Time-Varying Capacity Planning for Designing Large-Scale Homeless Care Systems

Graham Burgess[1], Dashi I. Singham[2], and Luke Rhodes-Leader[1]

[1]Management School, Lancaster University
[2]Operations Research Department, Naval Postgraduate School

September 30, 2024

Abstract

Many people in communities around the world are facing homelessness due to housing shortages. The San Francisco Bay Area has struggled to provide housing for thousands of people who are unsheltered. Permanent housing is the ideal solution for most people entering the system, but temporary shelter is also critical when there is not enough housing. Investment in housing and shelter is paramount to providing a long term solution to serve the current and future homeless population. We construct a queueing model for tracking the flow of people through shelter and housing based on Alameda County's coordinated entry system. In contrast to routing or allocation policies, we optimize the system through increasing shelter and housing server capacities. We formulate optimization problems to reduce the expected size of the unsheltered population given cost constraints by varying investment in housing and shelter over time. Additionally, there may be policy considerations that affect the feasibility of a proposed investment plan. We incorporate shape constraints to address the time-dependent nature of investment decisions to show how resources can be allocated between housing and shelter. This joint optimization approach can be broadly applied to capacity investment decisions for tandem queueing systems, for example, in healthcare settings.

Keywords: homelessness, capacity planning, fluid flow model, shape-constrained optimization, queueing systems

1 Introduction

In the San Francisco Bay Area, there are many communities which are struggling with unprecedented levels of homelessness. Directly east of San Francisco, Alameda County contains the cities of Berkeley and Oakland with over 1.6 million residents. There are high levels of homelessness, with approximately 8,000 people experiencing homelessness each night. In 2020, Alameda County formed an Office of Homeless Care and Coordination to conduct leadership and strategic planning with regards to the Continuum of Care (CoC) (Alameda County 2022). The CoC is defined by the Office of Housing and Urban Development (HUD) as “designed to assist individuals (including unaccompanied youth) and families experiencing homelessness and provide the services needed to help such individuals move into transitional and permanent housing, with the goal of long-term stability” (Office of Housing and Urban Development 2024). Counties are responsible for implementing support for the CoC within their geographical areas according to federal guidance from HUD, but may adjust their specific plan depending on the needs of their populations.

Alameda County has developed a particular focus on racial equity to drive their efforts. This is because a thorough examination of population data determined that some racial groups were overrepresented in the homeless populations. Details on the population analysis are contained in Oakland-Berkeley-Alameda County CoC (2020). The result of the analysis is that investment into certain types of housing is critical to alleviating homelessness, and also reducing racial disparity that exists in the housing market. This paper will address the critical problem of determining how to best invest in housing resources to address homelessness, and our approach will contribute to the general literature on capacity planning problems.

There are two main types of resources to support people experiencing homelessness. The first resource is access to permanent housing, which is defined as “community-based housing without a designated length of stay in which formerly homeless individuals and families live as independently as possible” (Office of Housing and Urban Development 2024). This can take many forms, for example, permanent supportive housing provides affordable housing in tandem with social services to allow the client to maintain successful housing. Dedicated affordable housing may be used to support households with extremely low incomes without the potential for salary increases. Rapid rehousing subsidies may allow clients to afford rent to remain in their current homes when the client has the potential to increase their income within an expected time period (Oakland-Berkeley-Alameda County CoC 2020).

The second main type of resource is transitional housing, or emergency shelter, which is "designed to provide homeless individuals and families with the interim stability and support to successfully move to and maintain permanent housing" (Office of Housing and Urban Development 2024). In the absence of a permanent housing solution, emergency shelter can provide temporary accommodations until a permanent housing solution can be established. While emergency shelter is an important part of a county's infrastructure, it should not be relied on as a sole substitute for housing.

In the rest of this paper, we will refer to permanent housing as "housing" and transitional housing or emergency shelter as "shelter". Our goal will be to optimize the choice of investment into building/acquiring these resources. Let h_t be the inventory of housing at time t, and s_t be the amount of shelter at time t. The decision variables h_t and s_t will be the key focus of our optimization model.

People in the system may occupy housing or shelter, or they may be unsheltered while waiting for county resources. The large number of unsheltered people in the Bay Area is what has driven increased attention and visibility around this crisis. The quality of the CoC performance will be driven by an objective function that depends in part on the number of unsheltered people in the system over time, u_t. The value of u_t is not a decision variable, but is calculated as an output of a queueing model. Figure 1 shows how one might model this system as a sequence of servers. A client arrives to the system, and if there is no housing or shelter available, they wait in the unsheltered queue. Shelter serves as a resource, but people only occupy shelter if housing is not available. Thus, people do not leave shelter until housing is available, so there exists a blocking mechanism between the servers with zero buffer.

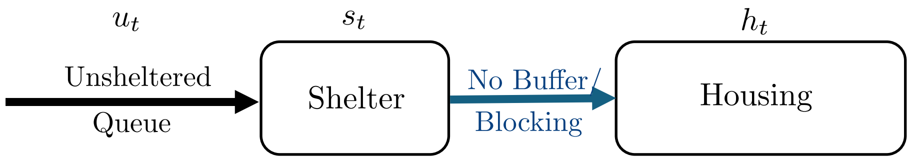


Figure 1: Tandem queue model of housing and shelter.

The lack of a defined service time at shelter combined with a blocking dynamic between shelter and housing means we can simplify the model to the setup in Figure 2. Thus, people in shelter are still in a queue for housing, they are just in a potentially better situation than those who are unsheltered by the county and are often (but not always) prioritized for limited housing resources. This means we can treat the decision variable s_t as an allocation of resources to shelter part of the queue, with housing being the sole server system. While housing is considered a permanent solution for those who are able to remain successfully housed, there may be a turnover rate of approximately 8% per year (Oakland-Berkeley-Alameda County CoC 2020). This allows us to model the system as an $M_t/G/h_t$ queue, given a non-homogeneous Poisson arrival process and general service time distribution.

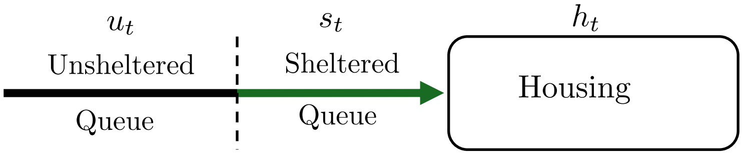


Figure 2: $M_t/G/h_t$ model of housing and shelter.

Increasing h_t and s_t will decrease u_t. Furthermore, in reality, people may move in and out of shelter while waiting for housing. Depending on the specific type of housing, people may move directly from being unsheltered into housing. Because shelter does not have its own service time distribution and simply holds clients until housing is available, the setup in Figure 2 is an equivalent model to Figure 1. This paper will construct a fluid flow model for the flow of clients through this system, and optimize the values of s_t and h_t over time to reduce the value of u_t, subject to budgetary and policy constraints.

We next describe some background that motivates our formulation. There are multiple actions that federal and local governments can take to address homelessness. At the federal level, the Office of Housing and Urban Development has established that each county must operate a Coordinated Entry system to ensure a standardized system of access points for clients experiencing homelessness to seek support (HUD Exchange 2022). However, each county has flexibility to adjust their approach to prioritizing and allocating resources based on the particular needs of their constituents. Thus, different policies and procedures may be employed by different locations.

