DISEASE MODELING AND RESOURCE SHARING

A Thesis
Presented to
The Academic Faculty

by

Yifan Wang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology
August 2020

DISEASE MODELING AND RESOURCE SHARING

Approved by:

Professor Pinar Keskinocak, Advisor
School of H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Institute of Technology

Professor Seong-Hee Kim, Co-advisor
School of H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Institute of Technology

Professor Julie Swann
School of H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Institute of Technology

Professor David Goldsman
School of H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Institute of Technology

Brian M. Gurbaxani, Ph.D.
Centers for Disease Control and Prevention

Date Approved: June 15, 2020

To my multiple personalities and all my imaginary friends.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Pinar Keskinocak, and co-advisor, Dr. Seong-Hee Kim, for their support, financially, intellectually, and emotionally. Their guidance helped me in all the time of research and writing of this thesis. They are truly the best mentors one could ask for.

I would like to thank my thesis committee members: Dr. Julie Swann, Dr. David Goldsman and Dr. Brian M. Gurbaxani for their kindness in evaluating my thesis and their valuable feedback. I would also like to thank my collaborators: Dr. Haengju Lee, Dr. Junzhuo Chen, Dr. Julie R. Gutman, Dr. Julie I. Thwing, Dr. Atul Vats ,and Tyler A. Perini for their excellent work. I would like to thank my previous advisor Dr. Eva K. Lee for her crucial help and Dr. Alan Erera for his guidance in this program.

I would like to thank my friends for accompanying me during my ups and downs, patiently listening to my random thoughts, and teaching me how to get a life.

Last but not least, I would like to thank my parents for their unconditional love and, more importantly, financial support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis makes contributions to three research topics: shared-resource allocation, disease modeling, and evaluation of intervention methods.

In Chapter 2, we consider a problem of allocating shared resources among multiple classes when a customer from a different class may require a different number of resources and give a different amount of rewards when leaving the system after service completion. A customer is rejected if the number of available resources at the time of her arrival is smaller than the number of resources required for the customer. In this chapter, we find a customer admission control policy that maximizes the long-run average total reward throughput with constraints on secondary performance measures. Our problem is different from the existing literature because we consider a deteriorating service speed depending on the total workload in the system, multiple classes with different reward amounts and different resource requirements, and constraints on secondary performance measures. For a small-scale problem, we calculate the long-run average reward throughput and other performance measures by solving balance equations directly from a multi-class M/G/C/C state dependent queueing model. For a large-scale problem, as balance equations cannot be solved analytically, we use simulation to estimate performance measures and use a Bayesian optimization algorithm based on the Gaussian process to find an optimal allocation among a large number of possible allocations quickly with and without constraints on secondary performance measures. We test the performance of our procedure on a highway access control problem and a server capacity allocation problem of an online retail store.

Agent-based simulation is a form of computer-based modeling that provides an intuitive and flexible approach to representing complex systems. It has been used in a wide range of health care applications. In Chapter 3, we develop an agent-based simulation with mosquito and human populations to model the spread of malaria in sub-Saharan Africa countries. We propose and test various strategies for allocating limited resources to evaluate and maximize the impact of proactive community case management (Pro-CCM). The simulation model utilizes ordinary differential equations

and incorporates temporal climate information, disease transmission from mosquitoes to humans, and the progression of the disease in infected humans. The model is validated using data from Senegal, a west African country in the Sahel with highly seasonal transmission. We test numerous scenarios to understand how the number and the frequency of sweeps impact the effectiveness of ProCCM.

In Chapter 4, we build a multi-water-source (MWS) agent-based simulation model, based on [**?**], to model the Guinea Worm (GW) disease transmission among dogs in Chad, as well as in each domestic clusters. This model contains three connected single-water-source (SWS) models, partitioning the majority part of Chad based on clustering results. Each SWS model adopts the general framework of the previous stochastic simulation model, which is used to simulate the life cycle of GW and daily interactions between the dogs, worms, and water source over multiple years. Each SWS model is validated using infection data within the corresponding cluster after parameter calibrations. Three SWS models were then connected into an MWS model using geographic information and local human characteristic data. Our MWS model was used to test the effectiveness and fairness of various intervention strategies. We searched for optimal solutions using the Cross-Entropy (CE) Method under various capacity and allocation constraints.

CHAPTER I

INTRODUCTION

1.1. Background and Literature Review for Resource Sharing Problem

Resource sharing considers a situation where there is a single resource pool of finite capacity, customers join the system if the number of available resources at the time of arrivals is more than the number of resources required for the customers, and rewards are given when the customers depart the system. If the number of currently available resources is smaller than the required number of resources for a customer, then she is rejected. Resource sharing among multiple classes has been studied extensively in the queueing community. [**?**] is one of survey papers for the resource sharing problem. The revenue management community has also studied either a pricing policy or an admission control policy among multiple classes for the resource sharing problem to maximize the expected long-run revenue rate. For example, see [**?**], [**?**], [**?**] and [**?**]. The existing literature considers applications in business such as call center service, car rental management and hotel management where the service rate V_i for class i is independent of both the arrival process and the state of the other classes.

In Chapter 2, we consider a problem of allocating shared resources among multiple classes when (i) a customer from a different class may occupy a different number of resources and give a different amount of a reward when leaving the system after service completion; (ii) a customer is rejected if the number of available resources at the time of her arrival is smaller than the number of resources required for the customer, and (iii) service speed gets slower as the current total workload in the system increases. The main difference between our work and the existing literature is that the service speed deteriorates as the total workload increases. For example, in highway traffic, as the number of vehicles in a lane increases, the traveling velocity decreases. When the number of vehicles in a lane is over its capacity limit, the traffic stops moving with a zero traveling velocity. Similar behavior is also observed in an online site. An online site often experiences slow-down when the number of connected users increases, which usually happens during big sale events such

as Black Friday, Christmas, or New Year sales. When the total workload exceeds the server's capacity, the site eventually experiences an outage.

For the traffic control problem, [**?**] proposes a token-based reservation system and a real-time scheduling algorithm to monitor the traffic condition and maintain the efficiency of the high priority lane. The token-based reservation is shown to increase the control accuracy. However, it requires a significant effort to monitor a large number of tokens, which makes it difficult to implement in practice. [**?**] presents the idea of using dynamic pricing as a control mechanism for the high-priority lane. Theoretically, by varying the price of using the priority lane, the demand could be controlled to avoid congestion. For a highway system, [**?**] models highway traffic as an M/G/C/C state dependent queueing model and provides analytical expressiones for various performance measures such as a probability of rejecting a car. This analysis can be useful in coming up with an admission control policy for highway traffic. However, this previous work has been performed assuming a single class. [**?**] shows that if there is only one class of vehicles with equal size, a traffic flow speed model is linear, and the incoming vehicle arrival rate is infinity, then the highest throughput is achieved when the maximum number of vehicles accepted in a highway lane is set to the half of the road's maximum physical capacity. In reality, vehicles in different sizes have different resource requirements in the sense that a larger vehicle occupy more spaces than a smaller vehicle. The traveling speed is determined by the total number of spaces occupied rather than the total number of vehicles currently in the highway lane and may not decrease linearly. In addition, if the decision-maker cares about the passenger throughput, a certain type of vehicles such as school buses will have a higher reward than, say, trucks. Thus the decision-maker may want to come up with an admission control policy among classes to maximize the reward rate for a shared resource, a special lane in a segment of highway.

Similarly, an online site admission can be viewed as a resource sharing problem with multiple classes and deteriorating service quality. Obviously, the server capacity is shared by customers who shop on the site, and the web-browsing speed gets slower as the number of customers currently online increases. Controlling e-server access to prevent overload and to guarantee Quality of Service (QoS) has been studied in the literature. [**?**] uses a feedback control theory to guarantee server performance by limiting only the utilization of the lower-class customers and differentiating service

qualities between customer classes. Higher class customers have access to more server capacity, while lower class ones may experience lower service quality, such as images with lower resolutions. [**?**] proposes kernel mechanisms for service differentiation to avoid e-server overload. In this chapter, we consider an online shopping server where service quality among customer classes is the same.

During a special sale event, a number of online retail stores come up with a strategy of limiting the total number of connected customers (admission control) to avoid a site crash. For example, some online department stores put potential shoppers in a virtual queue. This policy is implemented on the first-come-first-served (FCFS) basis. Many retail stores run membership programs, and by asking online users to log in, their membership levels can be identified. If a group with a higher membership status tends to spend more money and make fewer returns, it might be better to manage customers' admissions based on their membership levels rather than the FCFS basis because they bring higher rewards (revenues in this example). Thus, there should exist an optimal allocation among different classes that maximizes the reward rate. In reality, large retailers, like Amazon, could add additional servers when handling unusually higher traffic during a specific period to avoid the trade-off between degrading quality of service or rejecting customer requests. However, there are situations where a physical capacity is not easy to expand, and the resource allocation method we proposed would be useful in such scenarios.

1.2. Background and Literature Review for Malaria Transmission in sub-Saharan Africa

Globally, there were an estimated 228 million cases and $405,000$ deaths due to malaria in 2018, primarily among children under five years of age in sub-Saharan Africa [**?**] . Interventions for malaria prevention and control include insecticide-treated mosquito nets (ITNs), indoor residual spraying (IRS), intermittent preventive treatment for pregnant women (IPTp), seasonal malaria chemoprevention (SMC), diagnosis by malaria microscopy or rapid diagnostic test (RDT), and treatment with artemisinin-based combination therapy (ACT) [**?**] . Timely, accurate, and effective case-management, particularly diagnostic confirmation of suspected cases and appropriate treatment within 24 hours of symptom onset, is critical to malaria control. Early detection and treatment

reduces the severity of symptoms and might reduce the parasite reservoir in a community [**?**]. However, inadequate treatment-seeking behavior, e.g., due to barriers such as geographic distance to healthcare facilities, hinder coverage of malaria case-management [**?**]. Less than 20% of children with malaria in endemic zones are treated within the formal health system, and most of the deceased die at home without having receiving proper treatment [**?**].

Passive Community case management of malaria (CCMm) was introduced in Senegal to address these barriers to receiving prompt care in Senegal was reported in [**?**]. This CCMm program was evaluated based on data collected by the HCPs and at health facilities, which showed that intervention regions experienced a statistically significant decrease in all deaths and deaths attributed to malaria while no such decrease was seen in comparison regions.

Despite the scale-up of CCMm, poor care-seeking and access to care at the community level continued to be observed, prompting the development of a proactive model [**?**]. Proactive community case management (ProCCM) is a novel strategy for delivering malaria case management, in which a community health worker conducts visits of every household in the community (sweeps) regularly and frequently to identify individuals who display malaria symptoms, administer an RDT to test those with symptoms for malaria and administer an ACT to those who test positive. Patients who display severe symptoms, such as impairment of consciousness, prostration, or inability to eat or drink, are urgently referred to a nearby health facility. The potential advantages of ProCCM are early intervention for better health outcomes in malaria cases, and the potential to have an impact on the proportion of the population with gametocytes, which in turn has the potential to reduce future incidence of malaria and the overall proportion of the population infected.

ProCCM was implemented piloted in a pilot study in the Saraya district in the Kedougou region of Senegal, which has one of the highest childhood mortality rates (74 deaths per 1000 live births) and the highest prevalence of Plasmodium falciparum (15.3%) among children under 5 years in Senegal [**?**]. Community-case management of malaria (CCMm, known as PECADOM in Senegal) was introduced in 2008 in Saraya, and malaria-related deaths declined by 62.5% during $2008-2010$, compared with only 15.4% decline in other regions without CCMm [**?**]. However, care providers noted that care-seeking remained inadequate, resulting in preventable morbidity and

mortality. The ProCCM pilot began in one village in 2012 and was then expanded to 15 intervention villages and 15 comparison villages in a 2013 pilot in Senegal [**?**]. Built on the PECADOM platform, in the comparison group (Scenario I), only three home visits were conducted, during the first weeks in July, September, and November (see **??** for details) while in the intervention group (Scenario II), weekly sweeps conducted by community health workers during the entire malaria peak season (from July to November, for 21 weeks). At the end of the pilot, the proportion of the population with symptomatic, confirmed malaria infection was 16 times higher in the comparison villages than in the intervention villages. Adjusting for potential confounders, the ProCCM intervention was associated with a 30fold reduction in odds of symptomatic malaria in the intervention villages ($AOR = 0.033; 95\%CI : 0.017, 0.065$). The ProCCM strategy was adopted by the Senegal National Malaria Control Program following the pilot. Scale-up started in 2014 to include just over 150 villages in the Kedougou region and has continued, reaching over $2,000$ villages. Studies examining the potential impact of ProCCM, especially considering different implementation strategies in terms of sweep timing and frequency (depending on limited resources), are limited [**?**]. In Senegal, scale-up was rapid after the initial pilot study, and data collected for programmatic monitoring have lacked sufficient granularity to track impact.

1.3. Background and Literature Review for Guinea Worm Disease Transmission in Chad

Since 1986, The Carter Center had led the international campaign to eradicate Guinea worm (GW) disease, when the annual burden of this disease was estimated to be 3.5 million cases in 21 countries in Africa and Asia. In 2018, only 28 human cases were reported worldwide, 17 of them were from Chad. Chad also reported 1040 domestic dogs with GW infections this year. Guinea worm disease could become the second human disease in history, after smallpox, to be eradicated, as the ongoing efforts now focus on the four remaining endemic countries Chad, Ethiopia, Mali, and South Sudan. In 2019, Chad's national program intensified surveillance and broadly promoted cash incentives for reporting suspected infections, and as a result, rumors of human cases rose 137%, and rumors of animal infections rose 268% compared to 2018 [**?**].

Guinea worm disease (*dracunculiasis*) is a parasitic infection caused by the nematode roundworm parasite Dracunculus medinensis. It is contracted when *definitive host* (e.g., humans, dogs,

etc.) consume water from stagnant sources contaminated with Guinea worm larvae. Inside a host's abdomen, Guinea worm larvae mate and female worms mature and grow. After about a year of incubation, the female Guinea worm, one meter long, creates an agonizingly painful lesion on the skin and slowly emerges from the body. If the host enters/contacts a water source, the emerging worm will release its larvae into the water and begin the cycle of infection all over again.

In 2018, there were only 28 human GW infection cases reported in the world, 17 of which were from chad. Whereas human infections remained low (less than 20 per year), 1040 dog infections are reported in Chad in the same year [**?**]. GW infections among dogs are not fully understood. To our knowledge, there hasnt been extensive research done on the dynamics of GW transmission in dogs. Yet, this could be critical for achieving GW eradication [**?**].

Computer simulation models have been used to understand the epidemiology of diseases such as influenza [**?**], [**?**], HIV/AIDS [**?**], and malaria [**?**], [**?**]. [**?**] presented a simulation model that captures a reasonable representation of the natural transmission pathway(s) of GW in Chad dogs, where Chad is considered to be a single shared water source among all dogs. Their model is validated using integrated national data provided by the Chad GWEP and is used to test various intervention strategies for GW eradication purposes.

