Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

ESTIMATING QUANTILE FIELDS FOR A SIMULATED MODEL OF A HOMELESS CARE SYSTEM

Dashi I. Singham

Naval Postgraduate School
Operations Research Department
1411 Cunningham Road
Monterey, CA 93943 USA

ABSTRACT

We construct a simulation model of a homeless care system to determine the amount of new housing and emergency shelter needed to support the growing unsheltered population in Alameda County, California. To quantify the performance of the system, we assess the number of people having unmet need via an estimate of the quantile field using a recently developed batching method. This approach helps right-size the amount of housing and shelter resources needed to quickly provide services to the unsheltered population. We find that with a large investment in housing to help the system reach steady state, current levels of emergency shelter may be sufficient to serve those with unmet need.

1 INTRODUCTION

Many regions in California are facing a homelessness crisis due to a shortage of housing resources. For more than a decade, there has been a growing population of people unable to afford rapidly increasing housing prices. One such region with extreme levels of homelessness is Alameda County, which lies east of the San Francisco Bay. Alameda County contains the cities of Berkeley and Oakland, and currently has approximately 13,000 households seeking assistance each year. Due to limited amounts of physical capacity and funding to build new housing, many people requesting assistance are left with unmet housing needs.

One goal of the Office of Homeless Care and Coordination in Alameda County is to advocate for increases in housing and shelter resources. Modeling and simulation can play a critical role in helping evaluate the benefits of proposed investments into the homeless care system. We worked closely with experts in Alameda County to construct a simulation model based on their system, and note that similar models may apply to other geographic regions.

There are two major sets of resources monitored by the county. The first is housing, and the second is shelter (also called emergency shelter). Housing is intended to provide a home to those who are in need of long-term permanent accommodations. Emergency shelter, on the other hand, is intended as a stopgap to shelter people for weeks or a few months until a permanent solution can be found. However, due to a lack of available housing, people may remain in shelter for long periods of time. This further decreases the amount of shelter available for new arrivals to the system, leading to a large unsheltered population. This growing unsheltered population has led to visibility and increased attention to the homeless crisis.

The two resources, housing and shelter, can be modeled as a tandem queueing system with blocking. People cannot leave shelter until a housing resource is available, and hence the shelter server becomes blocked. This inspires a simulated queueing model to estimate the number of people having unmet need in the system. A person has unmet need if their request cannot be accommodated by housing or shelter resources (i.e., when they remain in queue for one of these servers).

Complex blocking logic makes analyzing this queueing model analytically difficult. Another challenge is that the system is currently unstable from a queueing perspective. The rate of arrivals is far greater than the rate that housing is becoming available, leading to a large queue. Thus, ramping-up of resources in the short term is needed to support the queue of thousands of unsheltered people with the hopes of eventually achieving a steady state system. We use discrete-event simulation to quickly estimate system performance under different input conditions while incorporating complex logic such as blocking, and prioritization of access to shelter according to relative need.

There exists research on simulation modeling of healthcare issues related to individual homeless shelters, for example disease spreading in Higgs et al. (2007), and resource management in Reynolds et al. (2010). Recent work by Singham et al. (2023) developed the first aggregate queueing simulation model of the flow of people through the continuum-of-care (CoC) to consider a county-wide system, as opposed to an individual shelter system. This work modeled complex "pathways" people take through the system to different types of housing resources. For example, some people may require dedicated affordable housing, while others may require permanent supportive housing with services to assist older/frail adults. This model was designed to help determine how much investment into future housing will be needed in these different categories to support the current unsheltered population and serve future arrivals to the system.

Singham et al. (2023) relied on a detailed systems modeling effort to predict the amount of each type of resource needed each year for the next five years to scale up to a system that achieves "functional zero". Functional zero exists when the system has enough resources to quickly rehouse people when they become homeless (the probability a person waits in the queue longer than some acceptable amount of time is small). The model details were validated by members of the Office of Homeless Care and Coordination in Alameda County, as well as other stakeholders and researchers involved in the systems modeling effort.

In this paper, we employ many of the same specific input details from Singham et al. (2023), but change the proposed implementation scheme which increases housing and shelter over time. We rely on first principles of queueing systems to design a two stage model to achieve steady state after a period of time. Based on current estimates of arrival rates, we estimate the amount of resources needed for each pathway to achieve steady state in the long run. In the first stage, we assume we can double these estimates to house the current queue, and in the second stage, we operate at a steady state to handle incoming requests for housing.