One goal of Alameda County is to reach “functional zero” in five years. The system is at functional zero when there is effectively no unmet need, meaning that the ‘expected’ waiting time for housing is under 90 days (Alameda County 2022). In queueing terms, this means that the probability that someone experiencing homelessness must wait in the queue (sheltered or unsheltered) more than 90 days for housing is small or close to zero. There are multiple strategies to achieve functional zero, including prevention of homelessness through early intervention, and investment in building housing and shelter to address the unique needs of the community. While building temporary shelter does not directly decrease the time until a client receives housing, a secondary goal is to provide shelter to those seeking health and safety benefits. Additionally, it may be easier to locate and provide resources to clients who are sheltered, compared to those who may be unsheltered.

The problem of limited financial and physical resources to address this crisis is obvious. Another constraint is time: it can take time for housing to be obtained or shelter to be constructed. Thus, a one-shot optimization may not be feasible to implement in a single time period. This motivates us to consider investment over time, with the state of the queueing system improving as housing and shelter inventory increases. It is better to invest early when possible since there is a human suffering cost to waiting years for shelter. Thus, one major way our model differs from other capacity sizing problems is that we allow the capacity variables s_t and h_t to change over time. Similarly, our objective function will integrate over time to penalize delays in obtaining housing for the unsheltered population.

Alameda County undertook a system’s modeling effort to model the flow of clients through the system and test the effects of different levels of investment in housing and shelter each year (Alameda County 2022). This allowed the stakeholders to determine an approximate cost needed to operate enough housing to reach functional zero. While the past efforts by the county delivered an excellent feasibility and cost analysis to an otherwise highly uncertain problem, they did not include queueing, uncertainty, or optimization directly. Singham et al. (2023) improved the county’s efforts by incorporating a simulation of a queueing model with uncertainty. However, the lack of tractability around this highly complex simulation model made determining an optimal solution difficult.

The present paper aims to address the problem of capacity sizing over time by constructing and optimizing a tractable fluid flow queueing model. To the best of our knowledge, this would be the first way capacity investment over time would be addressed. These results would allow Alameda County to determine the best allocation of resources between housing and shelter given a limited budget. The complexity in this model stems from the fact that we optimize the capacity of the system over time, so we are effectively estimating optimal service capacity functions over time. We accomplish this by discretizing time and adopting a fluid flow model which tracks queueing as capacity is added to the system. This model returns the number of housed, sheltered, and unsheltered people over time.

An additional feature that we incorporate to improve real-world feasibility is to include policy-based shape constraints to the decision variables. For example, it may be more feasible from a tax-raising standpoint to increase investment slowly over time, rather than requiring a large one-time investment up front, even though that may be the fastest resolution. Additionally, while shelter is critical to reducing unsheltered homelessness, communities may not want a large long-term reliance on shelter. One idea is to ramp up shelter in the short term while housing is still being built, and convert some of the shelter to housing in later years. This means that s_t would be unimodal over time with a peak partway through the model timeframe. Our flexible framework would allow for optimal solutions meeting feasibility constraints on what could be reasonably implemented.

Section 2 will briefly review the literature and place our contributions relative to past work. Section 3 describes the fluid flow queueing model, while Section 4 presents the optimization formulations and the corresponding numerical results are displayed in Section 5. Section 6 concludes.

2 Literature Review

This section will briefly review the literature related to our approach. One aspect we consider is the capacity sizing problem for queueing systems. Much of this research operates in a healthcare setting. Support for people experiencing homelessness often involves aligning homeless services closely with healthcare resources, so we will discuss research that approaches resolving homelessness from a healthcare perspective. We will conclude this section by discussing key simulation and optimization literature related to modeling of homelessness systems, including those related to runaway youths.

There are traditional tradeoffs between a quality driven regime, where the focus is on reducing the customer's waiting time, and an efficiency driven regime, where the focus is on having the servers always busy. In an efficiency driven regime, the probability that the customer must wait for a server converges to one. A balance between these regimes is the quality and efficiency driven (QED) regime. The well-known square root safety (SRS) rule determines a capacity that will achieve a QED regime. The SRS capacity is the sum of a base level to handle the mean arrival rate, plus a square root safety factor to accommodate variability. The effectiveness of the SRS hedging factor is measured relative to input parameter variability in Bassamboo et al. (2010). Besbes et al. (2022) also determine capacity planning rules for spatial contexts, whereby the SRS rule is insufficient to achieve a QED balance.

There has been much work in capacity planning for healthcare settings. Queueing models have been used to determine how many appointment slots to allow, for example, in planning for specialty clinics (Izady 2015). Our approach is concerned with capacity planning over longer time horizons, with thousands of clients spending years in the system. Additionally, it can take years to build enough housing and shelter, so this type of capacity planning requires large-scale modeling compared to many healthcare models which seek to plan daily appointment schedules. Optimal capacity planning in queueing systems is often performed by varying the arrival and service rates of a system (Bretthauer 1995, Stidham Jr 2009). Liu et al. (2011) combine analytical queueing methods and simulation modeling to determine capacity expansion plans for a semiconductor production manufacturing system. Izady and Worthington (2012) develop an approach for determining staffing of emergency departments over time subject to changing arrival rates and a probabilistic requirement on the sojourn time. Konrad and Liu (2023) use a simulation-based learning approach to balance exploration and exploitation in staffing models that seek a probabilistic tail delay limit.

The notion of tandem queues is widely present in healthcare settings, and is relevent to our modeling of housing and shelter systems. For example, emergency departments feed into hospitals, or acute term care facilities feed into long-term care facilities (Patrick 2011, Patrick et al. 2015). In many of these cases, if there is an issue with downstream capacity in the second server system, there will be long waiting times for the first server system. This is especially true if blocking exists between servers, so patients cannot leave the first server system until there is a spot available in the second server system. Methods for allocating resources across servers in zero-buffer systems for the purposes of optimizing throughput are studied in Yarmand and Down (2013, 2015).

The combination of healthcare modeling with homeless resource planning has become an important area of research for solving critical issues affecting the homeless population (Higgs et al. 2007, Reynolds et al. 2010, Ingle et al. 2021). There has also been research invested into determining the right level of detail or specification of a portfolio of services for homeless populations (Arora et al. 2021). Rahmattalabi et al. (2022) create a queueing system to study the effect of matching a client with resources according to an eligibility structure while taking fairness constraints into consideration. Optimization of equity in food redistribution for soup kitchens and homeless shelters is considered in Balcik et al. (2014).

There is a stream of research related to shortages of shelter beds for runaway youths in New York. Miller et al. (2022) analyzes alternatives by comparing improvements from increasing shelter capacity by optimizing benefit to cost ratios. Kaya and Maass (2022) develop a queueing model with abandonment to improve equitable access to youths with different types of shelter needs and priorities. Their formulation minimizes the number of servers needed subject to some constraints on the quality of service. Homelessness may also be tied to a higher risk of human trafficking, and optimization has been used to determine the allocation of shelters and the impact on societal value (Maass et al. 2020). Kaya et al. (2024) develop a complex optimization model to determine how to assign youth to shelter resources given particular profiles of the individual and the specific services offered by the shelters.

Singham et al. (2023) modeled the homeless system in Alameda County as a sequence of serial and parallel queues and used simulation to provide in-depth feasibility and cost analysis of different strategies under reasonable levels of uncertainty. While this simulation model provides a first approach to county-level modeling as a queue, it is not amenable to optimization given its complexity, high levels of uncertainty, and long runtimes. However, Singham (2023) does attempt to determine the appropriate long-term level of shelter using a batching-based quantile estimation method applied to highly dependent simulation output.

