Modelling and optimising the housing of homeless populations: ten month PhD review

Graham Burgess

October 2024

Abstract

Modelling and optimisation are popular tools for supporting resourcing and capacity decisions in healthcare and homeless settings. We show how deterministic optimisation with a fluid flow model can support long-term capacity planning for a homeless care setting in the San Francisco Bay Area, California. Models of multi fidelity, including stochastic simulation, are available in this setting, and the solution space is integer-ordered. We therefore explore both multi fidelity and integer-ordered simulation optimisation methods and discuss potential research contributions at the intersection of these active fields of research.

Todo:

- M(t)/M/s(t) -¿ $M_t/M/s_t$

- Format of Φ formulation

- reread section 4.3 and section 5

1 Introduction

Homelessness is a growing problem faced by communities worldwide. An example is Alameda County in the San Francisco Bay Area, California, where approximately $16,000$ people experienced homelessness in 2021 alone (Oakland-Berkeley-Alameda County Continuum of Care, 2021). Decision makers within these communities typically have some leverage over how relevant resources are allocated to help alleviate homelessness. In Alameda County, the local government must split their resources between emergency shelter and permanent social housing. Shelter is relatively cheap and quick to set up. It provides a safer alternative to unsheltered homelessness but does not provide a stable long-term living situation for its residents. Permanent social housing is more expensive than shelter and can take longer to set up, but it does offer the stable long-term living situation that homelessness people need to improve their quality of life. Building capacity to alleviate homelessness takes time, especially since funds are typically available in varying amounts from year to year. Decision makers in communities like Alameda County must therefore make good time-varying capacity planning decisions to reduce homelessness now and in the future.

Operational research methods offer helpful tools to support such public sector decision-making. Optimisation helps decision-making by looking for a feasible solution which performs best across a (potentially infinite) set of alternative feasible solutions. To do this, a model of the system's performance is needed, and the quality of the model affects the quality of the subsequent optimisation results. We can model the homeless care system as a queue. Individual clients arrive and if there is no space in shelter, they wait in an unsheltered queue. Once they are finished in shelter, they can only leave when housing becomes available. Once in housing they stay for some time before leaving for good or rejoining the queue for shelter. A solution for this system is a housing and shelter capacity plan.

The most accurate model of this queueing system is a high-fidelity stochastic simulation model. In this case, one can only estimate the performance of a given solution and the subsequent optimisation falls in the realm of simulation optimisation (SO). There are different SO methods for different types of problem (see Section 2.2) but the issue of limited computational resources pervades all SO methods. This issue stems from from the fact that a stochastic simulation is typically computationally expensive to run, and many simulation replications are required to be confident of the performance of a solution, given the associated uncertainty.

As is common in queueing systems, low-fidelity models such as analytical queueing models offer a computationally cheaper alternative to stochastic simulation. They also offer helpful alternative perspectives on the dynamics of a queueing system. The drawback is that these models are typically less accurate, given the necessary assumptions which must be made. If one only uses a low-fidelity deterministic model to evaluate system performance, optimisation falls in the realm of deterministic optimisation. Performing this deterministic optimisation can be a helpful first step in the decision-making process. However, there is more we can do with low-fidelity models. Multi-fidelity simulation optimisation (MFSO) enables low-fidelity models to be used alongside high-fidelity stocahstic simulation in a SO algorithm which reduces the computational burden and therefore enables an optimal solution to be found more efficiently. Novel MFSO methods will be the main topic of this PhD research.

This document is organised as follows: in Section 2 we review relevant literature on capacity planning in healthcare and homeless care settings. There are many similarities between these two settings but the former is more widely studied in the literature. We also review relevant SO methods including MFSO. In Sections 3, 4 and 5 we discuss the main content of the PhD research to date: Section 3 introduces three models of multi fidelity for homeless care systems; Section 4 introduces an optimisation formualation which addresses the time-varying capacity planning problem and we solve this problem in a determinsitc setting; in Section 5 we discuss how different types of uncertainty affect our optimisation problem. This final discussion motivates the need for simulation optimisation. In Section 6 we discuss relevant interesting gaps in the current MFSO literature which the remainder of this PhD research seeks to address.

2 Literature Review

2.1 Capacity planning in healthcare and homeless care settings

There are many similarities between queueing systems in healthcare and homeless care settings. We have already discussed a homeless care queueing system. In healthcare settings, queues build up in the form of waiting lists as patients wait for treatment. There may be several stages involved. For example, a consultation may be required in an out-patient clinic before specialist treatment is given at an in-patient clinic. Emergency care may first be required in hosptial followed by long-term care at another facility. Capacity planning can involve setting staffing levels or planning for extra facilities, beds or appointment slots. We now discuss several approaches to capacity planning in both healthcare and homeless care settings.

Scenario modelling is a popular tool in healthcare capacity planning. Due to the complex nature of healthcare systems, discrete-event simulation (DES) is popular for testing system performance. El Hage et al. (2021) model a busy COVID-19 testing clinic using DES, identifying bottlenecks based on the servers with high utilisation across a range of demand and testing-capacity scenarios. Reynolds et al. (2010) model a health clinic for homeless patients in Kentucky, USA using DES, testing the quality of care for different staffing levels. Fluid modelling (a.k.a. stock and flow modelling) is used when high-level strategic decisions are involved. Worthington (1991) shows how fluid flow models are particularly useful for addressing 'what if?' capacity questions in the context of hospital waiting lists.