While the previous model operated under a finite-horizon timeline for five years, we examine the steady state properties of the updated model using a recent quantile field estimation method. Comparing quantile fields for different input settings lets us right-size the investments into housing and shelter for long-term performance of 30 years, rather than for short-term results. The estimation of quantile fields is an active area of research, and a recent batching method developed by Pasupathy et al. (2023) produces infinite-dimensional confidence regions for quantile fields using dependent simulation output data. Given dependence in highly-utilized queueing models, naive estimation methods that assume independence will not perform well. We test the effects of different levels of shelter resources to show the potential effect on the number of people with unmet need by plotting quantile fields. Additionally, we evaluate the uncertainty associated with these estimates by estimating the volume of the confidence regions for these quantile fields.

The following outlines this paper. Section 2 describes the literature for modeling homeless care systems and summarizes past modeling efforts, while Section 3 presents the updated model used in this work. Section 4 presents the quantile field estimation method used to evaluate the simulation model output. Section 5 delivers the numerical results while Section 6 concludes.

2 REVIEW OF HOMELESS CARE SYSTEM SIMULATION

There are many data collection efforts underway to estimate the properties of homeless populations. The point-in-time count is a yearly effort to estimate the number of homeless people in a given region, and can be conducted using direct or indirect methods (Agans et al. 2014). The Homeless Management Information System (known as HMIS) allows counties to track confidential data about local populations (HUD Exchange 2023). Oakland-Berkeley-Alameda County CoC (2020) observed that some races and ethnicities may be

disproportionally affected by homelessness. Efforts on a smaller scale can be conducted to predict the usage of homeless shelters as in Ingle et al. (2021). The success of Project Roomkey during the pandemic, where unused hotels were converted into shelter, suggested that transitions from shelter to housing would be more effective in a non-congregate setting than in a congregate setting. In congregate settings with many people housed in the same space, occupants were less likely to stay and receive services needed to enable a successful transition to permanent housing (Zeger 2021).

Singham et al. (2023) develop two simulation models of the CoC in Alameda County. The first model, called the "aggregate" model, treats all people in the system equally. It is a simplified queueing model consisting of a tandem queue with blocking. The first server is shelter, and the second server models housing. People leave shelter for housing, but there is no separate queue for housing. The second model is called the "detailed" model because it takes into consideration the different types of housing needs. The eight types of need are modeled as pathways, ranging from people who need a place to stay for a few days or weeks between rentals, to those requiring permanent supportive housing which are aligned with services. These different tiers of service have varying lengths of stay, capacity limits, and costs. Some people require stays in emergency shelter if permanent housing is not available, so shelter is still modeled as a separate server with a queue. The types of need have different priorities to determine who could obtain first access to limited shelter. For example, older/frail adults who have medical conditions and cannot survive without shelter may be moved to the front of the line.

These two models attempted to replicate the process in Alameda County using available data. The model was run with the goal of obtaining functional zero after five years, whereby most housing needs could be resolved within a short period of time. A systems model was developed to predict the effect of changing arrival rates over time, and incrementally (year by year) update the amount of housing inventory needed for each pathway to reach functional zero in five years. These inventory estimates used in the systems model were used as input to the simulation to grow the capacity of the servers over time.

The arrival rate was modeled as a non-homogeneous Poisson process based on a predicted surge in demand to align with recent trends, followed by a decline back to pre-pandemic levels over time based on proposed prevention methods. Prevention methods are employed separately from the CoC and are designed to keep people in their current homes so that they do not require contact with the homeless care system.

The results of the simulation modeling analysis suggested that using the proposed arrival rates and growth rates for server capacities, the system could reach functional zero in five years. However, the results were extremely sensitive to the inputs because they were based on specific year by year predictions which in reality are subject to a lot of uncertainty. Additionally, some pathways could be overallocated while others were underallocated by projecting forward based on current usage which changes over time. This motivates us to construct a simpler queueing model which relies on basic principles to determine how to right-size the housing inventory levels.

3 UPDATED QUEUEING SIMULATION MODEL

We construct a new implementation scheme which relies on queueing theory to determine desired housing inventory levels. One simplification to the detailed model from Singham et al. (2023) is that the arrivals to the system are modeled as a homogenous Poisson process with rate λ rather than a non-homogeneous Poisson process. This allows for easy sensitivity testing and updating as new data arrives, rather than trying to predict future arrival rates with an absence of concrete information. A splitting Poisson process is then used to determine the arrival rate for each pathway, λ_i, using current estimates of proportional need, where $\sum_i \lambda_i = \lambda$.

The processing times which model the length of stay at each housing server are now modeled using exponential distributions rather than previously used triangular or uniform distributions. While this allows for less concrete information about limits on the length of stay, it lets us use estimates of the mean length of stay to inform processing rates μ_i. The exponential distribution also has an unbounded right tail to allow for the probability that a housing unit is occupied indefinitely, and we use conservative estimates of the mean to assume people will stay in housing as long as possible. While this is a worst-case estimate from

a queueing perspective because turnover in the server will be low, in practical terms, the goal is to have people remain in housing as long as possible, rather than exiting the system and potentially returning to the CoC later.