Finally, we discuss recent function estimation methods that can incorporate shape information as constraints. The ability to incorporate shape information into function estimation problems is an important area of research. While the literature on these types of models span a broad range of statistical literature that we omit here, we note shape constraints are an important part of structuring distributionally robust optimization problems over function spaces. In particular, a focus on unimodal functions may lead to tractable formulations (Lam et al. 2021, 2024), and we will employ unimodal shape constraints in our formulations.

3 Queueing model for a homeless care system

Given the current limits on homeless resources in Alameda County relative to demand, the system operates in an efficiency driven regime where the servers are always occupied due to queueing instability in the system. In the status quo the servers are not able to house people as quickly as needed because the arrivals to the system outpace the service rate, and there exists a large queue of people waiting from previous years. Not only are there not enough housing slots, but turnover may be low because people are staying in the system for long periods of time, and possibly permanently in some cases.

This motivates our use of a fluid flow model for the system. We use the terms fluid flow model and fluid flow queueing model interchangeably. Fluid flow models can be useful for evaluating waiting times in healthcare systems (Worthington 1991), for establishing the limits of complex queueing systems (Nov et al. 2022), and as a basis for understanding preliminary aspects of a simulation optimization (Jian and Henderson 2015). Background on the development of fluid flow models as the limits of queueing systems using the functional strong law of large numbers is available in Chapter 5 of Chen and Yao (2001). Because the assumption of an efficiency driven regime with a large inflow of customers in the short term relative to system outflow is realistic for many homeless systems in California, we employ a fluid flow model to allow for tractable optimization.

This section introduces a fluid flow queueing model which tracks the number of housed, sheltered, and unsheltered clients over time. The two main inputs to the model are a changing arrival rate over time, and a housing service rate which changes as housing is built. Additionally, the amount of shelter space available to support the queue for housing may change over time. This models the dynamics in Figure 2. Due to the current large queue for housing and continued inability for housing rates to keep up with arrivals to the system, the assumption that the servers will always be busy is not only reasonable, but significant enough to negate the usual assumptions of steady-state queueing behavior where the servers are idle with some positive probability.

In our fluid flow model we ignore the randomness in the arrival process and the service process for homeless people entering and leaving the homeless response system. Instead we assume that "fluid" flows into the system continuously at a rate $\lambda(t)$ and flows out at rate $\mu(t) = \mu_0 h(t)$ where μ_0 is the service rate of a single housing unit and $h(t)$ is the continuous-valued number of houses at time t. Given the initial number of people in the system X_0, at time t we can calculate the subsequent number of people in the system, $X(t)$, as

$$X(t) = X_0 + \int_0^t \lambda(t)dt - \int_0^t \mu_0 h(t)dt.$$

We split the queue for housing into an unsheltered and a sheltered part. We denote by $s(t)$ the continuous-valued number of shelters at time t. The size of the unsheltered queue $u(t)$ is then

$$\begin{aligned} u(t) &= X(t) - h(t) - s(t) \qquad (3.1) \\ &= X_0 + \int_0^t \lambda(t)dt - \int_0^t \mu_0 h(t)dt - h(t) - s(t), \qquad (3.2) \end{aligned}$$

where we assume that capacities $h(t)$ and $s(t)$ are sufficiently small compared to the given arrival rate $\lambda(t)$ so that these resources are always full, and the use of a fluid flow model remains appropriate. In other words, the number of people housed and the number in shelter are the same as the housing and shelter capacities $h(t)$ and $s(t)$, respectively. In reality, there may be some friction in the system in that housing may be idle while units are experiencing turnover and the next person in the queue is being located, but this time can be incorporated into the service time distribution.

When analyzing the dynamics of the fluid flow model over a modeling horizon, we discretize time into days. We now let λ_d, h_d^D, s_d^D and u_d for all $d \in \{1, ..., D\}$ be the discretized equivalents of $\lambda(t), h(t), s(t)$ and $u(t)$, respectively, where D is the modeling horizon in days and is used as a superscript where we must later distinguish between daily and annual capacities. In order to evaluate our various objective functions (which we describe in Section 4) we typically approximate (3.2) with the sum

$$u_d = X_0 + \sum_{d'=1}^{d} \lambda_{d'} \delta t - \sum_{d'=1}^{d} \mu_0 h_{d'}^D \delta t - h_d^D - s_d^D, \qquad (3.3)$$

where μ_0 is the daily service rate of a single housing unit and the stepsize $\delta t = 1$ day. In Figure 3 we give an illustrative example of the dynamics of u_d given by our fluid flow model, calibrated using realistic inputs for X_0, μ_0 and λ_d, h_d^D, s_d^D for all $d \in \{1, ..., D\}$ which we take directly from Singham et al. (2023).

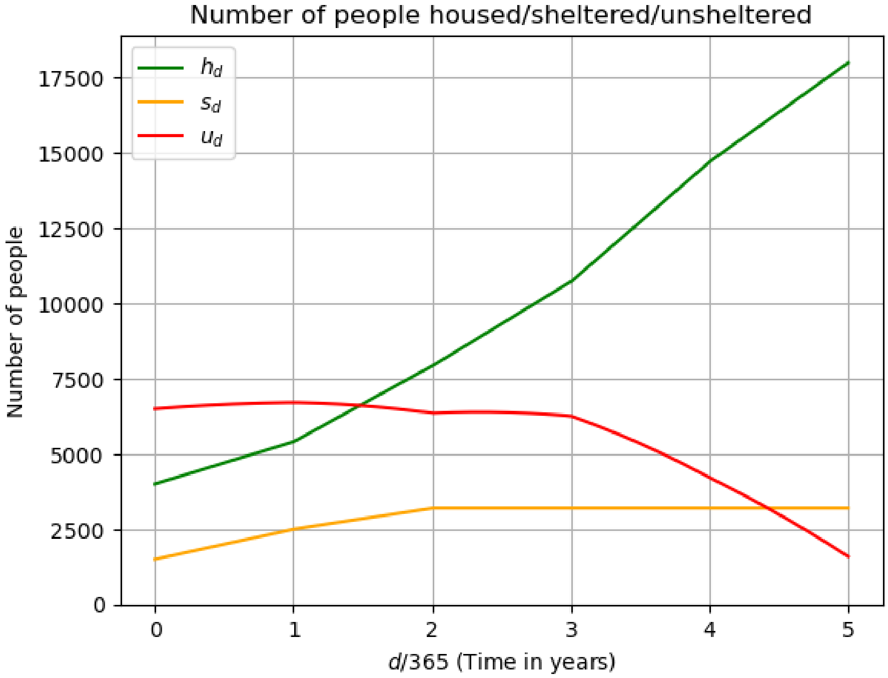


Figure 3: Dynamics of u_d, s_d^D and h_d^D. $X_0 = 12000$, $\mu_0 = 6.106 \times 10^{-4}$, Daily arrival rates λ_d in each year: $10.0, 11.9, 13.1, 13.1, 11.8$.

Figure 3 shows an example of how one might come close to reaching functional zero in five years. The level of housing investment steadily increases over time. There is some initial increase in shelter, though in general there is less investment in shelter over the long term than in housing. The unsheltered population is stabilized and then eventually decreases approaching zero. The arrival rates λ_d are projected based on a current estimate of 10/day, along with the assumption that arrivals will increase in the coming years due to repercussions of COVID-19. Eventually, prevention methods will take effect and the arrival rate will hopefully decline (Alameda County 2022). The daily service rate μ_0 is equivalent to the mean of a triangular distribution with lower limit 0 weeks, upper limit 400 weeks and mode 300 weeks.

