Multi-fidelity modelling and optimisation with applications in long-term public sector capacity planning: ten month PhD review

Graham Burgess

November 2024

1 Introduction

Long-term capacity planning is a challenge for organisations worldwide, especially for public sector bodies who must meet basic public needs with tight financial budgets. Examples of such problems arise in housing, healthcare and prisons. In housing, decision makers must split their resources between emergency shelter and permanent social housing to alleviate homelessness. In healthcare, planners must balance the need for critical and non-critical care services when designing new hospitals. In prisons, capacity must be sufficient to avoid needing to release prisoners early but not so high as to risk costly unused capacity.

Operational research methods offer helpful tools to support such public sector decision-making. Optimisation helps by looking for a feasible solution which performs best across a 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. The flow of people through public services can often be modelled as a queue. For example, in a homeless care system, 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 long-term capacity plan for housing and shelter.

The most accurate models of real-world queueing systems are often high-fidelity stochastic simulation models. 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 but the issue of limited computational resources pervades all SO methods. This issue stems from the fact that a stochastic simulation is typically computationally expensive to run, and multiple 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.

In order to capitalise on the strengths of high-fidelity stochastic simulation, we plan to develop a SO approach to tackle long-term public sector capacity planning problems. Integer-ordered SO is naturally applicable since decision variables are often the number of units (or the number of multiple units) of resource to make available. Multi-fidelity methods are also attractive given the availability of suitable multi-fidelity queueing models of public sector systems. In multi-fidelity SO, low-fidelity models are used alongside high-fidelity stochastic simulation which reduces the computational burden and therefore enables an optimal solution to be found more efficiently. We therefore propose tackling technical challenges at the interface of multi-fidelity and integer-ordered SO, which we elaborate on in Section 6.

Applying SO methods to long-term public sector capacity planning problems poses further challenges. If service times are long (as in housing or in prisons), uncertainty in the input models can overshadow stochastic uncertainty. An extension of our research will be to appropriately incorporate this input uncertainty. Secondly, with long planning horizons, decision makers typically revisit and adjust their plans on a regular basis, in light of new information. This dynamic element to the decision-making process is rarely addressed in the SO literature and will motivate another extension to our 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. 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 including a low-fidelity fluid flow model; Section 4 introduces an optimisation formulation which addresses a capacity planning problem for homeless care systems. We solve this problem in a determinsitc setting using the fluid flow model. Addressing long-term capacity planning by constructing and optimising a fluid flow model (Burgess et al., 2024) is a novel development. In Section 5 we discuss how different types of uncertainty affect optimisation in long-term capacity planning problems. Finally, in Section 6 we discuss possible research contributions at the intersection of integer-ordered and multi-fidelity SO.

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. Lentle et al. (2024) use 20-year long judgemental demand forecasts and DES modelling to support long-term bed capacity planning for a new hospital. Forecasts use a truncated normal distribution to sample annual growth. Each sample is dissaggregated to a 3-hourly demand projection which then acts as a model input. They test a range of scenarios to identify capacity levels which are unlikely to be full more than 5% of the time. 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 stationary queues 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 trafficking. 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 benefit 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 tackling long-term capacity planning problems. This is partly because SO is not a well known OR technique amongst practitioners. It is also because 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. One might therefore consider SO overkill for public sector problems with a small number of feasible solutions. However, experience supporting public sector decisions suggests that modelling system performance on a fine-grained solution space and recommending solutions therein are valuable aids. These analyses can help planners to establish new solutions for consideration and can support a decision on a general direction of travel in capacity planning. This PhD research will explore to what extent novel MFSO methods can give meaningful analytical support to public sector decision makers in long-term capacity planning problems.

The rest of this section is organised as follows. Section 2.2.1 summarises several high-level categories of SO methods, particularly those for discrete-valued solution spaces. Sections 2.2.2 and 2.2.3 focus on the detail of integer-ordered methods and multi-fidelity methods, respectively.

2.2.1 Overview of SO methods

In SO, as described by Nelson and Pei (2021), one is typically interested in solving the following optimisation problem:

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

where $f(\boldsymbol{x})$ 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. Here we discuss both categories with a focus on discrete problems.

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. Indifference-zone (IZ) R&S methods start with δ, set by the user to define the smallest difference between objective values which is deemed important. IZ methods provide a guarantee that if the true best solution is at least δ better than the second best, the probablity of selecting this solution is at least $1-\alpha$ where α is set by the user and $0 < \alpha < 1$. A popular IZ R&S method is that of Kim and Nelson (2001) who initially simulate each solution a relatively small number of times. A decision is then made iteratively about which solutions to discard and which to keep simulating.