We observed that the intensity of the GW infections, the number of GW infection cases per 1000 dog population per year, are different among regions in Chad. Furthermore, regional infection data suggested that the intervention coverages are not uniform all over the country [**?**]. Moreover, we verified that using a share infectivity curve or a uniformed tethering coverage among clusters is inaccurate in capturing regional GW infections. Thus, it is critical to design a multi-water-source simulation model. We presented an MWS agent-based simulation model incorporating [**?**]s model dynamic. Our model is composed of multiple interactive SWS models, each as presented in the previous work.

CHAPTER II

SHARED-RESOURCE ALLOCATION AMONG MULTIPLE CLASSES WITH DIFFERENT REWARDS WHEN SERVICE SLOWS DOWN DEPENDING ON TOTAL WORKLOAD

In this chapter, we formulate a general resource sharing problem when limited resources are shared by customers from different classes and customers are rejected when there are no available resources; each class has a different resource requirement; processing times get slow as the total current workload increases; and the reward of a completed job is different depending its class. We also consider constraints on secondary performance measures. While the reward rate is important, it is possible that a decision maker consider constraints on other (secondary) performance measures such as the probability of rejecting customers with a higher membership status less than or equal to a threshold. Constraints on secondary performance measures make the allocation problem even more difficult especially when feasibility needs to be determined based on stochastic simulation estimates of the performance measures. We propose a Bayesian optimization algorithm based on the Gaussian process (GP) to solve our formulated problem because each observation is from an expensive simulation run and the Bayesian optimization is known to be useful when observations are expensive. The proposed Bayesian algorithm can handle stochastic constraints. Note [**?**] consider a traffic control problem for multiple classes with different rewards. However, they consider cases where long-run average can be analytically calculated (thus no need for simulation) and use a cross-entropy method ([**?**]) for an optimization tool .

2.1. Problem Formulation

In this section, we define notation and formulate our problem.

2.1.1 Problem

Consider a single reusable resource pool of (real-valued) capacity $0 < C_{\max} < \infty$ and m classes of customers. We assume customers of class i for $i = 1, \ldots, m$ arrive following an independent Poisson

process with respective arrival rate λ_i. A vector $\boldsymbol{N}(t) = (N_1(t), N_2(t), \cdots, N_m(t)) = (n_1, n_2, \cdots, n_m)$ represents the current system state at time t where n_i represents the number of class i customers in the system. A class i customer requires s_i units of the resource to complete a job with size J_i. Job size could be deterministic, e.g. the length of a highway segment, or stochastic, e.g. the number of web-pages a customer views. The total workload of the resource pool w is defined as $w = \sum_{i=1}^{m} s_i n_i$. A class i customer uses the seized s_i units of the resource for a constant or random amount of time with mean service rate $v(w)$ where $v(w)$ is a function of workload w. We assume that $v(w)$ is a non-increasing function of w where $v(C_{\max}) = 0$ and $v(1)$ is the fastest possible mean service rate. When a customer's service is completed, the customer leaves, the seized units of the resource become available to serve next customers and the system receives a reward whose value can be different depending on the customer's class.

Let $\boldsymbol{x}$ denote an allocation. We consider two different allocation types: the pooled allocation and the dedicated allocation. In the pooled allocation, we determine a constant, $\boldsymbol{x} = x_p < C_{\max}$ and customers are admitted as long as the number of available units out of total x_p units is more than the customer's requirement s_i. In the dedicated allocation, we determine $\boldsymbol{x} = (x_1, x_2, \cdots, x_m)$ and x_i units of the resource are exclusively assigned to class i and thus the system can accept up to $\lfloor \frac{x_i}{s_i} \rfloor$ number of class i customers where $\lfloor \cdot \rfloor$ denotes a floor function. This policy makes sense if the decision maker wants to leave resource units available to customers from more profitable classes.

When a class i customer's service is completed and she leaves the system when the current allocation is $\boldsymbol{x}$, the system collects a reward r_i and the long-run average reward rate collected from class i with allocation $\boldsymbol{x}$ is $\mathrm{tr}_i(\boldsymbol{x})$. The reward r_i can be the number of passengers in a vehicle for a highway traffic model or a dollar amount a customer spends for online retailers. As there is no waiting buffer, rejected customers are lost and leave the system.

Then our problem for the pooled policy is

$$\begin{aligned} \max_{x_p} \quad & \sum_{i=1}^{m} \mathrm{tr}_i(x_p) \\ \text{subject to} \quad & \sum_{i=1}^{m} s_i N_i(t) \le x, \text{ for any } t > 0; \\ & 0 \le x < C_{\max}. \end{aligned} \tag{1}$$

and for the dedicated policy, we get the following formulation:

$$
\begin{aligned}
\max_{\boldsymbol{x}=(x_1,x_2,\cdots,x_m)} \quad & \textstyle\sum_{i=1}^{m} \mathrm{tr}_i(\boldsymbol{x}) \\
\text{subject to} \quad & N_i(t) \le \lfloor \tfrac{x_i}{s_i} \rfloor, \quad i = 1,\cdots,m, \text{ and any } t > 0; \\
& \textstyle\sum_{i=1}^{m} x_i < C_{\max}; \\
& x_i \ge 0, i = 1,\cdots,m.
\end{aligned} \tag{2}
$$

It is possible that the decision maker is also interested in a secondary performance measure from each class, such as: the rejection probability of each class, and wants to keep the measure less than some threshold value b_i. Let $q_i(\boldsymbol{x})$ represent the secondary performance measure for class i with allocation $\boldsymbol{x}$. More generally, the secondary performance measure $q(\boldsymbol{x})$ could be a function across all classes, such as: the overall rejection probability among all classes. For the cases discussed in this chapter, we limit the secondary performance measure function only to a particular class of customer. Then these additional constraints change the above problem formulations as follows:

$$
\begin{aligned}
\max_{x_p} \quad & \textstyle\sum_{i=1}^{m} \mathrm{tr}_i(x_p) \\
\text{subject to} \quad & q_i(x_p) \le b_i, \quad i = 1,\ldots,m; \\
& \textstyle\sum_{i=1}^{m} s_i N_i(t) \le x_p \text{ for any } t > 0; \\
& 0 \le x_p < C_{\max}.
\end{aligned} \tag{3}
$$

and

$$
\begin{aligned}
\max_{\boldsymbol{x}=(x_1,x_2,\cdots,x_m)} \quad & \textstyle\sum_{i=1}^{m} \mathrm{tr}_i(\boldsymbol{x}) \\
\text{subject to} \quad & q_i(\boldsymbol{x}) \le b_i, \quad i = 1,\ldots,m; \\
& N_i(t) \le \lfloor \tfrac{x_i}{s_i} \rfloor, \quad i = 1,\ldots,m \text{ and any } t > 0; \\
& \textstyle\sum_{i=1}^{m} x_i < C_{\max}; \\
& x_i \ge 0, i = 1,\cdots,m.
\end{aligned} \tag{4}
$$

Table **??** provides a list of notation used for problem formulation.

2.1.2 Simulation

Under some conditions, it may be possible to calculate the objective analytically. For example, if $C_{\max}$ is small and $\mathrm{tr}(\boldsymbol{x})$ is a long-run average, then it is possible to formulate our problem as a

Table 1: Notation for Problem Formulation

Notation	Definition
$C_{\max}$	(physically) the maximum number of resource units
m	the number of customer classes (dimension of alternatives)
i	class index
k	total number of possible allocations
$N_i(t)$	number of class i customers at time t
s_i	number of resources units required to process a class i customer (E.g. in traffic case, passenger vehicle takes 1 unit space in the lane while pick-up truck takes 2, to finish the same job with size of 1 mile.)
J_i	the job size required for a class i customer (E.g. in online shopping case, premium customer requires 20 MB of data ,while regular customer requires 10 MB, both taking 1 unit space in the same server.)
r_i	unit reward collected from a completed class i customer
x_p	allocation for a pooled policy
$(x_1, x_2, \ldots, x_m)$	allocation for a dedicated policy
$\boldsymbol{x}$	generic allocation either x_p or $(x_1, x_2, \ldots, x_m)$
λ_i	arrival rate of class i customers
$v(1)$	fastest mean service rate
$v(w)$	service rate for class i customer when system workload is w
$\mathrm{tr}_i(\boldsymbol{x})$	the (partial) reward rate collected from a class i with allocation $\boldsymbol{x}$
$\mathrm{tr}(\boldsymbol{x})$	$\sum_{i=1}^{m} \mathrm{tr}_i(\boldsymbol{x})$, the reward rate collected from all classes with allocation $\boldsymbol{x}$
$q_i(\boldsymbol{x})$	secondary performance measure for class i with allocation $\boldsymbol{x}$
b_i	constraint threshold for $q_i(\boldsymbol{x})$

queuing model and analytically calculate a long-run average reward rate and possibly $q_i(\boldsymbol{x})$ for a given allocation. Then a deterministic optimization method can find an optimal solution with or without constraints on $q_i(\boldsymbol{x})$. However, when the problem size is too large to solve analytically due to large $C_{\max}$ or m or $\mathrm{tr}(\boldsymbol{x})$ and $q_i(\boldsymbol{x})$ cannot be analytically calculated, we use simulation to obtain observations corresponding to each performance measure and use them for the estimation of the performance measures.

For a steady-state simulation, after a proper warm-up period, $Y_i^j(\boldsymbol{x})$ and $Z_i^j(\boldsymbol{x})$ represent either batch means from batch j or within-replication averages from replication j corresponding to the reward rate measure and secondary performance measure from class i, respectively, when allocation $\boldsymbol{x}$ is used. We are interested in

$$\lim_{n\to\infty} \bar{Y}_i(\boldsymbol{x}) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} Y_i^j(\boldsymbol{x}) \quad \text{and} \quad \lim_{n\to\infty} \bar{Z}_i(\boldsymbol{x}) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} Z_i^j(\boldsymbol{x})$$

Similarly, for a terminating simulation, $Y_i^j(\boldsymbol{x})$ and $Z_i^j(\boldsymbol{x})$ represent within-replication averages from replication j corresponding to the reward rate measure and secondary performance measure and the decision maker is interested in $\mathrm{E}[\bar{Y}_i(\boldsymbol{x})]$ and $\mathrm{E}[\bar{Z}_i(\boldsymbol{x})]$. Then we calculate $\mathrm{tr}_i(\boldsymbol{x})$ and $q_i(\boldsymbol{x})$ from a finite number of observations for both steady-state and terminating simulations as follows:

$$\mathrm{tr}_i(\boldsymbol{x}) = \bar{Y}_i(\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{n} Y_i^j(\boldsymbol{x}) \quad \text{and} \quad q_i(\boldsymbol{x}) = \bar{Z}_i(\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{n} Z_i^j(\boldsymbol{x}).$$

Customer arrivals from class i are assumed to follow a Poisson arrival process (PP) with rate λ_i. In many applications, the arrival process is often non-homogeneous meaning that the rate changes over time. However, the allocation becomes the most important when the system is busiest. Thus we consider the arrival process during the most busiest period of a system which can be assumed to follow a homogeneous Poisson process.

For example, for a traffic case, one runs multiple replications of a traffic simulation with a run length from 6:30 AM to 9 AM with an appropriate warm-up. Then $Y_i^j(\boldsymbol{x})$ from replication j can be obtained by the total rewards for class i observed during the run length divided by 2.5 hours and $Z_{ij}(\boldsymbol{x})$ is the total number of class i customers rejected during the run length divided by the total number of class i customers who attempted to enter the system.

In the traffic case, we consider a segment of highway with length of 1 mile (i.e., $J_i = 1$ for all i). Thus, an admitted customer who sees the system state $N(t) = (n_1, n_2, \ldots, n_m)$ uses s_i of the resource for a duration of $J_i/v(w) = 1/v(w)$. For a single lane case, service time for each incoming vehicle is deterministic because its speed only depends on cars ahead of it once it is admitted. So we assume that the vehicle travels at the same speed as it entered the system until it exits. In other words, a state change after a car is admitted does not affect service speeds of vehicles that are already in the system. On the other hand, in online shopping case, an admitted class-i customer who sees system state $N(t) = (n_1, n_2, \ldots, n_m)$ uses one unit of resources to complete a job whose service time is exponentially distributed with mean of J_i. Thus, at a given server speed, the duration of usage is also a random variable that is exponentially distributed with mean $J_i/v(w)$. Note that all customers using the system experience the same service speed at any given time. Moreover, each time there is an customer admitted or leaving the system, the remaining shopping times for all customers that are still in the system become longer or shorter due to service speed change. For instance, a customer

wanting to download 10Mb data from a website would take 1 minute if service speed is 10Mb/sec and this duration would increase to 10 minutes if the system admits customers large enough to drop the service speed to 1Mb/sec. Therefore, the simulation model for online shopping cases, e.g. e-commerce service, streaming service, will be more complicated.

Although we consider only a homogeneous Poisson arrival process and constant or exponential service times in this chapter, it is fine to use more general arrival processes or service times as long as $Y_i^j(\boldsymbol{x})$ and $Z_i^j(\boldsymbol{x})$ can be obtained from simulation.

2.2. Bayesian Optimization

In this section, we first present a Bayesian optimization algorithm without stochastic constraint. Then we present a new algorithm to handle stochastic constraints.

2.2.1 Sampling Model and Posterior Distribution for Unconstrained Bayesian Optimization Algorithm

Bayesian optimization (BO) constructs a probabilistic model for performance measures over a large search space and then exploits the probabilistic model to make decision on where to evaluate next and which solution achieves the maximum or minimum performance measure. BO is shown to be useful when the evaluation of each solution $\boldsymbol{x}$'s performance is expensive requiring either heavy calculation or simulation. In our case, the performance measure is the expected total reward, $\mathrm{tr}(\boldsymbol{x})$. To perform BO, we need to choose a prior over the function to be optimized and an acquisition function that guides the search for the optimum. For the former one, we choose a Gaussian Process with a squared exponential kernel. A Gaussian process is able to express a rich distribution on functions by using different types of covariance functions ([**?**]). For an acquisition function, we choose to the expected improvement (EI). The algorithm is explained in details as following.

When $\mathrm{tr}(\boldsymbol{x})$ can be analytically calculated such as the long-run average throughput from the M/G/C/C/ queueing model in the traffic case, we have a deterministic problem. On the other hand, if $\mathrm{tr}(\boldsymbol{x})$ can only be estimated by observations from a stochastic simulation, then we have a stochastic problem. In the stochastic case, we estimate $\mathrm{tr}(\boldsymbol{x})$ by $\bar{Y}(\boldsymbol{x})$. We assume that $\mathrm{tr}(\{\boldsymbol{x}\})$, defined over the entire search space $\{\boldsymbol{x}\}$, has a multivariate normal distribution with mean θ and covariance matrix

Λ. For clarification, at alternative $\boldsymbol{x}$, its mean is $\theta(\boldsymbol{x})$ and its variance is $\Lambda(\boldsymbol{x},\boldsymbol{x})$.