In Section 4 we will evaluate (3.3) using annual housing and shelter capacity vectors $\boldsymbol{h} = \{h_t \, \forall t \in 0, ..., T\}$ and $\boldsymbol{s} = \{s_t \, \forall t \in 0, ..., T\}$ where T is a time horizon in years. In this case we assume that any annual increase or decrease in capacity is spread evenly throughout the year, and

(3.3) becomes

$$u_d(\boldsymbol{h}, \boldsymbol{s}) = X_0 + \sum_{d'=1}^{d} \lambda_{d'} \delta t - \sum_{d'=1}^{d} \mu_0 h_{d'}^D(\boldsymbol{h}) \delta t - h_d^D(\boldsymbol{h}) - s_d^D(\boldsymbol{s}), \tag{3.4}$$

where

$$h_d^D(\boldsymbol{h}) = h_{\lfloor \frac{d}{365} \rfloor} + \frac{d - \lfloor \frac{d}{365} \rfloor}{365} (h_{\lceil \frac{d}{365} \rceil} - h_{\lfloor \frac{d}{365} \rfloor}) \tag{3.5}$$

and

$$s_d^D(\boldsymbol{s}) = s_{\lfloor \frac{d}{365} \rfloor} + \frac{d - \lfloor \frac{d}{365} \rfloor}{365} (s_{\lceil \frac{d}{365} \rceil} - s_{\lfloor \frac{d}{365} \rfloor}). \tag{3.6}$$

4 Optimization formulations

The fluid flow model represents a new approach for quickly assessing the feasibility and effectiveness of different investment plans h_t, s_t. High levels of investment earlier in our horizon will more quickly decrease the unsheltered queue. However, there are obvious cost and implementation limitations, which motivates a constrained optimization approach.

In this section, we will present different optimization formulations applied to our fluid flow queueing model. These formulations will optimize the levels of housing and shelter to be built over time, and the objective functions will attempt to minimize the unsheltered and sheltered population according to different metrics. First, we present the basic notation associated with the terms in our formulation. Section 4.1 will present a linear formulation, while Sections 4.2 and 4.3 will present more complex nonlinear formulations. The associated numerical results will be presented in Section 5. We define the following terms:

- Let subscript d denote time in days and subscript t denote time in years.
- Let T_a be the horizon (in years) over which we model the dynamics of the system while altering housing and shelter capacities, where $T_a \in \mathbb{N}$.
- Let T_b be the additional horizon (in years) over which we continue to model the dynamics of the system without altering housing or shelter capacities, where $T_b \in \mathbb{N}$. We do this in order to allow increased housing capacity to have a meaningful effect on the system over a long

period of time beyond a finite investment period.

- Let $D = (T_a + T_b) \times 365$ be the total modeling horizon in days.
- The vectors $\boldsymbol{h} = \{h_t \; \forall t \in 0, ..., T_a + T_b\}$ and $\boldsymbol{s} = \{s_t \; \forall t \in 0, ..., T_a + T_b\}$ are the model decision variables which contain continuous-valued annual housing and shelter capacities, respectively. The fluid flow model spreads annual changes in capacity equally over each day in the year, as detailed in equations (3.5) and (3.6).
- C is the total budget for building housing and shelter.
- Let c_h and c_s be the costs of increasing h_t and s_t, respectively, by 1, at any time.
- Let H_0 and S_0 be the initial housing and shelter capacities, respectively.
- Let B^h and B^s be baseline minimum annual housing and shelter build rates, respectively.
- Define $w \in (0, 1)$ as a weight between two objective terms which ensures that a sheltered queue is not penalized more than an unsheltered queue of the same size.

4.1 Linear formulation/objective 1

Our first formulation Φ_0 minimizes a linear combination of the unsheltered and sheltered queues subject to minimum build constraints and a total budget constraint. Recall that u_d and s_d are the output of the fluid flow model reporting the unsheltered and sheltered populations each day, respectively. Let $y_0(\boldsymbol{h}, \boldsymbol{s})$ be a deterministic linear objective function, evaluated using the fluid flow model equations (3.4), (3.5) and (3.6).

$$y_0(\boldsymbol{h}, \boldsymbol{s}) = \frac{1}{D} \sum_{d=1}^{D} u_d(\boldsymbol{h}, \boldsymbol{s}) + \frac{w}{D} \sum_{d=1}^{D} s_d^D(\boldsymbol{s}) \tag{4.1}$$

The following linear formulation Φ_0 is

$$
\begin{aligned}
\Phi_0 = \min_{\boldsymbol{h},\boldsymbol{s}}\ & y_0(\boldsymbol{h},\boldsymbol{s}) && && (4.2)\\
\text{s.t.}\ & \sum_{t=1}^{T_a} c_h[h_t - h_{t-1}] + c_s[s_t - s_{t-1}] \leq C && && (4.3)\\
& h_0 = H_0 && && (4.4)\\
& h_t \geq h_{t-1} + B^h && \forall t \in \{1, ..., T_a\} && (4.5)\\
& h_t = h_{T_a} && \forall t \in \{T_a + 1, ..., T_a + T_b\} && (4.6)\\
& s_0 = S_0 && && (4.7)\\
& s_t \geq s_{t-1} + B^s && \forall t \in \{1, ..., T_a\} && (4.8)\\
& s_t = s_{T_a} && \forall t \in \{T_a + 1, ..., T_a + T_b\}. && (4.9)
\end{aligned}
$$

Constraint (4.3) ensures the total budget is not exceeded. Constraints (4.4) and (4.7) enforce the initial housing and shelter capacities. Constraints (4.5) and (4.8) ensure levels of capacity are always increasing by a baseline amount, needed to ensure a sensible amount of building takes place throughout the horizon T_a. This equates to a shape constraint that says s_t and h_t must be monotonically increasing over time. Finally, constraints (4.6) and (4.9) fix h_t and s_t during the horizon T_b after the building horizon has occurred.

4.2 Nonlinear formulation/objective 2

Here we introduce a quadratic objective function to reflect the fact that neither the unsheltered nor the sheltered queue should become excessively long. Finding this balance involves a careful trade-off between building shelter (which quickly reduces the unsheltered queue) and building housing (which gives long term relief to the system, at the expense of initially large unsheltered queues). Furthermore, as seen in Alameda County, long waiting times can increase subsequent service times as people's situations may deteriorate. This further motivates the quadratic penalty on both parts of the queue. We keep the same budget constraint and constraints on increasing capacity by some minimum amount. Let $y_1(\boldsymbol{h}, \boldsymbol{s})$ be a deterministic quadratic objective function, evaluated using the fluid flow model:

$$y_1(\boldsymbol{h}, \boldsymbol{s}) = \frac{1}{D} \sum_{d=1}^{D} u_d(\boldsymbol{h}, \boldsymbol{s})^2 + \frac{w}{D} \sum_{d=1}^{D} s_d^D(\boldsymbol{s})^2. \tag{4.10}$$

Our first nonlinear formulation Φ_1 has objective (4.10) with the same constraints as in the linear formulation:

$$\begin{aligned}
\Phi_1 = \min_{\boldsymbol{h},\boldsymbol{s}} \; & y_1(\boldsymbol{h}, \boldsymbol{s}) && && (4.11)\\
\text{s.t.} \; & \sum_{t=1}^{T_a} c_h[h_t - h_{t-1}] + c_s[s_t - s_{t-1}] \le C && && (4.12)\\
& h_0 = H_0 && && (4.13)\\
& h_t \ge h_{t-1} + B^h && \forall t \in \{1, ..., T_a\} && (4.14)\\
& h_t = h_{T_a} && \forall t \in \{T_a + 1, ..., T_a + T_b\} && (4.15)\\
& s_0 = S_0 && && (4.16)\\
& s_t \ge s_{t-1} + B^s && \forall t \in \{1, ..., T_a\} && (4.17)\\
& s_t = s_{T_a} && \forall t \in \{T_a + 1, ..., T_a + T_b\}. && (4.18)
\end{aligned}$$

4.3 Nonlinear formulation/objective 3

Here we introduce different shape constraints. Instead of ensuring capacity increases by a minimum amount each year, we ensure that the rate of capacity increase must stay the same or increase over T_a to reflect the fact that the budget available for housing capacity expansion may typically grow over time and not all be available immediately. This not only requires housing to increase over time, but the rate of change must increase as well, which amounts to an increasing derivative shape constraint.