Figure 1 shows the different pathways in the model as servers. Details of the pathways are described in Singham et al. (2023) and omitted here for brevity. Three of the pathways (Youth, Rapid Resolution, and Rapid Rehousing without Shelter) do not begin with stays in emergency shelter. We model a separate queue for each server using red arrows preceding the server, and each of these three pathways can be represented as an $M/M/c$ system.

The five remaining pathways typically involve first an arrival to shelter, then a transition to the appropriate housing resource when it becomes available. As in Singham et al. (2023), shelter will be modeled as a server which is subject to blocking, whereby people cannot leave shelter until there is space for them in their specific housing pathway. Thus, shelter serves as a stopgap while people are waiting for housing. The dashed arrows represent the fact that there is no separate queue for housing (i.e., no input buffer capacity), and the shelter server becomes blocked when there is no housing downstream. This leads to a large queue of unsheltered people in the red arrow leading to emergency shelter, which is the visible source of the major crisis in Alameda County.

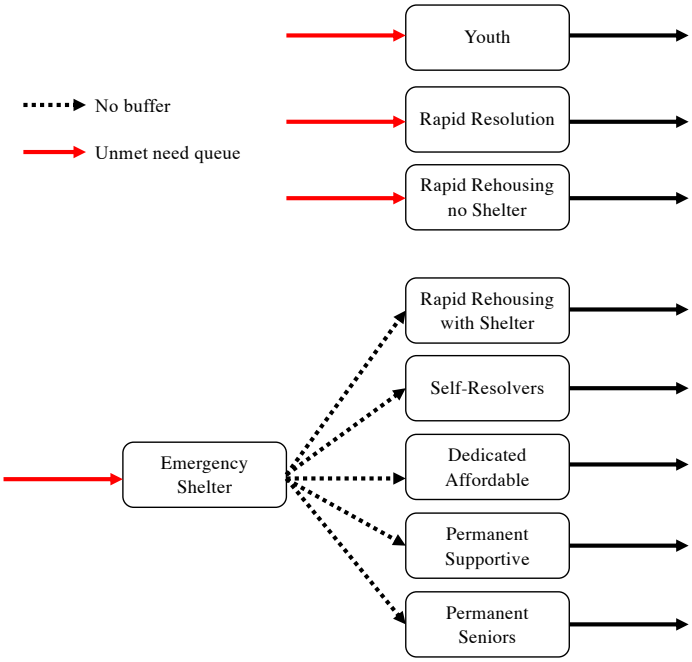

Figure 1: Layout of major pathways through the model. Dashed arrows represent the absence of a buffer, and red arrows correspond to queues of people with unmet need.

Using exponential interarrival times and service times yields easy calculations of utilization values if we model each pathway through the housing servers separately as an $M/M/c$ queue. Based on the arrival rate λ_i and service rate μ_i of each pathway, we calculate the number of servers c_i needed to achieve some desired level of utilization ρ. Given limited resources available to invest in housing, we choose the value of $\rho = 0.95$ to permit small levels of availability for turnover, but generally we expect the housing servers to be highly utilized. For the five pathways that require emergency shelter, blocking complicates the queueing logic, but numerical results suggest that the same utilization calculation can be used to approximate the number of servers needed in each housing pathway. The shelter server effectively supports some of the people that are in the queue for these housing pathways, reducing the unsheltered population.

While the model in Singham et al. (2023) was a finite-horizon simulation designed to reach a specific objective in five years, the model in this paper is designed to operate over an infinite-horizon under steady state conditions. The current system in Alameda County is unstable from a queueing perspective, in that the arrival rate exceeds the service rate and thus there is an increasing population of unsheltered people. This model proposes a two-stage investment process. The goal is to reach a second stage whereby things can operate in the steady state according to calculated values of c_i needed to keep utilization at 0.95.

However, in order to reach steady state, we must have a temporary surge in housing and shelter to help the current large queue of unsheltered people. We propose to model this surge using double the steady state c_i values for the first three years. Such a surge is needed to help those currently unhoused, and the model suggests that maintaining the system at double the capacity for three years will allow for those currently not served to enter and circulate through the system. This warms up the system past its initial conditions before it reaches steady state, and we will employ a truncation method to only use data after five years in the steady state quantile estimation methods.

Because the model includes complexities such as blocking, entity priorities, and entity matching with the appropriate server, the system is not truly an $M/M/c$ system. Thus, we cannot guarantee that our calculations for stability using an $M/M/c$ model will result in sufficient shelter to help mitigate the queue for housing. For example, the shelter could be full of people who require one particular pathway which is not available, and they block the server even though other pathways are available to people in the queue. In reality, people in other pathways may be matched with available housing and bypass the shelter server. But because they are unsheltered this matching may not be efficient, and for simplicity this option is not modeled here.