The normality assumption is commonly used in statistical selection and simulation optimization ([**?**]), and in Gaussian process regression ([**?**]). In Bayesian formulation, a multivariate normal prior distribution is used to describe uncertainty about unknown value θ,

$$\theta \sim \mathcal{N}(\mu_0, \Sigma_0).$$

In our case, we adopt a commonly used Gaussian kernel for Σ_0:

$$\Sigma_0(\boldsymbol{x},\boldsymbol{x}') = \sigma_0^2 \exp\Big\{ -\sum_{i=1}^{d} \alpha_i [\zeta(\boldsymbol{x}_i - \boldsymbol{x}_i')]^2 \Big\},$$

where σ_0^2 is the homogeneous prior variance of the unknown means, $\vec{\alpha} = \{\alpha_i\}_i^d$ is a vector of scaling parameters, d is the dimension of the alternatives, ($d = 1$ or $d = m$) and ζ is the distance measure for the ith dimension of the alternative ($\boldsymbol{x}_i$). We also let η be a parameter for the mean in this model and let $\vec{1}$ be a vector of k ones, and thus $\mu_0 = \eta\vec{1}$ defining the prior distribution at the very beginning. If unknown, we need to make initial assumptions, then estimate μ_0, Σ_0 and Λ using observations from first n_0 iterations. After that we update these parameters using Maximum Likelihood Estimations (MLEs) during each iteration.

Next, we explain how we update a posterior distribution. At each iteration after initial sampling, $n = n_0 + 1, n_0 + 2, ...$, we choose an alternative $\boldsymbol{x}^n$ to sample according to the selected allocation rule, and take β number of replications using simulation. We observe β total rewards ,$[Y^1(\boldsymbol{x}^n), Y^2(\boldsymbol{x}^n), ..., Y^\beta(\boldsymbol{x}^n)]$. Note that the total rewards is the sum of partial rewards over all classes: $Y^i(\boldsymbol{x}^n) = \sum_{j=1}^{m} Y_j^i(\boldsymbol{x}^n)$. After that we take the mean of total rewards from β replications and get $\bar{Y}(\boldsymbol{x}^n)$. The conditional distribution of $\bar{Y}(\boldsymbol{x}^n)$ given $\boldsymbol{x}^n$ and θ is assumed to be Gaussian and independent of any previous observation, $(\boldsymbol{x}_l, \bar{Y}(\boldsymbol{x}^l) : l < n)$

$$\bar{Y}(\boldsymbol{x}^n)|\theta, \boldsymbol{x}^n, (\boldsymbol{x}_l, \bar{Y}(\boldsymbol{x}^l) : l < n) \sim \mathcal{N}(\theta(\boldsymbol{x}^n), \Lambda(\boldsymbol{x}^n, \boldsymbol{x}^n)/\beta).$$

Then we update our knowledge on the distribution of the mean at this observed alternative, $\theta(\boldsymbol{x}^n)$, using Bayes' rule:

$$p(\theta(\boldsymbol{x}^n)|\bar{Y}(\boldsymbol{x}^n)) \propto p(\bar{Y}(\boldsymbol{x}^n)|\theta(\boldsymbol{x}^n)) p(\theta(\boldsymbol{x}^n))$$

where $p(\bar{Y}(\boldsymbol{x}^n)|\theta(\boldsymbol{x}^n))$ is the likelihood function for the mean of current observations and $p(\theta(\boldsymbol{x}^n))$ is the prior for the test mean at the same alternative, assuming current β observations are normally distributed with mean of $\theta(\boldsymbol{x}^n)$ and variance of $\Lambda(\boldsymbol{x}^n,\boldsymbol{x}^n)$.

We assume that the sampling covariance Λ is in following form: sampling variance is $\sigma^2(\boldsymbol{x}^i)$ for alternative $\boldsymbol{x}^i$ and all off-diagonal elements are 0s. We make further simplification that sampling variances are the same at all alternatives: $\sigma(\boldsymbol{x}^1) = \sigma(\boldsymbol{x}^2) = .. = \sigma(\boldsymbol{x}^k) = \sigma_\varepsilon$ to reduce the computational workload.

$$\begin{bmatrix} \sigma^2(\boldsymbol{x}^1) & 0 & ... & 0 \\ 0 & \sigma^2(\boldsymbol{x}^2) & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \sigma^2(\boldsymbol{x}^k) \end{bmatrix}$$

After updating posterior distribution, we use MLEs to update the hyper-parameters, η, σ_0^2, $\vec{\alpha} = \{\alpha_i\}_{1\le i\le d}$ and σ_ε^2, which determine Λ, μ_0 and Σ_0 according the previous definitions. The derivation is based on [**?**]. Let $\mathscr{X}_n$ and $\mathscr{Y}_n$ represent all alternatives and corresponding means of total rewards that have been observed through iteration n. Since we only sample one alternative at each iteration, $|\mathscr{X}_n| = n$. Set $g = \sigma_0^2/\sigma^2$, $\sigma^2 = \sigma_0^2 + \sigma_\varepsilon^2$. $\mathscr{Y}_n$ follows normal distribution with mean$\eta\vec{1}$ and covariance matrix $\sigma^2 R$, where R is defined as:

$$R(i,j) := \begin{cases} 1, & \text{if} i = j \\ g\exp\{-\sum_{l=1}^{d} \alpha_l[\zeta(\mathscr{X}_{n,l}^i - \mathscr{X}_{n,l}^j)]^2\}, & \text{if} i \neq j \end{cases} \tag{5}$$

Note that $\mathscr{X}_{n,l}^i$ is the lth dimension of $\mathscr{X}_n^i$.

The MLEs are then used to calculate: $\arg\max_{\eta,\sigma_0^2,\vec{\alpha},\sigma_\varepsilon^2} \log p(\mathscr{Y}_n|\eta,\sigma_0^2,\vec{\alpha},\sigma_\varepsilon^2)$. We reparameterize this by replacing $(\sigma_0^2,\sigma_\varepsilon^2)$ with (g,σ^2):

$$\arg\max_{\eta,\sigma_0^2,\vec{\alpha},\sigma_\varepsilon^2} \log p(\mathscr{Y}_n|\eta,\sigma_0^2,\vec{\alpha},\sigma_\varepsilon^2) = \arg\max_{\eta,\sigma_0^2,\vec{\alpha},\sigma_\varepsilon^2} \log p(\mathscr{Y}_n|\eta,\sigma_0^2,R)$$

We solve this optimization problem in two steps, first analytically optimizing over σ^2 and η with the other parameters fixed, and then numerically optimizing the resulting value over the set of

the R matrices that can be achieved with the remaining parameters $g, \vec{\alpha}$.

The likelihood function of $\mathscr{Y}_n$ is:

$$p(\mathscr{Y}_n|\eta,\Lambda,\Sigma_0) = \frac{1}{(2\pi)^{n/2}|\Lambda+\Sigma_0|^{1/2}} e^{-1/2(\mathscr{Y}_n-\eta\vec{1})^T(\Lambda+\Sigma_0)^{-1}(\mathscr{Y}_n-\eta\vec{1})}$$

The maximum log-likelihood over η and σ^2 with R fixed is:

$$\log p(\mathscr{Y}_n|\hat{\eta},\hat{\sigma}^2,R) = \max_{\eta,\sigma^2}\log p(\mathscr{Y}_n|\eta,\sigma^2,R) = -\frac{1}{2}(n\log\hat{\sigma}^2+\log|R|) - \frac{n}{2}(1+\log 2\pi)$$

where $\vec{1}$ denote a length-n column vectors of ones, $|R|$ is the determinant of R, and

$$\hat{\sigma}^2 = \frac{1}{n}(\mathscr{Y}_n-\hat{\eta}\vec{1})^T R^{-1}(\mathscr{Y}_n-\hat{\eta}\vec{1}), \text{ and } \hat{\eta} = (\vec{1}^T R-1\vec{1})^-\vec{1}^T R-1\mathscr{Y}_n.$$

We then solve g and $\vec{\alpha}$ using the Steepest Descent Algorithm with $\hat{\sigma}$ and $\hat{\eta}$ calculated from above. To find the next point to sample at iteration $n+1$, we used maximum expected improvement (EI) as allocation rule:

$$\begin{aligned}
\boldsymbol{x}^{n+1} &= \arg\max_{\boldsymbol{x}\in\{\boldsymbol{x}\}} \mathrm{EI}(\boldsymbol{x}) \\
&= \arg\max_{\boldsymbol{x}\in\{\boldsymbol{x}\}} \mathbb{E}[(\theta(\boldsymbol{x})-\mu_n^*)^+] \\
&= \arg\max_{\boldsymbol{x}\in\{\boldsymbol{x}\}} (\mu_n(\boldsymbol{x})-\mu_n^*))\Phi(\frac{\mu_n(\boldsymbol{x})-\mu_n^*}{\sigma(\boldsymbol{x})}) + \sigma(\boldsymbol{x})\phi(\frac{\mu_n(\boldsymbol{x})-\mu_n^*}{\sigma(\boldsymbol{x})})
\end{aligned}$$

where μ_n^* denote the largest posterior mean at iteration n and $\sigma(\boldsymbol{x}) = \sqrt{\Lambda(\boldsymbol{x},\boldsymbol{x})/\beta}$.

Without stochastic constraints, the generic algorithm is generalized in Algorithm **??**. Table **??** provides a list of notation used for Bayesian Optimization Algorithm.

2.2.2 Constrained Bayesian Optimization Algorithm

If a constraint (c_i) is included, we assume that the corresponding secondary measurements, $q_i\{\boldsymbol{x}\}$, also has a multivariate normal distribution with mean θ^{c_i} and covariance matrix Λ^{c_i}. In addition to documenting the average of total reward $\bar{Y}(\boldsymbol{x}^n)$ during each iteration, we also document the average of objective for each constraint c_i: $\bar{Z}_{c_i}(\boldsymbol{x}^n)$. After each iteration, we update posterior distribution for each constraint objectives using the same method but with $\mathscr{X}_n$ and $\mathscr{Z}_n^{c_i}$. Additionally we update the hyper-parameters for μ_0^{ci}, $\Sigma_0^{c_i}$ and Λ^{c_i}, if unknown, using the same MLEs.

Algorithm 1: BO algorithm for optimality search without constraints

[**Setup**:]
Select a stopping rule: when either maximum improvement is smaller than target, max(EI) $<= \bar{E}$, or total number of iterations is at maximum target, $n = \text{Iter}_{max}$, is met .
[**1. Initialization**:]
Initialize $n = 0$ and set $\mathscr{X}_0$, $\mathscr{Y}_0$ to be empty vectors. If μ_0, Σ_0,Λ_0 are known, specify them. Otherwise, make initial assumptions, ex: $\mu_0 = 0$, $\sigma_0 = 100$ $\vec{\alpha} = \vec{1}$, $\sigma_\varepsilon = 0$, for deterministic case and $\sigma_\varepsilon = 100$ for stochastic case.
Make n_0 initial iterations: randomly select n_0 alternatives, each to sample β replications. Update $\mathscr{X}_n$ and $\mathscr{Y}_n$. Set $n = n_0$.
[**2. Update parameters**:]
If μ_0, σ_0, $\vec{\alpha}$ and σ_ε are unknown, update them using MLEs with all current training data ($\mathscr{X}_n$ and $\mathscr{Y}_n$).
[**3. Check stopping criteria**:]

IF stopping criteria is not met:

select the next alternative $\boldsymbol{x}^{n+1}$ to sample:

$$\boldsymbol{x}^{n+1} = \arg\max_{\boldsymbol{x}\in\{\boldsymbol{x}\}} \text{EI}(\boldsymbol{x}).$$

ELSE: go to step 5.
[**4. Sample**:]
Simulate β replications at $\boldsymbol{x}^{n+1}$ and take the average of throughput $\bar{Y}(\boldsymbol{x}^{n+1})$. Append $\boldsymbol{x}^{n+1}$ to $\mathscr{X}_n$ to get $\mathscr{X}_{n+1}$. Append $\bar{Y}(\boldsymbol{x}^{n+1})$ to $\mathscr{Y}_n$. Increase n by 1. Go to step 2.
[**5. Selection rule**:]
Find the alternative with the largest posterior mean: $\arg\max_{\boldsymbol{x}\in\mathscr{X}_n} \mu_n(\boldsymbol{x})$, as the optimal solution.

Next, we construct two sets of confident sets for each constraint: one with higher tolerance for selecting the next alternative to observe and the other one stricter for selecting the optimal solution. For simplicity, we assume there's only one constraint c_1: $q_i(\boldsymbol{x}) <= q_i$ for class i customers. After n iteration, secondary performance measures have posterior mean $\mu_n^{c_1}$ and the covariance matrix Λ^{c_1}. We construct two confident sets as following:

$$CSI_n^{c_1} = \{\boldsymbol{x} | \mu_n^{c_1}(\boldsymbol{x}) - 1.96\sqrt{\Lambda^{c_1}(\boldsymbol{x},\boldsymbol{x})/\beta} \leq q_i\}$$

$$CSII_n^{c_1} = \{\boldsymbol{x} | \mu_n^{c_1}(\boldsymbol{x}) + 1.96\sqrt{\Lambda^{c_1}(\boldsymbol{x},\boldsymbol{x})/\beta} \leq q_i\}$$

If there's multiple constraints, we would construct two confident sets for each one and select the point with maximum expected improvement that is included in all confident set Is. Similarly, we select the optimal solution that is included in all confident set IIs.

Table 2: Notation for Bayesian Optimization

Notation	Definition
n_0	number of iterations for initial random sampling
$\boldsymbol{x}^n$	alternative selected to observe on iteration n
β	number of replications taken when evaluating an alternative
$\{c_1, c_2, c_3, \cdots\}$	set of constraints for secondary performance measure
$\mathscr{X}_n = [\boldsymbol{x}^1, \boldsymbol{x}^2, \cdots, \boldsymbol{x}^n]$	all alternatives observed through iteration n
$Y^j(\boldsymbol{x}^n)$	total rewards (sum of partial rewards among all classes) from replication j on the nth iteration
$\bar{Y}(\boldsymbol{x}^n)$	mean of total rewards from β replications on the nth iteration
$\mathscr{Y}_n = [\bar{Y}(\boldsymbol{x}^1), \bar{Y}(\boldsymbol{x}^2), \cdots, \bar{Y}(\boldsymbol{x}^n)]$	all mean total rewards observed through iteration n
$\bar{Z}_{c_i}(\boldsymbol{x}^n)$	mean of secondary performance measures from β replications on the nth iteration
$\mathscr{Z}_n^{c_i} = [\bar{Z}_{c_i}(\boldsymbol{x}^1), \bar{Z}_{c_i}(\boldsymbol{x}^2), \cdots, \bar{Z}_{c_i}(\boldsymbol{x}^n)]$	all mean secondary performance measures observed through iteration n
T_w	number of solutions searched in the end using method w
Iter_{max}	maximum number of iterations allowed
$\text{EI}(\boldsymbol{x})$	expected improvement at alternative $\boldsymbol{x}$
$\bar{E}$	expected improvement threshold
$CSI_n^{c_i}, CSII_n^{c_i}$	confident sets constructed for constraint c_i after iteration n

With stochastic constraints, we make the following modifications in Algorithm **??**.