Staffing rules are a common approach to healthcare capacity planning when arrival rates to a healthcare facility are non-stationary. The simple stationary approximation (Green et al., 1991) and the pointwise stationary approximation (Green and Kolesar, 1991) approximate a non-stationary queue with a stationary one and staffing levels can then be set to meet a performance target using the corresponding steady-state results. The resulting staffing levels, however, often do not meet the performance target for the non-stationary queue due to the approximations made. The infinite-server approximation of Jennings et al. (1996) estimates the effective demand on an $M_t/M/s_t$ system at some time t using the utilisation of the equivalent $M_t/M/\infty$ queue. The resulting time-dependent staffing rule includes a base level to accommodate the effective demand plus a square-root term to accommodate the demand variability over time. This is known as the square-root staffing rule and shows considerable improvement on meeting desired performance targets. Izady and Worthington (2012) use the square-root staffing rule alongside stochastic simulation to meet requirements on waiting times in Accident and Emergency settings. Konrad and Liu (2023) use a reinforcement learning approach to set staffing levels which meet requirements for the tail probability of delay.

Capacity planning in homeless care settings is not as widespread in the literature as healthcare settings. Here we mention previous modelling of the homeless care system in Alameda County which is the motivating example for this PhD research. Alameda County performed a systems modelling exercise which estimated future arrivals to the homeless care system and future departures from the system for a number of investment strategies to increase housing and shelter capacity (Alameda County, 2022). Singham et al. (2023) incorporated stochasticity and used DES to model the system as a tandem queue with blocking between shelter and housing. They also modelled a number of investment strategies to identify the capacity needed to bring down the level of unsheltered homelessness to zero within 5 years.

Several studies have looked at capacity planning for services to support runaway homeless youths (RHYs). Kaya et al. (2022b) use DES to model capacity expansion scenarios for a RHY shelter which offers accommodation and support such as counselling and medical treatment. Other studies use optimisation frameworks, each using a different format of objective function. Kaya et al. (2022a) use integer programming (IP) to optimally match demand for support services from RHYs, with supply. They take a cost minimisation approach, where existing capacity is cheaper than potential new capacity. Maass et al. (2020) use IP to optimise the location of new shelters for RHYs and others involved in human traficking. They maximise benefit minus cost, using several metrics for societal benefit such as criminal justice costs avoided and labour productivity gained. Miller et al. (2022) tackle a similar problem to Kaya et al. (2022a) but they incorporate societal benefit and maximise a benfit to cost ratio, using a fractional programming method known as Dinkelbach's algorithm.

2.2 Simulation optimisation (SO)

Simulation optimisation is not widely used in the management of healthcare or homeless care settings. This is likely due to the fact that decision-makers often face a small number of practical alternatives when considering capacity planning or would like to see how a plan fares under a small number of hypothetical scenarios. When stochastic simulation is used to model these alternatives, a simple comparison of simulation outputs is often helpful and sufficient, especially when there are multiple competing criteria to consider, such as performance and cost.

SO is aimed at decision problems where there are a large number of alternative solutions so allocating substantial computational resource for simulation to each solution is not practical. SO methods offer intelligent systematic procedures to find optimal solutions efficiently using a finite computational budget for simulation. The time-varying capacity planning problem in homeless care settings does involve a large solution space since housing and shelter capacity may take a wide range of values year on year. In this PhD research we would like to explore to what extent novel efficient MFSO methods can give meaningful analytical support to decision makers in this setting. With this in mind, we now review current SO methods, in particular integer-order and multi-fidelity methods which have particular relevance to our problem.

2.2.1 Overview of SO methods

In SO, one is typically interested in solving the following optimisation problem:

$$\min \theta(\boldsymbol{x}), \ \ \boldsymbol{x} \in C \tag{1}$$

where $\theta(\boldsymbol{x})$ is the true cost of solution $\boldsymbol{x}$ which can only be estimated with stochastic simulation and C is the set of feasible solutions, or the solution space. SO methods can be categorised into those which deal with discrete-valued solution spaces and those which deal with contiuous-valued solution spaces. Stricly speaking, the decision variables in our problem are discrete (the number of housing units or shelter units) however it can be convenient to approximate these variables as continuous-valued (as we do in Section 4) so here we discuss both discrete and continuous SO methods.

In discrete problems, with a sufficiently small number of feasible solutions, it may be realistic to simulate every solution at least once. In this case, **ranking and selection** (R&S) methods are appropriate. Initially each solution is simulated a relatively small number of times, and then

a decision is made iteratively about which solutions to discard and which to keep simulating. By the end of the procedure, one has a statistical guarantee on the probabilty that a solution within an 'indifference zone' of the true best has been selected (in a frequentist setting).

In discrete problems where it is not realistic to simulation every solution, **adaptive random search** (ARS) methods are appropriate. At each iteration, a set of feasible solutions is randomly selected for simulation (or further simulation) based on a probability distribution across the feasible region. The nature of the probability distribution is different for different ARS methods, but they informally exploit the idea that good solutions tend to be clustered together. ARS methods converge asymptotically to local or global optima (depending on the method) and rules are used to decide when to stop the procedure and return the sample best solution.

In many discrete problems, including the homeless care problem, each decision variable is defined on an integer-ordered scale. **Integer-ordered** SO methods can exploit the spatial relationships between solutions when looking for a optimal solutions. Some integer-ordered methods work with gradients, while others use a neighbourhood structure. We discuss several key integer-ordered methods in Section 2.2.2.

We now briefly discuss two popular approaches for continuous problems. **Stochastic approximation** methods search across a continuous solution space for locally optimal solutions where the true cost $\theta(\boldsymbol{x})$ is continuous and differentiable in $\boldsymbol{x}$. If we can estimate the gradient $\nabla\theta(\boldsymbol{x})$ using simulation, then we can move along the gradient to a region where we expect better solutions. There are several methods for esimating the gradient, generally classified as 'indirect' if they estimate an approximation of the gradient using simulation at points a small finite difference apart, or 'direct' if they attempt to estimate the true gradient, by, for example, differentiating the simulation code itself.