Even in steady state, we expect there will not be enough room in the shelter for everyone at all times, so our output measure of performance is the number of people with unmet need. This includes people waiting for shelter, and also people waiting for housing in pathways that do not require shelter, as seen in the red arrows of Figure 1. Our primary input to the model that we wish to change is the steady state amount of shelter needed. Right-sizing shelter levels would allow for the system to operate at functional zero and keep the number of people with unmet need at low levels.

4 QUANTILE FIELD ESTIMATION USING BATCHING

One measure for evaluating the resilience of systems is to estimate a quantile (or quantiles) from simulation output. In evaluating the performance of a homeless care system, stakeholders may be interested in the probability that the number of people with unmet need exceeds a certain amount. Currently, community resources like food and showers may be provided to those with unmet need, and knowledge of the demand for these services can help with planning.

When evaluating a single quantile, such as a 90%, 95% or 99% quantile, the value delivered gives the analyst some measure of tail risk to understand potential extreme cases. However, evaluating the entire quantile field for all probabilities $u \in (0,1)$ gives a complete picture of the underlying random variable. For systems that are highly uncertain, looking at both the tails and the center of the distribution can be more useful than only observing an arbitrary tquantile.

There has also been interest in developing confidence intervals to assess the uncertainty associated with quantile estimates. Simultaneous confidence regions for multiple quantile estimates are now available using consistent methods which may attempt to estimate the variance or density of the underlying random variable as in Lei et al. (2020) and Lei et al. (2022) or cancellation methods (Calvin and Nakayama 2013; Pasupathy et al. 2022). The quantile field is an infinite-dimensional object over all probabilities $u \in (0,1)$ and confidence regions for estimates of the quantile field are now available in Pasupathy et al. (2023). This method relies on batching, whereby quantile estimates can be estimated for each of many batches of data sliced from a series of simulation output. These multiple estimates can be combined to estimate the variance and construct a confidence interval (or confidence region for multiple quantiles).

To define this problem, let $\{X_j, j \geq 1\}$ be a real-valued discrete-time stationary stochastic process, where F is the cumulative distribution function (cdf) of X_j, $j = 1,2,...,n$. The cdf F is assumed to be continuous

on $(-\infty, \infty)$ and f is the corresponding density function. Allow $\epsilon_0 \in (0, 1/2)$, to be arbitrarily small, and define the u-quantile

$$Q(u) := \inf\{x : F(x) \geq u\}, \quad u \in I_{\epsilon_0} := [\epsilon_0, 1 - \epsilon_0].$$

Define the infinite-dimensional quantile field

$$\mathbb{Q} := \{Q(u), u \in I_{\epsilon_0}\} \subset C(I_{\epsilon_0}),$$

where $C(I_{\epsilon_0})$ is the space of continuous functions on I_{ϵ_0}. Let $\alpha \in (0, 1)$ be the Type I error of a confidence interval procedure used to construct a $(1 - \alpha)$ confidence region on $\mathbb{Q}$. The confidence region $C_{n,\infty}$ should be calculated such that

$$\lim_{n\to\infty} P(\mathbb{Q} \in C_{n,\infty}) = 1 - \alpha. \tag{1}$$

In order to construct (1), we use a finite-dimensional method which estimates confidence regions for some vector of probabilities $\mathbf{u} = (u_1, u_2, \ldots, u_d)$ using as an estimate $Q(\mathbf{u}) := (Q(u_1), Q(u_2), \ldots, Q(u_d))$. This estimate will be used to produce a confidence region C_n such that $\lim_{n\to\infty} P(\mathbb{Q} \in C_n) = 1 - \alpha$.

The work in Pasupathy et al. (2023) delivers many extensions to the current quantile estimation literature using a batching method. In addition to considering the infinite-dimensional quantile field, it allows for overlapping batching (as opposed to most methods employing non-overlapping batching), varying batch sizes, considers two different centerpoint estimates for intervals (batching and sectioning), and different interval shapes (elliptical, rectangular, etc). For simplicity in this paper, we will consider one type of confidence region implementation that appears to work well in practice — using sectioning to determine interval centerpoints while forming elliptical confidence regions using non-overlapping batching. When non-overlapping batches are used, a time series of length n is divided into b batches of length m, so that $n = mb$. We use $b = 5$ non-overlapping batches of length $m = 0.2n$. This corresponds to a value of $\beta = 0.2$ in Pasupathy et al. (2023) which performed favorably in many numerical tests.