Other frequentist R&S methods work by optimally allocating computing budget to the simulation of different solutions. The goal is to reduce the probability of selecting a sub-optimal solution to zero at the fastest rate (i.e. with as few simulation replications as possible). Glynn and Juneja (2004) show such a rate-optimal allocation rule for both normally distributed output data and for general output data distributions. Bayesian R&S methods can incorporate prior belief about solution performance and make a probabilistic posterior assessment which can be used iteratively to decide where to simulate next. For example, at each iteration of their algorithm, Chen and Ryzhov (2019) choose to simulate either the sample best solution or the solution with the best expected improvement relative to the sample best.

In discrete problems where it is not realistic to simulate 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. The adaptive hyperbox algorithm of Xu et al. (2013) places positive probability on feasible solutions in or on a hyperbox around the sample best solution. The boundaries of the hyperbox are drawn at the solutions which have been simulated and are closest to the sample best solution in one or more coordinate direction. 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, decision variables are defined on an integer-ordered scale. For example, capacity planning problems often involve integer units of capacity, such as a number of beds. **Integer-ordered** SO methods exploit the spatial relationships between solutions when looking for optimal solutions. Some integer-ordered methods work with gradients, while others use a neighbourhood structure. We discuss key integer-ordered methods in Section 2.2.2.

We now briefly discuss relevant methods for continuous problems, since discrete versions of these methods are becoming popular. **Stochastic approximation** methods (Fu, 2006) use gradient estimators to search for locally optimal solutions where the true objective function $f(\boldsymbol{x})$ is continuous and differentiable in $\boldsymbol{x}$. In **sample-average approximation** (Kim et al., 2015) one fixes the random numbers used for the simulation of each feasible solution. The goal is then to minimise the deterministic sample-average objective function. **Meta-model** approaches capitalise on the low computational cost of an alternative model of the system in question. Local meta-models usually involve polynomial regression models, as described by Barton and Meckesheimer (2006). The gradient of the polynomial meta-model is used in the search. Global meta-models, such as Gaussian process (GP) models, approximate performance across the solution space. Bayesian optimisation (BO) (Frazier, 2018) uses a GP meta-model and maximises an acquisition function to determine where to perform more simulation.

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 to an objective function $\tilde{f}$ over a continuous domain. Stochastic approximation, as discussed in Section 2.2.1, is performed on the continuous domain to move from point to point down the gradient of $\tilde{f}$. The gradient of $\tilde{f}$ at a point $\boldsymbol{x}$ is estimated by performing multiple simulation replications at each integer point on a simplex surrounding $\boldsymbol{x}$ and taking differences between the resulting estimates of f. 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). 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. It 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 continuous domain and the 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 estimated gradient, 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 one unit away in any dimension. Within each iteration of SPLINE, 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) (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 solution. The Markov property holds if one can predict the objective function of any solution only with objective function information from the neighbourhood. That is, each random variable is conditionally independent of all random variables outside of the corresponding neighbourhood, conditional on the random variables in the neighbourhood. A GMRF is defined by its mean vector and precision matrix (inverse of the covariance matrix). Zeros in a precision matrix correspond to conditional independence between random variables, given all other random variables (see Rue and Held (2005) for proof). 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.

Salemi et al. (2019) introduce the Gaussian Markov improvement algorithm (GMIA). The conditional distribution for the GMRF (conditional on all previous simulation results) is updated as more simulation is performed. Following each update, a decision is made about where to simulate next using a complete expected improvement (CEI) criterion. The CEI criterion is similar to the expected improvement (EI) criterion discussed in Section 2.2.1, but it treats both the objective function at the current best solution and that at each candidate solution as random variables. Once the maximum CEI across all candidate solutions is below some threshold δ, the algorithm terminates and by using CEI to decide when to stop, one has similar statistical confidence on the selected best as is provided by indifference-zone R&S (see Section 2.2.1). The GMIA algorithm converges with probability 1 to the globally optimal solution based on the fact that without the stopping rule, as the number of iterations goes to infinity, each solution will be simulated infinitely often.

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, which we now discuss.

One approach is to **perform a deterministic optimisation (DO) first** using a deterministic low-fidelity model. This helps develop a better understanding of the dynamics of the system in question and of the nature of good solutions. Also, an optimal solution from a DO procedure can be used as a initial solution for a SO procedure, giving the latter a ‘warm start’. This approach is explained by Jian and Henderson (2015) in the context of managing the supply of bicycles 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/departures of bicycles to/from stations, using a fluid flow model of the system. They use this to find simple rules of thumb for identifying the optimal number of bicycles at 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 full/empty stations. Finally, they capture more system complexity using a discrete-event simulation model. They attempt to solve the resulting SO problem using a random search. By starting their search at the optimal solution from the queueing model optimisation, they find good solutions with less simulation effort.