In **sample average approximation** the optimisation problem is transformed to a deterministic problem by fixing the random numbers used in the stochastic simulation of each feasible solution. The goal is then to minimise the sample average objective function, whose optimal solution is a good approximation of the true optimal solution to the original problem provided the samples are big enough. Of course, one still has to perform the deterministic optimisation, but this can take advantage of some structural problems of the sample average problem such as convexity.

Finally, we discuss the use of **meta models** in solving discrete and continuous problems. Meta-model approaches capitalise on the low computational cost of an alternative model of the system in question. This meta model is used to aid the search for a local or global solution. For continuous problems, local meta models usually involve polynomial regression models. These perform well locally but poorly over a large solution space. The gradients of the polynomial meta model is used in the search. Global meta models such as Gaussian process (GP) models, work well in low-dimensional settings where the cost of simulation is high. In Bayesian optimisation (BO), a GP prior distribution of the simulation output is placed over the entire continuous-valued solution space. At each iteration in BO, a sampling decision is made to choose the solution with the best expected improvement on the current best solution and the posterior distribution is subsequently updated. In Section 2.2.2 we discuss a discrete version of the GP approach. In **multi-fidelity simulation optimisation**, one or more low-fidelity meta models of the system are used to aid in the optimisation process. As we discuss in Section 3, in queueing systems multi-fidelity models are typically available subject to certain assumptions.

2.2.2 Integer-ordered SO methods

In **discrete stochastic approximation** (DSA) (Lim, 2012), the objective function f over the integer-ordered discrete domain is extended an objective function $\tilde{f}$ over a continuous domain. In one-dimensional settings, this extension is done with a simple piecewise linear interpolation. In multiple dimensions, the continuous domain is partitioned into a set of convex hulls covering all integer points. The extended objective function $\tilde{f}(\boldsymbol{x})$ at point $\boldsymbol{x}$ in the continuous domain is constructed as a linear combination of f at each integer point on the convex hull surrounding $\boldsymbol{x}$. Stochastic approximation, as discussed in Section 2.2.1, is performed on the continuous domain to move from point to point along the direction of steepest gradient descent. The gradient at $\boldsymbol{x}$ is estimated by performing multiple simulation replications at each integer point on the convex hull surrounding $\boldsymbol{x}$ and taking differences between the resulting estimates of the objective function. At the end of the algorithm, the best solution is rounded to the nearest integer point. The authors define the property of multi-modularity for the objective function f on the discrete domain. Multi-modularity for functions on a discrete domain is the counterpart to convexity for functions on a continuous domain. They show that their algorithm converges to the optimal solution for a multi-modular f and they describe how many queueing systems exhibit this property, making their algorithm particularly suitable for queueing problems.

Another promising method for solving integer-ordered SO problems using gradient information is the retrospective search with piecewise-linear interpolation and neighborhood enumeration **(R-SPLINE)** (Wang et al., 2013). While DSA is a stochastic approximation procedure for discrete problems, R-SPLINE can be considered a sample-average approximation procedure for discrete problems. Each iteration of R-SPLINE solves a sample path problem retrospectively (R) using the SPLINE procedure, with a larger number of samples at each simulated solution than the previous iteration. The procedure is retrospective in the sense that each iteration of SPLINE uses the solution from the previous iteration as a 'warm-start'. The SPLINE procedure repeatedly performs a search with piecewise linear interpolation (SPLI) and a neighbourhood enumeration (NE). The SPLI step moves from one integer point to another, but to do that the search is 'perturbed' onto a point in the continous domain and a gradient at that point is estimated using the same approach as gradient-estimation in DSA. Once the search has travelled in the continuous domain along the direction of steepest gradient descent, the search returns to the nearest integer point and the process is repeated until one of several stopping criteria is met. The NE step checks for any better solution which is separated from the current best by one unit in any dimension. The resulting best solution is taken to the next iteration of SPLINE as a starting solution. Within each iteration of the SPLINE procedure, any solution which is simulated is done using the same number of replications and common random numbers. From one SPLINE iteration to the next, new random numbers are used and an increasing sample size is used, as the algorithm tends to focus on a smaller number of solutions. R-SPLINE converges asymptotically to the set of local optimal solutions and performs competitively in finite-time for a number of real-world problems.

In recent years, the use of **Gaussian Markov random fields** (GMRFs) (L. Salemi et al., 2019) for integer-ordered SO problems has become popular. GMRF-based methods can exploit spatial structure across a multi-dimensional solution space using a neighbourhood approach. A GMRF is a Gaussian random vector where each vector element models the objective function of a feasible solution. The GMRF is defined on an undirected graph comprised of nodes and edges. Each node corresponds to a feasible solution and the edges connecting nodes define

the neighbourhood of each feasible solution and the corresponding Markov property. The neighbourhood of a given node is defined as all the nodes which are connected to that node with an edge. The Markov property holds if one can predict the objective function of any feasible solution only with information about the objective functions of the feasible solutions in its neighbourhood. That is, the random variable modelling the objective function of a feasible solution is conditionally independent of all feasible solutions outside of its neighbourhood, conditional on the feasible solutions in its neighbourhood.

A GMRF is defined by its mean vector and precision matrix. The precision matrix is the inverse of the covariance matrix. Whereas zeros in a covariance matrix correspond to independence between random variables, zeros in a precision matrix correspond to conditional independence between random variables, given all other random variables (see Rue and Held (2005) for details). Assuming that the neighbourhood of each solution is small compared to the total size of the solution space, the precision matrix will be sparse, simplifying computations within this framework.