2.3. Numerical Examples

In this section, we first compare the performances between multiple searching methods on traffic simulation for deterministic cases. We then show test cases for stochastic cases for allocating traffic resource using Bayesian Optimization. Finally, we applied the same method for e-server resource allocation cases.

2.3.1 Cross-entropy Method

In this section, we present a brief overview of the Cross-entropy Method, which could be used for optimization search over a large number of possible allocations without constraints.

The optimal allocation can be searched by the exhaustive method when the total number of possible allocations is small. Exhaustive method is not desirable due to computation time. In online shopping cases, this method is inapplicable due to the magnitude of the problem. Cross-entropy method is known to be efficient for searching for the best solution when the search space is large

Algorithm 2: BO algorithm for optimality search with constraints

[**Setup**:]
Select a stopping rule: when either maximum improvement is smaller than target, max(EI) $<= \bar{E}$, or total number of iterations is at maximum target, $n = \text{Iter}_{max}$, is met. Define constraints: $c_1, c_2, ...$
[**1. Initialization**:]
Initialize $n = 0$ and set $\mathscr{X}_0$, $\mathscr{Y}_0$, $\mathscr{Z}_0^{c_i}$ to be empty vectors. If μ_0, Σ_0, Λ, $\mu_0^{c_i}$, $\Sigma_0^{c_i}$ and Λ^{c_i} are known, specify them. (Each constraint have unique set of parameters.) Otherwise, make initial assumptions for hyper-parameters.
Make n_0 initial iterations: randomly select n_0 alternatives, each to sample β replications. Update $\mathscr{X}_n$, $\mathscr{Y}_n$ and $\mathscr{Z}_n^{c_i}$. Set $n = n_0$.
[**2a. Update parameters**:]
If μ_0, σ_0, $\vec{\alpha}$ and σ_ε are unknown, update them using MLEs with all current training data ($\mathscr{X}_n$ and $\mathscr{Y}_n$). Update $\mu_0^{c_i}$, $\sigma_0^{c_i}$, $\vec{\alpha}^{c_i}$ and $\sigma_\varepsilon^{c_i}$ for each constraint c_i using $\mathscr{X}_n$ and $\mathscr{Z}_n^{c_i}$.
[**2b.Check feasibility**:]
Construct confident sets, $CSI_n^{c_i}$, for each constraint.
[**3. Check stopping criteria**:]

IF stopping criteria is not met, and $\{CSI_n^{c_1} \cap CSI_n^{c_2} \cap ... \cap CSI_n^{c_i}\} \neq \emptyset$:

Select the next alternative $\boldsymbol{x}^{n+1}$ to sample:

$$\boldsymbol{x}^{n+1} = \underset{\boldsymbol{x} \in \{CSI_n^{c_1} \cap CSI_n^{c_2} \cap ... \cap CSI_n^{c_i}\}}{\arg\max} \text{EI}(\boldsymbol{x}).$$

ELSE: go to step 5a.
[**4. Sample**:]
Simulate β replications at $\boldsymbol{x}_{n+1}$ and take the average of throughput $\bar{Y}(\boldsymbol{x}^{n+1})$ and $\bar{Z}_{c_i}(\boldsymbol{x}^{n+1})$.
Append $\boldsymbol{x}_{n+1}$ to $\mathscr{X}_n$ to get $\mathscr{X}_{n+1}$. Append $\bar{Y}(\boldsymbol{x}^{n+1})$ to $\mathscr{Y}_n$ and $\bar{Z}_{c_i}(\boldsymbol{x}^{n+1})$ to $\mathscr{Z}_n^{c_i}$. Increase n by 1.
Go to step 2a.
[**5a.Check feasibility**:]
Construct confident sets, $CSII_n^{c_i}$, for each constraint.
[**5b. Selection rule**:]
Find the alternative with the largest posterior mean:

$$\underset{\boldsymbol{x} \in \{CSII_n^{c_1} \cap CSII_n^{c_2} \cap ... \cap CSII_n^{c_i}\}}{\arg\max} \mu_n(\boldsymbol{x}),$$

as the optimal solution. End if $\{CSII_n^{c_1} \cap CSII_n^{c_2} \cap ... \cap CSII_n^{c_i}\} = \emptyset$.

and evaluation of the objective function takes time. The main idea of the Cross-entropy method is to convert a deterministic optimization problem to a rare-event probability estimation problem. Initially, a 'flat' sampling distribution is used for generating the n number of possible solutions. Then each solution's performance is calculated and the sampling distribution is updated by giving higher probabilities to solutions with good performances and lower probabilities to solutions with

bad performances. These steps are repeated until the sampling distribution converges and a solution with the largest probability is returned as the best solution. A full description of our cross-entropy method is described next.

Initialization: Pick a number of sampled allocations n, an integer for a stopping condition c, and a weight for updating a sampling distribution $0 < \alpha < 1$. Set a value ρ, a quantile probability. Set iteration counter $k = 0$ and the maximum possible values for each class M_1 and M_2 are $M_1 = C_{\max} - 1$ and $M_2 = \lfloor C_{\max} - 1 \rfloor$ where $\lfloor x \rfloor$ returns the greatest integer less than or equal to x.

The initial sampling distribution $\mathbb{P}^{(0)}$ is a 2 by $(M_1 + 1)$ matrix of $[P_{rj}^{(0)}]$ such that

$$P_{rj}^{(0)} = \begin{cases} \frac{1}{M_1+1}, & r = 0, j = 0, \ldots, M_1; \\ \frac{1}{M_2+1}, & r = 1, j = 0, \ldots, M_2; \\ 0, & \text{otherwise}, \end{cases}$$

Sample Allocations: Sample n number of possible allocations $\boldsymbol{A}^{(i)} = (x_1^{(i)}, x_2^{(i)})$ for $i = 1, \ldots, n$ using $\mathbb{P} = \mathbb{P}^k$ as follows:

1. Generate a random permutation (w_1, w_2) of $\{1, 2\}$.
2. Sample $A_{w_1}^{(i)} \in \{0, 1, \ldots, M_{w_1}\}$ with pmf $(P_{w_1,0}, \ldots, P_{w_1,M_{w_1}})$.
3. Set $U = \left\lfloor C_{\max} - s_{w_1} A_{w_1}^{(i)} \right\rfloor$.
4. Sample $A_{w_2}^{(i)} \in \{0, \ldots, U\}$ with pmf $\left(\frac{P_{w_2,0}}{\sum_{i=0}^{U} P_{w_2,i}}, \ldots, \frac{P_{w_2,U}}{\sum_{i=0}^{U} P_{w_2,i}}, 0, \ldots, 0 \right)$.

Performance Calculation: Calculate $\boldsymbol{\pi}$ and the long-run total reward $\mathrm{TH}(\boldsymbol{A}^{(i)})$ for $i = 1, \ldots, n$ by solving balance equations and order these from largest to smallest, $\mathrm{TH}_1 \geq \cdots \geq \mathrm{TH}_n$. Define $\gamma = \mathrm{TH}_{\lfloor \rho n \rfloor}$.

Update: Using the same sample, update $\mathbb{P}^{(k+1)} = [P_{r,j}^{(k+1)}]$ as

$$P_{r,j}^{(k+1)} = \alpha \frac{\sum_{i=1}^{n} \mathbb{I}\left\{ \mathrm{TH}(\boldsymbol{A}^{(i)}) \geq \gamma \right\} \cdot \mathbb{I}\left\{ A_r^{(i)} = j \right\}}{\sum_{i=1}^{n} \mathbb{I}\left\{ \mathrm{TH}(\boldsymbol{A}^{(i)}) \geq \gamma \right\}} + (1 - \alpha) P_{r,j}^{(k)},$$

where $\mathbb{I}(A)$ is an indicator function of event A.

Set $k \leftarrow k+1$ and find ξ_r the index of the maximal element of the rth row of $\mathbb{P}^{(k)}$ for $r = 0, 1$.

Return (ξ_1, ξ_2) as the current best allocation.

Stopping Rule: Stop if (ξ_1, ξ_2) has not changed for c consecutive times.

Note that **Sample Allocations** and **Update** steps are from [**?**]. Also, it is recommend that $0.7 \leq \alpha \leq 0.9$, $0.05 \leq \rho \leq 0.2$ and $c = 5$. We use $\alpha = 0.8$, $\rho = 0.2$ and $c = 5$ in this chapter.

The drawback of this method is not being able to consider constraints: sample probability $\mathbb{P}$ is updated based on performance only. We introduce Bayesian Optimization method as an alternative to overcome this problem. Furthermore, we compare the performance of both methods in next section.

2.3.2 Traffic Simulation

In this chapter, we consider two classes of vehicles and the overall arrival rate λ is set to $\lambda = 3960$ vehicles per hour for most settings but varies from 3960 to 7200. We tested two scenarios where 80% and 50% of the traffic comes from class 1 vehicles, i.e. $[\lambda_1, \lambda_2] = [3168, 792]$ or $[\lambda_1, \lambda_2] = [1980, 1980]$. The size of vehicles in each class s_i is assumed to be $[1, 2]$ and the expected number of passengers (reward) carried in a single vehicle is assumed to be 1 and 1.5 for customers from class 1 and 2 respectively, $[r_1, r_2] = [1, 1.5]$.

For traffic flow speed models, we consider linear and exponential models modified from [**?**]. Suppose that there is only one class with size 1. A linear model assumes that $v(N_T(t))$ is calculated as follows:

$$v(N_T(t)) = v(1) - \frac{v(1)}{C_{\max}} N_T(t).$$

where $v(1)$ is referred to the free flow speed for a lone occupant and $C_{\max}$ is the jam density, also referred as the maximum physical density for the selected road segment.

An exponential model represents traffic flow speed as :

$$v(N_T(t)) = v(1) \exp\left[-\left(\frac{N_T(t) - 1}{\beta} \right)^{-\phi} \right],$$

where

$$\phi = \ln \frac{\ln(v(N_a)/v(1))}{\ln(v(N_b)/v(1))} \Big/ \ln \frac{(N_a-1)}{(N_b-1)}$$

$$\beta = \frac{N_a-1}{[\ln(v(1)/v(N_a))]^{1/\phi}} = \frac{N_b-1}{[\ln(v(1)/v(N_b))]^{1/\phi}}$$

and $(N_a, v(N_a))$ and $(N_b, v(N_b))$ are two points of a number of vehicles and its corresponding traffic flow speed which are selected from the empirical data to define traffic flow model.

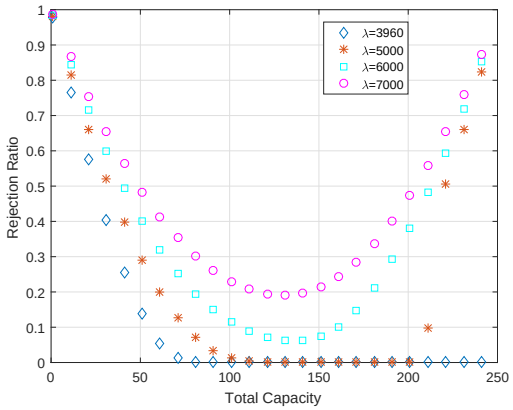

Figure 1: Rejection probability for allocation x_1 (or C_T) for various arrival rates λ.

For single one class customer traffic cases, Figure **??** showed that there are unique optimal solutions where rejection ratios are at minimum and total rewards are at maximum. In these cases, the total reward is deterministic and we can calculate using M/G/C/C queuing model at each alternative.

2.3.2.1 Two Classes Traffic Cases using CE Method

Now we consider two classes. A linear traffic flow model is used with $v(1) = 75$ miles per hour. At this point, all the test cases are deterministic. The total reward is again calculated by solving M/G/C/C queuing model.

[**?**] demonstrated the efficiency of the cross-entropy (CE) method when solving Problem (**??**) for $C_{\max} = 110, 160$ and 220 with $[\lambda_1, \lambda_2] = [3168, 792]$. The CE method uses parameters $\alpha = 0.8, \rho = 0.2$ and $\mathscr{N} = 400$. In all cases tested, the CE method stops after 13 iterations visiting 10%

to 20% of the alternatives comparing with exhaustive method, while both methods return the same allocation as the best.

[**?**] showed that CE Method could also be used to find the optimal allocation under various traffic flow speed models for Problem (**??**).

2.3.2.2 Cross-entropy method v.s. Bayesian Optimization method

Previous test cases showed that Cross-entropy method finds the optimal solution by visiting few number of solutions for deterministic cases. We now show that Bayesian Optimization method performs better under the same situation. Without loss of generality, we set $C_{\max} = 220, [\lambda_1, \lambda_2] = [1980, 1980]$, $v(1) = 75$ mph and adopt linear traffic model. We used the same stopping criteria for Cross-entropy method. For Bayesian Optimization method, we start with 10 random samples, $n_0 = 10$, and set the stopping criteria to be $\max(EI) < 0.1$ or $z_{B0} = 500$. Initially, we set σ_0 to be 100 and α to be $[10, 10]$. Then we update them during each iteration after initial sampling. Since the test cases are deterministic, sampling variance, σ_ε at each alternative is 0 and we only need to take one replication at each sample. We test both algorithm five times and compare the outcomes with true optimal solution find using exhaustive methods. The results are shown in Table **??**.

Based on the results, we observe that Bayesian Optimization method is able to find solutions that are close to the true optimal solutions as the Cross-Entropy method, while evaluating a much smaller number of alternatives. Since Bayesian Optimization method is also applicable to problems with constraints, we adopt it as our method for finding optimal solution for all the rest cases.

2.3.2.3 Deterministic Cases v.s. Stochastic Cases

In this case, we use the simulation model to capture the dynamic of highway traffic to represent problems with stochastic outputs. The purpose of the experiment is two-folded: on one hand, we test the performance of our proposed method while sampling variance σ_ε is not zero at any alternatives; on the other hand, using simulation to calculate output is more applicable to cases such as online shopping server, where the speed model is more complicated and the number of alternatives is much larger.

Table 3: Optimal solutions returned by the cross-entropy method and the bayesian optimization method.

