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

Graham C. M. Burgess [a], Dashi I. Singham [b] and Luke A. Rhodes-Leader [c]

[a] STOR-i Centre for Doctoral Training, Lancaster University, Lancaster, UK
[b] Operations Research Department, Naval Postgraduate School, Monterey, USA
[c] Department of Management Science, Lancaster University, Lancaster, UK

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. 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 single adults 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 size of the unsheltered population given cost constraints by varying investment in housing and shelter over time. Additionally, we impose policy-based shape constraints to reflect the time-dependence and feasibility constraints associated with planning decisions. We thus show how resources can be allocated between housing and shelter over time. While this joint optimization approach can be used to analyze homeless populations outside of Alameda County, it also 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 in Alameda County 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. Shelter is often congregate, in that many people will be grouped together in a shared space. Permanent housing may include social services like drug rehabilitation, mental health support, and other health and medical services. This makes it more expensive than shelter, but also

more desirable because it is more likely to lead to long-term success in terms of clients remaining successfully housed and having better health outcomes. The goal is to successfully house clients as quickly as possible, with shelter serving as a backstop when housing is limited.

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. Increasing h_t and s_t will decrease u_t. 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.

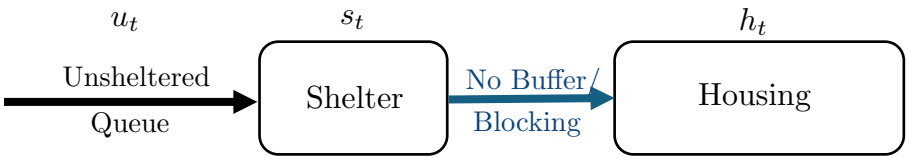

Figure 1: Tandem queue model of housing and shelter.

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.

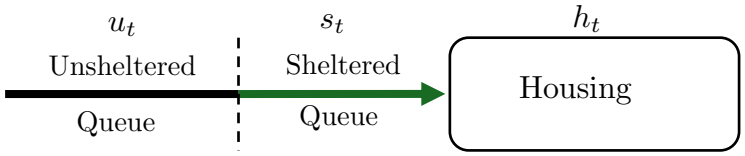

Figure 2: $M_t/G/h_t$ model of housing and shelter.

In reality, people may move in and out of shelter while waiting for housing. However, because Alameda County faces a high unsheltered population, we model the shelters as always full using a fluid flow model and the effect of individual transfers in and out of shelter is negligible. 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 series 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 for achieving 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 developing 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 approach 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. We then formulate objective functions which model the performance of a given capacity plan using the fluid flow model.

An additional feature that we incorporate to improve real-world feasibility is to include policy-based shape constraints on 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.

Our modeling choices are motivated by numerous discussions with leadership of the Alameda Office of Homeless Care and Coordination, and the San Francisco Department of Homelessness and Supportive Housing. One author joined bi-weekly virtual data analysis meetings with Alameda County during 2022. The group was analyzing the results of survey work performed with stakeholders across housing organizations and focus groups including people with lived homeless experience. Two authors met in person in 2023 with the now-Director of Alameda County's Office of Homeless Care and Coordination, as well as leadership and analysts in San Francisco's Homelessness and Supportive Housing Office. Our desire to find optimal capacity functions over time using a queueing model is directly motivated by planning needs faced by both counties.

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 presents the results of sensitivity analysis for the optimal results, while Section 7 concludes.

2 Literature Review

This section will briefly review the literature related to our approach. We will first consider the capacity sizing problem for queueing systems. Much of this research operates in a healthcare setting and support for people experiencing homelessness often involves aligning homeless services closely with healthcare resources. Next, we will discuss research that approaches resolving homelessness from a healthcare perspective. Finally, 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). In the context of homelessness, we cannot necessarily control the arrival and service rates, but we can potentially control the levels of housing and shelter. This motivates us to consider varying server capacities over time. 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.

To the best of our knowledge, there are very few examples of long-term capacity planning studies which strive for not only an optimal future capacity, but also optimal intermediate steps. Mohammadi Bidhandi et al. (2019) optimize future capacity across a network of community health services but only *model* the change from the status-quo to the optimal capacity, rather than optimize it. Zhang et al. (2012) propose the use of a simulation optimization approach called bisection search for setting long-term capacity for beds in care facilities. From year to year they find the minimum number of beds required to have a high probability of meeting set service levels. Each search step requires multiple runs of a discrete-event simulation model. Lin et al. (2024) use this approach to optimize annual care capacity over a twenty year planning horizon for an elderly care system in China. In contrast to simulation optimization, we aim to use a fast deterministic model to optimize

long-term capacity plans.

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). While the homeless housing system can be thought of as a tandem queue (as in Figure 1), because clients only stay in shelter if housing is not available, we are able to reformulate the tandem queue to the format in Figure 2.

The application of operations research approaches to modeling specific homelessness solutions is an active area of research. 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).

In particular, there have been recent efforts to model homelessness as a queueing problem. 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. Coordinated entry has also emerged as an important federally-mandated mechanism for streamlining client entry into county managed homeless systems to enable efficient matching with available resources, and tracking wait-lists for various types of shelter and housing. Clients experiencing homelessness can enter the system at coordinated entry access points, where their needs will be managed as part of a centralized system to avoid attempting to find housing at separate individually managed facilities (U.S. Department of Housing and Urban Development, 2024). Managing routing of clients through a queueing system using coordinated entry in San Francisco is simulated in Singham et al. (2025). While these papers attempt to model specific policies and system behaviors, we are motivated to construct a simplified model that can directly be optimized to suggest long-term planning solutions to guide policymakers.

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. In particular, we will suggest that investment into housing increases monotonically over time to encourage increased investment into the system. Additionally, we will test the effects of a unimodal shelter function, whereby shelter will initially increase to serve the current unsheltered population, but is allowed decrease after additional permanent housing has been constructed.

3 Queueing model for a homeless care system

The previous section discussed many papers that model and simulate the specific movements of people through the homeless system (Kaya and Maass, 2022; Kaya et al., 2024; Singham et al., 2023, 2025). In contrast to this work, we develop a simplified queueing model that will allow us to optimize long-term planning investments in housing and shelter. This analysis will enable Alameda County to estimate the overall level of investment needed to reduce the unsheltered queue, and will help determine the relative allocation of investment into housing versus shelter over time. Because there are thousands of people in the system, a fast model that considers the general flow of clients through the system as in Figure 2 could approximate aggregate behavior effectively. A corresponding optimization formulation could then suggest an optimal capacity plan that could then be applied to a more detailed simulation model, such as the one in (Singham et al., 2023).