The basic outline of a SO algorithm with GMRFs, as detailed by L. Salemi et al. (2019), involves updating the conditional mean vector and precision matrix of the GMRF as more information is collected about specific solutions via stochastic simulation. Following each update, a decision is made about where to simulate next (from a set of candidate solutions) in a manner which balances the exploitation of promising areas of the solution space, and exploring new areas of the solution space. L. Salemi et al. (2019) make this decision using a complete expected improvement (CEI) criterion, which is evaluated for each candidate solution. The CEI criterion is similar to the expected improvement (EI) criterion discussed in Section 2.2.1 for Bayesian optimisation, but it treats both the objective function at the current best solution and that at the candidate solution as a random variable (instead of just the candidate solution as in EI). The algorithm is stopped when the maximum CEI across all solutions is repeatedly below some threshold value and the solution with the lowest conditional mean is returned as the optimal solution.

2.2.3 Multi-fidelity SO methods

As discussed in Section 2.2.1, one approach to SO with meta modelling is to use models of multi fidelity in the search. In queueing systems, an analytical queueing model can act as a low-fidelity meta model, as we discuss further in Section 3. There are several established approaches for incorporating multi-fidelity models into SO, we now discuss three such approaches.

One approach to incorporating multi-fidelity models into SO is to **perform a deterministic optimisation (DO) first** using a deterministic low-fidelity model. Informally, this helps develop a better understanding of the dynamics of the system in question and of the nature of good solutions. Formally, an optimial solution from a DO procedure can be used as a initial solution for a SO procedure, giving the latter a 'head start'.

This approach is explained by Jian and Henderson (2015) in the context of managing bicycle station capacity in a bicycle sharing system at the start of a morning rush hour. Their objective is to minimise customer disatisfaction due to full/empty stations. They first ignore randomness in arrivals and departures of bikes to and from stations and find simple rules of thumb for deciding on the morning capacity of a station depending on whether the station experiences net in-flow or net out-flow. This gives helpful intuition but has serious flaws from ignoring

stochasticity, such as assigning no bikes to net in-flow stations. The authors then incorporate stochasticity by modelling the number of bikes at each station i as an $M/M/1/r_i$ queue and including the expected number of full/empty stations in the objective function. This gives more intuitive solutions in situations where the fluid flow model would have preferred a full/empty station. Finally, they capture more system complexity using a discrete-event simulation model. They attempt to solve the resulting SO problem using a random search approach and find that by starting their search at the optimal solution from the queueing model optimisation, they find good solutions with less simulation effort.

Another multi-fidelity modelling approach to SO is known as **multi-fidelity optimisation with ordinal transformation and optimal sampling** (MO2TOS) and was first proposed by Xu et al. (2016) and is appropriate for discrete problems. The idea is that low-fidelity evaluations of the entire multi-dimensional solution space can be used to 'cheaply' transform solutions onto a new one-dimensional scale where they are ordered from good to bad. Grouping solutions based on their position in the transformed space results in lower group variances and higher inter-group distances than randomly grouping solutions without transformation. These characteristics enables the following optimal sampling scheme to more efficiently search for an optimal solution. The optimal sampling strategy is inherited from the well-known optimcal computing budget allocation rule (OCBA). Subsequent papers have improved the original MO2TOS method by, for example, applying clustering techniques to group solutions in the transformed space to further reduce group variances and further increase inter-group distances (Cao et al., 2023). The downside of MO2TOS-based methods is that all feasible solutions must be evaluated with the low fidelity model which may be problematic if the solution space is very large and/or the low-fidelity evaluation is not very cheap.

- Modelling the error of a low-fidelity model
    - Polynomial error terms e.g. Chong and Osorio (2018).
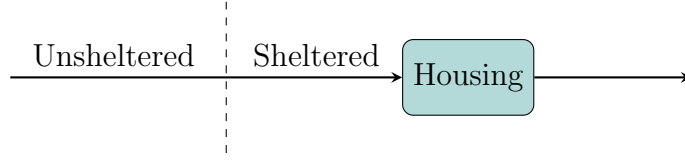   - Gaussian Process error terms e.g. Huang et al. (2006).

Figure 1: Simple queueing system

3 Models of multi fidelity

Here we introduce three models which we have developed for homeless care systems from low-to high-fidelity. Our low-fidelity model is a fluid flow model, which assumes housing servers are always busy and treats flow in and out of the system like a liquid with a continuous-valued volume. Our medium-fidelity model is an $M_t/M/h_t$ queueing model which relaxes the server-always-busy assumption. We incorporate stochasticity and using a Markov chain analysis we can compute the expected number of people housed, sheltered and unsheltered at some point in time, given initial conditions. Our high-fidelity model is a discrete-event simulation model. Here, as well as caputuring stochasticity we are able to model additional real-world complexities such as a tandem queueing system with shelter and housing servers, partial re-entry to the system and non-Markovian service time distributions. We now discuss each model in turn.

3.1 Fluid flow model

We start by making some assumptions to allow a simple model to be appropriate. Given that the main purpose of shelter is to accommodate people waiting for housing, rather than to offer a particular service which is needed to proceed, it is reasonable to treat shelter as part of the queue for housing, with houses being the sole set of servers in the queueing system. We therefore would like to model the system illustrated in Figure 1. In an overloaded queueing system such as the homeless care system in Alameda County, it is reasonable to assume that housing servers are always busy. With this assumption, the service process becomes independent of the arrival process and it is then straightforward to separately compute the number of arrivals and service completions in some time period, given arrival and service rates, which may be non-stationary.