Test case	True Optimal Solution	Cross-Entropy Method Results	Bayesian Optimization Method Results
C_{max}	$[(x_1,x_2),\text{tr},T_{EX}]$	$[(x_1,x_2),\text{tr},T_{CE}]$	$[(x_1,x_2),\text{tr},T_{BO}]$
110	$[(57,6),2006,3136]$	$[(57,6),2006,788]$	$[(57,6),2006,38]$
		$[(57,6),2006,718]$	$[(52,8),1998,17]$
		$[(58,6),2006,755]$	$[(57,6),2006,26]$
		$[(57,6),2006,740]$	$[(57,6),2006,29]$
		$[(57,6),2006,710]$	$[(57,6),2006,50]$
160	$[(67,28),2730,6561]$	$[(67,28),2730,935]$	$[(66,28),2730,31]$
		$[(67,28),2730,718]$	$[(62,28),2730,23]$
		$[(67,28),2730,755]$	$[(65,28),2727,35]$
		$[(67,28),2730,740]$	$[(63,28),2731,21]$
		$[(68,28),2730,710]$	$[(67,28),2730,44]$
220	$[(70,60),3586,12321]$	$[(70,60),3586,1028]$	$[(74,60),3586,53]$
		$[(73,60),3586,1024]$	$[(70,60),3586,131]$
		$[(69,60),3586,966]$	$[(71,60),3586,277]$
		$[(74,60),3586,930]$	$[(77,60),3586,68]$
		$[(70,60),3586,942]$	$[(71,60),3586,20]$

We set the simulation time as 3.5 hours including warm-up period of an hour, due to the assemblance to traffic condition in morning rush, which begin at 6 a.m. and peak at 7 a.m. until 9 a.m.. The disadvantage against short simulation time is system does not always reaches steady state. For example, in the previous test case, where $C_{max} = 220$, and $[\lambda_1,\lambda_2] = [1980,1980]$, when $(x_1,x_2) = (164,28)$, the analytic total rewards is 73 with rejection ratios as $[97.0,99.5]$; the simulation throughput is 3044, and rejection ratios are $[0,46.7]$. We sample five replications at this alternative and only once is the steady state reached. We prolong the simulation to 24 hours and the probability of reaching steady state is still not 100%. Although this is not ideal, this problem does not create issues when finding optimal solution using Bayesian Optimization in this test case since the true optimal solution has a total reward that's significantly higher than the total reward at this alternative. When the performances between the two are closer, optimal solutions found using Bayesian Optimization method could be far away from the one in deterministic case. This could be rationalized if the operation in real life does not reach steady state either. In other words, "theoretically bad" decision points are not always bad in actual performance.

Table 4: Optimal solutions returned by the bayesian optimization method for stochastic case.

Deterministic Optimal Solution $[(x_1,x_2),\text{tr}]$	Optimal Solution found using BO $[(x_1,x_2),\text{tr}$	Deterministic Total Reward at Corresponding Alternative $[(x_1,x_2),\text{tr}$	Number of Iterations and Maximum Expected Improvement $[T_{BO},\max(EI)]$
$[(70,60),3586]$	$[(71,58),3589]$	$[(71,58),3532]$	$[41,2.2]$
	$[(68,60),3587]$	$[(68,60),3544]$	$[27,0.59]$
	$[(70,58),3595]$	$[(70,58),3533]$	$[27,0.35]$

We test the performance of Bayesian Optimization for stochastic cases when $C_{max} = 220$, $[\lambda_1,\lambda_2] = [1980,1980]$, $v(1) = 75$ mph and adopt linear traffic model. We also test different sampling rules and stopping criteria. We change the stopping criteria from $\max(EI) < 0.1$ to $\max(EI) < 3$ to avoid excessive iterations. Simulation results show that in the former case, Bayesian Optimization will stop at maximum iteration of 500 while not achieving the expected improvement requirement. Note that we only take one replication at each evaluation to shorten the running time. The Optimal solutions shown in Table **??** are very close to the optimal solution in deterministic case.

2.3.3 Online Shopping Sites

2.3.3.1 Deterministic Shopping Time v.s. Stochastic Shopping Time

In this section, we apply our method to online shopping website admission control. Different from highway traffic, the structure of online server is more complicated, where admission control could be done on different level. In this chapter we consider the following simplified case: there are two type of customers, premium and regular, both type request same "space" from the server, $s_1 = s_2$, but with different job sizes and unit rewards. The actual server capacity of an online shopping website is extremely large, containing numbers of identical servers. If we could evenly distribute the incoming requests to all the servers, the searching process for optimal allocation strategy is much faster. In our test cases, we consider a server with 32 core CPU. The size of server and the amount of incoming traffic could be adjusted proportionally, so that our method could be used to other shopping site server as well.

For the speed model of the internet server, we assume the following: when system contains only

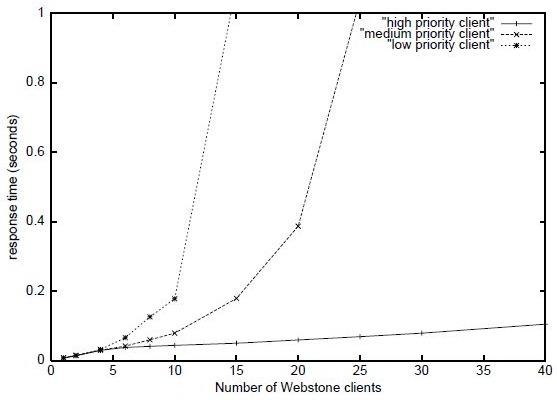

Figure 2: Internet speed model

one user, the service speed is $1Mb$. time of class 1 and 2 are exponentially distributed with mean $1/\mu_1$ and $1/\mu_2$. Since server response time increase exponentially as the total number of customers in system increases, shown in [**?**], we assume the service time of both classes behave in similar manners. Figure **??** showed an example of server respond time as a function of simultaneous users.

Since the speed model for internet service is exponential, the maximum capacity for a server, where speed reaches zero, is in theory infinity. However, maximum capacity is not needed for formulating the exponential speed model. Furthermore, we could adjust the searching space to a subset of maximum capacity for realistic reasoning. Given a 32-CPU server, assuming the respond time is 1.5s, and the user click frequency is 1 to 2 per minutes, it could handle 960 number of users simultaneously. Note that we assumed that all pages have uniform sizes and such for responding time, for simplicity. We further assume that the server could respond within 1 second at fastest, when only one user in system, and the speed follows a simple exponential model. The inflation factor for this server is approximately $e^{-1/2000}$. That is each additional customer admitted to the system, the remaining shopping time for all existing customers inflated by around 0.05%. This is certainly a simplification for the actual e-server performance. But in this setting, a customer that shops for 10 minutes when he is the only one using the server would spend more than an hour when number of customers in system is about 4000. In reality, server speed could keep deteriorate until system crush when number of customers reached a larger amount. Since we wanted to avoid the crushing situation, C_{max} is ignored in the online shopping settings. Instead we uses $C_T = 4000$ in

the following test cases as a boundary for solution searching.

Now we consider this example: Firm A's website has a monthly visit of 6.8 million customers. We assume 30 days a month, and 50% of the visits happened during peak hours, from 5p.m. to 11p.m.. Thus, approximately, Firm A's website has $19,000$ visiting requesting per hour during peak. We also assume that among these customers, 6 out of 10 are premium customers, who provide higher rewards, $(m_1, m_2) = (1.3, 1)$. We test two set of shopping rate means: $[\mu_1, \mu_2] = [6, 4]$per hour, or $[\mu_1, \mu_2] = [6, 12]$. The first scenarios representing regular customers shopping for longer time and the second the opposite. We also compare the outcome when the shopping times are deterministic with rates μ_1, μ_2, with outcome when the shopping times are exponentially distributed with rates μ_1, μ_2.

For Bayesian Optimization method, we choose to start with 10 random samples, $n = 10$ and take 3 replications at each alternative, $\beta = 3$. We set the stopping criteria to be when either $\max(EI) < 15$ or $z = 300$ is satisfied. We limit the maximum number of iterations to 300 due to the heavy computational workload of Gaussian Process Regression when number of alternatives is big. We start with test cases with different incoming traffic density, when no rejection ratio constraints are considered. Rejection ratios at each solution are obtained by running 100 replications after the optimal solutions is returned by BO. The results are shown in Table **??**.

In the first case, incoming e-traffic density is so high that the optimal solution is adopting class selection policy where the class with higher profit per unit resource used occupies all the resource and the other class was completely rejected. Similarity in optimal solution is seen between deterministic and stochastic cases, yet the results are slightly different. The reason lies in the randomness. In the deterministic case, a customer from class *i* entering the system earlier is definitively going to finish before a later one from same class. The "update" in server speed only varies the difference between the two finish times. This is not true for stochastic case. Moreover, in stochastic case, an customer who shops for longer time is going to utilize the server for a even longer time after service speed deflation. In the first test case, the average service time for customers left the system is 1648 seconds and the average remaining service time for customers in the system is 1056 seconds. While, both number are higher in stochastic case, 1676 seconds and 1668 seconds, part of the remaining customers go through significant delays. Test case 1 is considered to be unstable

Table 5: Online shopping website numerical examples 1

Scenario Setting	Optimal Solution Using BO (Deterministic Shopping time)	Optimal Solution Using BO (Stochastic Shopping time)
$\lambda = 19000$/hr $[\mu_1, \mu_2] = [6,4]$/hr	$x^* = [2050, 0]$ $TH = 5977$ $[\mu_{r_1}, \mu_{r_2}] = [0.596, 1.000]$ $[\sigma_{r_1}, \sigma_{r_2}] = [0.003, 0.000]$ $T_{BO} = 40$	$x^* = [2040, 0]$ $TH = 5721$ $[\mu_{r_1}, \mu_{r_2}] = [0.613, 1.000]$ $[\sigma_{r_1}, \sigma_{r_2}] = [0.005, 0.000]$ $T_{BO} = 30$
$\lambda = 19000$/hr $[\mu_1, \mu_2] = [6,12]$/hr	$x^* = [290, 1900]$ $TH = 8398$ $[\mu_{r_1}, \mu_{r_2}] = [0.940, 0.002]$ $[\sigma_{r_1}, \sigma_{r_2}] = [0.004, 0.001]$ $T_{BO} = 30$	$x^* = [280, 1920]$ $TH = 8349$ $[\mu_{r_1}, \mu_{r_2}] = [0.947, 0.001]$ $[\sigma_{r_1}, \sigma_{r_2}] = [0.002, 0.001]$ $T_{BO} = 128$
$\lambda = 9500$/hr $[\mu_1, \mu_2] = [6,4]$/hr	$x^* = [1950, 0]$ $TH = 5767$ $[\mu_{r_1}, \mu_{r_2}] = [0.220, 1.000]$ $[\sigma_{r_1}, \sigma_{r_2}] = [0.004, 0.000]$ $T_{BO} = 30$	$x^* = [1930, 0]$ $TH = 5728$ $[\mu_{r_1}, \mu_{r_2}] = [0.235, 1.000]$ $[\sigma_{r_1}, \sigma_{r_2}] = [0.005, 0.000]$ $T_{BO} = 257$
$\lambda = 9500$/hr $[\mu_1, \mu_2] = [6,12]$/hr	$x^* = [1340, 1060]$ $TH = 7089$ $[\mu_{r_1}, \mu_{r_2}] = [0.565, 0.000]$ $[\sigma_{r_1}, \sigma_{r_2}] = [0.003, 0.000]$ $T_{BO} = 32$	$x^* = [1300, 1100]$ $TH = 7046$ $[\mu_{r_1}, \mu_{r_2}] = [0.575, 0.000]$ $[\sigma_{r_1}, \sigma_{r_2}] = [0.006, 0.000]$ $T_{BO} = 245$

Table 6: Online shopping test considering redundant constraint

Without Constraint $[(x_1,x_2),\text{TH}]$	With Constraint $[(x_1,x_2),\text{TH}]$
$[(1930,0),5728]$	$[(1940,0),5734]$
$[(1940,0),5726]$	$[(1940,0),5718]$
$[(1940,0),5735]$	$[(1960,0),5720]$
$[(1960,0),5727]$	$[(1960,0),5751]$
$[(1960,0),5742]$	$[(1960,0),5738]$

from queuing perspective since $\lambda_1(11400) > \mu_1 x_1(2.10 * 2100)$. Using queuing model, we derived that when assigning 2000 spaces to class 1 customers, the rejection ratio would be $r_1 = \frac{\mu_1 x_1}{\lambda_1} = 0.60$, which matches our result in the deterministic case. For stochastic case, r_1 is also calculated to be 0.60 using queuing equations. This is lower than simulation result due to residue customers in the system shop for longer time, leaving less available resource for upcoming customer while slowing down the system performance for all customers in the system. This effect of delay is more significant when considering rejection ratio constraints, see Table **??**. Knowing that class selection policy is used for case 1, we could also verify the accuracy of the optimal solution found by using Bayesian Optimization method by comparing with queuing model solutions considering only class 1 traffic. Using simply derivation of $e^{(-x/2000)}x$, we found $x_1 = 2000$ to be the solution point that maximize the system throughput. For more complicated test cases, we validate the output from multiple runs in the next section and the results should be close to each other as they did in traffic case. For the rest of the chapter, we decide to use stochastic shopping time distributions to capture the real-life shopping experience.

2.3.3.2 Stochastic Shopping Time with Rejection Ratio Constraint

Without loss of generality, we consider the third test case. We use the same stopping criteria but take 10 replications when evaluating any alternative to improve the accuracy. We first verify the accuracy of our outputs by adding redundant constraint: $r_2 \geq 0.00$. The results are shown in Table **??**.

The optimal solutions show consistency between runs and between both cases. We further validate our method by replacing the redundant constraint with a constraint that would result in empty

feasible region: $q_2 < 0$. BO method stopped after 10 initial sampling and returned empty solution.

We then test cases with meaningful constraint: $q_2 \leq 0.9$. We run our algorithm five times and record the optimal solutions. We check if these solutions are truly feasible by taking 100 simulation replications at these points and record the mean and standard deviation of throughput and rejection ratio. The results are listed in Table **??**. It is shown that our method is able to output feasible solutions, and both the estimated throughput and rejection ratio are close to the sample mean of 100 iterations.

As mentioned previously, this test case is considered unstable. Using queuing theory, the minimum number of space reserved for class 2 to meet the rejection ratio constraint is:

$$x_2^{min} = (1 - \bar{q_2})\lambda_2/\mu_2$$

.

For verification purpose, we assume $X_T = 1650 + 380 = 2030$. Then, $\mu_2 = 1.09$ after adjustment. Thus, $x_2^{min} = 350$. Our simulation results showed that we need more than these spaces to guarantee rejection ratio constraint due to system performance degraded by long-shopping-time users. Thus, we cannot simply use queuing equation to find allocations matching rejection ratio constraint for unstable cases.

2.3.3.3 Pooled v.s. Dedicated

Finally, we compare the optimal allocation using pooled and dedicated methods for online shopping case. We consider the following case where $\lambda = 9500$, $C_T = 4000$ and $(p_1, p_2) = (0.6, 0.4)$. We test two set of shopping rates: $[\mu_1, \mu_2] = [6, 4]$ and $[\mu_1, \mu_2] = [6, 9]$. We still consider shopping time to be exponentially distributed. We also use the previous parameter settings, where 10 initial samples are taken and then 3 replications are taken at each evaluation. Table **??** showed results.