Another approach, aimed at discrete problems, is **multi-fidelity optimisation with ordinal transformation and optimal sampling** (MO2TOS), first proposed by Xu et al. (2016). 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 enable the subsequent optimal sampling scheme to more efficiently use high-fidelity simulation to search for an optimal solution. The optimal sampling scheme inherits from the well-known optimal computing budget allocation (OCBA) rule. The performance of the algorithm does naturally depend on the quality of the low-fidelity used, which is measured using the correlation between low- and high-fidelity outputs. 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 if the low-fidelity evaluation is not very cheap.

We now discuss multi-fidelity SO approaches which **model the error of low-fidelity models**. Chong and Osorio (2018) optimise a traffic system using an expensive high-fidelity traffic simulator. Their meta-model has a physical component (an analytical queueing model) and a polynomial error term. They iteratively fit their meta-model using high-fidelity simulation output and solve a trust-region problem using the meta-model to identify a candidate solution for more simulation.

Huang et al. (2006) present a Gaussian process (GP) approach to modelling low-fidelity model error. Their approach involves multi-fidelity models, where the fidelity levels place them in order of increasing accuracy. Fidelity level is denoted by l where $l \in \{1, ..., m\}$ and m is the total number of models of multi fidelity. The objective function at point $\boldsymbol{x}$ with model l is modelled as $f_l(\boldsymbol{x}) = f_{l-1}(\boldsymbol{x}) + \delta_l(\boldsymbol{x})$ where $\delta_l(\boldsymbol{x})$ is the systematic error of model $l-1$ at point $\boldsymbol{x}$ with respect to model l. $\delta_l(\boldsymbol{x})$ is modelled as the sum of a linear model, systematic bias and noise:

$$\delta_l(\boldsymbol{x}) = \beta_l^T \boldsymbol{b}_l(\boldsymbol{x}) + Z_l(\boldsymbol{x}) + \epsilon_l \qquad \forall l \in \{1, 2, ..., m\}, \tag{2}$$

where $\boldsymbol{b}_l(\boldsymbol{x})$ is a vector of basis functions of $\boldsymbol{x}$ and β_l is a vector of coefficients for the linear model. $Z_l(\boldsymbol{x})$ is modelled as a zero-mean Gaussian process and ϵ_l is random noise. The multi-fidelity sequential kriging optimisation procedure of Huang et al. (2006) sequentially fits $\hat{f}_l(\boldsymbol{x})$, a linear predictor of $f_l(\boldsymbol{x})$, to simulation output and then decides where to simulate next and with which model. These decisions are made using an augmented expected improvement (EI) criterion. The augmented EI criterion for model l at point $\boldsymbol{x}$ does not only capture the expected improvement from simulating at point $\boldsymbol{x}$ using the high fidelity model. It also captures the reduced reward from using a lower-fidelity model l, the computational cost associated with model l and the diminishing returns from additional replicates from model l.

Finally, we summarise **multi-fidelity batch BO** where solution-fidelity pairs in a batch are simulated in parallel at each iteration. Information-theoretic acquisition functions are used to efficiently search for an optimal solution. In information theory, for a random variable X, the differential entropy $h(X)$ is a measure of the uncertainty in X and is defined as $h(X) = -\mathbb{E}_X[\log(p(X))]$ where $p(X)$ is the probability density function for X. In max-value entropy search (MES) (Wang and Jegelka, 2017), one works with $h(f^*)$, where f^* is the unknown true best objective value. Here one seeks to simulate using the solution(s) and fidelity level(s) which most reduce the differential entropy $h(f^*)$. This is done using an acquisition function based on the expected reduction in $h(f^*)$, a quantity known as the mutual information (MI) between candidate solution-fidelity evaluations and f^*. The main problem with MES is the expression for $p(f^*)$ which is typically lacking in closed-form but needed to compute differential entropies. Moss et al. (2021) introduce GIBBON by defining information gain (IG) as the reduction in entropy associated with knowing the best objective value from a candidate batch of solution-fidelity pairs. The authors approximate IG using a lower bound which has a simple analytical form. The acquisition function is the expected IG where the expectation is taken across f^* using a Monte-Carlo estimator. Samples of f^* are generated using a Gumbel distribution which is fit to empirical data. GIBBON has been shown to perform well versus other acquisition functions based on differential entropy.

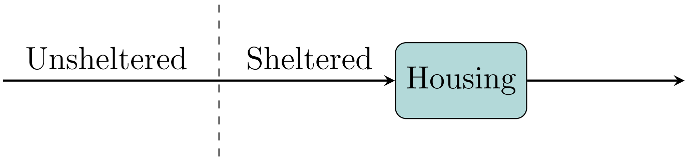


Figure 1: Simple queueing system

3 Models of multi fidelity

Here we introduce three models which we have developed for homeless care systems from low- to high-fidelity. We have used each one to model the flow of people through the homeless care system in Alameda County, California. 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 tandem queues, 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 here acknowledge that service times are long in this setting compared to a typical modelling horzion, a feature which is unusual in queueing problems. We discuss the implications of this on uncertainty in Section 5.