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 assumption that there is always a queue motivates the use of a fluid flow model, described next.

3.1 Fluid flow model

The overloaded state of this efficiency driven regime motivates our use of a fluid flow model. Since servers are always occupied, the outflow from the system is independent of the inflow. The mean queue length can then be easily approximated using fluid flow equations. This involves modeling a continuous-valued number of people which is reasonable when the numbers are large. Because the assumption of an efficiency driven regime with a large inflow of customers relative to system outflow is realistic for many homeless systems in California, we employ a fluid flow model to allow for tractable optimization. This optimization is fast in part due to the low computational cost of our fluid flow model. 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).

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 care 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

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

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

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

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 (3.1) with the sum

$$u_d = X_0 + \sum_{d'=1}^{d} \lambda_{d'} \delta t - \sum_{d'=1}^{d} \mu_0 h_{d'}^D \delta t - h_d^D - s_d^D, \tag{3.2}$$

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 estimates for X_0, μ_0 and λ_d, h_d^D, s_d^D for all $d \in \{1, ..., D\}$ based on Singham et al. (2023). The arrival rate first increases and then subsequently decreases as independent prevention efforts take effect (Alameda County, 2022). The daily service rate μ_0 is equivalent to a service time of 4 years.

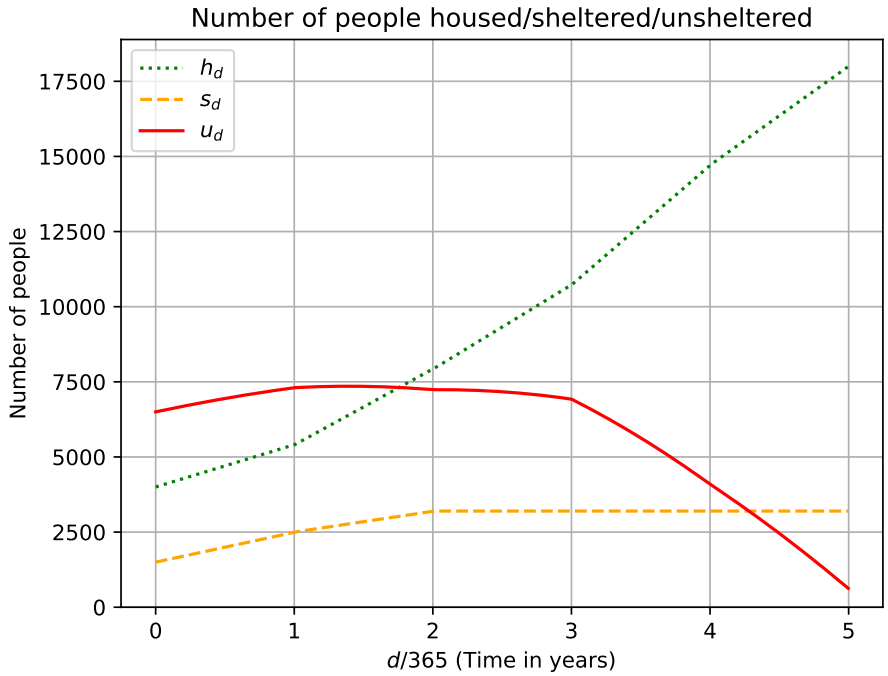

Figure 3: Dynamics of u_d, s_d^D and h_d^D. $X_0 = 12000$, $\mu_0 = \frac{1}{4 \times 365 \text{ days}}$, Daily arrival rates λ_d in each year: $12.0, 13.2, 13.2, 11.9, 10.7$.

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. We emphasize that Figure 3 displays the results for optimistic current estimates which cannot necessarily be extrapolated to future years. The user may use our optimization model to recompute the figure using adjusted arrival and service rates representing actual conditions in future years if more information becomes available.

In Section 4 we will evaluate (3.2) 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 (3.2) becomes

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

where

$$h_d^D(\boldsymbol{h}) = h_{\lfloor \frac{d}{365} \rfloor} + \left(\frac{d}{365} - \left\lfloor \frac{d}{365} \right\rfloor \right) (h_{\lceil \frac{d}{365} \rceil} - h_{\lfloor \frac{d}{365} \rfloor}) \tag{3.4}$$

and

$$s_d^D(\boldsymbol{s}) = s_{\lfloor \frac{d}{365} \rfloor} + \left(\frac{d}{365} - \left\lfloor \frac{d}{365} \right\rfloor \right) (s_{\lceil \frac{d}{365} \rceil} - s_{\lfloor \frac{d}{365} \rfloor}). \tag{3.5}$$

3.2 Modeling Assumptions and Extensions

This section describes the assumptions required for our model, and also the ways in which our model can be extended to other settings by changing the inputs to our code.

3.2.1 Queue structure

The main assumption motivating the use of a fluid flow model is that the housing servers are always busy and there exists a queue. This motivates the use of an aggregate model, with which we are looking to optimize a capacity plan for the entire system. In the aggregate model we consider a single queue for the entire system, which does not necessarily represent an actual physical list of clients waiting for housing, but rather the county-wide count of unsheltered people that require resources.

We do not consider individual needs of clients for differentiated types of housing services. We do allow for a non-homogeneous arrival process, but assume the service rate for each housing unit stays constant for the duration of the model run (the overall service rate will change with changing levels of housing). We also assume that the housing and shelter capacity changes occurring in a given year are spread equally throughout the year.

Additionally, we do not consider abandonment from the queue. In reality, there will be a lot of friction in the system preventing a simple first-in-first-out execution. For example, clients need to be eligible sobriety-wise for some types of housing, or priority will be given to clients with certain physical or mental health needs. Since our model considers planning at an aggregate scale which assumes there will always be clients available to fill empty housing spots, we can absorb friction from abandonment and other implementation issues into the service time at housing. These simplifying assumptions are all directly in-line with Alameda County's Home Together Plan (Alameda County, 2022) which attempts to assess the overall costs and housing needs for the coming year.