In both test cases, dedicated method performs better. In online shopping cases, adopting pooled method would result in customer from both class having same probabilities of being rejected, which would violate the idea of being "premier" for class 1 customer. Yet, the numbers of iterations using pooled method are only 10% of that of dedicated method since in the former case, the number of

Table 7: Online shopping test considering constraint

Optimal Solution Using BO $[(x_1,x_2),\text{TH}]$	Simulation Output $[\mu_{TH},s_{TH}]$ $[\mu_{q_1},\mu_{q_2}]$ $[\sigma_{q_1},\sigma_{q_2}]$
$[(1650,370),5106]$ $T_{BO}=27$	$[5120,55]$ $[0.362,0.894]$ $[0.023,0.003]$
$[(1650,380),5103]$ $T_{BO}=35$ $[(1650,380),5103]$ $T_{BO}=42$	$[5102,60]$ $[0.365,0.893]$ $[0.020,0.003,]$
$[(1680,380),9037]$ $T_{BO}=36$	$[5106,61]$ $[0.366,0.893]$ $[0.017,0.004]$
$[(1690,380),9037]$ $T_{BO}=33$	$[5105,61]$ $[0.893,0.366]$ $[0.021,0.004]$

Table 8: Online shopping website for pooled v.s. dedicated method

Shopping Rate $[\mu_1,\mu_2]$	Optimal Allocation $[x^*,TH]$ $(r_1,r_2),T_{BO}$	Optimal Allocation $[(x_1^*,x_2^*),TH]$ $(r_1,r_2),T_{BO}$
$[6,4]$	$[2340,3674]$ $[0.678,0.679],36$	$[(1930,0),5562]$ $[0.235,1.000],257$
$[6,12]$	$[2700,6528]$ $[0.434,0.435],33$	$[(1300,1100),7047]$ $[0.575,0.000],245$

alternatives is significantly smaller.

2.4. Conclusions and Future Work

In this chapter, we considered resource allocation problems among multiple class customers in traffic and e-server setting. We presented Bayesian Optimization methods for optimal solution search to maximize system throughput in both settings. Our test results showed that Bayesian Optimization method is able to find solutions that are closed to true optimal while evaluating a much smaller set of alternatives comparing with exhaustive search and Cross-entropy Method. Furthermore, we

are able to adopt this method in more complicated settings where system performance is stochastic, searching space is large, and secondary constraints are included.

CHAPTER III

MODELING THE IMPACT OF PROACTIVE COMMUNITY CASE MANAGEMENT ON REDUCING MALARIA TRANSMISSION IN SUB-SAHARAN AFRICAN COUNTRIES

In this chapter, our goal is to identify essential factors influencing the impact of ProCCM and develop a methodology to predict the effectiveness of various implementation strategies considering the resources required. We built an agent-based simulation that models the mosquito population, the transmission from mosquitoes to humans, and transitions between different stages of infection in humans. The model was validated using the data from the pilot [**?**]. The model enables the user to test various strategies for implementing ProCCM (e.g., with different combinations of coverage and visit frequency), and compare outcomes.

3.1. Methods

We developed an agent-based simulation model that incorporates the dynamics of the mosquito population over time, the progression of malaria within humans, and transmission dynamics between humans and mosquitos.

3.1.1 Data

The majority of the data for the simulation model (see Appendix, **??**, Table **??**) comes from the Saraya Health District, with a total population of $52,590$ full-time residents [**?**]. We estimate the treatment seeking rate for fever in children under 5 to be the same as that of adults and 30% of the population to be immune to malaria [**?**] at the beginning of the simulation. To capture the impact of seasonality on transmission patterns, we use temperature and precipitation data from 2010 to 2013 obtained from NNDC Climate Data Online [**?**] (see Appendix, **??**,Figure **??**). The malaria transmission model proposed in this chapter is built comprehensively on existing literature.

3.1.2 Agent-Based Simulation Model

3.1.2.1 Mosquito Population Model

The mosquito (*Anopheles*) population model [**?**] is based on a system of ordinary differential equations (see Appendix, **??**, Equation **??**) and incorporates the impact of temperature and precipitation [**?**], [**?**]. Table **??** and Figure **??** in the Appendix (**??**) depict the stages in the life cycle of Anopheles: aquatic stages (E, eggs; L, Larvae; P, pupae), emerging adult stage (A_{em}), nulliparous stages (A_{1h}, A_{1g}, A_{1o}) and parous stages (A_{2h}, A_{2g}, A_{2o}) [**?**]. Note that only the female mosquitoes are represented in the adult stage. The parameters and functions are defined in Tabl **??** and Table **??** in the Appendix (**??**). The adult female mosquito population (N_v) is divided into three groups: clean mosquitos (S_v), incubating mosquitos (C_v), and infectious mosquitos (I_v), i.e., $N_v = S_v + C_v + I_v$.).

3.1.2.2 Human Population Model

Our simulation model is based on [**?**] capturing the same progression process between human infection states while being less detailed. We made simplifications such as classifying humans into categories based on age, and including only antiparasite immunity, which confers protection against high-density parasitemia and the attendant risk of severe disease. These modifications simplify the age-based simulation model while still capturing the essence of the malaria transmission process. We separately model children under 5 (C) and human older than 5 (A). The transition processes between disease states follows the same patterns for both groups, but with different parameters. At a given time, a human is in one of the following states based on his/her infectious condition [**?**], [**?**], [**?**]: clean non-immune (S_h), incubating non-immune (C_h), infectious non-immune (I_h), clean immune (SR_h), incubating immune (CR_h), infectious immune (IR_h), for $h \in \{A, C\}$. After a successful infection by a mosquito, the infected human also goes through (some of) the following phases, based on the severity of their symptoms [**?**]:

$$\begin{aligned} \text{Infection} \rightarrow \text{Asymptomatic Parasitemia}(A_h) \quad &\rightarrow \text{Uncomplicated Illness}(M_h^m) \\ &\rightarrow \text{Severe Malaria}(M_h^S) \rightarrow \text{Death}. \end{aligned} \tag{6}$$

Model details are included in the Appendix, **??**.

3.1.2.3 Interaction between Human and Mosquito Populations

A clean non-immune human is susceptible to malaria and could be infected (with probability I_v) after a bite from an infectious mosquito. We assume that every human has equal chance of being contacted by any mosquito [**?**] and children under 5 years old do not have immunity and thus, always show symptoms if infected ($P_{gi} = 0$) [**?**]. We adopt the idea of effective contact ratio that determines daily numbers of newly infected humans (NI_h) and newly infected mosquitoes (NI_v) [**?**] and incorporate the individual infectiousness level, $i_H(t) \in [0,1]$ (see Appendix, **??**, Equation **??**).

3.1.2.4 Treatments and Recovery Processes

A human with uncomplicated malaria or severe malaria has the possibilities of self-recovering, e.g., self-treatment, seeing a traditional healer, taking herbal medication [**?**] or simply recovering as the immune system responds. If a symptomatic human seeks treatment, at any stage after developing symptoms, we assume that he/she will receive artemisinin-based combination treatments (ACTs) [**?**], and the recovery process starts immediately after the first day [**?**]. Treatment outcome depends on the how long one has been infected as well as the severity of his/her symptoms. Details are explained in the Appendix, **??**. After recovery, a non-immune human can gain immunity with a certain probability ($P_g i$) [**?**]. An immune human will become infected after a successful bite from an infected mosquito with a smaller chance [**?**]. After being infected, the immune human will then go through the asymptomatic parasitemia phase with a longer duration [**?**]. The majority of the immune humans with asymptomatic parasitemia will self-recover. Some who develop symptoms after the asymptomatic parasitemia phase may recover if they seek treatments. Losing immunity after recovery (P_{ii}) is also very unlikely under consistent heavy malaria exposure [**?**]. Details are explained in the Appendix, **??**. Traditionally, an infected human would only actively seek treatment (P_{st}) if he/she developed symptoms. Treatment-seeking possibilities are different between humans in the adult group and children. Those who dont seek treatment will keep infecting clean mosquitos. ProCCM could actively detect infected humans with symptoms, give them treatments and terminate the infection loop.

3.1.2.5 Model Validation

We first validated our simulation model by comparing the model results with the data collected during the ProCCM pilot under Scenarios I and II (see Table **??** in the Appendix, **??**, for details of each scenario) [**?**]. The simulation results, including the baseline case of no sweeps, are shown in Figure **??** and Figure **??**, representing scenarios I and II, respectively.

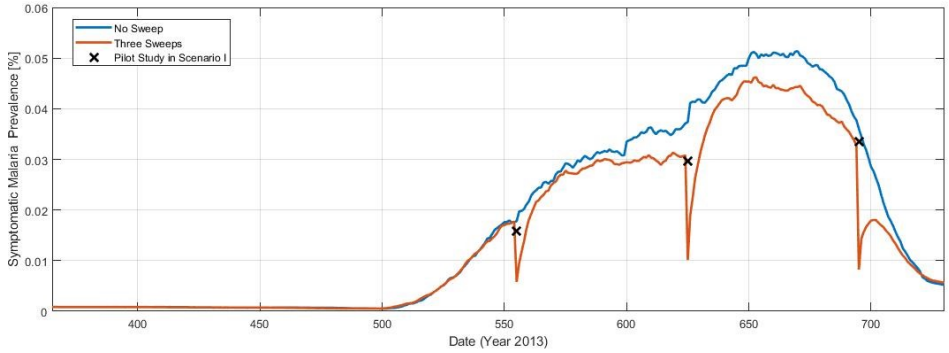

Figure 3: Symptomatic Mmalaria prevalence simulation results for scenario I (3 sweeps)

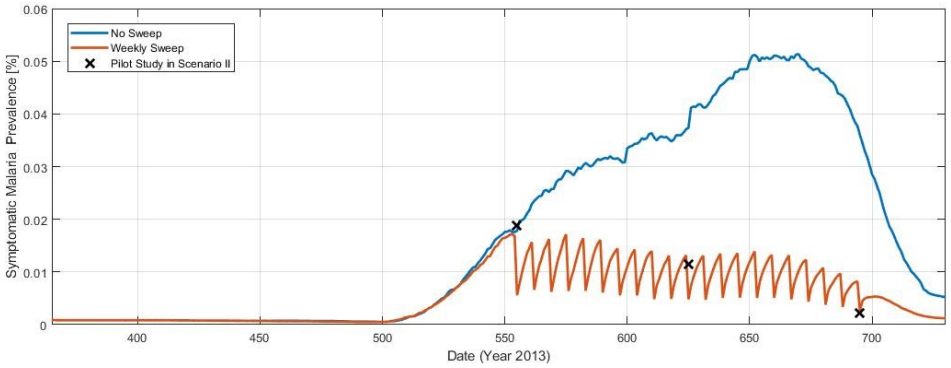

Figure 4: Symptomatic Mmalaria prevalence simulation results for scenario II (weekly sweeps $*21$ weeks)

Based on the mean of 50 replications, our model shows consistency with the malaria infection data from the pilot study [**?**]. The data points from the pilot study lie within the 95% confident intervals of our simulation results. Validation details are included in the Appendix **??**.

3.1.2.6 Simulation Model Runs

After validation, we ran our simulation model considering various ProCCM implementation strategies (see Table **??**) during the 21 weeks of the malaria transmission season. For each strategy, we performed 50 replications and recorded the mean number of year-round symptomatic malaria infection cases, peak percentage of humans with symptomatic malaria infections (peak symptomatic malaria infection prevalence), the total number of RDT tests used during ProCCM sweeps and the corresponding standard deviations of the means (reported in Table **??** to **??** in the Appendix **??**. Each simulation run spans 2 years, including a warm-up period of one year. To test the robustness of model results, we also conducted a sensitivity analysis on ProCCM sweep coverage and simulation year (see **??** in the Appendix).

Table 9: List of all simulated ProCCM strategies

ProCCM Strategy	Coverage	*No.*Sweeps	Sweep Starting Date	Sweep Ending Date	*No.* Replications (*No.* strategies $*$50 replications $*$ *No.* years simulated)
A. No Sweep	-	0	-	-	$50*4$
B. 3 Sweeps	100%	3	July 8th	Nov 25th	$50*4$
C. Weekly Sweeps	100%	21	July 8th	Nov 25th	$50*4$
C^*. Weekly Sweeps	80 - 95%	21	July 8th	Nov 25th	$4*50*4$
D. Weekly Sweeps	50%	21	July 8th	Nov 25th	$50*4$
E. Bi-weekly Sweeps	100%	11	July 8th	Nov 25th	$50*4$
F. Twice per Week	100%	41	July 8th	Nov 25th	$50*4$
G. Weekly Sweeps for 7 weeks	100%	7	July 8th	Nov 11th	$13*50*4$
H. Bi-weekly Sweeps for 14 weeks	100%	7	July 8th	Nov 11th	$7*50*4$
I. Weekly and bi-weekly combined	100%	14	July 8th	Nov 18th	$50*4$
I^*. Weekly and bi-weekly combined	80 - 95%	14	July 8th	Nov 18th	$4*50*4$
J. Weekly and bi-weekly combined	100%	11	July 22nd	Oct 21st	$50*4$

3.2. Results

We compared the number of year-round symptomatic malaria infection cases and peak symptomatic malaria infection prevalence under different sweep strategies (see examples in Table **??** to **??** in the Appendix **??**). Figure **??** shows the year-round percent reduction (compared to the base case of no sweeps) for symptomatic malaria infection cases and Figure **??** shows the average number of symptomatic malaria infection cases detected per sweep under different sweep strategies.

Figure 5: Reduction of Symptomatic cases under various strategies

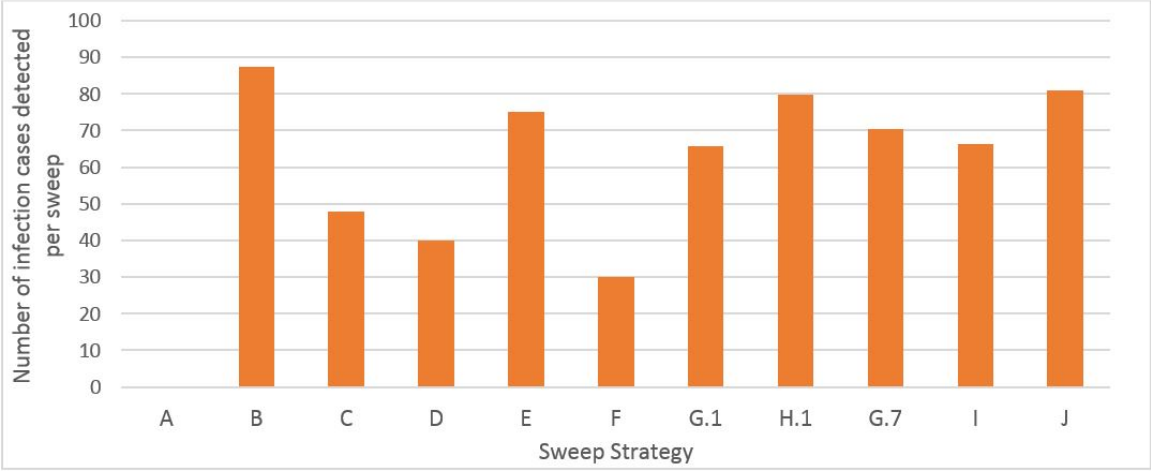

Figure 6: Number of infection cases detected per sweep under various strategies

In Scenario I, shown in Figure **??**, we observe that having 3 sweeps at the beginning, middle, and end of the malaria peak season does not provide ideal infection control, only reducing 12.8%

of symptomatic infection cases per year. Symptomatic malaria infection cases significantly decline right after each ProCCM sweep. Yet, this number quickly returns to the level where no sweeps are conducted, after seven to ten days. Due to the high infection prevalence within the mosquito population during the peak season, simply identifying and treating infected humans, once, or even at a very low frequency, would not effectively control malaria infection from the community level. To prolong the effect of ProCCM, conducting weekly sweeps for 21 consecutive weeks (Scenario II), shown in Figure **??**, provides a much more promising result: 37.8% of symptomatic infection cases are reduced per year. Moreover, the peak level of symptomatic malaria infection prevalence is less than half compared to no intervention. When adopting strategy I, combining weekly and biweekly sweeps for a total of 14 sweeps during peak season, 34.2% symptomatic infection cases are reduced per year.