The empirical distribution formed from the entire time series is

$$F_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}_{(-\infty,x]}(X_k);$$

where $\mathbb{I}_A(X)$ is the indicator event which is 1 when $X \in A$. The empirical distribution formed from an individual batch of data is denoted $F_{i,m}(x)$ where only data in batch i of size m is used:

$$F_{i,m}(x) = \frac{1}{m}\sum_{k=(i-1)m+1}^{(i-1)m+m} \mathbb{I}_{(-\infty,x]}(X_k).$$

A so-called sectioning estimator can be used to estimate the infinite-dimensional quantile $\mathbb{Q} = \{Q(u), u \in (0, 1)\}$ using

$$\mathbb{Q}_{S,n} := \{Q_{S,n}(u), u \in I_{\epsilon_0}\},$$

where $Q_{S,n}(u) := F_n^{-1}(u)$. Additionally, an estimate can be calculated for each batch using the batch quantile estimates $Q_{j,m}(u) := F_{j,m}^{-1}(u)$, $j = 1, 2, \ldots, b$. Straightforward extensions can be made to multiple dimensions for $Q_{S,n}(\mathbf{u})$ and $Q_{j,m}(\mathbf{u})$. Rates of error at which quantile estimates decay for time series are found in Bahadur (1966) with Sen (1972) extending to ϕ-mixing data. The method in Pasupathy et al. (2023) relies on well-known convergence results of the function of a quantile process to a Kiefer process (Csorgo and Revesz 1978; Deheuvels 1998).

Definition 1 (Kiefer process) The Kiefer process $\{K(y, t), 0 \leq y \leq 1, 0 \leq t < \infty\}$ is a two-parameter Gaussian process such that

(a) $K(y, 0) = 0$ for all $y \in [0, 1]$;
(b) $\mathbb{E}[K(y, t)] = 0$ for all $(y, t) \in [0, 1] \times [0, \infty)$;

(c) $\mathbb{E}[K(y_1,t_1)K(y_2,t_2)] = t_1 \wedge t_2\,(y_1 \wedge y_2 - y_1 y_2)$ for all $(y_j,t_j) \in [0,1]\times[0,\infty)$, $j = 1,2$.

For additional details on Kiefer processes, see Csörgo and Révész (1981), Section 1.15. The convergence results in Pasupathy et al. (2023) are used to establish a statistic for constructing estimates of finite-dimensional quantile vectors, which are then interpolated to estimate the entire quantile field. To establish confidence intervals for a quantile, we estimate the marginal variance of the quantile field estimator. For each $u \in (0,1)$,

$$S_n^2(u) := \frac{1}{1-\beta}\frac{m}{b}\sum_{j=1}^{b}(Q_{j,m}(u) - Q_{S,n}(u))^2;$$

and for a finite dimensional vector $\mathbf{u}$

$$S_n^2(\mathbf{u}) := \frac{1}{1-\beta}\frac{m}{b}\left(\sum_{j=1}^{b}(Q_{j,m}(\mathbf{u}) - Q_{S,n}(\mathbf{u}))\odot(Q_{j,m}(\mathbf{u}) - Q_{S,n}(\mathbf{u}))\right).$$

where $1-\beta$ is a bias-correction factor and $\odot$ is Hadamard element-wise multiplication. A t-statistic can be constructed using

$$T_n(\mathbf{u}) := \sqrt{n}\,(Q_{S,n}(\mathbf{u}) - Q(\mathbf{u}))\oslash S_n(\mathbf{u}) \tag{2}$$

where $\oslash$ is Hadamard element-wise division. Under some assumptions, Theorem 1 of Pasupathy et al. (2023) establishes that the sequences $\{S_n^2(\mathbf{u}), n \geq 1\}, \{T_n(\mathbf{u}), n \geq 1\}$ satisfy for any $\epsilon_0 \in (0,1)$ and fixed $\mathbf{u} = (u_1, u_2, \ldots, u_d)$,

$$S_n^2(\mathbf{u}) \xrightarrow{\text{d}} \chi^2 \oslash (f(Q(\mathbf{u}))\odot f(Q(\mathbf{u})));$$
$$T_n(\mathbf{u}) \xrightarrow{\text{d}} \chi^{-1}\odot K(\mathbf{u},1) =: T(\mathbf{u}),$$

where

$$\chi^2(u_i) := \frac{1}{1-\beta}\frac{1}{\beta}\frac{1}{b}\sum_{j=1}^{b}\Big(K(u_i, c_j+\beta) - K(u_i,c_j) - \beta K(u_i,1)\Big)^2 \qquad b \in [2,\infty)$$
$$\chi^{-1}(u_i) := \frac{1}{\sqrt{\chi^2(u_i)}},$$

and $c_j := (j-1)(1-\beta)/(b-1)$. By the continuous mapping theorem (Billingsley 1999) on (2) with a mapping function $\|x\|^2$, we have for $p \geq 1$ the following convergence in distribution:

$$\sqrt{n}\left\|(Q_{S,n}(\mathbf{u}) - Q(\mathbf{u}))\oslash S_n(\mathbf{u})\right\|_p \xrightarrow{d} \|T(\mathbf{u})\|_p\,. \tag{3}$$

The limiting distribution in (3) can be numerically calculated using code shared in Pasupathy et al. (2023), and the appropriate t-value chosen to construct an $(1-\alpha)$ confidence region. The volume of this confidence region can be calculated as a measure of uncertainty in the estimate.