We can also require shelter investment to follow a unimodal function, whereby it increases for a given time period, and then decreases. This shape constraint has been suggested by Alameda County as a way of encouraging an initial ramp-up of shelter, but eventually excess shelter could be converted to housing to avoid permanent large shelters once the queue has been reduced. To implement this unimodality constraint on s_t, we introduce a mode T_c for the shelter capacity function over time, where $T_c \leq T_a$ and $T_c \in \mathbb{N}$. We ensure that the shelter capacity monotonically increases before T_c and monotonically decreases subsequently. Decreases in the shelter capacity correspond to shelter being decomissioned - in this case the money saved may be spent on housing. The non-linear formulation including this unimodal shape constraint and rate of change constraint is:

$$
\begin{aligned}
\Phi_2 = \min_{\boldsymbol{h},\boldsymbol{s}} \ & y_1(\boldsymbol{h},\boldsymbol{s}) && && (4.19)\\
\text{s.t. } & \sum_{t=1}^{t'} c_h[h_t - h_{t-1}] + c_s[s_t - s_{t-1}] \leq C && \forall t' \in \{1, ..., T_a\} && (4.20)\\
& h_0 = H_0 && && (4.21)\\
& h_t \geq h_{t-1} && \forall t \in \{1, ..., T_a\} && (4.22)\\
& h_t = h_{T_a} && \forall t \in \{T_a + 1, ..., T_a + T_b\} && (4.23)\\
& h_{t+1} - h_t \geq h_t - h_{t-1} && \forall t \in \{1, ..., T_a - 1\} && (4.24)\\
& s_0 = S_0 && && (4.25)\\
& s_t \geq s_{t-1} && \forall t \in \{1, ..., T_c\} && (4.26)\\
& s_t \leq s_{t-1} && \forall t \in \{T_c + 1, ..., T_a\} && (4.27)\\
& s_t \geq s_0 && \forall t \in \{T_c + 1, ..., T_a\} && (4.28)\\
& s_t = s_{T_a} && \forall t \in \{T_a + 1, ..., T_a + T_b\}. && (4.29)
\end{aligned}
$$

Constraints (4.20) ensure the total budget is never exceeded. Here a single budget constraint as in our previous formulations is not enough, since then the total budget could be exceeded in one year as long as a saving was subsequently made from decommissioning shelter. With this set of constraints we ensure that at no point can the total expenditure to that point exceed the total budget, so any savings from decommissioning shelter cannot be spent before they are made. Constraints (4.22) ensure the housing capacity monotonically increases, while constraints (4.24) ensure the rate of change of housing capacity also monotonically increases from year to year. Constraints (4.26) ensure the shelter capacity monotonically increases up to the mode T_c and constraints (4.27) ensure it subsequently decreases monotonically. Finally, constraints (4.28) ensure the shelter capacity never drops below its initial capacity.

Table 1: Model parameters

Parameter	**Value (Φ_0)**	**Value (Φ_1)**	**Value (Φ_2)**
T_a	5 years	5 years	5 years
T_b	5 years	5 years	5 years
T_c	-	-	3 years
λ_d	$\frac{10}{day}\ \forall d \in \{1, ..., T_a \times 365\}$ $\frac{6}{day}\ \forall d \in \{T_a \times 365 + 1, ..., D\}$	$\frac{10}{day}\ \forall d \in \{1, ..., T_a \times 365\}$ $\frac{6}{day}\ \forall d \in \{T_a \times 365 + 1, ..., D\}$	$\frac{10}{day}\ \forall d \in \{1, ..., T_a \times 365\}$ $\frac{6}{day}\ \forall d \in \{T_a \times 365 + 1, ..., D\}$
X_0	12,000 people	12,000 people	12,000 people
h_0	4,000 units	4,000 units	4,000 units
s_0	1,500 units	1,500 units	1,500 units
c_h	30,000 USD/unit	30,000 USD/unit	30,000 USD/unit
c_s	10,000 USD/unit	10,000 USD/unit	10,000 USD/unit
C	200,000,000 USD	200,000,000 USD	200,000,000 USD
B^h	500 units	500 units	-
B^s	500 units	500 units	-
μ_0	6.106×10^{-4}/day	6.106×10^{-4}/day	6.106×10^{-4}/day
w	0.3	0.3	0.3

5 Numerical Results

In Table 1 we list the model parameters we used when optimising formulations Φ_0, Φ_1 and Φ_2. These approximate values are taken from Alameda County (2022) and Singham et al. (2023). We choose $B^h, B^s = 500$ units in order to allocate half of the budget C to meeting the minimum build constraint and allow the remaining half to be spent in an optimal way. We choose w to be sufficiently high to give a meaningful penalty to shelter but without undermining its advantage over an unsheltered queue. Additionally, while we use a current estimate of the arrival rate of 10/day for the first T_a years of the modeling horizon, we anticipate with major local prevention efforts (Regional Impact Council 2021), the arrival rate could potentially drop significantly to an estimate of 6/day.

In Figures 4, 5 and 6 we illustrate the model dynamics for the optimal solutions to Φ_0, Φ_1 and Φ_2, respectively. For Φ_0, the optimal values of h_t, s_t and the corresponding fluid flow model output u_t is displayed in Figure 4. We prefer to spend all surplus budget (beyond what is needed for the baseline capacity) in the first year on housing. There is no incentive to spend the surplus budget later when the effect would be diminished. The benefit (on the objective value) per USD spent on housing in the first year is greater than the equivalent benefit of shelter. If we were to increase c_h,

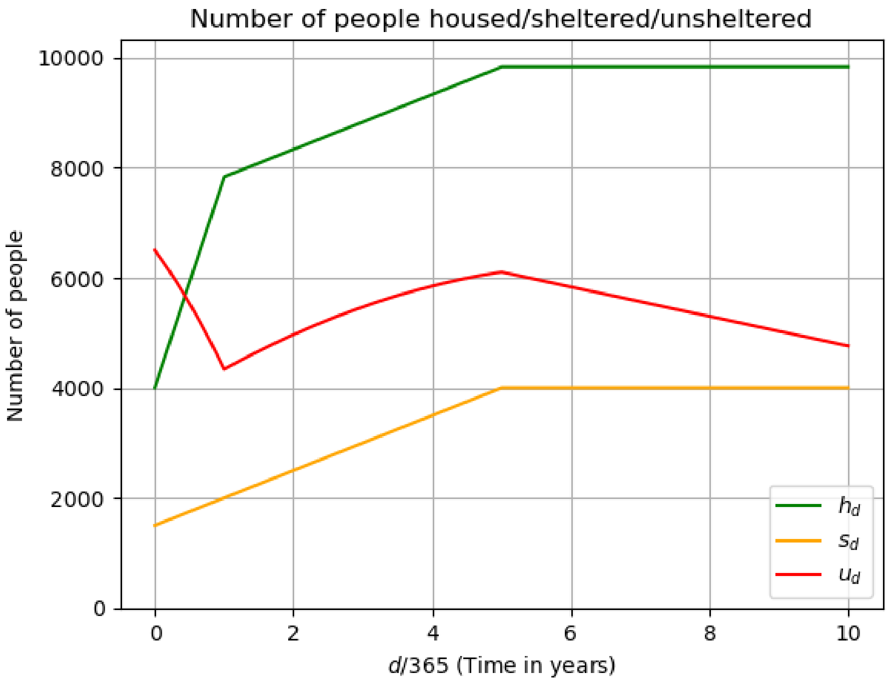


Figure 4: Optimal solution for Φ_0.