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

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

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

$$\begin{aligned} u(t) &= n(t) - h(t) - s(t) && (3)\\ &= n_0 + \int_0^t \lambda(s)ds - \int_0^t \mu_0 h(s)ds - h(t) - s(t), && (4) \end{aligned}$$

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

When analyzing the dynamics of the fluid flow model over a modeling horizon, we 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 (4) with the sum

$$u_d = n_0 + \sum_{i=1}^{d} \lambda_i \delta t - \sum_{i=1}^{d} \mu_0 h_i^D \delta t - h_d^D - s_d^D, \tag{5}$$

where μ_0 is the daily service rate of a single housing unit and the stepsize $\delta t = 1$ day.

In Figure 2(a) we give an illustrative example of the dynamics of u_d given by our fluid flow model for the homeless care system in Alameda County. We use 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 the DES model of Singham et al. (2023). Note: in this DES model the numbers of people and housing/shelter units are scaled down by a factor of 100 from realistic estimates to reduce the computational burden of simulation. This is just one example of a long-term capacity planning problem which we use to illustrate how our model works. The plot 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.

In Section 4 we will evaluate (5) 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 (5) becomes

$$u_d(\boldsymbol{h}, \boldsymbol{s}) = n_0 + \sum_{i=1}^{d} \lambda_i \delta t - \sum_{i=1}^{d} \mu_0 h_i^D(\boldsymbol{h}) \delta t - h_d^D(\boldsymbol{h}) - s_d^D(\boldsymbol{s}), \tag{6}$$

where

$$h_d^D(\boldsymbol{h}) = h_{\lfloor \frac{d}{365} \rfloor} + \left(\frac{d}{365} - \left\lfloor \frac{d}{365} \right\rfloor \right) (h_{\lceil \frac{d}{365} \rceil} - h_{\lfloor \frac{d}{365} \rfloor}) \tag{7}$$

and

$$s_d^D(\boldsymbol{s}) = s_{\lfloor \frac{d}{365} \rfloor} + \left(\frac{d}{365} - \left\lfloor \frac{d}{365} \right\rfloor \right) (s_{\lceil \frac{d}{365} \rceil} - s_{\lfloor \frac{d}{365} \rfloor}). \tag{8}$$

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 6.

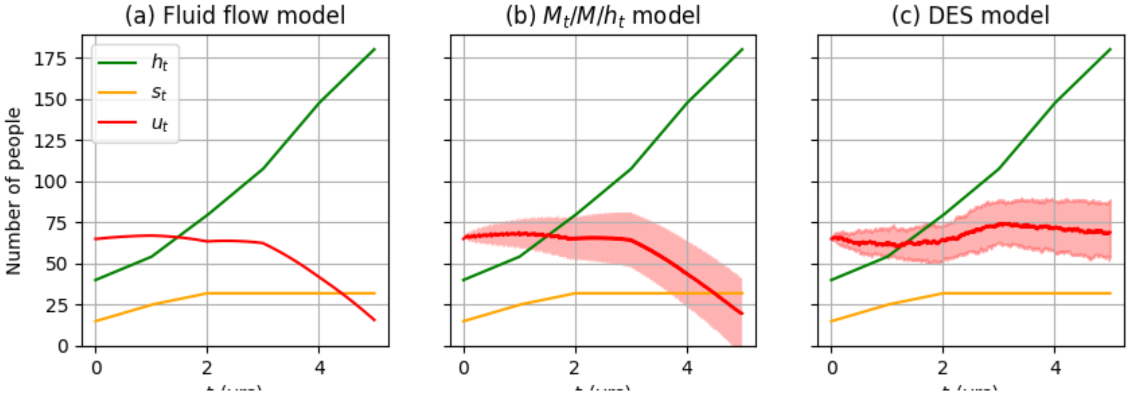


Figure 2: Multi-fidelity modelling of homeless care system in Alameda County, California. Plots include dynamics of u_t (unsheltered), s_t (sheltered) and h_t (housed). Common modelling inputs: Initial # people: 120, Annual service rate per house ($\frac{1}{\text{yr}}$): 0.22, Annual arrival rates ($\frac{1}{\text{yr}}$): $36.6, 43.5, 47.8, 47.8, 43.1$.

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 discrete units. We introduce stochasticity and would like to compute the probability distribution for the number of people housed, sheltered and unsheltered, throughout the time horizon.

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. We note here that working with the exponential distribution in discrete time is equivalent to using a geometric distribution, which also enjoys a memoryless property. We may later use this equivalent framework, but for now continue with the exponential distribution.

In our $M_t/M/h_t$ queueing model, 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. Let p_t^n be the probability of being in state n at time t.