3.3. Discussion

Using an agent-based simulation model, we predict that ProCCM is very effective in reducing malaria incidence at the community level in this setting with moderate, highly seasonal transmission. Annual symptomatic malaria cases could be reduced by as much as 41.6% by conducting twice a week sweeps during the peak infection transmission season. Early start, long-term intervention, and high sweep frequency are critical factors in improving the ProCCM outcome, although these all increase costs. By varying weekly and biweekly sweeps during peak season (with higher frequency in mid-season), we propose an alternative strategy that could reach a similar level of infection malaria incidence reductioncontrol, while reducing the intervention cost by 33.3%.

Conducting ProCCM is very effective in reducing symptomatic malaria infection control. This can be explained on the individual and community level. Early detection allows for identification of infection at a lower parasite density, which could lead to a shorter recovery time after treatment, less chance of treatment failure, and a lower mortality rate. Furthermore, an infected human who is less likely to be gametocytemic at the time of treatment is likely to be less infectious during and after treatment, although the relationship is not linear. By reducing the chance of infecting clean mosquitoes and shortening the duration of infectiousness, malaria prevalence is reduced at the community level. Ideally, if ProCCM could detect all infected humans during the incubation

period, before they become infectious, the transmission between human and mosquito populations would end, though this level of achievement is highly unlikely. Based on our simulation results, the most obvious factor influencing the ProCCM outcome is sweep frequency. Yet, the relationship between level of reduction for malaria infection cases and sweep frequency is not linear, as shown in Table 9. Biweekly sweeps during peak season, 11 sweeps per year, could reduce malaria infection cases by 31.5% per year. If conducted weekly during peak season, the infection cases are reduced by 37.8%. This number can be increased even further to 41.6% by doubling the sweep frequency again to twice weekly. Higher sweep frequency results in more effective malaria infection control, as one would expect. However, a clear asymptote is reached when the frequency is above weekly. Besides frequency, our simulation results suggest that starting date of intervention, sweep duration and sweeping coverage are other critical factors for ProCCM in this setting. A longer intervention period is also shown to be beneficial as one would expect. As mentioned previously, malaria peak season starts around early July in Senegal (around the 27th week). Our simulation model shows a significant increasing trend in the mosquito population, which provides more opportunities for malaria transmission. Thus, early intervention could delay the outbreak of malaria in September. If not able to conduct sweeps starting at the beginning of the transmission season and throughout the season, deciding when to start interventions would be critical. Simulation results indicate that the optimal starting point for intervention is within the first five weeks after the beginning of peak season if only 7 consecutive sweeps are conducted. More details are included in the Appendix (Section 4). Since the mosquito population is significantly smaller out of the peak season, the effect of ProCCM sweeps is predicted to be limited during this time. After the peak season, as there aren't any new emerging mosquitoes joining the adult population, transmission wanes. Our results confirm that conducting fortnightly sweeps out of the peak season is inefficient, reducing the year-round symptomatic infection cases by around 2%. Assuming ProCCM implementation cost is directly related to number of sweeps conducted, two alternative strategies appear potentially most cost-effective. Compared with a weekly sweep strategy, strategies I and J (where biweekly and weekly sweeps are combined with higher frequency during mid-peak season) result in only 6.3% and 3.6% increases in the total number of year-round symptomatic malaria infection cases, but result in costs of only 52.3% and 66.6% of the cost of annual weekly sweeps. Overall, strategy *J*

(where a total combination of 11 biweekly and weekly sweeps are conducted with higher frequency during mid-peak season) is the more cost-efficient while strategy I (where a total of 14 biweekly and weekly sweeps are conducted with higher frequency during mid-peak season) is more effective, although both provide relatively good control of infections while reducing costs.

3.4. Limitations

There are parameters in our human infection transition model that are estimated within certain ranges. Given that each infected human responds to malaria differently, it is hard to specify these ranges to smaller intervals or even describe them using particular distributions. Hence, our simulation model outcomes have high variance. To solve this problem, we used 50 replications for each simulation case to minimize the variance. Additionally, we adopted parameter ranges that fit our model the best and provided the closest result to existing data in [**?**].

The model used in this chapter uses, among other input variables, temperature, precipitation and human characteristics, which can be adapted to other societies and climates. With some calibration on parameters, the model could be adopted to other malaria endemic regions. Our model also enables decision makers to evaluate ProCCM strategies for a given budget.

Disclaimer: The findings and conclusions in this chapter are those of the authors and do not necessarily represent the views of the U.S. Centers for Disease Control and Prevention or the US Agency for International Development.

CHAPTER IV

MODELING THE TRANSMISSION OF GUINEA WORM DISEASE IN DOGS AMONG MULTIPLE WATER SOURCES AND EVALUATING THE EFFECTIVENESS OF INTERVENTION METHODS IN CHAD.

In this chapter, we built a multi-water-source(MWS) agent-based simulation model, based on [**?**]., to model the GW disease transmission among dogs in Chad, as well as in each domestic clusters. This model contains three connected single-water-source (SWS) models, partitioning the majority part of Chad based on clustering results. Each SWS model adopted the general framework of the previous stochastic simulation model, which is used to simulate the life cycle of GW and daily interactions between the dogs, worms, and water source over multiple years. Each SWS model was validated using infection data within the corresponding cluster after parameter calibrations. Three SWS models were then connected into a MWS model using geographic information and local human characteristic data.

We examined strategies for allocating tethering and abating resources among three clusters in Chad. Using our MWS simulation model, we found the level of intervention coverages needed to reach nationwide eradication when interventions are performed uniformly is lower than the level suggested by [**?**] due to natural division among clusters. Next, we conducted sensitivity analysis and found more effective allocation schemes while maintaining the same resource capacity. In particular, allocating more tethering resource in the Central cluster and more abating in the West cluster results in significantly better infection control. Finally, we presented a simulation optimization algorithm that enables the decision-maker to find the optimal intervention strategy under any resource capacity and allocation constraints. We examined the effectiveness of intervention strategies based on the percentage of dog population infected after years of intervention, and fairness of intervention strategies based on the maximum difference in intervention coverages between clusters.

4.1. Data and Methods

4.1.1 Data

The Chad GWEP provided data about the number of worms emerging from dogs per month and the number of dogs with an emerging worm per month for the years 2016-2018 within 1674 village [**?**]. Data on daily precipitation and temperature measurements from all 17 weather stations in Chad were collected from the National Centers for Environmental Informations Climate Data Online database [**?**]. We used the average monthly rainfall and temperature over the years 2016-2018.

4.1.2 Agent-Based Simulation Model

The simulation model presented in [**?**] simulates the life cycle of GW and daily interactions between the dogs, worms, and water source over multiple years. With some probability, dogs acquire infection from the water-source (WS) via consumption of water or (short-term) transport hosts, and dogs with patent infections (emergent worms) infect the water source. Dogs remain susceptible to new infections, whether they are already infected or not, and one or more GWs emerge from an infected dog.

The agent-based simulation model presented in this work inherited the framework of the previous one with the same dynamics of GW transmission in dogs. Instead of considering Chad as one shared WS, we first partitioned Chad into multiple clusters and then considered Chad as an integrated model of three connecting WSs.

4.1.3 Regional Clustering

Our data documented GW infections within 5 regions, 19 districts, 88 zones, and 1674 villages of Chad. We considered "district" as the smallest units for clustering to balance the level of complexity and data quality (of GW infection in each unit). We used the relative worm emergence per month, latitude, longitude, position along river, and elevation of each district center as inputs for K-means to generate clustering results. The clustering result is shown in Figure **??**.

Our model considered three clusters in Chad (Table **??**), including 76.4% of the susceptible dog population and 92.7% of the infected dog population documented in the data. Details are included in the Appendix, Section **??**.

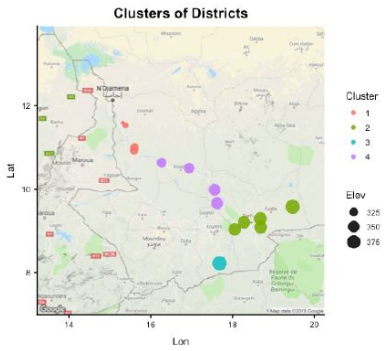

Figure 7: Clustering of Chad districts

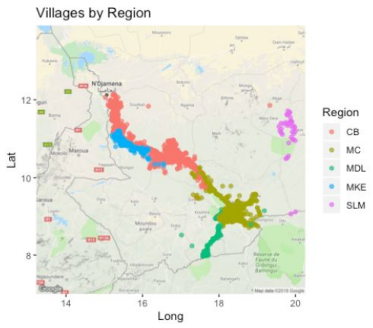

Figure 8: Administrative regions in Chad

Table 10: Clustering result for Chad

Cluster (Figure**??**)	Districts	Susceptible Population	Infected Population
1 West Chari	Dourbali, Guelendeng, Mandelia, Massenya, Ndjamena-Sud	8223(14.3%)	1219(36.0%)
2 East Chari	Banda, Bedaya, Biobe, Danamadji, Haraze, Kyabe, Sarh	23,855(41.4%)	1,245(36.7%)
4 Central Chari	Bailli, Bousso, Korbol, Kouno	11,910(20.7%)	679(20.0%)

4.1.4 Parameter Calibration

We first integrate the number of worms emerging from dogs per month and the number of dogs with an emerging worm per month for the years 2016-2018 for each cluster (Table **??**). We calibrated infectivity parameters and seasonality parameters for each cluster using integrated data from the corresponding cluster. Parameter searches are done using the Cross Entropy method. Details are included in the Appendix, Section **??**.

4.1.5 Multi-Water-Source Model

In this section, we present the final simulation model integrating the three single-water-source models discussed above. Each single-water-source model contains distinct infectivity parameters, seasonality parameters, dog population, and intervention coverages from 2016 to 2018. Traveling behavior of dogs, which defines the level of interactions between clusters, are determined based on McDonalds research on the ecology of Guinea Worm infection in dogs [**?**]. We validate this

model using data collected from Chad (Figure **??**). Additional details are included in the Appendix, Section **??**.

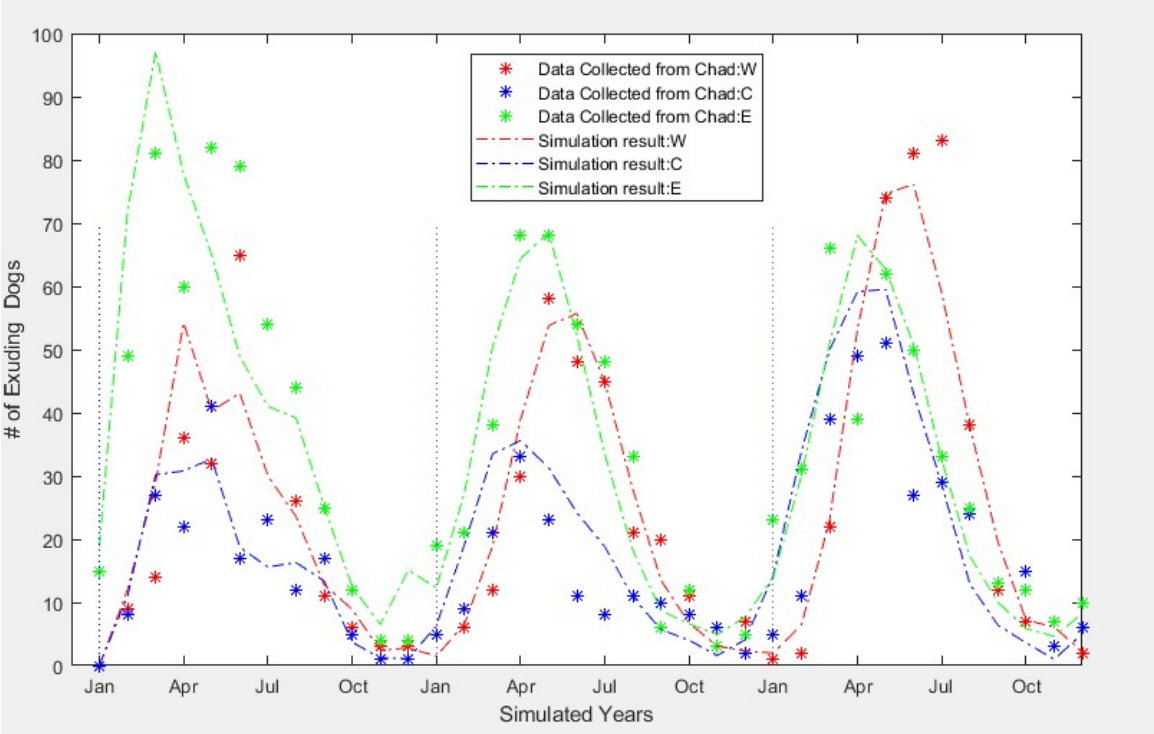

Figure 9: Simulation results for MWS model

4.1.6 Intervention Methods

Five intervention methods are examined in the previous work: tethering, abating, vaccine, alternative water provision, and preventive tethering [1]. It is shown that a combination of high level (96%) tethering and moderate-high level (75%) of abating is most likely to reach eradication within ten years. Additionally, vaccines, alternative water provision, and preventive tethering work in similar ways in our model by limiting the successful transmission rate between uninfected dogs and contaminated water sources. Greedy-based allocation strategy would be efficient for these intervention methods: under a given intervention resource capacity, cover dog groups from the most likely to be infected to the least. Thus, we only consider tethering and abating in this work.

4.1.7 Simulation Optimization Methods

We use the Cross-Entropy Method presented in Wang et al. to search for optimal intervention strategies under different tethering and abating resource capacities [**?**]. We used the percentage of dogs newly infected in the end year as the primary objective. To avoid imbalanced strategies, we also added a penalty on the maximum difference in tethering coverages among clusters. We further adapted the algorithm of the CE method with performance constraints presented in [**?**]. Model details are included in the Appendix, Section **??**. All test cases and simulation results (mean of 20 replications) are listed in Table 3 to Table 8. Test cases are run for 9 to 24 years. The standard error is less than 2% of the mean for all cases.

4.2. Results

We first examine intervention strategies, with a combination of uniform tethering and uniform abating. We analyze the effectiveness of each strategy using use percentage of dogs newly infected in the end year and the duration needed to achieve eradication in all three clusters. As shown in the heat map (Table **??**), we simulated tethering coverage from 45% to 95% and abating coverage from 15% to 75%. We limit the upper bound of abating coverage to be less than four times higher than the national average between 2016 to 2018.

Table 11: Percentage of dog infection after uniform intervention for 5 years.