3.2.2 Client type

The model is designed to treat clients arriving to the system as homogeneous. The homeless population in Alameda County consists largely of single adults, with 91% of clients falling into this category (Oakland-Berkeley-Alameda County CoC, 2020). Alameda County has developed parallel modeling analyses for single adults, families and youths, by using the same modeling approach but varying the data inputs. In this paper, we focus on a model for individual homelessness so that one person should be allocated one housing unit. However, by adjusting the arrival rates and service rates for a different type of customer (i.e., families, couples, or youths) the same formulations could be applied as long as there was a one-to-one matching between a client and a type of housing unit. We calibrate our numerical results using the single adult category which has greatest amount of housing need in Alameda County and supports the assumptions required for using a fluid flow model.

3.2.3 Numerical results and extensions

We make certain assumptions in our displayed numerical results that are not required to extend the model and formulations to other settings. For example, our numerical results will show a particular sequence of arrival rates over time. The arrival rates can be modified for each year by changing the input files to the optimization, while we assume the service rate for a housing unit stays constant

(though this constant value can be varied by the user). Section 6 will demonstrate the optimal objective sensitivity from varying the arrival and service rate inputs. Finally, while our numerical formulations contain constraints requiring housing capacity to increase over time, this constraint can be removed allowing housing capacity to decrease. We suggest some shape constraints in the formulations to be presented in Section 4, and additional shape constraints may be included as desired.

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 with sensitivity analysis in Section 6. 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 (3.4) and (3.5).

- 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, where $B^h, B^s > 0$.
- 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 (3.3), (3.4) and (3.5):

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

The following linear formulation Φ_0 is

$$\Phi_0 = \min_{\boldsymbol{h},\boldsymbol{s}} \; y_0(\boldsymbol{h}, \boldsymbol{s}) \tag{4.2}$$

$$\text{s.t.} \sum_{t=1}^{T_a} c_h[h_t - h_{t-1}] + c_s[s_t - s_{t-1}] \leq C \tag{4.3}$$

$$h_0 = H_0 \tag{4.4}$$

$$h_t \geq h_{t-1} + B^h \qquad \forall t \in \{1, ..., T_a\} \tag{4.5}$$

$$h_t = h_{T_a} \qquad \forall t \in \{T_a + 1, ..., T_a + T_b\} \tag{4.6}$$

$$s_0 = S_0 \tag{4.7}$$

$$s_t \geq s_{t-1} + B^s \qquad \forall t \in \{1, ..., T_a\} \tag{4.8}$$

$$s_t = s_{T_a} \qquad \forall t \in \{T_a + 1, ..., T_a + T_b\}. \tag{4.9}$$

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. This constraint is needed to ensure the optimization does not allocate all resources towards building in time 0, which would not be feasible from a long-term implementation planning perspective since continual investment is easier to budget and justify to stakeholders. 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 (though we do not model state-dependent service times here). 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:

$$\Phi_1 = \min_{\boldsymbol{h},\boldsymbol{s}} \; y_1(\boldsymbol{h}, \boldsymbol{s}) \tag{4.11}$$

$$\text{s.t.} \sum_{t=1}^{T_a} c_h[h_t - h_{t-1}] + c_s[s_t - s_{t-1}] \leq C \tag{4.12}$$

$$h_0 = H_0 \tag{4.13}$$

$$h_t \geq h_{t-1} + B^h \qquad \forall t \in \{1, ..., T_a\} \tag{4.14}$$

$$h_t = h_{T_a} \qquad \forall t \in \{T_a + 1, ..., T_a + T_b\} \tag{4.15}$$

$$s_0 = S_0 \tag{4.16}$$

$$s_t \geq s_{t-1} + B^s \qquad \forall t \in \{1, ..., T_a\} \tag{4.17}$$

$$s_t = s_{T_a} \qquad \forall t \in \{T_a + 1, ..., T_a + T_b\}. \tag{4.18}$$

4.3 Nonlinear formulation/objective 3

Here we introduce different shape constraints to show the flexibility of our framework. 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 not decrease as well, which amounts to a non-decreasing derivative shape constraint. We note that this assumption is not realistic in practice, in that planners cannot usually guarantee increasing rates of investment levels over time due to fluctuating budgets and resource availability. However, planners may wish to show the effects of an ideal investment situation where there is continually increasing investment into housing and shelter. Thus, this constraint demonstrates the feasibility of adding custom time-varying shape constraints to our formulations.

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 decommissioned, which simply means the objective function shelter penalty decreases. Furthermore, there is a financial saving when shelter capacity decreases, in which case the second term in the summation in our budget constraint is negative. The planned implementation is for excess shelter to be converted to housing, but we don't model that conversion explicitly here. The non-linear formulation including this unimodal shape constraint and rate of change constraint is:

$$\Phi_2 = \min_{\boldsymbol{h},\boldsymbol{s}} \; y_1(\boldsymbol{h}, \boldsymbol{s}) \tag{4.19}$$

$$\text{s.t.} \sum_{t=1}^{t'} c_h[h_t - h_{t-1}] + c_s[s_t - s_{t-1}] \leq C \qquad \forall t' \in \{1, ..., T_a\} \tag{4.20}$$

$$h_0 = H_0 \tag{4.21}$$

$$h_t \geq h_{t-1} \qquad \forall t \in \{1, ..., T_a\} \tag{4.22}$$

$$h_t = h_{T_a} \qquad \forall t \in \{T_a + 1, ..., T_a + T_b\} \tag{4.23}$$

$$h_{t+1} - h_t \geq h_t - h_{t-1} \qquad \forall t \in \{1, ..., T_a - 1\} \tag{4.24}$$

$$s_0 = S_0 \tag{4.25}$$

$$s_t \geq s_{t-1} \qquad \forall t \in \{1, ..., T_c\} \tag{4.26}$$

$$s_t \leq s_{t-1} \qquad \forall t \in \{T_c + 1, ..., T_a\} \tag{4.27}$$

$$s_t \geq s_0 \qquad \forall t \in \{T_c + 1, ..., T_a\} \tag{4.28}$$

$$s_t = s_{T_a} \qquad \forall t \in \{T_a + 1, ..., T_a + T_b\}. \tag{4.29}$$

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.