If we know the initial probabilities p_0^n 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 $p_{t+\Delta t}^n$ as:

$$\begin{aligned} p_{t+\Delta t}^n &= p_t^n(1 - \lambda_t \Delta t - \mu_0 m_t^n \Delta t) \\ &+ p_t^{n+1}(\mu_0 m_t^{n+1} \Delta t) \\ &+ p_t^{n-1}(\lambda_t \Delta t) \\ &+ \ o(\Delta t) \qquad 1 \leq n \leq N - 1 \end{aligned}$$

$$\begin{aligned} p_{t+\Delta t}^0 &= p_t^0(1 - \lambda_t \Delta t) \\ &+ p_t^1(\mu_0 m_t^1 \Delta t) \\ &+ \ o(\Delta t) \end{aligned}$$

$$\begin{aligned} p_{t+\Delta t}^N &= p_t^N(1 - \mu_0 m_t^N \Delta t) \\ &+ p_t^{N-1}(\lambda_t \Delta t) \\ &+ \ o(\Delta t) \end{aligned}$$

where λ_t is the arrival rate at time t and $m_t^n = \min(n, h_t)$ 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 λ_t and h_t vanishes as $\Delta t \to 0$. There is also bias associated with using finite N. This can be mitigated by choosing N to be large, but not so large that computation is slow. The probabilities p_t^n can 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:

$$\mathbb{E}[u_t] = \sum_{n=0}^{N} p_t^n \max(0, n - h_t - s_t), \tag{9}$$

where s_t is the time-dependent number of shelters.

In Figure 2(b) we show outputs of our $M_t/M/h_t$ model of the homeless care system in Alameda County. Where applicable, we have used the same model inputs as we used with the fluid flow model. It can be seen that the expected size of the unsheltered population matches what we found using the fluid flow model. Our $M_t/M/h_t$ model also gives us distributional information on our outputs. For example, with the given inputs, we can see that there is a small chance that the unsheltered population vanishes at the end of the modelling horizon.

The $M_t/M/h_t$ model is more computationally costly than the fluid flow model, but still cheap compared to using a discrete-event simulation model. We have introduced stochasticity and relaxed the server-always-busy assumption, 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. A single run of a DES model can be computationally cheaper than that of an analytical queueing model such as our $M_t/M/h_t$ model. The former proceeds from event to event whereas the latter must proceed in small time steps. However, one must run a DES model many times to obtain an output distribution, which means using such a model is computationally expensive.

We have developed a DES model for homeless care systems, which is based upon the DES model of Singham et al. (2023). In this model, shelter acts as a server, giving rise to a tandem queueing system, as illustrated in Figure 3. In Figure 2(c) we show outputs of our DES model of the homeless care system in Alameda County. Where applicable, we have used the same model inputs as we used with the fluid flow model and the $M_t/M/h_t$ model. We use the following additional model inputs for the DES model:

- The service time at shelter is modelled as exponential with mean 3 months, reflecting a short time in comparison to time spent in housing.
- A triangular distribution (lower limit 0 years, upper limit 8 years, mode 6 years) is used to model the service time at housing. This bounded unimodal distribution was chosen by Singham et al. (2023) to ensure a realistic number of people left housing during the simulation horizon of 5 years.
- 17% of those leaving housing re-enter the queue for shelter, reflecting estimates from Alameda County.

To sample the remaining housing service time for a customer already in housing at the start of the simulation, we use the procedure outlined in Algorithm 1. We here assume that we do not know how long those currently in housing have already been in service.

Algorithm 1 Sample a customer's remaining housing service time using Triangle distribution

```
Set Triangle distribution parameters: $a$, $b$, $c$
Set Done = False
Sample $x_0 \sim \text{Triangle}(a, b, c)$. This represents service time already completed.
while not Done do
    Sample candidate $x \sim \text{Triangle}(a, b, c)$. This represents total service time.
    if $x \geq x_0$ then
        $x_1 = x - x_0$. This represents remaining service time.
        Done = True
    end if
end while
Return remaining service time $x_1$
```

The DES model shows longer unsheltered queues than the lower-fidelity models. This is because our DES model includes a non-zero service time at shelter and a non-zero proportion of those leaving housing rejoining the queue for shelter. 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

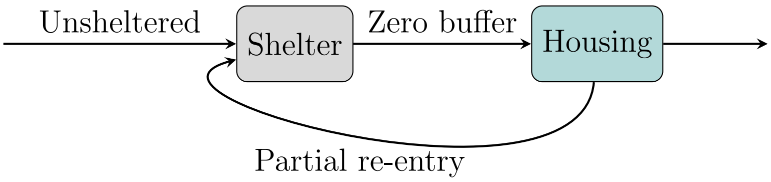


Figure 3: Tandem queueing system with re-entries

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 wilfl 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 there 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. To the best of our knowledge, this is the first time long-term capacity planning is addressed by constructing and optimisng a tractable fluid flow model.