then the benefit per USD of housing would decrease as fewer houses could be built. If we were to decrease the housing service rate μ_0, the benefit would also get worse. In either case, with sufficient change, the benefit per USD spent on shelter may surpass that of housing, and building shelter would become preferable. This would also happen if we were to reduce the cost c_s sufficiently, enabling more shelters to be built per USD. Due to the linearity of the objective function, it would never be preferential to spend the surplus budget on a mixture of housing and shelter.

The results for Φ_1 are dispayed in Figure 5. With this nonlinear formulation, we still prefer to spend all surplus budget in the first year, but there is now a preference for a mixture of extra housing and extra shelter. This is because the quadratic penalty associated with a high unsheltered population encourages shelter which quickly reduces the size of the unsheltered queue. However, the quadratic penalty of having a large sheltered population encourages early investment in housing above the minimum. This housing investment in time also has a meaningful effect on reducing the unsheltered queue, since sufficient houses may be built to have a total service rate higher than the arrival rate, thus bringing stability to the system.

With Φ_2 (results in Figure 6), we can see the effect of shape constraints. We note that the initial ramp up of shelter is able to bring the unsheltered queue down in the short term. The rate

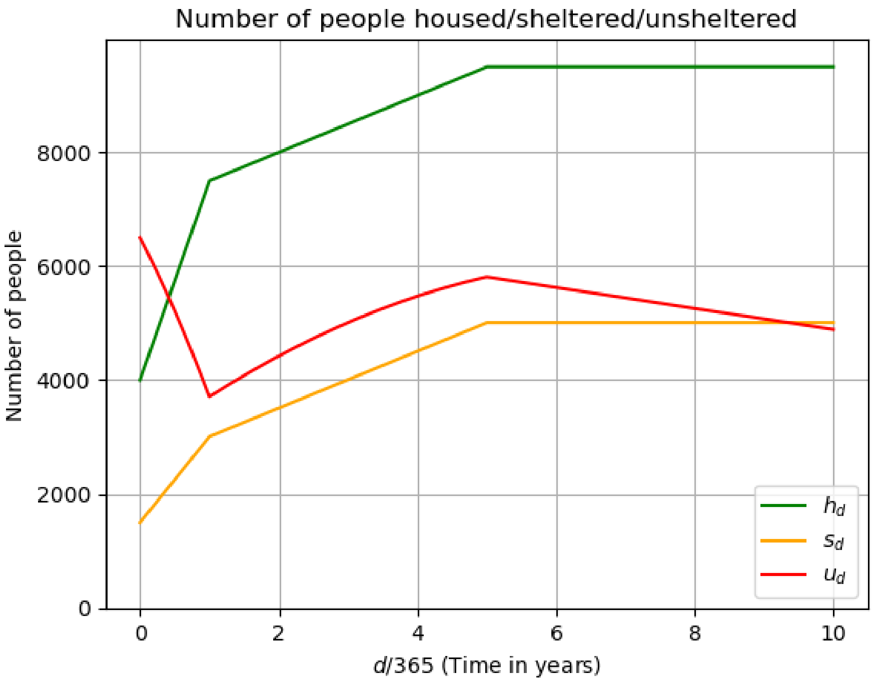


Figure 5: Optimal solution for Φ_1.

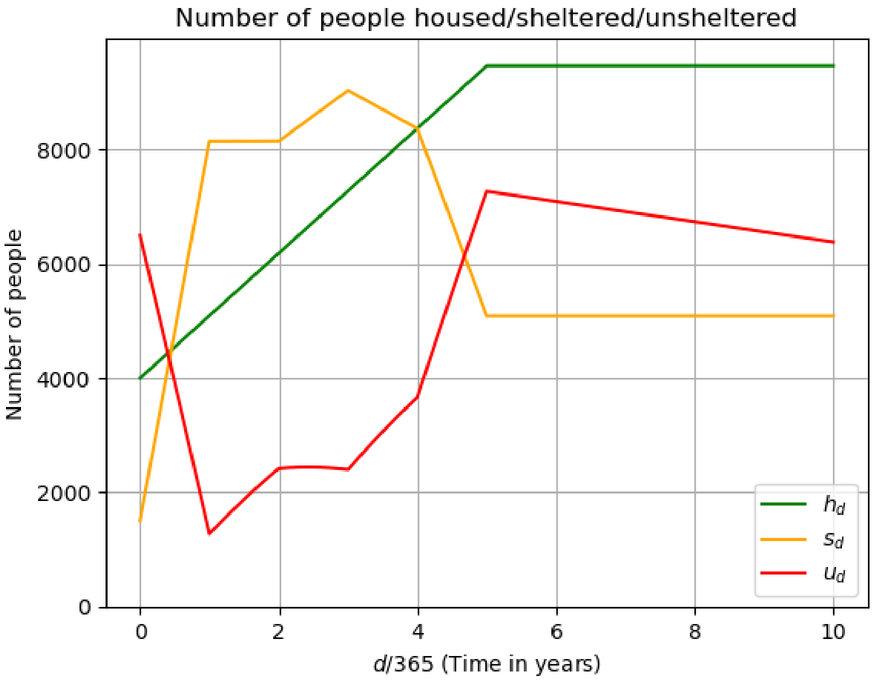
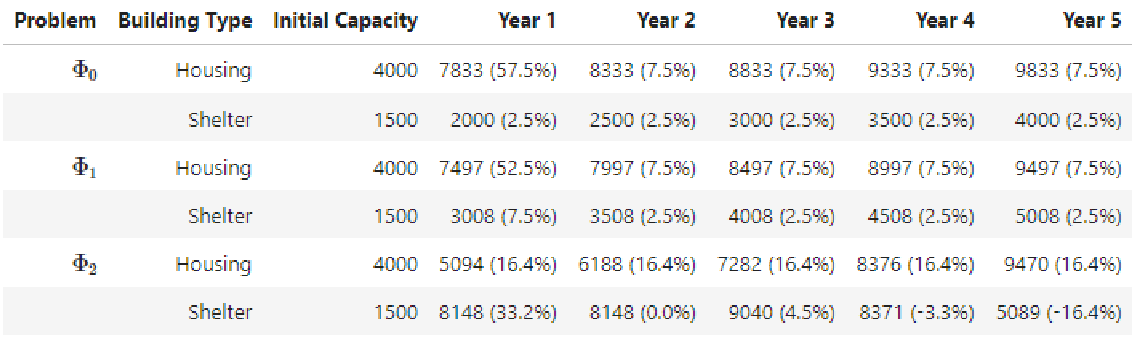


Figure 6: Optimal solution for Φ_2.

of increase in the housing capacity must not decrease over time so we see a more steady increase in housing compared to previous solutions. The total amount of housing we can build is affected by the fact that after the shelter mode at $t = 3$ years, decommissioning shelter makes more budget available for housing. Thus we are able to achieve sufficient housing for a stable system in the long-term, while affording immediate relief to the system via shelter. We note that with this formulation, for every 3 shelters decommissioned, 1 house may be acquired, resulting in 2 people immediately rejoining the unsheltered part of the queue. Although this enables the housing capacity to increase which is good for long-term relief to the system, the immediate effect is undesirable in practice and we see that after 5 years the unsheltered queue is again very large. An alternative formulation may enforce a more controlled decommissioning process by, for example, including a shape constraint on the total number of housing and shelter units.



