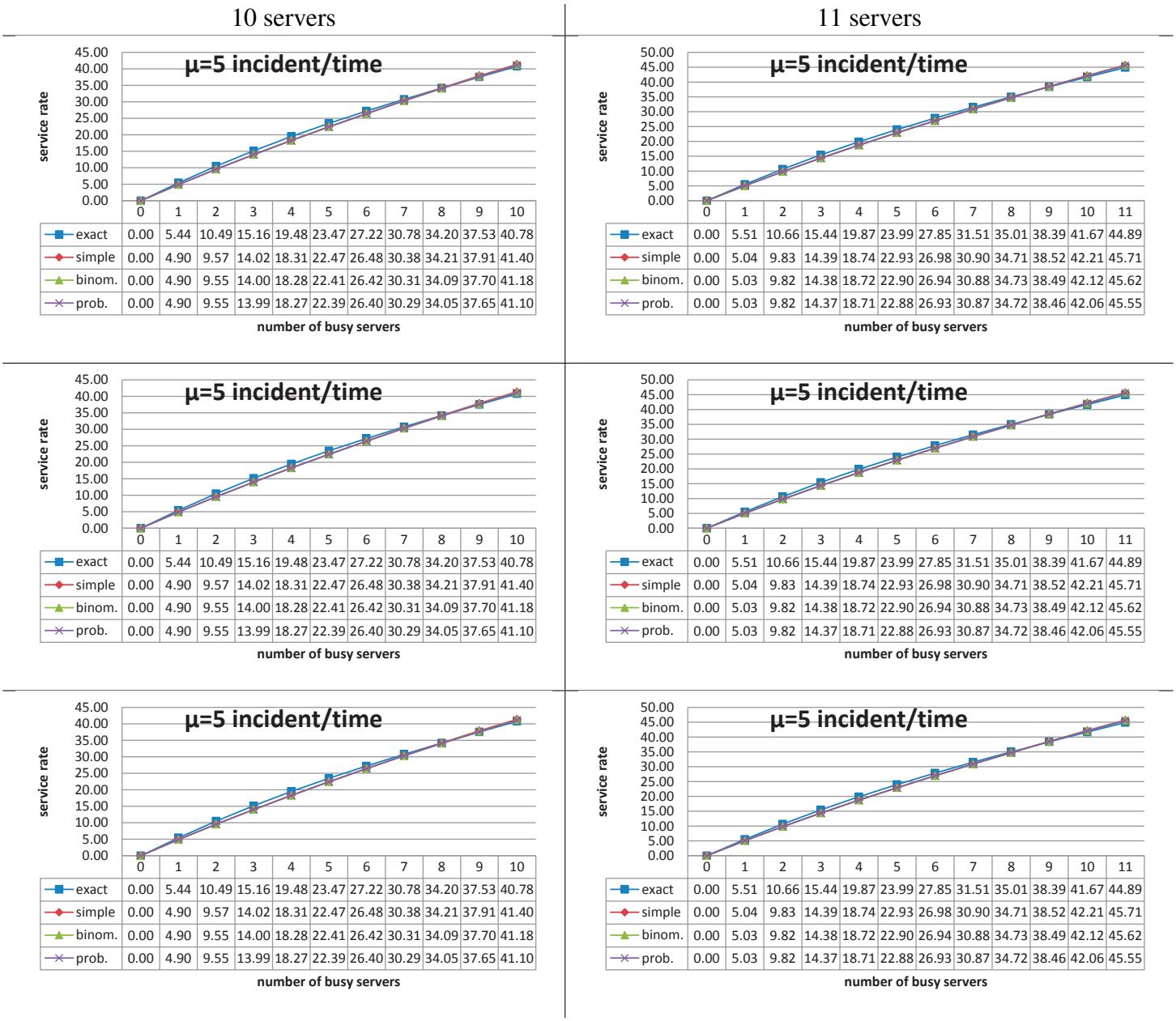
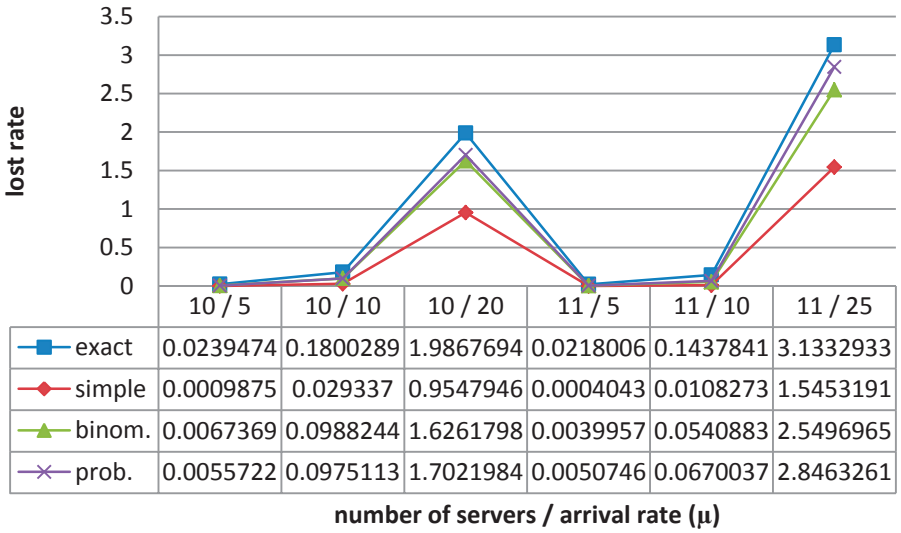


Figure 8: Comparison of service rates per busy server of exact values and approximate methods

quite close to each other. However, this is not the case for the lost rates. In Figure 9, approximate methods always underestimates the exact lost rate. However, this error is lowest for the probabilistic approach and less than 10% for the high demand cases and around 30% for the medium demand cases.