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 (7) and (8).
- 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 $f(\boldsymbol{h}, \boldsymbol{s})$ be a deterministic quadratic objective function, evaluated using the fluid flow model equations (6), (7) and (8). 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.

$$f(\boldsymbol{h}, \boldsymbol{s}) = \frac{1}{D}\sum_{d=1}^{D} u_d(\boldsymbol{h}, \boldsymbol{s})^2 + \frac{w}{D}\sum_{d=1}^{D} s_d^D(\boldsymbol{s})^2. \tag{10}$$

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:

$$\begin{aligned}
\Phi = \quad & \min_{\boldsymbol{h},\boldsymbol{s}} f(\boldsymbol{h}, \boldsymbol{s}) && && (11)\\
\text{s.t.} \quad & \textstyle\sum_{t=1}^{t'} c_h[h_t - h_{t-1}] + c_s[s_t - s_{t-1}] \leq C && \forall t' \in \{1, ..., T_a\} && (12)\\
& h_0 = H_0 && && (13)\\
& h_t \geq h_{t-1} && \forall t \in \{1, ..., T_a\} && (14)\\
& h_t = h_{T_a} && \forall t \in \{T_a + 1, ..., T_a + T_b\} && (15)\\
& h_{t+1} - h_t \geq h_t - h_{t-1} && \forall t \in \{1, ..., T_a - 1\} && (16)\\
& s_0 = S_0 && && (17)\\
& s_t \geq s_{t-1} && \forall t \in \{1, ..., T_c\} && (18)\\
& s_t \leq s_{t-1} && \forall t \in \{T_c + 1, ..., T_a\} && (19)\\
& s_t \geq s_0 && \forall t \in \{T_c + 1, ..., T_a\} && (20)\\
& s_t = s_{T_a} && \forall t \in \{T_a + 1, ..., T_a + T_b\}. && (21)
\end{aligned}$$

Constraints (12) 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 (13) and (17) enforce the initial housing and shelter capacities. Constraints (14) ensure the housing capacity monotonically increases, while constraints (16) ensure the rate of change of housing

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	$\frac{0.22}{365} = 6.106 \times 10^{-4}$/day
w	0.3

capacity also monotonically increases from year to year. Constraints (18) ensure the shelter capacity monotonically increases up to the mode T_c and constraints (19) ensure it subsequently decreases monotonically. Constraints (20) ensure the shelter capacity never drops below its initial capacity. Finally, constraints (15) and (21) 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 formulation 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 steady increase in housing over the horizon T_a. 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

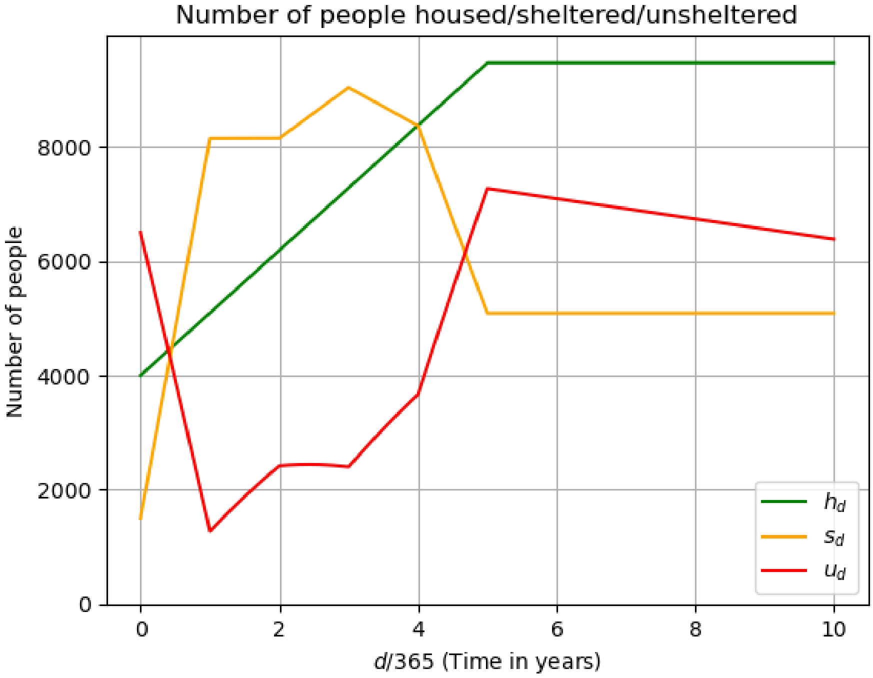
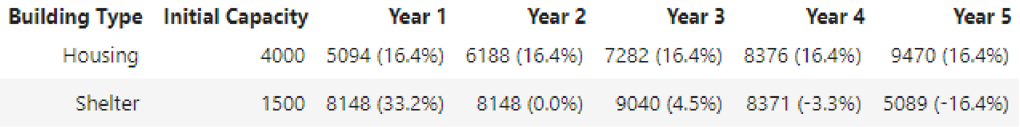


Figure 4: Optimal solution for Φ.