5 Numerical Results

In Table 1 we list the model parameters we used when optimizing formulations Φ_0, Φ_1 and Φ_2. These approximate values are based on 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 considered a range of weights w for the quadratic shelter penalty in Equation 4.10. A higher weight corresponds to a preference for long-term housing solutions at the expense of a larger unsheltered queue in the short-term. The choice of weight should in practice be guided by the decision-makers preference for a long- or short-term solution. We settled on $w = 0.3$ to give a meaningful penalty to shelter but without undermining its advantage over an unsheltered queue. We use a current estimate of the arrival rate of 10/day for the first T_a years of the modeling horizon. We anticipate that with major local prevention efforts (Regional Impact Council, 2021), which take place independently of capacity expansion, the arrival rate might potentially drop significantly to an estimate of 6/day in the long run. However this assumption is based on optimistic predictions from discussions with Alameda County and so is not well-justified with data. In Section 6 we test a variety of possible long-term arrival rates to analyze the sensitivity of the optimization results to this updated arrival rate, including the case where the arrival rate does not change at all.

Our framework is flexible to alternative arrival rate functions, so we recommend planners adjust their input values accordingly as data is updated. In practice, actual housing resources may not match anticipated levels, so the input values to the model and code should be updated using recent data as it becomes available. While we employ values from Alameda County's single adult population which has extremely high queues, our code and approach can be used to help with capacity planning for other populations. The main resource for data is HMIS (Homeless Management Information System) which can be used to estimate of the number of people arriving to the system. These values can then be used to calculate arrival rates. HMIS also tracks the number served, as well as current housing and shelter inventory levels, and these values can be used to calibrate housing server parameters. Additionally, counties are responsible for tracking clients as they arrive using Coordinated Entry. We used aggregated data from Alameda County and housing planning discussions to determine the values of T_a, λ_d, X_0, h_0, s_0, c_h, c_s and μ_0 in Table 1. We notionally chose the values of T_b, T_c, C, B_h, B_s and w as proposed parameters for our optimization problem, but these could change based on varying preferences of the system planners.

Table 1: Model parameters

	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	$\frac{1}{4 \times 365 \text{ days}}$	$\frac{1}{4 \times 365 \text{ days}}$	$\frac{1}{4 \times 365 \text{ days}}$
w	0.3	0.3	0.3

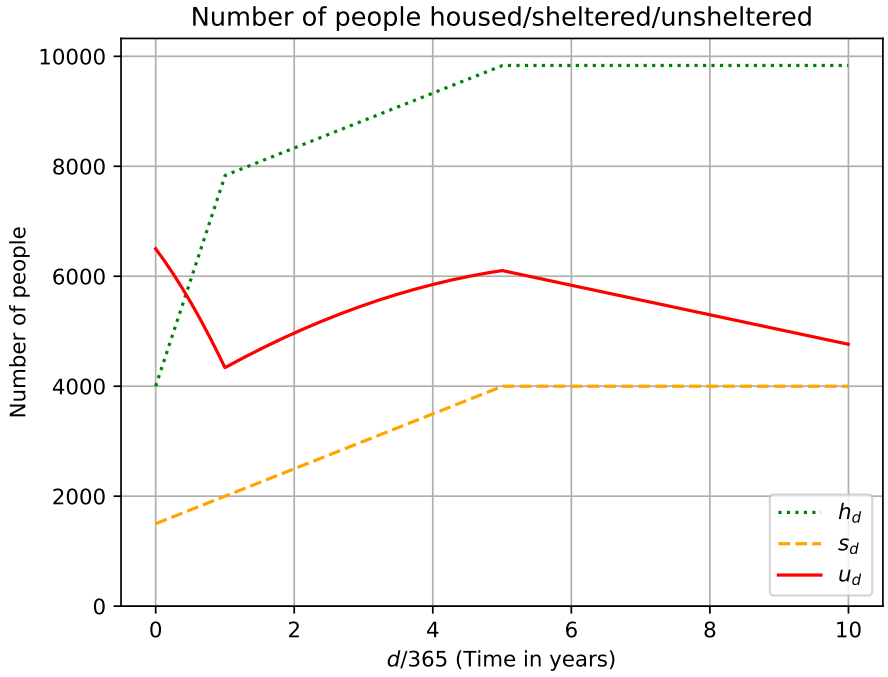

Figure 4: Optimal solution for Φ_0.

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 daily quantities h_d, s_d and u_d corresponding to the optimal values of h_t and s_t are 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, 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.

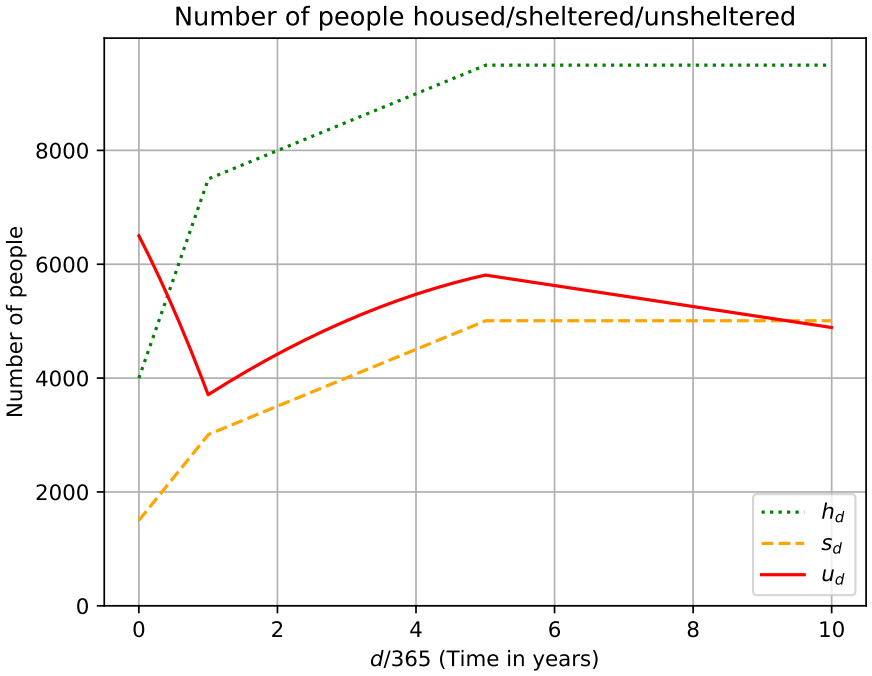

Figure 5: Optimal solution for Φ_1.