5 NUMERICAL RESULTS

We run a series of numerical experiments to derive quantile field estimates for the amount of unmet need. The simulation produced daily estimates of unmet need to estimate the varying number of people who are not served by the system each night. Using simple utilization calculations for $M/M/c$ systems, we approximate the desired capacity c_i for each pathway by setting the utilization $\rho = \lambda_i/(c_i\mu_i)$ to 0.95. Based on the most recent observed data available, the overall arrival rate to the CoC is approximated as 0.4/hour, or approximately 10/day.

Table 1 shows the splitting percentages broken down by pathways, as well as conservative estimates of the mean average stay values. The current levels of inventory (based on the most recent 2021 data published in Home Together (2022)) are also provided. The last column shows the desired steady state housing capacity values needed to obtain 95% utilization levels. One exception is for self-resolvers who find housing outside the county CoC, hence they stay in shelter for an average of 12 weeks until they find their own housing. Emergency shelter is treated separately and people have different distributions for their length of stay based on their pathway type. The desired amount of emergency shelter is an input variable that will be changed in the model. Table 1 reveals that while some pathways have enough capacity currently to manage steady state inflow, other areas are severely lacking. The lack of investment in pathways such as dedicated affordable housing and permanent supportive housing is what has contributed the large amount of unmet need.

Table 1: Relative proportion of arrivals, average stay lengths, current capacity, and desired steady state capacities for each pathway.

Resource	Pct Arrivals	Avg Stay (weeks)	Current Cap.	Desired Cap.
Youth	2%	75	153	110
Rapid Resolution	10%	12	53	88
Rapid Rehousing - No Shelter	10%	400	0	2,947
Rapid Rehousing - With Shelter	15%	75	427	829
Self-Resolvers	10%	12	N/A	N/A
Dedicated Affordable	28%	400	0	8,252
Permanent Supportive	15%	400	2,736	4,421
Permanent Seniors	10%	400	0	2,947
Emergency Shelter	N/A	N/A	1,648	Input variable

We use Simio simulation software build and analyze the homeless care system, and MATLAB to batch the output data and calculate the quantile fields. The simulation experiments for the homeless care system took approximately 17 hours to run on a single machine with four cores. We ran six scenarios varying the amount of emergency shelter units from 700 to 1200 in increments of 100 units. Each scenario was run using 100 replications, and each replication was run for 1,700 weeks (over 32 years). The first three years of the experiment use double the steady state capacity to accommodate the current overflow population, and a total simulation warm-up time of five years was used to incorporate two additional years of warmup to steady state capacity levels. This resulted in over 10,000 days of observations of unmet need produced by each replication.

Figure 2 shows the mean of the quantile field estimates across the 100 replications for different levels of emergency shelter. In order to approximate an infinite-dimensional quantile field, we use a finite-dimensional approximation across probabilities u_i from 0.005 to 0.995 in increments of 0.005. This results in a 199 dimensional field which we interpolate to estimate the quantile field.

The current number of people with unmet need is estimated to be over 8,000 people (Home Together 2022). Figure 2 shows that under the proposed simulation model, the amount of unmet need increases as the number of shelter units decrease, which makes sense because there will be more people unable to access shelter. For 700 shelter units, the amount of unmet need can approach 250, which is considerably less than the current 8,000 people, but could imply instability in the system. With more than 1,000 shelter units there are diminishing returns in that the quantile fields are essentially the same. Note that there will always be some people with unmet need because three of the pathways do not involve emergency shelter and will have queues, hence the distribution of unmet need will be the same when shelter levels are high enough. If the arrival rate changes, the model should be rerun to adjust the server capacities accordingly and choose the number of shelter units.