We would like to track the number of housed, sheltered, and unsheltered people 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. 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 server and $h(t)$ is the continuous-valued number of houses at time t. Given the initial number of people in the system n_0, at time t we can calculate the subsequent number of people in the system, $n(t)$, as

$$n(t) = n_0 + \int_0^t \lambda(t)dt - \int_0^t \mu_0 h(t)dt.$$

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

$$u(t) = n(t) - h(t) - s(t) \tag{2}$$

$$= n_0 + \int_0^t \lambda(t)dt - \int_0^t \mu_0 h(t)dt - h(t) - s(t), \tag{3}$$

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 discretise 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 discretised 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 an objective functions (which we describe in Section 4) we approximate (3) with the sum

$$u_d = n_0 + \sum_{d'=1}^d \lambda_{d'}\delta t - \sum_{d'=1}^d \mu_0 h_{d'}^D \delta t - h_d^D - s_d^D, \tag{4}$$

where μ_0 is the daily service rate of a single housing unit and the stepsize $\delta t = 1$ day. In Figure 2 we give an illustrative example of the dynamics of u_d given by our fluid flow model, calibrated using realistic inputs for n_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 2 shows an example of how one might come close to reaching zero unsheltered homelessness 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 (4) 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 (4) becomes

$$u_d(\boldsymbol{h}, \boldsymbol{s}) = n_0 + \sum_{d'=1}^d \lambda_{d'}\delta t - \sum_{d'=1}^d \mu_0 h_{d'}^D(\boldsymbol{h})\delta t - h_d^D(\boldsymbol{h}) - s_d^D(\boldsymbol{s}), \tag{5}$$

where

$$h_d^D(\boldsymbol{h}) = h_{\lfloor \frac{d}{365} \rfloor} + \frac{d - \lfloor \frac{d}{365} \rfloor}{365}(h_{\lceil \frac{d}{365} \rceil} - h_{\lfloor \frac{d}{365} \rfloor}) \tag{6}$$
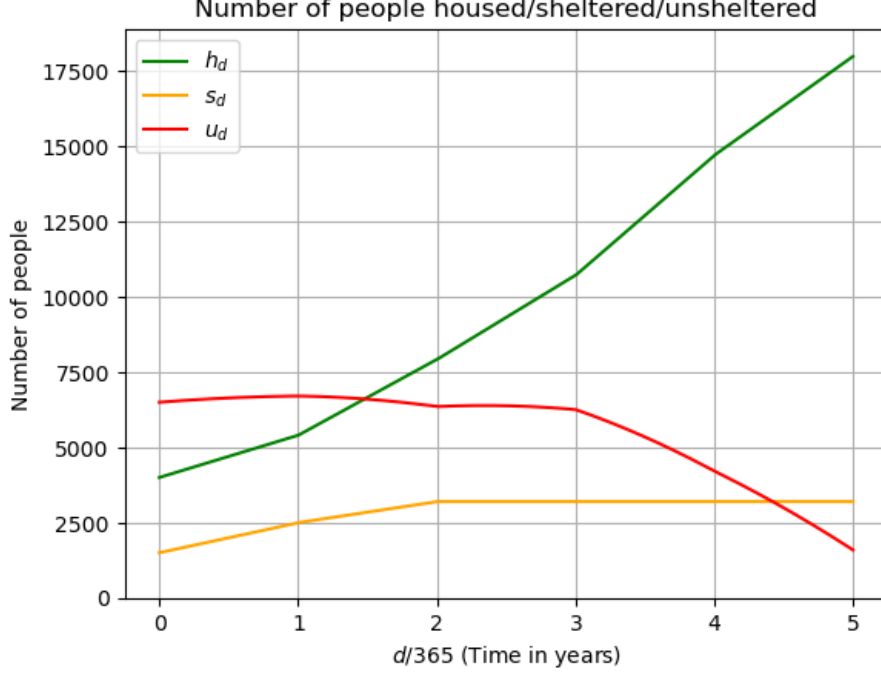
10

Figure 2: Dynamics of u_d, s_d^D and h_d^D. $n_0 = 12000$, $\mu_0 = 6.106 \times 10^{-4}$, Daily arrival rates λ_d in each year: $10.0, 11.9, 13.1, 13.1, 11.8$.

and

$$s_d^D(\boldsymbol{s}) = s_{\lfloor \frac{d}{365} \rfloor} + \frac{d - \lfloor \frac{d}{365} \rfloor}{365} \left(s_{\lceil \frac{d}{365} \rceil} - s_{\lfloor \frac{d}{365} \rfloor} \right). \tag{7}$$

This fluid flow model has the advantage of being computationally cheap to run. It is also flexible to incorporate certain real-world system complexities. For example, we could model heterogeneous customer groups in different compartments with different servers, arrival rates and service rates. We could also incorporate re-entries to the system by applying an appropriate discount factor to the third term in Equation 5.

3.2 $M(t)/M/h(t)$ queueing model

With an $M(t)/M/h(t)$ queueing model, where $h(t)$ is the time-dependent number of housing servers, we are still modelling the system as illustrated in Figure 1 but we no longer treat arrivals and departures as a fluid, but as individual people. We introduce stochasticity and would like to compute the probability distribution for the number of people housed, sheltered and unsheltered, throughout the time horizon. In doing so we allow a non-zero probability of their being 0 people in a housing server, thus relaxing the server-always-busy asumption we made with the fluid flow model.