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.



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: `https://github.com/grahamburgess3/psor-paper-housing`.

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 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. 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 consider the effects of uncertainty on our problem.

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. These aspects of uncertainty pervade all long-term capacity planning problems and it is therefore important that we consider them as we move forward to consider simulation optimisation.

As discussed in Section 3, a stochastic simulation can offer a high-fidelity model of a queueing system. It incorporates the stochasticity in inter-arrival and service times, and it allows us to capture certain important real-world complexities. 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 simulating a finite amount of random variates from the input models. The latter can be mitigated by performing many simulation replications. The former can in some cases be quantified 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. In our case, however, bootstrapping is not appropriate for quantifying input uncertainty. Uncertainty in our input models does not come from a lack of historic data, but from the fact that the past is not necessarily a good indication of the future.

The relative importance of input and stochastic uncertainty varies when modelling different systems. If service times are long, such as in housing or prison models, then the stochastic uncertainty may be less important; one is more interested in high input uncertainty stemming from knowing neither future population demands nor service rates even for today's customers. If service times are short, such as in many healthcare applications, service time input models are typically more accurate and the stochastic uncertainty becomes more important.

Simulation optimisation suffers from both input uncertainty and stochastic uncertainty as a result of using stochastic simulation to model system performance. The main focus of this PhD research, at least initially, will be integer-ordered multi-fidelity simulation optimisation where we do not consider input uncertainty. We do, of course, consider stochastic uncertainty, which is a defining feature of stochastic simulation and simulation optimisation. This is an appropriate first step for a couple of reasons. Firstly, the initial research challenges (discussed in Section 6) are challenging enough without input uncertainty. Secondly, stochastic uncertainty is important in many long-term capacity planning problems, especially when service times are short.

We acknowledge that for the homeless care system, there are extremely high levels of input uncertainty. Even if we fitted input models for inter-arrival and service times using extensive historic data, we would still be highly uncertain of whether these models would accurately predict future inter-arrival and service times. This is because they are connected with other highly unpredictable events such as house prices, political decisions and global health and climate emergencies. Input uncertainty may be more important than stochastic uncertainty in this particular setting which motivates further research of input uncertainty in SO.

6 Potential research contributions

The next steps of this PhD research will address current challenges at the intersection of integer-ordered and multi-fidelity SO. An integer-ordered SO approach is naturally applicable to long-term public sector capacity planning problems since decision variables are often the number of units (or the number of multiple units) of resource to make available. Multi-fidelity methods are also attractive in this area given the availability of suitable multi-fidelity queueing models. We discuss below some initial ideas for contributing to the literature at the intersection of these active fields of research. We conduct this reseach in the spirit of exploring to what extent these methods can give meaningful analytical support to decision makers in the public sector.

Gradient-based integer-ordered SO methods such as DSA and R-SPLINE expend substantial simulation effort in estimating gradients on the continuous domain. For a given point on the d-dimensional continuous domain, to estimate the gradient, all $(d+1)$ integer points on the surrounding simplex must be simulated multiple times. This could be more efficient by using information from low-fidelity models. One could, for example, use solely a low-fidelity model for objective function estimation at points on the simplex. Alternatively one could use a meta-model defined as a combination of a low-fidelity model and an error term, using some simulation effort to fit this meta-model.

GMRF-based integer-ordered SO methods are popular in the literature but are hampered by the computational expense of inverting the conditional precision matrix which is needed to update the conditional distribution. Some efforts to address this problem aim to reduce the number of times a full inversion is required. For example, the 'rapid' GMIA algorithm (rGMIA) of (Semelhago et al., 2021) constructs a 'promising set' of solutions and efficiently searches in this set for some time by updating only the relevant parts of the conditional distribution. This leads to considerable efficiencies, but it does not involve the use of multi-fidelity model information.

The multi-fidelity GMIA algorithm (MFGMIA) of (Li et al., 2022) adds a new 'layer' of nodes to the GMRF graph for each low-fidelity model available. Each high-fidelity solution is connected to all of its low-fidelity counterpart nodes with an edge. The precision matrix and the resulting conditional distribution are designed to control the influence given to each low-fidelity model for each solution, based on the quality of each low-fidelity model and on the amount of high-fidelity simulation effort expended for each solution. These are good ideas, however there are still problems and room for improvement. Firstly, it is assumed that low-fidelity information for all feasible solutions is available from the start without cost, which is problematic if low-fidelity models do incur cost and/or if the solution space is very large. Secondly, the prior distribution for the GMRF is still a constant across the entire solution space for each model which does not use multi-fidelity information to its full potential.