Table 2 summarizes the key outputs for each experiment averaging across replications. The mean amount of unmet need decreases as the number of shelter units increases, and converges to the amount in

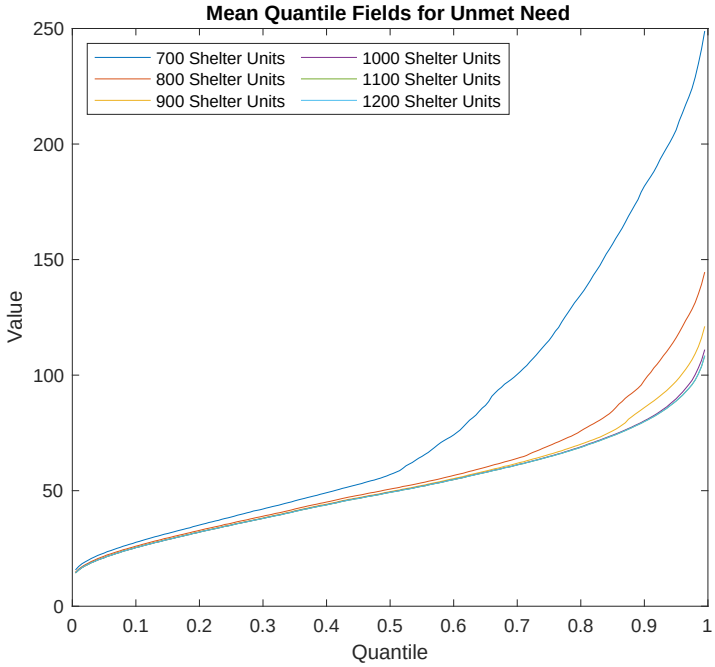

Figure 2: Quantile fields for different numbers of emergency shelter units.

the queue for the three pathways without shelter. The half-volumes of the confidence region also decreases as the number of people in the queue decreases, implying greater variability in the levels of unmet need when there are fewer resources available.

Table 2: Mean unmet need and scaled confidence region half-volumes for quantile fields.

Shelter Units	Mean Unmet Need	Confidence region volume
700	82.72	152.35
800	56.36	59.31
900	52.62	46.29
1,000	51.33	42.44
1,100	51.14	42.09
1,200	51.14	42.09

What is interesting about these results is that there currently are 1,648 shelter units (see Table 1). This means if there was enough county investment in housing to reach steady state, the current shelter capacity would be enough to handle most unmet need.

6 CONCLUSIONS

We construct and analyze a simulation model to determine the number of shelter units needed to support a homeless care system operating under steady state conditions. The model is calibrated using real input data and detailed knowledge of the CoC in Alameda County, CA. The steady state capacities for housing are chosen using queueing approximations, while the amount of shelter is varied through a series of discrete-event simulation experiments. The effect of modifying shelter levels on the amount of daily unmet need is assessed using quantile field estimates. These quantile field estimates are calculated using a recently developed batching method designed to work for infinite-dimensions.

The model suggests that current shelter levels of around 1,000 units may be sufficient to handle system overflow when housing levels are ramped up to the necessary levels to operate in steady state given current arrival conditions. Of course, because current housing levels are severely below these desired steady state levels needed for a stable system, the queue has been growing rapidly and shelter is serving essentially as permanent housing rather than as a stopgap. The main challenge is obtaining enough investment in the housing servers to keep up with demand. Because housing is costly and can take many years to build, one potential solution is to ramp up shelter in the short term while housing is being built, and convert this shelter to housing after unmet need has been reduced.

Future work will explore two main areas. One is optimizing the allocation between shelter and housing across the various pathways given uncertainty in the input arrival rate. A second area involves estimating the steady state probability of the system by focusing on blocking aspects of the queueing system. While there are some analytical methods available to model blocking (Balsamo, de Nitto Personé, and Onvural 2001), developing theoretical approaches for this particular setting is part of ongoing research. We expect many research developments towards using simulation to address this important humanitarian crisis.

ACKNOWLEDGMENTS

The author is grateful for the support of Jennifer Lucky of the Office of Homeless Care and Coordination in Alameda County, CA for her detailed knowledge of the system and its data components, and Stephanie Reinauer of Abt Associates for her systems modeling expertise which inspired this simulation model. Many discussions with them contributed to fine tuning the details of the model.

REFERENCES

Agans, R. P., M. T. Jefferson, J. M. Bowling, D. Zeng, J. Yang, and M. Silverbush. 2014. "Enumerating the Hidden Homeless: Strategies to Estimate the Homeless Gone Missing From a Point-in-Time Count". *Journal of Official Statistics (JOS)* 30(2).

Bahadur, R. R. 1966. "A Note on Quantiles in Large Samples". *The Annals of Mathematical Statistics* 37(3):577–580.

Balsamo, S., V. de Nitto Personé, and R. Onvural. 2001. *Analysis of Queueing Networks with Blocking*, Volume 31. Springer Science & Business Media.

Billingsley, P. 1999. *Convergence of Probability Measures*. 2nd ed. ed. Wiley Series in Probability and Statistics. New York: Wiley.