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

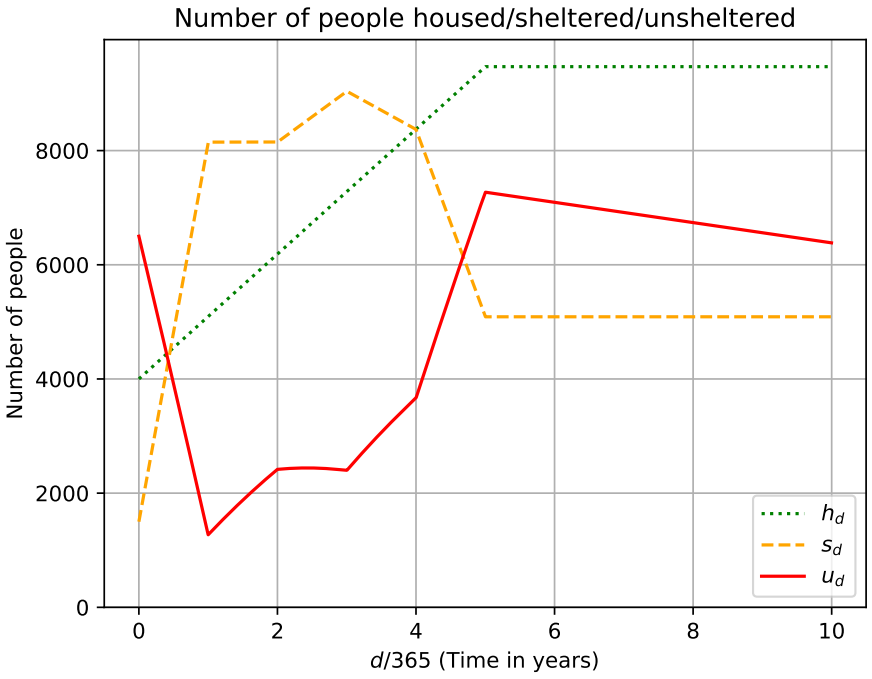

Figure 6: Optimal solution for Φ_2.

initial ramp up of shelter is able to bring the unsheltered queue down in the short term. The rate 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.

In Table 2 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

Table 2: Optimal capacities at the end of each year (proportion of total budget spent)

	Type	Initial	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%)

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 7 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. If the arrival rate does not decrease as anticipated, the long-term unsheltered population may not stabilize given current resource levels. 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.

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/tvcp

We note that the input values used to display numerical results in this paper are subject to change, and emphasize that the model user should update the input values as additional data becomes available. The optimal solutions presented in the figures are designed to show how the objectives and constraints could affect planned investment in housing and shelter for our given set of inputs. Alameda County was searching for an investment plan that would reach functional zero in five years, and we demonstrate some conditions under which the number of unsheltered people stabilizes over the long term. In reality, arrival rates may be higher than expected and service rates lower than anticipated. Thus, it is important to consider a range of possible inputs to the model,

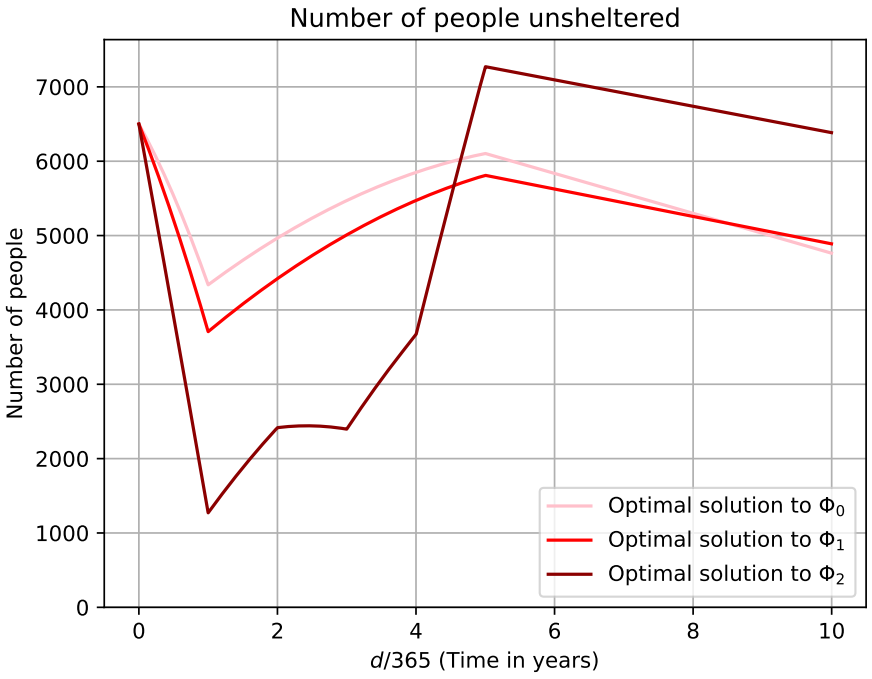

Figure 7: Unsheltered queue for each optimal solution

motivating the sensitivity analysis in the next section which reveals situations when the system never stabilizes.

6 Sensitivity analysis

As discussed in Section 3, the two main inputs to the fluid flow model are the arrival rate and the service rate. There is uncertainty associated with these inputs, especially over planning horizons such as those faced by Alameda County. In this section we shed light on the sensitivity of the optimization results to changes in these model inputs. In Section 5 we used an arrival rate of 10/day which reduced to 6/day at the end of the 5-year decision horizon. We maintain the initial arrival rate of 10/day during the 5-year decision horizon. However, we now consider alternative assumptions for the additional 5-year modeling horizon. We give particular attention to the case where the arrival rate remains at 10/day during this period. We also let the arrival rate during the additional 5-year modeling horizon take values from 5/day to 13/day. For simplicity we model a step change in the arrival rate at the end of the decision horizon. In Section 5 the service time for a client in housing was 4 years. We now let this service time take values from 3.5 to 5.5 years. These ranges were chosen to reflect the uncertainty in the real-world quantities, purposefully avoiding

model inputs which would violate the assumption of houses and shelters always being occupied. We note that it is very difficult to predict future arrival rates which may change every year (as opposed to changing only after 5 years), and so it is important for the user to update the model input code and rerun the planning model as new information becomes available.

6.1 Optimization results for a range of model inputs

In this section we solve the optimization problem Φ_2 for each element of a 9×9 input space. We construct this input space using the aforementioned ranges in arrival rates and service times. The objective function Φ_2 is a quadratic combination of the unsheltered and sheltered queues on each day of the 10-year modeling horizon, as given in Equation 4.10. Figure 8 displays the objective values for the resulting optimal solutions. As the arrival rate increases, and as the service time at housing increases, the objective value of the optimal solution becomes worse. In Figure 8 we indicate with white squares where the optimal solution exhibits queueing stability after the end of the decision horizon, i.e., where the total service rate given the housing capacity exceeds the arrival rate. This illustrates that given our optimization framework, only certain arrival rate and service time combinations will result in an optimal solution which will lead to a long-term decrease in the unsheltered queue.