6 Conclusion

In this paper we have investigated an approximation method for spatial queueing systems which have wide range of usage in the urban systems such as deciding the response areas of ambu-



	10 / 5	10 / 10	10 / 20	11 / 5	11 / 10	11 / 25
exact	0.0239474	0.1800289	1.9867694	0.0218006	0.1437841	3.1332933
simple	0.0009875	0.029337	0.9547946	0.0004043	0.0108273	1.5453191
binom.	0.0067369	0.0988244	1.6261798	0.0039957	0.0540883	2.5496965
prob.	0.0055722	0.0975113	1.7021984	0.0050746	0.0670037	2.8463261

Figure 9: Comparison of lost rates of exact values and approximate methods

lances or paratransit vehicles. Although there are some hypercube queueing models exist in the literature, they are not applicable to real life problems because of their computational complexity. The new method that we propose has higher efficiency with acceptable accuracy. However, we still believe that we can improve accuracy by applying an extension to this new method in such a way that the problem will be solved sequentially. In addition to that, both current and sequential methods are suitable for dynamic programming approach if candidate locations for servers are given in advance.

One of the other points missing in this work is the partitioning algorithm for the regions. Although the experiments on worked instances showed that partitioning has small effect on the results, we need to investigate more on that as well. Last but not least is the limitations for the exact problems. As it is stated before, the maximum size of a problem that can be solved by the 3^n hypercube model is 11. However, we can also solve larger cases by using simulation. In order to see the performance of the approximation method for larger instances, we are planning to calculate exact results by simulation.

References

Atkinson, J., I. Kovalenko, N. Kuznetsov and K. Mykhalevych (2008) A hypercube queueing loss model with customer-dependent service rates, *European Journal of Operational Research*, **191** (1) 223 – 239, ISSN 0377-2217.

Boyacı, B. and N. Geroliminis (2012) Extended hypercube models for large scale spatial queueing systems, paper presented at *90th Annual Meeting of the Transportation Research Board*, Washington D.C.

Church, R. and C. ReVelle (1974) The maximal covering location problem, *Papers in Regional Science*, **32**, 101–118.

Daskin, M. (1983) A maximum expected covering location model: Formulation, properties and heuristic solution, *Transportation Science*, **17** (1) 48–70.

Daskin, M. and E. Stern (1981) A hierarchical objective set covering model for emergency medical service vehicle deployment, *Transportation Science*, **15** (2) 137–152.

Galvão, R. and R. Morabito (2008) Emergency service systems: The use of the hypercube queueing model in the solution of probabilistic location problems, *International Transactions in Operational Research*, **15** (5) 525–549, ISSN 1475-3995.

Gendreau, M., G. Laporte and F. Semet (1997) Solving an ambulance location model by tabu search, *Location Science*, **5** (2) 75–88.

Geroliminis, N., M. Karlaftis and A. Skabardonis (2009) A spatial queuing model for the emergency vehicle districting and location problem, *Transportation Research Part B: Methodological*, **43** (7) 798 – 811, ISSN 0191-2615.

Geroliminis, N., K. Kepaptsoglou and M. Karlaftis (2011) A hybrid hypercube - genetic algorithm approach for deploying many emergency response mobile units in an urban network, *European Journal of Operational Research*, **210** (2) 287–300, ISSN 0377-2217.

Hakimi, S. (1964) Optimum locations of switching centers and the absolute centers and medians of a graph, *Operations Research*, **12** (3) 450–459, ISSN 0030364X.

Iannoni, A., R. Morabito and C. Saydam (2008) A hypercube queueing model embedded into a genetic algorithm for ambulance deployment on highways, *Annals of Operations Research*, **157**, 207–224, ISSN 0254-5330. 10.1007/s10479-007-0195-z.

Iannoni, A. P. and R. Morabito (2007) A multiple dispatch and partial backup hypercube queuing model to analyze emergency medical systems on highways, *Transportation Research Part E: Logistics and Transportation Review*, **43** (6) 755 – 771, ISSN 1366-5545. Challenges of Emergency Logistics Management.

Karlaftis, M., K. Kepaptsoglou and A. Stathopoulos (2004) Genetic algorithm-based approach for optimal location of transit repair vehicles on a large urban network, *Transportation Research Record: Journal of the Transportation Research Board*, **1879**, 41–50.

Larson, R. (1974) A hypercube queuing model for facility location and redistricting in urban emergency services, *Computers & Operations Research*, **1** (1) 67 – 95, ISSN 0305-0548.

Larson, R. (1975) Approximating the performance of urban emergency service systems, *Operations Research*, **23** (5) 845–868, September-October 1975.

Larson, R. and A. Odoni (1981) *Urban Operations Research*, Prentice-Hall, Englewood Cliffs, N.J.

Manne, A. (1961) Capacity expansion and probabilistic growth, *Econometrica*, **29** (4) 632–649, ISSN 00129682.

Marianov, V. and C. ReVelle (1996) The queueing maximal availability location problem: A model for the siting of emergency vehicles, *European Journal of Operational Research*, **93** (1) 110–120, ISSN 0377-2217.

Mladenović, N., J. Brimberg, P. Hansen and J. Moreno-Pérez (2007) The p-median problem: A survey of metaheuristic approaches, *European Journal of Operational Research*, **179** (3) 927 – 939, ISSN 0377-2217.

ReVelle, C. and K. Hogan (1989) The maximum availability location problem, *Transportation Science*, **23** (3) 192–200, August 1989.

Toregas, C., R. Swain, C. ReVelle and L. Bergman (1971) The location of emergency service facilities, *Operations Research*, **19** (6) 1363–1373.