Calvin, J. M., and M. K. Nakayama. 2013. "Confidence Intervals for Quantiles with Standardized Time Series". In *Proceedings of the 2013 Winter Simulations Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl., 601–612. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Csorgo, M., and P. Revesz. 1978. "Strong Approximations of the Quantile Process". *The Annals of Statistics* 6(4):882–894.

Csörgo, M., and P. Révész. 1981. *Strong Approximations in Probability and Statistics*. Academic Press.

Deheuvels, P. 1998. "On the Approximation of Quantile Processes by Kiefer Processes". *Journal of Theoretical Probability* 11(4):997–1018.

Higgs, B. W., M. Mohtashemi, J. Grinsdale, and L. M. Kawamura. 2007. "Early Detection of Tuberculosis Outbreaks Among the San Francisco Homeless: Trade-offs Between Spatial Resolution and Temporal Scale". *PLOS One* 2(12):e1284.

Home Together 2022. "Home Together 2026: 5-Year Plan to End Homelessness in Alameda County". https://homelessness.acgov.org/homelessness-assets/docs/reports/Home-Together-2026_Report_051022.pdf, accessed 14th August 2023.

HUD Exchange 2023. "HUD Exchange". https://www.hudexchange.info, accessed 14th August 2023.

Ingle, T. A., M. Morrison, X. Wang, T. Mercer, V. Karman, S. Fox, and L. A. Meyers. 2021. "Projecting COVID-19 Isolation Bed Requirements for People Experiencing Homelessness". *PLOS One* 16(5):e0251153.

Lei, L., C. Alexopoulos, Y. Peng, and J. R. Wilson. 2020. "Confidence Intervals and Regions for Quantiles using Conditional Monte Carlo and Generalized Likelihood Ratios". In *Proceedings of the 2020 Winter Simulation Conference*, edited by K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Theising, 2071–2082. Piscataway, New Jersey: IEEE: Institute of Electrical and Electronics Engineers, Inc.

Lei, L., C. Alexopoulos, Y. Peng, and J. R. Wilson. 2022. "Estimating Confidence Intervals and Regions for Quantiles by Monte Carlo Simulation". https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3959456, accessed 14th August 2023.

Oakland-Berkeley-Alameda County CoC 2020, December. "Centering Racial Equity in Homeless System Design". https://everyonehome.org/wp-content/uploads/2021/02/2021-Centering-Racial-Equity-in-Homeless-System-Design-Full-Report-FINAL.pdf, accessed 14th August 2023.

Pasupathy, R., D. Singham, and Y. Yeh. 2022. "Overlapping Batch Confidence Regions on the Steady-State Quantile Vector". In *Proceedings of the 2022 Winter Simulation Conference*, edited by B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C. Corlu, L. Lee, and P. Lendermann. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Pasupathy, R., D. Singham, and Y. Yeh. 2023. "Overlapping Batch Confidence Regions on the Steady-State Quantile Field". *Under review*.

Reynolds, J., Z. Zeng, J. Li, and S.-Y. Chiang. 2010. "Design and Analysis of a Health Care Clinic for Homeless People Using Simulations". *International Journal of Health Care Quality Assurance* 23(6):607–620.

Sen, P. K. 1972. "On the Bahadur Representation of Sample Quantiles for Sequences of φ-Mixing Random Variables". *Journal of Multivariate Analysis* 2(1):77–95.

Singham, D., J. Lucky, and S. Reinauer. 2023. "Discrete-Event Simulation Modeling for Housing of Homeless Populations". *PLOS One* https://doi.org/10.1371/journal.pone.0284336.

C. Zeger 2021. "Evaluating Project Roomkey in Alameda County: Lessons from a Pandemic Response to Homelessness". Report for the Alameda County Office of Homeless Care and Coordination, https://homelessness.acgov.org/homelessness-assets/img/reports/Final%20PRK%20Report%20Summary.pdf, accessed 14th August 2023.

AUTHOR BIOGRAPHIES

DASHI I. SINGHAM is a Research Associate Professor at the Naval Postgraduate School in the Operations Research Department where she conducts research, teaches, and advises students in simulation modeling and analysis. She serves as Director of the SEED (Simulation Experiments & Efficient Designs) Center and has served as a past Treasurer and Council Member of the INFORMS Simulation Society. Dr. Singham is a recipient of the Menneken Award for Highly Meritorious Research (2020) and an NSF Graduate Research Fellowship. She has worked in numerous methodological areas related to simulation, as well as applied work in healthcare, energy, and defense domains. She received a Ph.D. in Industrial Engineering & Operations Research and an M.A. in Statistics from the University of California, Berkeley, and a B.S.E. in Operations Research & Financial Engineering from Princeton University. Her email address is dsingham@nps.edu and her homepage is https://faculty.nps.edu/dsingham/.