To compare the specific case when the arrival rate drops to 6/day with the case when the arrival rate remains at 10/day, we plot the dynamics of the corresponding optimal solutions in Figure 9. When the arrival rate remains at 10/day, the optimal solution includes less housing but more shelter compared to when the arrival rate drops to 6/day, however these differences are relatively small. When the arrival rate remains at 10/day, the unsheltered population grows during the additional 5-year modeling horizon. This is in stark contrast to when the arrival rate drops to 6/day. Here, the unsheltered population decreases in the long term. It is interesting to note that while the objective values of the optimal solutions are very different, the optimal solutions themselves (the capacity plans) are reasonably similar. It is also worth noting that the instability of the queue with the arrival rate of 10/day is a result of the budget constraint of this problem. An increase in the budget would help to achieve stability at higher arrival rates.

6.2 The effect of making a "wrong" decision

In this section we explore the regret associated with choosing a capacity plan using model inputs which do not materialize in reality. To this end, first consider the optimal solutions from Section

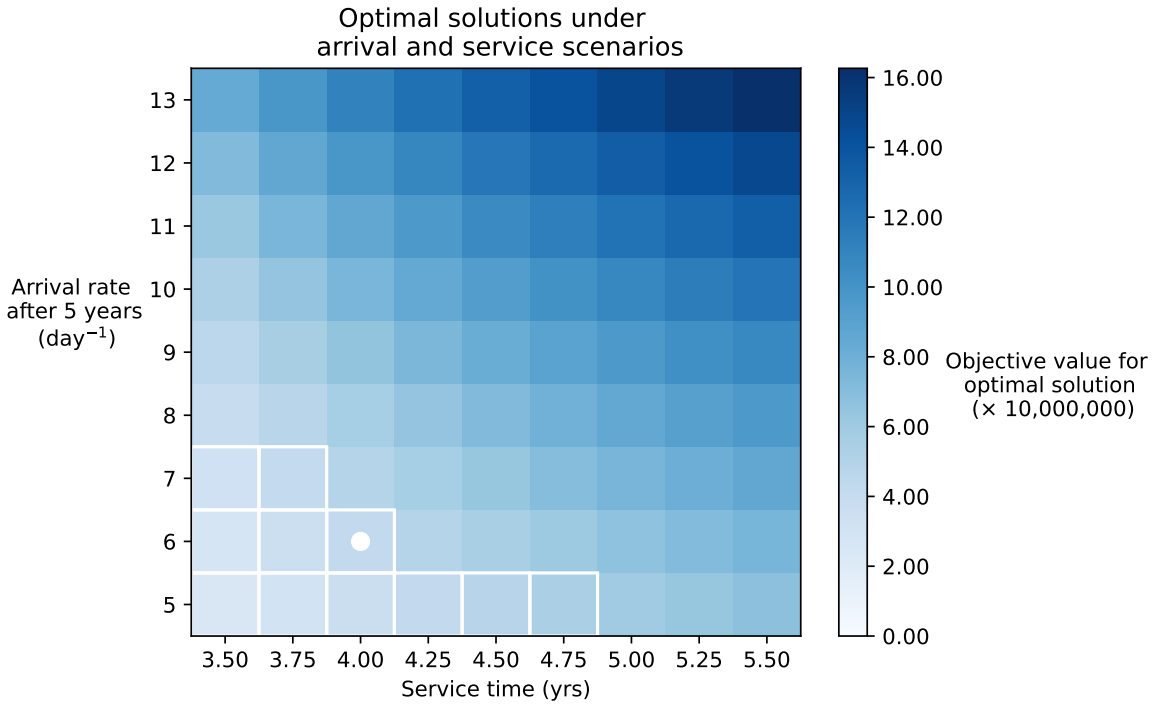

Figure 8: Objective values for optimal solutions to Φ_2 (quadratic objective including sheltered and unsheltered queues) with a range of arrival and service scenarios. White squares mark solutions where queueing stability is reached by the end of the decision horizon. The white dot indicates the scenario analyzed in Section 5.

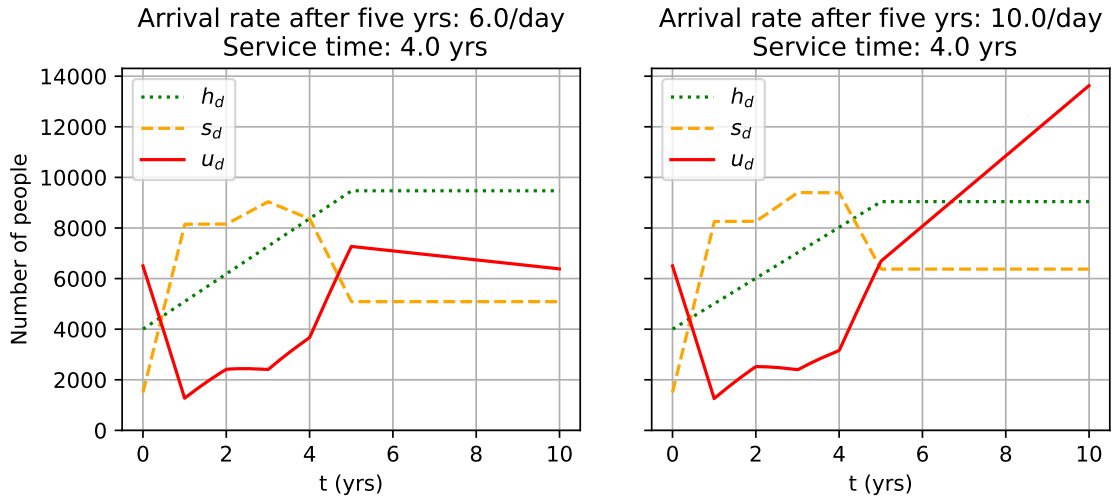

Figure 9: Number of people housed/sheltered/unsheltered in optimal solutions to Φ_2 for the cases when the arrival rate drops to 6/day and remains at 10/day.