Given the memoryless property of the exponential inter-arrival and service times in an $M(t)/M/h(t)$ queue, we can discretise time and treat this queue as a discrete-time Markov chain where the state space represents the number of people in the system and some large but finite N is the largest possible number in the system. With sufficiently small time intervals of length Δt, we can assume that within each interval the arrival rate is constant, the number of servers is

constant and at most one state change can occur. Therefore, if we know the initial probabilities $p_n(t = 0)$ of being in all states $n \in \{0, 1, ..., N\}$ then for all times $t \in \{0, \Delta t, ..., T - \Delta t\}$, we can calculate the probabilities $p_n(t + \Delta t)$ as:

$$
\begin{aligned}
p_n(t + \Delta t) = \; & p_n(t)[1 - \lambda(t)\Delta t - \mu_0 m_n(t)\Delta t] \\
& + p_{n+1}(t)[\mu_0 m_{n+1}(t)\Delta t] \\
& + p_{n-1}(t)[\lambda(t)\Delta t] \\
& + o(\Delta t) \qquad 1 \le n \le N - 1
\end{aligned}
$$

$$
\begin{aligned}
p_0(t + \Delta t) = \; & p_0(t)[1 - \lambda(t)\Delta t] \\
& + p_1(t)[\mu_0 m_1(t)\Delta t] \\
& + o(\Delta t)
\end{aligned}
$$

$$
\begin{aligned}
p_N(t + \Delta t) = \; & p_N(t)[1 - \mu_0 m_N(t)\Delta t] \\
& + p_{N-1}(t)[\lambda(t)\Delta t] \\
& + o(\Delta t)
\end{aligned}
$$

where $\lambda(t)$ is the arrival rate at time t, μ_0 is the service rate of a single housing server and $m_n(t) = \min(n, h(t))$ is the number of busy housing servers when the system is in state n at time t. The bias in the results introduced by using constant values in each time interval for $\lambda(t)$ and $h(t)$ vanishes as $\Delta t \to 0$. The probabilities $p_n(t)$ can then be analysed to estimate the expected number of people housed, sheltered and unsheltered, throughout the time horizon. For example, the expected number unsheltered at time t is given by:

$$
\mathbb{E}[u(t)] = \sum_{n=0}^{N} p_n(t) \max(0, n - h(t) - s(t)), \tag{8}
$$

where $s(t)$ is the time-dependent number of shelters. This model is more computationally costly than the fluid flow model, but still cheap compared to a discrete-event simulation model. We have introduced stochasticity and allowed housing servers to be empty with some positive probability, but we are still limited in what real-world complexity we can model. Our choice of distribution for inter-arrival times and service times is also limited to the exponential distribution.

3.3 Discrete-event simulation model

Discrete-event simulation (DES) is a form of stochastic simulation which models the evolution of a complex system according to a chronological event list which is updated throughout the simulation. DES is a powerful modeling tool given its ability to incorporate bespoke system complexities. It also naturally accomodates stochasticity by using a stream of Uniform($\min = 0, \max = 1$) pseudo-random numbers to drive the generation of random variates for model variables such as inter-arrival and service times. The drawback of DES is its computational cost which arises from the need to step through each event which occurs within the modelled system over the time horizon.

We have developed a DES model of our homelessness care system, which is a Python version of the DES model of Singham et al. (2023), built in Simio. In this model, shelter acts as
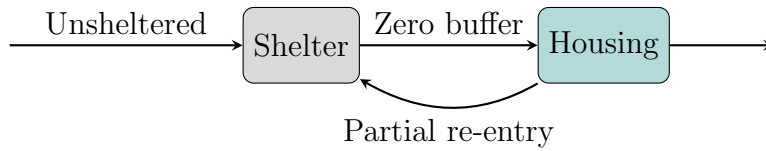
Figure 3: Tandem queueing system with re-entries

a server, giving rise to a tandem queueing system, as illustrated in Figure 3. A proportion of those leaving housing re-enter the system, as illustrated, and the exponential and triangle distributions are used for inter-arrival and housing service times, respectively.

While this is still a relatively simple DES model, it provides the framework for us to add many more system complexities should we so desire. For example, we could model the non-zero time needed to occupy a house with a new resident following the departure of its previous resident. We could also model the process of conversion of shelter to housing, which is a strategy proposed by Alameda County to improve services in the long term. We thus treat our DES model as a high-fidelity model, capable of the highest level of modelling accuracy, in comparison to our low- and medium-fidelity models, albeit at a computational cost.

4 Deterministic optimisation with low-fidelity model

In this section, we will present an optimisation formulation applied to our deterministic fluid flow model. This formulation will optimise the levels of housing and shelter to be built over time, and the objective function will attempt to minimise the unsheltered and sheltered population over the modelling horizon. Aside from the obvious need for cost constraints, we improve the real-world feasibility of our solutions by incorporating policy-based shape constraints on the housing and shelter capacities, which are functions of time. For example, though it would be better from a queueing standpoint to spend all financial resources upfront, the reality is that investment from the taxpayer will likely ramp up over a period of years. Furthermore, though their is political appetite for the use of emergency shelter in the short term, this is not seen as a sustainable long-term solution, so we introduce a unimodal shape constraint on the shelter capacity function.

4.1 Optimisation formulation

First, we present the basic notation associated with the terms in our formulation. The associated numerical results will be presented in Section 4.2. 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 (6) and (7).

- 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.

- 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.

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 $y(\boldsymbol{h}, \boldsymbol{s})$ be a deterministic quadratic objective function, evaluated using the fluid flow model equations (5), (6) and (7). We use 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.

$$y(\boldsymbol{h}, \boldsymbol{s}) = \frac{1}{D} \sum_{d=1}^{D} u_d(\boldsymbol{h}, \boldsymbol{s})^2 + \frac{w}{D} \sum_{d=1}^{D} s_d^D(\boldsymbol{s})^2. \tag{9}$$