Problem	Building Type	Initial Capacity	Year 1	Year 2	Year 3	Year 4	Year 5
Φ_0	Housing	4000	7833 (57.5%)	8333 (7.5%)	8833 (7.5%)	9333 (7.5%)	9833 (7.5%)
	Shelter	1500	2000 (2.5%)	2500 (2.5%)	3000 (2.5%)	3500 (2.5%)	4000 (2.5%)
Φ_1	Housing	4000	7497 (52.5%)	7997 (7.5%)	8497 (7.5%)	8997 (7.5%)	9497 (7.5%)
	Shelter	1500	3008 (7.5%)	3508 (2.5%)	4008 (2.5%)	4508 (2.5%)	5008 (2.5%)
Φ_2	Housing	4000	5094 (16.4%)	6188 (16.4%)	7282 (16.4%)	8376 (16.4%)	9470 (16.4%)
	Shelter	1500	8148 (33.2%)	8148 (0.0%)	9040 (4.5%)	8371 (-3.3%)	5089 (-16.4%)

Figure 7: Optimal capacity per year (Proportion of total budget spent per year)

In Figure 7 we list the optimal solutions to Φ_0, Φ_1 and Φ_2 in terms of the capacity at the end of each year and the proportion of the total budget spent on building in that year. Negative budget spent corresponds to a saving made by decommissioning shelter. All optimal solutions spend the maximum possible budget of 200,000,000 USD. The solution to Φ_0 sees the biggest early investment in housing as it is preferable to shelter according to the given linear objective function. With the quadratic objective function, the solution to Φ_1 sees a slightly smaller early investment in housing along with a slightly bigger early ramp up of shelter, compared to Φ_0. The solution to Φ_2, in contrast, sees a large early ramp up of shelter and a steady investment in housing over time. In years 4 and 5 we see decommissioning of shelter in the solution to Φ_2 to enable the continued investment in housing.

In Figure 8 we compare the dynamics of the unsheltered queue for each optimal solution. All solutions see an initial drop in the unsheltered queue as early investment is made, followed by a subsequent rise as capacity slowly catches up with demand. Then there is a decrease as the arrival rate drops and enough houses have been built to bring stability to the system. In comparison to Φ_0, the solution to Φ_1 gives greater immediate relief to the system via shelter but less long-term relief via housing. The solution to Φ_2 gives substantial short-term relief to the system. Long-term relief to the system is slower here with a more realistic gradual increase in housing capacity, enforced by the shape constraints.

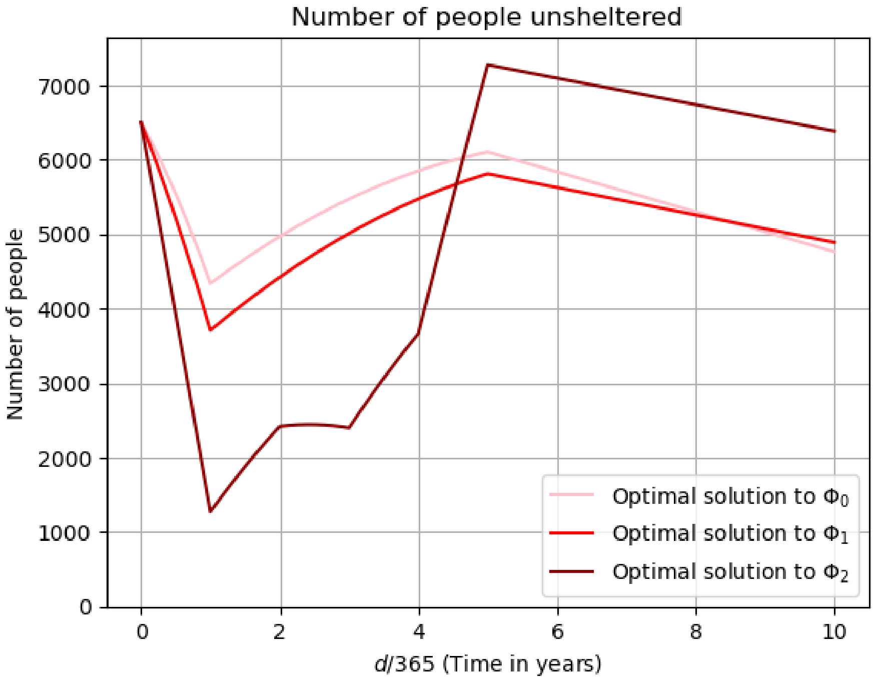


Figure 8: Unsheltered queue for each optimal solution

We solved all problems in Pyomo using the GLPK solver for Φ_0 and the IPOPT solver for Φ_1 and Φ_2. Problems Φ_0, Φ_1 and Φ_2 were solved in 0.662, 0.760 and 0.759 seconds, respectively. All code used for this analysis is publicly available at `https://github.com/grahamburgess3/psor-paper-housing`.

6 Conclusions

Most capacity planning formulations we have reviewed in the literature consider capacity expansion from a single-stage perspective, in that the decision-maker has one shot to choose and optimize a fixed capacity to accommodate the queueing system. In reality, most public sector services do not have the resources to instantaneously ramp up to the ideal capacity, as there may be budgetary or time constraints that control this rate. A model that accounts for these limits in capacity expansion over time will provide a more realistic and executable plan, hence we attempt to provide a method for determining how to allocate resources over time. While housing is the primary resource and is modeled as the main server system, we also model investment into shelter, which supports some of the people in the queue, while not modeled as a server.

Few models exist for modeling the flow of the homeless population through a CoC, especially for locations like Alameda County where there is clearly a major lack of resources compared to demand. We develop a fluid flow queueing model to track the unsheltered population over time given an investment policy into housing and shelter. This model is uniquely poised to account for the instability the system and the currently high queueing backlog. Our model is amenable to optimization, so we construct different formulations to balance the desire for high levels of housing at high cost against cheaper shelter options. In addition to budgetary constraints, we employ shape constraints as a means of ensuring our investment function output is feasible from a policy-making and implementation standpoint. The idea of a unimodal function for shelter investment has been suggested by Alameda County, and such shape constraints can easily be implemented in our framework.

There are many opportunities for future work. Exploring alternative objective functions and constraints would reveal many alternative formulations. For example, smoothness constraints on the unimodal shelter capacity function may give more practical solutions that appear reasonable to constituents. Further constraints to control the decommissioning of shelter may also be appropriate. A bi-objective formulation would likely give further insight into the trade-off between short-term relief to the system via shelter and long-term relief via housing. A natural next step would be to develop a stochastic optimization model to account for parameter uncertainty.