A key potential research contribution is to explore the possibility of using low-fidelity information to bring structure to a prior distribution of system performance, without requiring low-fidelity evaluation of the entire solution space. Our research may also build upon the ideas of MFGMIA by transferring concepts from multi-fidelity sequential kriging optimisation (MFSKO) (Huang et al., 2006) where the cost of low-fidelity information is accounted for and there is heirarchy amongst the low-fidelity models. We may also merge ideas from rGMIA and MFGMIA to use low-fidelity information to help identify promising sets. We may also draw on information-theoretic concepts from multi-fidelity batch BO such as the use of entropy-based acquisition functions. Finally, we may expand the application of GMRFs to include modelling the error of low-fidelity models, again using concepts from MFSKO but on a discrete solution space.

We plan to extend our SO research by considering one or both of the following research challenges which arise when applying SO methods to long-term public sector capacity planning problems. First is the question of input uncertainty which can overshadow stochastic uncertainty as discussed in Section 5. Standard methods of bootstrapping from input data are not suitable since input uncertainty in this setting does not arise due to a lack of data but due to genuine uncertainty regarding the complex factors which could affect future arrival and service rates. An optimisation approach which considers a range of possible input models may prove appropriate. We might consider a range of input distribution classes as well as a range of input model parameters.

Secondly, with long planning horizons, decision makers typically revisit and adjust their plans on a regular basis, in light of new information. A dynamic optimisation approach may therefore be appropriate to consider. For example, suppose we require a T-year plan where $T > 1$. We might first use SO with the known current system state to propose a plan for the next L years where $1 < L < T$. We might then simulate the system for a year assuming we follow the proposed plan for the first year, and observe the resulting simulated system state. We could then optimise again with SO for the next L years based on the simulated system state after one year. By repeatedly simulating for a year and re-optimising we could obtain a T-year plan which incorporates the dynamic element of the decision-making process. Of course many replications would be needed to establish a good plan with a method like this, given the randomness associated with simulating the system with the first year of a proposed plan.

References

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

Barton, R. R. and Meckesheimer, M. (2006). Metamodel-based simulation optimization. *Handbooks in operations research and management science*, 13:535–574.

Burgess, G., Singham, D. I., and Rhodes-Leader, L. (2024). Time-varying capacity planning for designing large-scale homeless care systems. *Under Review.*

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.

Chen, Y. and Ryzhov, I. O. (2019). Complete expected improvement converges to an optimal budget allocation. *Advances in Applied Probability*, 51(1):209–235.

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.

Frazier, P. I. (2018). Bayesian optimization. In *Recent advances in optimization and modeling of contemporary problems*, pages 255–278. INFORMS.

Fu, M. C. (2006). Gradient estimation. *Handbooks in operations research and management science*, 13:575–616.

Glynn, P. and Juneja, S. (2004). A large deviations perspective on ordinal optimization. In *Proceedings of the 2004 Winter Simulation Conference, 2004.*, volume 1. IEEE.

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.

Kim, S., Pasupathy, R., and Henderson, S. G. (2015). *A Guide to Sample Average Approximation*, pages 207–243. Springer New York, New York, NY.

Kim, S.-H. and Nelson, B. L. (2001). A fully sequential procedure for indifference-zone selection in simulation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 11(3):251–273.

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.

Lentle, D., Sachser, V., Incze, E., Tako, A., Rostami-Tabar, B., Spencer, C., and Morgan, J. (2024). Using simulation for long-term bed modelling in critical care. *Journal of Simulation*, pages 1–17.

Li, D., Liu, H., Jin, X., Li, H., Chew, E. P., Tan, K. C., and Lin, Y. H. (2022). Multi-fidelity discrete optimization via simulation. In *2022 Winter Simulation Conference (WSC)*, pages 3170–3181. IEEE.

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*.

Moss, H. B., Leslie, D. S., Gonzalez, J., and Rayson, P. (2021). Gibbon: General-purpose information-based bayesian optimisation. *Journal of Machine Learning Research*, 22(235):1–49.

Nelson, B. and Pei, L. (2021). *Foundations and methods of stochastic simulation.* Springer.

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.

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.

Semelhago, M., Nelson, B. L., Song, E., and Wächter, A. (2021). Rapid discrete optimization via simulation with gaussian markov random fields. *INFORMS Journal on Computing*, 33(3):915–930.

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.

Wang, Z. and Jegelka, S. (2017). Max-value entropy search for efficient bayesian optimization. In *International Conference on Machine Learning*, pages 3627–3635. PMLR.

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

Xu, J., Nelson, B. L., and Hong, L. J. (2013). An adaptive hyperbox algorithm for high-dimensional discrete optimization via simulation problems. *INFORMS Journal on Computing*, 25(1):133–146.

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.