We now introduce policy-based shape constraints. First, we ensure that the rate of housing capacity increase must stay the same or increase over T_a to reflect the fact that 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:

$$\Phi = \min_{\boldsymbol{h}, \boldsymbol{s}} \quad y(\boldsymbol{h}, \boldsymbol{s}) \tag{10}$$

$$\text{s.t.} \quad \sum_{t=1}^{t'} c_h[h_t - h_{t-1}] + c_s[s_t - s_{t-1}] \leq C \quad \forall t' \in \{1, ..., T_a\} \tag{11}$$

$$h_0 = H_0 \tag{12}$$

$$h_t \geq h_{t-1} \quad \forall t \in \{1, ..., T_a\} \tag{13}$$

$$h_t = h_{T_a} \quad \forall t \in \{T_a + 1, ..., T_a + T_b\} \tag{14}$$

$$h_{t+1} - h_t \geq h_t - h_{t-1} \quad \forall t \in \{1, ..., T_a - 1\} \tag{15}$$

$$s_0 = S_0 \tag{16}$$

$$s_t \geq s_{t-1} \quad \forall t \in \{1, ..., T_c\} \tag{17}$$

$$s_t \leq s_{t-1} \quad \forall t \in \{T_c + 1, ..., T_a\} \tag{18}$$

$$s_t \geq s_0 \quad \forall t \in \{T_c + 1, ..., T_a\} \tag{19}$$

$$s_t = s_{T_a} \quad \forall t \in \{T_a + 1, ..., T_a + T_b\}. \tag{20}$$

Constraints (11) ensure the total budget is never exceeded. A single budget constraint 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 (12) and (16) enforce the initial housing and shelter capacities. Constraints (13) ensure the housing capacity monotonically increases, while constraints (15) ensure the rate of change of housing capacity also monotonically increases from year to year. Constraints (17) ensure the shelter

Table 1: Model parameters

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

capacity monotonically increases up to the mode T_c and constraints (18) ensure it subsequently decreases monotonically. Constraints (19) ensure the shelter capacity never drops below its initial capacity. Finally, constraints (14) and (20) fix h_t and s_t during the horizon T_b after the building horizon has occurred.

4.2 Numerical results

In Table 1 we list the model parameters we used when optimising formulations Φ. These approximate values are taken from Alameda County (2022) and Singham et al. (2023). 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 Figure 4 we illustrate the model dynamics for the optimal solution to Φ, which involves the building of a mixture of extra housing and extra shelter. 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 investment in housing. 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.

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 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
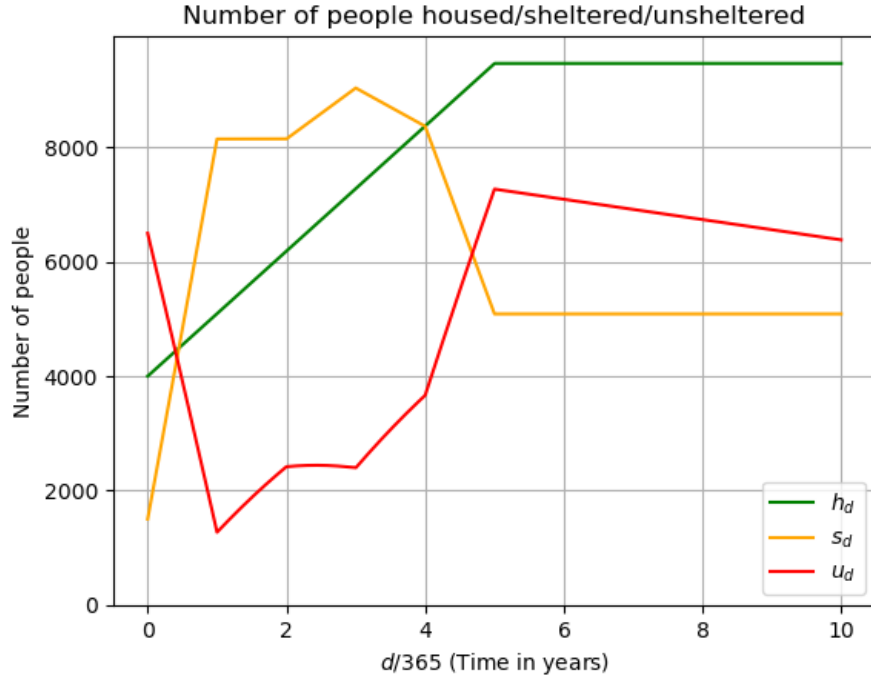
Figure 4: Optimal solution for Φ.

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.

Building Type	Initial Capacity	Year 1	Year 2	Year 3	Year 4	Year 5
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 5: Optimal capacity per year (Proportion of total budget spent per year)

In Figure 5 we detail the optimal solution to Φ 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. The optimal solution spends the maximum possible budget of 200,000,000 USD. The solution 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 to enable the continued investment in housing.

We solved this problem in 0.759 seconds using the IPOPT solver in Pyomo. All code used for this analysis is publicly available on GitHub.

17

4.3 Determinsitic optimisation: discussion

Most capacity planning optimisation formulations we have reviewed in the literature (e.g. in runaway homeless youth systems) 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. Our quadratic objective function is able 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 alternative formulations. Smoothness constraints on the uni-modal 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. Alternatively, a goal programming formulation which penalised deviations from a time-dependent goal on the unsheltered population would be interesting to explore.

We have here performed deterministic optimisation using a low-fidelity fluid flow model. This has enabled us to develop and explore an optimisation formulation which captures the key objectives, trade-offs and constraints. It has also enabled us to quickly get a feel for the nature of good solutions to the time-dependent capacity planning problem using a simple model. The natural next step is to develop a stochastic optimisation model to account for parameter uncertainty.

5 Discussion of uncertainty

Our optimisation in Section 4 assumes that we know the future arrival rates of homeless people into the homeless care system, and that we know their service times in housing. The fluid flow model projects forward the resulting unsheltered queue based on these exact quantities. In reality, actual inter-arrival times (or service times) are realisations of some unknown true distribution which we can only estimate using information available to us. Even if we knew the true distributions, we would not know what the next inter-arrival time (or service time) would be, due to the random nature of these processes. It is important that we consider these aspects of uncertainty as we move forward to consider simulation optimisation.