Acknowledgements

This paper is based on work completed while Graham Burgess was part of the EPSRC funded STOR-i centre for doctoral training (EP/S022252/1). The authors want to acknowledge the help of Dr. Rob Shone and Dr. Dave Worthington in discussions for writing up this paper. Additionally, we thank the Alameda County Office of Homeless Care and Coordination, and the San Francisco

Department of Homelessness and Supportive Housing for valuable discussions motivating this work.

References

Alameda County. Home Together 2026 Community Plan: A 5-year Strategic Framework Centering Racial Equity to End Homelessness in Alameda County. April 2022.

Priyank Arora, Morvarid Rahmani, and Karthik Ramachandran. Doing Less to do More? Optimal Service Portfolio of Non-Profits that Serve Distressed Individuals. *Manufacturing & Service Operations Management*, 2021.

Burcu Balcik, Seyed Iravani, and Karen Smilowitz. Multi-Vehicle Sequential Resource Allocation for a Nonprofit Distribution System. *IIE Transactions*, 46(12):1279–1297, 2014.

Achal Bassamboo, Ramandeep S Randhawa, and Assaf Zeevi. Capacity Sizing Under Parameter Uncertainty: Safety Staffing Principles Revisited. *Management Science*, 56(10):1668–1686, 2010.

Omar Besbes, Francisco Castro, and Ilan Lobel. Spatial Capacity Planning. *Operations Research*, 70(2): 1271–1291, 2022.

KM Bretthauer. Capacity Planning in Networks of Queues with Manufacturing Applications. *Mathematical and Computer Modelling*, 21(12):35–46, 1995.

Hong Chen and David D Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*, volume 4. Springer, 2001.

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

HUD Exchange. HUD Exchange, 2022. URL `https://www.hudexchange.info`.

Tanvi A Ingle, Maike Morrison, Xutong Wang, Timothy Mercer, Vella Karman, Spencer Fox, and Lauren Ancel Meyers. Projecting COVID-19 Isolation Bed Requirements for People Experiencing Homelessness. *PLOS One*, 16(5):e0251153, 2021.

Navid Izady. Appointment Capacity Planning in Specialty Clinics: A Queueing Approach. *Operations Research*, 63(4):916–930, 2015.

Navid Izady and Dave Worthington. Setting Staffing Requirements for Time Dependent Queueing Networks: The Case of Accident and Emergency Departments. *European Journal of Operational Research*, 219 (3):531–540, 2012.

Nanjing Jian and Shane G Henderson. An Introduction to Simulation Optimization. In L. Yilmaz, W. K. V.

Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, editors, *Proceedings of the 2015 Winter Simulation Conference.*, pages 1780–1794, Piscataway, New Jersey, 2015. IEEE, Institute of Electrical and Electronics Engineers, Inc.

Yaren Bilge Kaya and Kayse Lee Maass. Leveraging Priority Thresholds to Improve Equitable Housing Access for Unhoused-At-Risk Youth. *arXiv preprint arXiv:2212.03777*, 2022.

Yaren Bilge Kaya, Kayse Lee Maass, Geri L Dimas, Renata Konrad, Andrew C Trapp, and Meredith Dank. Improving Access to Housing and Supportive Services for Runaway and Homeless Youth: Reducing Vulnerability to Human Trafficking in New York City. *IISE Transactions*, 56(3):296–310, 2024.

Kurtis Konrad and Yunan Liu. Achieving Stable Service-Level Targets in Time-Varying Queueing Systems: A Simulation-Based Offline Learning Staffing Algorithm. In C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, editors, *Proceedings of the 2023 Winter Simulation Conference.*, pages 327–338, Piscataway, New Jersey, 2023. IEEE, Institute of Electrical and Electronics Engineers, Inc.

Henry Lam, Zhenyuan Liu, and Xinyu Zhang. Orthounimodal Distributionally Robust Optimization: Representation, Computation and Multivariate Extreme Event Applications. *arXiv preprint arXiv:2111.07894*, 2021.

Henry Lam, Zhenyuan Liu, and Dashi Singham. Shape-Constrained Distributional Optimization via Importance-Weighted Sample Average Approximation. *https://arxiv.org/pdf/2406.07825*, 2024.

Jingang Liu, Feng Yang, Hong Wan, and John W Fowler. Capacity Planning Through Queueing Analysis and Simulation-Based Statistical Methods: A Case Study for Semiconductor Wafer Fabs. *International Journal of Production Research*, 49(15):4573–4591, 2011.

Kayse Lee Maass, Andrew C Trapp, and Renata Konrad. Optimizing Placement of Residential Shelters for Human Trafficking Survivors. *Socio-Economic Planning Sciences*, 70:100730, 2020.

Frederick Miller, Yaren Bilge Kaya, Geri L Dimas, Renata Konrad, Kayse Lee Maass, Andrew C Trapp, et al. On the Optimization of Benefit to Cost Ratios for Public Sector Decision Making. *arXiv preprint arXiv:2212.04534*, 2022.

Yuval Nov, Gideon Weiss, and Hanqin Zhang. Fluid Models of Parallel Service Systems Under FCFS. *Operations Research*, 70(2):1182–1218, 2022.

Oakland-Berkeley-Alameda County CoC. Centering Racial Equity in Homeless System Design, December 2020. URL `https://everyonehome.org/wp-content/uploads/2021/02/2021-Centering-Racial-Equity-in-Homeless-System-Design-Full-Report-FINAL.pdf`.

Office of Housing and Urban Development, June 2024. URL `https://www.hudexchange.info/programs/coc/coc-program-eligibility-requirements/`.

Jonathan Patrick. Access to Long-Term Care: The True Cause of Hospital Congestion? *Production and Operations Management*, 20(3):347–358, 2011.

Jonathan Patrick, K Nelson, and Dan Lane. A Simulation Model for Capacity Planning in Community Care. *Journal of Simulation*, 9(2):111–120, 2015.

Aida Rahmattalabi, Phebe Vayanos, Kathryn Dullerud, and Eric Rice. Learning Resource Allocation Policies from Observational Data with an Application to Homeless Services Delivery. In *022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, New York, NY, 2022. ACM. doi: https://doi.org/10.1145/3531146.3533181.

Regional Impact Council. Regional Action Plan: A Call to Action from the Regional Impact Council, 2021. URL `https://www.allhomeca.org/wp-content/themes/allhome/library/images/plan/210413_Regional_Action_Plan_Final.pdf`.

Jared Reynolds, Zhen Zeng, Jingshan Li, and Shu-Yin Chiang. Design and Analysis of a Health Care Clinic for Homeless People Using Simulations. *International Journal of Health Care Quality Assurance*, 2010.

Dashi I Singham. Estimating Quantile Fields for a Simulated Model of a Homeless Care System. In C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, editors, *Proceedings of the 2023 Winter Simulation Conference.*, Piscataway, New Jersey, 2023. Institute of Electrical and Electronics Engineers, Inc.

Dashi I Singham, Jennifer Lucky, and Stephanie Reinauer. Discrete-Event Simulation Modeling for Housing of Homeless Populations. *PLOS One*, April 2023.

Shaler Stidham Jr. *Optimal Design of Queueing Systems.* Chapman and Hall/CRC, 2009.

David Worthington. Hospital Waiting List Management Models. *Journal of the Operational Research Society*, 42:833–843, 1991.

Mohammad H Yarmand and Douglas G Down. Server Allocation for Zero Buffer Tandem Queues. *European Journal of Operational Research*, 230(3):596–603, 2013.

Mohammad H Yarmand and Douglas G Down. Maximizing Throughput in Zero-Buffer Tandem Lines with Dedicated and Flexible Servers. *IIE Transactions*, 47(1):35–49, 2015.