6.1 where for each element of the 9×9 input space we solve the optimization problem Φ_2. In the left plot of Figure 10 we display the average unsheltered queue over the 10-year modeling horizon

using these optimal solutions. This can be interpreted by a decision-maker as the range of optimal outcomes given the range of model inputs. Next consider the optimal solution to Φ_2 from Section 5 where the arrival rate reduces to 6/day and service time at housing is 4 years. We call these model inputs $\boldsymbol{x}'$ and the corresponding optimal solution $(\boldsymbol{h}', \boldsymbol{s}')$. In the center plot of Figure 10 we display the average unsheltered queue over the 10-year modeling horizon using this optimal solution $(\boldsymbol{h}', \boldsymbol{s}')$ but with the range of different model inputs. This can be interpreted by a decision-maker as the range of outcomes in practice following this specific decision. For a given set of model inputs $\boldsymbol{x}$ (a given grid cell) by running the fluid model with the optimal solution $(\boldsymbol{h}', \boldsymbol{s}')$ which was optimized using model inputs $\boldsymbol{x}'$, we expect to see a worse outcome compared to running the model with optimal solution $(\boldsymbol{h}, \boldsymbol{s})$ obtained by optimizing with model inputs $\boldsymbol{x}$. In the right plot of Figure 10 we observe the extent to which the outcomes are worse for this reason. We call this the regret. This can be interpreted by a decision-maker as the effect of making the "wrong" decision i.e. choosing a capacity plan which is optimal for a set of model inputs which does not materialize in reality. Note: for a meaningful comparison, we plot the average unsheltered queues as opposed to the objective values (see Φ_2) which are quadratic in both the average unsheltered and sheltered queues. Figure 10 only plots the unsheltered queue as opposed to the full objective function from Φ_2, which includes sheltered queues. Therefore, the outcome for the average unsheltered queue can appear improved when we have optimized with a different set of model inputs. This apparent improvement comes at the expense of a worse outcome for the average sheltered queue. Therefore, in the right plot of Figure 10 we omit the cases where the average unsheltered queue reduces because it is misleading to consider these cases an improvement when we are not considering the effect on the sheltered queue.

Figure 10 illustrates that if only the arrival rate is worse in reality than expected, the outcome having optimized with a different arrival rate would not be much worse than optimizing with the actual arrival rate. This agrees with insights from our experiments illustrating that an optimal solution structure does not change much when we increase only the arrival rate, as suggested in Figure 9. However, if the service time in housing is longer in reality than expected, the outcome having optimized with a different service time could be considerably worse than optimizing with the actual service time. From detailed explorations into the solution structure, this is because the optimal solution for a shorter service time would involve fewer shelter units but more houses. When the service time is actually longer, those extra houses are less effective than shelter in reducing the unsheltered queue.

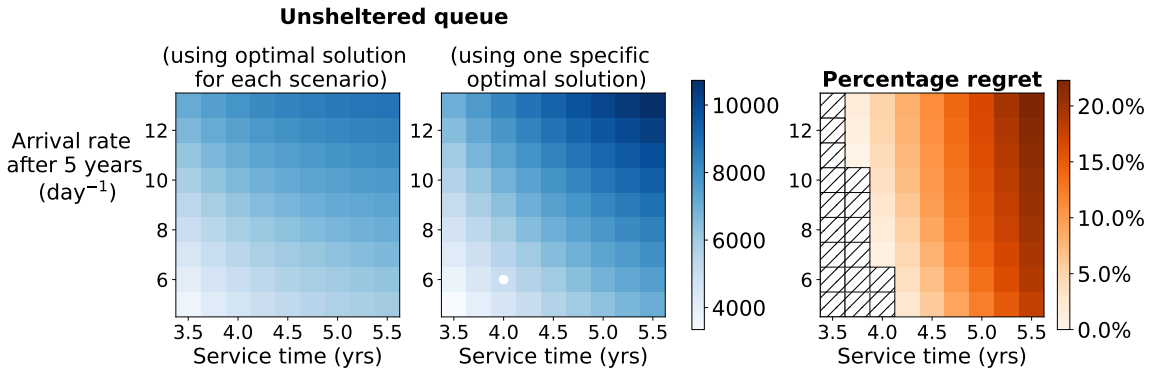

Figure 10: The effect of making a "wrong" decision. **Left**: Average unsheltered queue over a 10-year modeling horizon when running each scenario with its own optimal solution. **Center**: Average unsheltered queue over a 10-year modeling horizon when running each scenario with the optimal solution to Φ_2 from Section 5 (white dot). **Right**: The percentage increase to the average unsheltered queue (the regret) when running each scenario with the optimal solution to Φ_2 from Section 5 as opposed to its own optimal solution. White boxes with diagonal lines omit cases where the average unsheltered queue reduces as these improved results do not consider the potential effects on the sheltered queue.

7 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 of 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.

While our fluid queueing model is simple and makes many assumptions that may not be fully realistic, it provides an aggregate picture of the flow of clients through the system that allows for broad insights into the optimal allocations between housing and shelter. An aggregate model is critical to demonstrating the effects of inadequate resources leading to thousands of people being unsheltered over a long period time – additional model detail is not necessary to prove this point or demonstrate that thousands of additional housing units are needed. Certainly queueing models with deeper levels of complexity exist to accurately model the system, but they aren't amenable to optimization and hence rely on trial and error to perform comparisons between plans. While delivering a highly specialized optimization plan could be useful, Alameda County is unable to fully control implementation due to numerous resource constraints (for example limited funds, slow inventory build rates, and zoning ordinances). Thus, we develop an approach for providing general insights on the allocation of resources between housing and shelter, and how this allocation should change over time. Additionally, our formulation could be used to optimize capacity expansion for any highly-congested queueing system beyond the homelessness domain.

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 goal programming formulation which penalized deviations on a time-dependent goal on the unsheltered population would be interesting to explore.

We have also worked with discrete-event simulation (DES) models of homeless care systems to test capacity plans without optimization. A natural next step is to apply optimal planning to a DES. While our fluid flow model captures expected queue lengths quickly and effectively with a simple setup, with DES we can model a range of system complexities which are beyond the scope of the fluid flow model. Examples of such complexities include tandem queueing, the conversion of shelter to housing and the non-zero time needed to occupy a house with a new resident following the departure of the previous resident. Optimization with such a model would then fall in the realm

of simulation optimization, and this is the subject of our ongoing work in this area.

Acknowledgements

This work was supported by EPSRC under Grant EP/S022252/1. We acknowledge Dave Worthington and Rob Shone of Lancaster University for many useful discussions and suggestions that contributed to construction of this model. We are also extremely grateful to the Alameda Office of Homeless Care and Coordination for sharing their data and providing detailed insights into their decisions and challenges. Finally, we are indebted to two referees whose suggestions greatly contributed to the clarity and presentation of this paper.