As discussed in Section 3, a stochastic simulation model offers a high-fidelity model of the homeless care system. It incorporates the stochasticity in inter-arrival and service times, and it allows us to capture certain important real-world complexities, such as the conversion of shelter to housing. There are two types of uncertainty in stochastic simulation: input uncertainty and stochastic uncertainty. Input uncertainty arises from not knowing the true input models for, say, inter-arrival and service times. Stochastic uncertainty arises from only ever simulating a finite amount of random variates from the input models used. The latter can be mitigated by performing many simulation replications. The former is typically mitigated by bootstrapping from the data used to fit the input models, to fit new input models which can be used to generate new random variates in new simulation replications.

Simulation optimisation suffers from both input uncertainty and stochastic uncertainty as a direct result of using stochastic simulation to model system performance. The main focus of this PhD research, at least initially, will be multi-fidelity simulation optimisation where we do not consider the effects of input uncertainty. I.e. we assume that we know the true distributions for inter-arrival times and service times. We do, of course, consider stochastic uncertainty: i.e. given known input models, with a finite comptuational budget we can only estimate the performance of a solution. We will endeavour to use multi-fidelity models alongside stochastic simulation so as to identify optimal or close to optimal solutions in a computationally efficient way, given the inherent stochastic uncertainty.

We do here acknowledge that for the homeless care system, there are extremely high levels of input uncertainty. Even if we fitted an input model for inter-arrival times using an extensive dataset of inter-arrival times of homeless people in recent years, we would still be highly uncertain of whether this model would accurately predict future inter-arrival times because the future is full of unknowns. It is simply impossible to predict future inter-arrival times because they are connected with many other events which are highly unpredictable, such as house prices, political decisions and global health and climate emergencies. Our decision to not consider input uncertainty is therefore not without due consideration. We anticipate that incorporating multi-fidelity models into simulation optimisation is challenging enough without input uncertainty. Once we have made progress with this, we will consider how our methodology could be adapted to consider input uncertainty. To this end, we may first consider simulation optimisation with high input uncertainty, without incorporating multi-fidelity models. This in itself is an active area of research.

6 Potential contributions at intersection of integer-ordered and multi fidelity SO

- Using low-fidelity models to quickly compute gradients in RSPLINE/DSA.
- Adding prior information to GMRF using low-fidelity model.
- Modelling errors of low fidelity models using GMRF.

7   Conclusion

References

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

Cao, Z., Li, H., Chew, E. P., Wang, H., and Tan, K. C. (2023). Cluster-based sampling allocation for multi-fidelity simulation optimization. In *2023 Winter Simulation Conference (WSC)*, pages 3448–3459. IEEE.

Chong, L. and Osorio, C. (2018). A simulation-based optimization algorithm for dynamic large-scale urban transportation problems. *Transportation Science*, 52(3):637–656.

El Hage, J., Gravitt, P., Ravel, J., Lahrichi, N., and Gralla, E. (2021). Supporting scale-up of Covid-19 RT-PCR testing processes with discrete event simulation. *PLoS One*, 16(7):e0255214.

Green, L. and Kolesar, P. (1991). The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, 37(1):84–97.

Green, L., Kolesar, P., and Svoronos, A. (1991). Some effects of nonstationarity on multiserver Markovian queueing systems. *Operations Research*, 39(3):502–511.

Huang, D., Allen, T. T., Notz, W. I., and Miller, R. A. (2006). Sequential kriging optimization using multiple-fidelity evaluations. *Structural and Multidisciplinary Optimization*, 32:369–382.

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.

Jennings, O. B., Mandelbaum, A., Massey, W. A., and Whitt, W. (1996). Server staffing to meet time-varying demand. *Management Science*, 42(10):1383–1394.

Jian, N. and Henderson, S. G. (2015). An introduction to simulation optimization. In *2015 winter simulation conference (wsc)*, pages 1780–1794. IEEE.

Kaya, Y. B., Maass, K. L., Dimas, G. L., Konrad, R., Trapp, A. C., and Dank, M. (2022a). Improving access to housing and supportive services for runaway and homeless youth: Reducing vulnerability to human trafficking in new york city. *IISE Transactions*, pages 1–15.

Kaya, Y. B., Mantell, S., Maass, K. L., Konrad, R., Trapp, A. C., Dimas, G. L., and Dank, M. (2022b). Discrete event simulation to evaluate shelter capacity expansion options for lgbtq+ homeless youth. In *2022 Winter Simulation Conference (WSC)*, pages 1033–1044. IEEE.

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

L. Salemi, P., Song, E., Nelson, B. L., and Staum, J. (2019). Gaussian markov random fields

for discrete optimization via simulation: Framework and algorithms. *Operations Research*, 67(1):250–266.

Lim, E. (2012). Stochastic approximation over multidimensional discrete sets with applications to inventory systems and admission control of queueing networks. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 22(4):1–23.

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 preprint arXiv:2212.04534*.

Oakland-Berkeley-Alameda County Continuum of Care (2021). Centering racial equity in homelessness system design. `https://everyonehome.org/wp-content/uploads/2021/02/2021-Centering-Racial-Equity-in-Homeless-System-Design-Full-Report-FINAL.pdf`.

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.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC.

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

Wang, H., Pasupathy, R., and Schmeiser, B. W. (2013). Integer-ordered simulation optimization using r-spline: Retrospective search with piecewise-linear interpolation and neighborhood enumeration. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 23(3):1–24.

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

Xu, J., Zhang, S., Huang, E., Chen, C.-H., Lee, L. H., and Celik, N. (2016). Mo2tos: Multi-fidelity optimization with ordinal transformation and optimal sampling. *Asia-Pacific Journal of Operational Research*, 33(03):1650017.