References

Alameda County (2022). Home together 2026 community plan: A 5-year strategic framework centering racial equity to end homelessness in Alameda County.

Arora, P., Rahmani, M., and Ramachandran, K. (2021). Doing less to do more? Optimal service portfolio of non-profits that serve distressed individuals. *Manufacturing & Service Operations Management*, 24(2):883–901.

Balcik, B., Iravani, S., and Smilowitz, K. (2014). Multi-vehicle sequential resource allocation for a nonprofit distribution system. *IIE Transactions*, 46(12):1279–1297.

Bassamboo, A., Randhawa, R. S., and Zeevi, A. (2010). Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science*, 56(10):1668–1686.

Besbes, O., Castro, F., and Lobel, I. (2022). Spatial capacity planning. *Operations Research*, 70(2):1271–1291.

Bretthauer, K. (1995). Capacity planning in networks of queues with manufacturing applications. *Mathematical and Computer Modelling*, 21(12):35–46.

Chen, H. and Yao, D. D. (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer, New York.

Higgs, B. W., Mohtashemi, M., Grinsdale, J., and Kawamura, L. M. (2007). Early detection of tuberculosis outbreaks among the San Francisco homeless: Trade-offs between spatial resolution and temporal scale. *PLOS One*, 2(12):e1284.

HUD Exchange (2022). HUD Exchange.

Ingle, T. A., Morrison, M., Wang, X., Mercer, T., Karman, V., Fox, S., and Meyers, L. A. (2021). Projecting COVID-19 isolation bed requirements for people experiencing homelessness. *PLOS One*, 16(5):e0251153.

Izady, N. (2015). Appointment capacity planning in specialty clinics: A queueing approach. *Operations Research*, 63(4):916–930.

Izady, N. and Worthington, D. (2012). Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments. *European Journal of Operational Research*, 219(3):531–540.

Jian, N. and Henderson, S. G. (2015). An introduction to simulation optimization. In *2015 Winter Simulation Conference*, pages 1780–1794, Huntington Beach, CA, USA. IEEE.

Kaya, Y. B. and Maass, K. L. (2022). Leveraging priority thresholds to improve equitable housing access for unhoused-at-risk youth. *arXiv:2212.03777*.

Kaya, Y. B., Maass, K. L., Dimas, G. L., Konrad, R., Trapp, A. C., and Dank, M. (2024). 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.

Konrad, K. and Liu, Y. (2023). Achieving stable service-level targets in time-varying queueing systems: A simulation-based offline learning staffing algorithm. In *2023 Winter Simulation Conference*, pages 327–338, San Antonio, TX, USA. IEEE.

Lam, H., Liu, Z., and Singham, D. (2024). Shape-constrained distributional optimization via importance-weighted sample average approximation. *arXiv:2406.07825*.

Lam, H., Liu, Z., and Zhang, X. (2021). Orthounimodal distributionally robust optimization: Representation, computation and multivariate extreme event applications. *arXiv:2111.07894*.

Lin, M., Tong, W., Chen, J., Zeng, T., Ma, L., Feng, J., Sun, R., and Liang, Z. (2024). Optimal multiresource planning for the elderly care system: A case study. *International Journal of Production Research*, 63(9):3091–3116.

Liu, J., Yang, F., Wan, H., and Fowler, J. W. (2011). 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.

Maass, K. L., Trapp, A. C., and Konrad, R. (2020). Optimizing placement of residential shelters for human trafficking survivors. *Socio-Economic Planning Sciences*, 70:100730.

Miller, F., Kaya, Y. B., Dimas, G. L., Konrad, R., Maass, K. L., Trapp, A. C., et al. (2022). On the optimization of benefit to cost ratios for public sector decision making. *arXiv:2212.04534*.

Mohammadi Bidhandi, H., Patrick, J., Noghani, P., and Varshoei, P. (2019). Capacity planning for a network of community health services. *European Journal of Operational Research*, 275(1):266–279.

Nov, Y., Weiss, G., and Zhang, H. (2022). Fluid models of parallel service systems under FCFS. *Operations Research*, 70(2):1182–1218.

Oakland-Berkeley-Alameda County CoC (2020). Centering racial equity in homeless system design.

Office of Housing and Urban Development (2024).

Patrick, J. (2011). Access to long-term care: The true cause of hospital congestion? *Production and Operations Management*, 20(3):347–358.

Patrick, J., Nelson, K., and Lane, D. (2015). A simulation model for capacity planning in community care. *Journal of Simulation*, 9(2):111–120.

Rahmattalabi, A., Vayanos, P., Dullerud, K., and Rice, E. (2022). Learning resource allocation policies from observational data with an application to homeless services delivery. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1240–1256, Seoul, South Korea. ACM.

Regional Impact Council (2021). Regional action plan: A call to action from the Regional Impact Council.

Reynolds, J., Zeng, Z., Li, J., and Chiang, S.-Y. (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.

Singham, D. I. (2023). Estimating quantile fields for a simulated model of a homeless care system. In *2023 Winter Simulation Conference*, pages 1054–1064, San Antonio, TX, USA. IEEE.

Singham, D. I., Lucky, J., and Reinauer, S. (2023). Discrete-event simulation modeling for housing of homeless populations. *PLOS One*, 18(4):e0284336.

Singham, D. I., McDonald, M., and Elliot, R. (2025). Tradeoffs between equity and efficiency in coordinated entry of homeless housing systems. *Socio-Economic Planning Sciences*, 99:102212.

Stidham Jr, S. (2009). *Optimal Design of Queueing Systems*. Chapman and Hall/CRC, New York.

U.S. Department of Housing and Urban Development (2024). Coordinated entry core elements.

Worthington, D. (1991). Hospital waiting list management models. *Journal of the Operational Research Society*, 42(10):833–843.

Yarmand, M. H. and Down, D. G. (2013). Server allocation for zero buffer tandem queues. *European Journal of Operational Research*, 230(3):596–603.

Yarmand, M. H. and Down, D. G. (2015). Maximizing throughput in zero-buffer tandem lines with dedicated and flexible servers. *IIE Transactions*, 47(1):35–49.

Zhang, Y., Puterman, M. L., Nelson, M., and Atkins, D. (2012). A simulation optimization approach to long-term care capacity planning. *Operations Research*, 60(2):249–261.