		Abating Coverage (Uniform)												
		0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75
Tethering Coverage (Uniform)	0.45	6.19%	5.63%	5.05%	4.57%	4.10%	3.64%	3.10%	2.59%	2.12%	1.54%	1.04%	0.79%	0.50%
	0.5	5.84%	5.37%	4.86%	4.31%	3.89%	3.38%	2.90%	2.35%	1.81%	1.32%	0.94%	0.66%	0.49%
	0.55	5.52%	5.06%	4.59%	4.07%	3.59%	3.08%	2.44%	1.96%	1.51%	1.02%	0.79%	0.60%	0.44%
	0.6	5.25%	4.71%	4.22%	3.70%	3.19%	2.59%	2.00%	1.51%	1.22%	0.91%	0.75%	0.55%	0.42%
	0.65	4.76%	4.33%	3.79%	3.17%	2.50%	1.98%	1.56%	1.24%	0.99%	0.79%	0.65%	0.51%	0.40%
	0.7	4.03%	3.40%	2.77%	2.33%	1.90%	1.48%	1.30%	1.02%	0.89%	0.77%	0.61%	0.50%	0.36%
	0.75	2.95%	2.48%	2.01%	1.72%	1.44%	1.27%	1.08%	0.94%	0.81%	0.69%	0.56%	0.46%	0.34%
	0.8	1.96%	1.74%	1.54%	1.35%	1.30%	1.10%	0.97%	0.86%	0.74%	0.61%	0.49%	0.42%	0.31%
	0.85	1.58%	1.42%	1.34%	1.21%	1.09%	0.99%	0.84%	0.77%	0.64%	0.55%	0.42%	0.32%	0.26%
	0.9	1.36%	1.25%	1.16%	1.07%	0.94%	0.85%	0.72%	0.61%	0.53%	0.41%	0.37%	0.22%	0.17%
	0.95	0.95%	0.87%	0.82%	0.70%	0.58%	0.53%	0.39%	0.30%	0.23%	0.16%	0.03%	0.01%	0.00%

We used the Cross-Entropy (CE) method to search for optimal intervention strategy under the resource capacities that can cover 70% tethering and 20% abating if conducted uniformly (purple cell in Table **??**). Additionally, we added allocation upper bounds (maximum intervention coverage in any cluster) for tethering and abating coverages in any clusters. Initially, we limited the upper bounds for tethering to be 95% and abating to be 50% at any cluster for realistic reasons. We further lowered the upper bounds of tethering from 95% to 75% and the upper bounds of abating from 50% to 25%. The percentages of infected dogs after five years of intervention with the optimal strategy under any set of upper bounds are shown in Table **??**. The true optimal solution is shown in Table **??**.

Table 12: Percentage of dog infection using optimal intervention strategy (under different allocation upper bounds) for 5 years.

Resource Level: [0.70, 0.20]		Upper Bound of Abating						
		0.2	0.25	0.3	0.35	0.4	0.45	0.5
Upper Bound of Tethering	0.7	3.40%	-	-	-	-	-	-
	0.75	-	2.18%	2.03%	1.95%	1.90%	1.89%	1.88%
	0.8	-	1.85%	1.82%	1.80%	1.78%	1.73%	1.69%
	0.85	-	1.68%	1.67%	1.65%	1.62%	1.55%	1.52%
	0.9	-	1.59%	1.52%	1.51%	1.49%	1.48%	1.46%
	0.95	-	1.54%	1.46%	1.41%	1.40%	1.35%	1.31%

Table 13: Percentage of dog infection using optimal intervention strategy (under different allocation upper bounds) for 5 years.

Resource Level: [0.80, 0.30]		Upper Bound of Abating				
		0.3	0.35	0.4	0.45	0.5
Upper Bound of Tethering	0.8	1.35%	-	-	-	-
	0.85	-	1.33%	1.31%	1.27%	1.25%
	0.9	-	1.28%	1.24%	1.16%	1.02%
	0.95	-	1.10%	0.97%	0.91%	0.88%

Beside intervention effectiveness, fairness is another focus. Lowering the upper bound for tethering or abating coverage in any cluster is one way to maintain decision fairness (equality of tethering/abating coverages between clusters). Alternatively, we added penalty weight on the maximum tethering coverage difference between and adopted the same CE algorithm for optimality search.

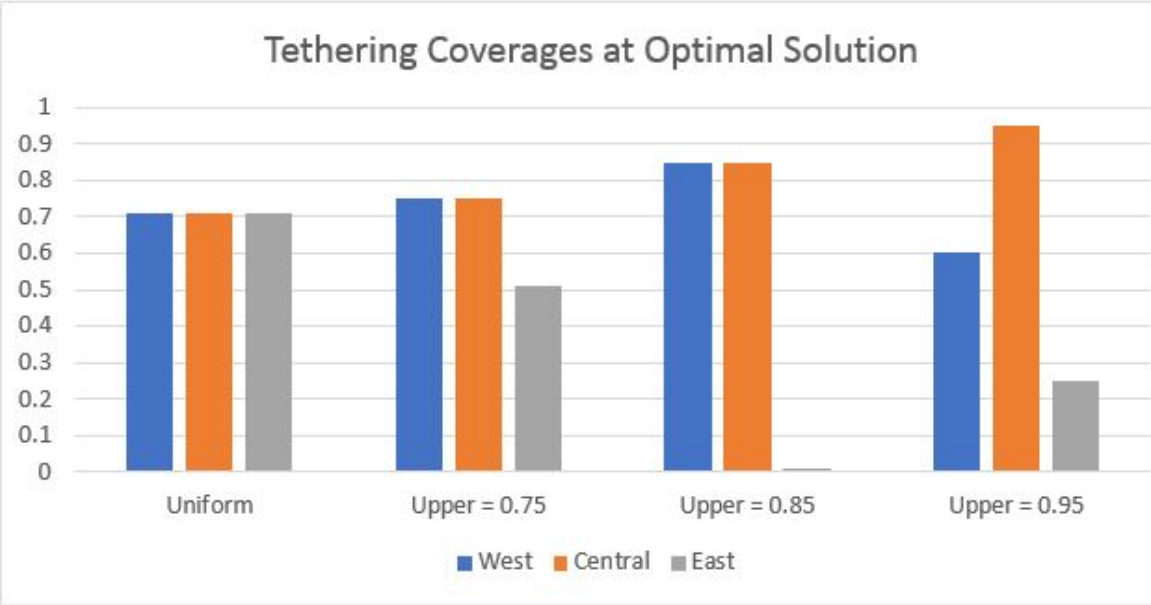

Figure 10: Tethering coverages in all clusters under different allocation upper bounds

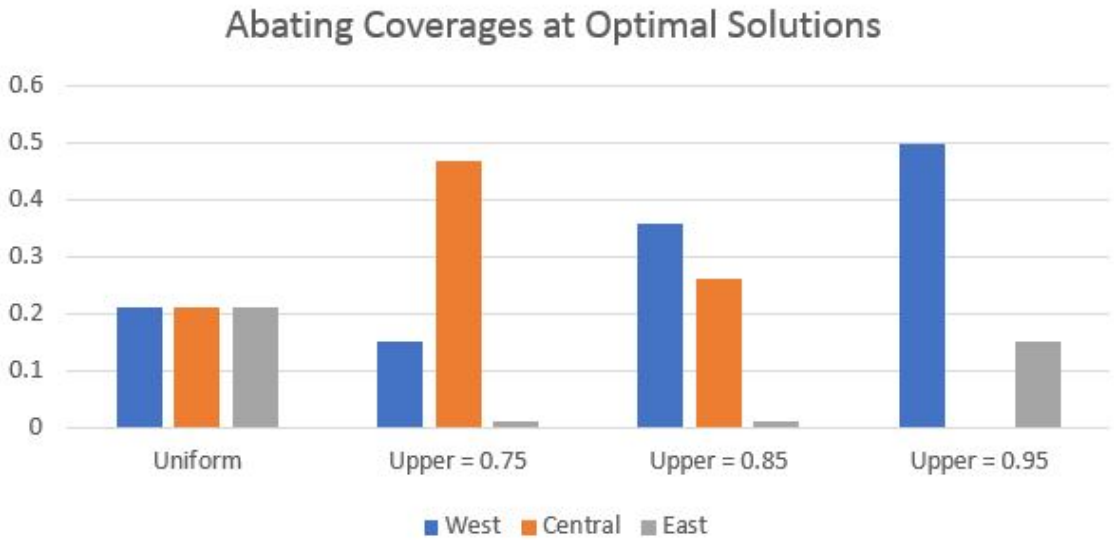

Figure 11: Abating coverages in all clusters under different allocation upper bounds

Table 14: Optimal intervention strategy under resource capacity of [0.70,0.20].

Intervention	West	Central	East
Tethering	0.60	0.95	0.25
Abating	0.50	0.00	0.10
Percentage of Infected dogs	1.33%	0.62%	1.71%

Table 15: Optimal intervention strategy under resource capacity of [0.80,0.30].

Intervention	West	Central	East
Tethering	0.75	0.95	0.45
Abating	0.60	0.00	0.30
Percentage of Infected dogs	0.70%	0.30%	1.25%

We added penalty weight, between 0 to 20, on the maximum tethering coverage difference between clusters while keeping the allocation upper bounds to be 95% and 50% for tethering and abating respectively. The results are shown in Table **??**. We also considered outcome fairness (equality of interventions between clusters). Specifically, we find the optimal intervention strategy (Table **??**) under which all clusters have lower infection percentages compared with a uniform intervention strategy.

Finally, we examined the optimal intervention strategies under a different resource capacity.

Table 16: Optimal intervention strategy under different penalty weight.

Penalty Weight	T_w	T_c	T_e	A_w	A_c	A_e	Percentage of infected dogs (Total, [Regional])
0 (True Optimal Solution)	0.6	0.95	0.25	0.5	0.00	0.10	1.31% $[1.33\%, 0.62\%, 1.72\%]$
2	0.6	0.9	0.3	0.40	0.00	0.20	1.45% $[2.01\%, 0.94\%, 1.55\%]$
4	0.63	0.9	0.34	0.37	0.00	0.23	1.52% $[2.37\%, 1.01\%, 1.46\%]$
6	0.53	0.9	0.54	0.39	0.00	0.21	1.63% $[2.55\%, 1.16\%, 1.52\%]$
20 (Uniform)	0.70	0.70	0.70	0.20	0.20	0.20	3.23% $[2.96\%, 6.92\%, 1.51\%]$

Table 17: Optimal intervention strategy with minimum performance(outcome fairness) constraint.

Intervention	West	Central	East
Tethering	0.77	0.88	0.14
Abating	0.17	0.01	0.42
Percentage of Infected dogs	2.76%	1.18%	1.13%

We increase the resource amount from 70% to 80% for uniform tethering, and from 20% to 30% for uniform abating (pink cell in Table **??**). The optimal strategies under various allocation upper bounds are shown in Table **??**. The true optimal solution is shown in Table **??**.

4.3. Discussion

We first examine intervention strategies, with a combination of uniform tethering and uniform abating. After five years of intervention, eradication can only be achieved when tethering coverage is higher than 95%, and abating coverage is higher than 65%. Eradication takes longer when tethering coverage is 95% while abating coverage is lowered to 60%. When tethering coverage is below 90%, and abating coverage is below 60%, more dogs are infected each year. Thus eradication is impossible. Specifically, simulation results showed that 3.23% of dogs would be infected in 2024 when tethering and abating are conducted uniformly with 71% and 21% coverages, respectively. Data collected from Chad showed that 2.32% of dogs are infected in 2018 with the same average intervention coverages. The rest of the strategies, blue cells in Table 2, decrease the percentage of dogs infected per year, yet eradication is not achievable in the foreseeable future. Our results showed consistency with results presented in [**?**]: eradication can only be achieved with an extremely high level of tethering and a moderately high level of abating. Modeling Chad as a single shared WS significantly increases the level of interactions between clusters. Therefore, the intervention coverages needed to reach eradication is slightly lower than [**?**].

We perform sensitivity analysis on intervention coverage changes. When abating coverage is low ($\leq 25\%$), increasing tethering from 65% to 75% results in the most significant reduction (1.78% - 1.84%) in the percentage of infected dogs. When abating coverage is moderate (30% -55%), increasing tethering from 55% to 65% results in the most significant reduction (0.52% - 1.10%) in

the percentage of infected dogs. Additionally, it is shown that abating is more effective than tethering when having the same coverages: 55% tethering combined with 75% abating results in better intervention outcome than 75% tethering combined with 55% abating. When tethering coverage is higher than 75%, and abating coverage is higher than 35%, a bottleneck is reached where increasing either coverage would not result in a significant reduction in dog infections. Thus, if interventions are performed uniformly across clusters, maintaining a minimum of 75% tethering and 35% abating is recommended.

Next, we examined various options of allocating one resource while keeping the other one uniform at a given resource capacity. We concluded that interventions are the most effective in the Central cluster, and the least in the East cluster. The reason is two-folded. On the one hand, the Central cluster has the highest number of infected dogs when no intervention is adopted. On the other hand, comparing with the other two, the Central cluster is the one that's most connected to the others: the highest number of dogs going in and out of the cluster for WS. Furthermore, it is shown that moving resource, tethering or abating, from the East cluster to the Central cluster, from the East cluster to the West or from the East cluster and the West cluster to the Central cluster, are all beneficial. However, diminishing marginal reduction in percentage of dogs infected is observed in all three directions. Details are included in the Appendix, Section **??**.

The optimal intervention strategy (Table **??**), adopts maximum tethering coverage in the Central, and maximum abating coverage in the West. Using this strategy, only 1.31% of dogs would be infected after five years of intervention, compared to 3.23% if interventions are performed uniformly while both require the same resource capacities. This strategy provides a similar level of infection control as 80% uniform tethering and 35% uniform abating. When allocation upper bounds are lowered, the optimal solutions are also more balanced (Figure **??**, Figure **??**).

We noted that the optimal solutions, under different resource capacities, allocate the majority of the tethering resource, 59% to 67%, to the Central cluster and most of the abating resource, 67% to 83%, to the West cluster. The reason is third folded. First, interventions are most effective in the Central cluster and least in the East cluster, since the East cluster has the lowest percentage of infected dogs. Second, tethering alone, without abating, isnt very effective unless we cover a cluster with a high percentage of dogs, while abating is effective even with low coverage in comparison.

Third, at the given capacity, there is enough resource to tether up to 100% of dogs in any cluster but only enough to abate up to 63% WS in any cluster. Thus, when tethering upper bounds is low, the optimal strategy assigns a high level of tethering and abating in Central to reach close-to-eradication in that cluster. When tethering upper bounds is high, the optimal strategy assigns an extremely high level of tethering in Central to reach close-to-eradication in that cluster while assigns a high level of abating in the West.

Besides intervention effectiveness, fairness is another focus. Decision fairness can be improved by either lowering the allocation upper bounds or adding penalty weight. As we expected, when we add a higher penalty weight, the optimal solution is more balanced. We can also maintain outcome fairness while improving the intervention outcome (Table **??**). Our algorithm enables decision-makers to find the optimal allocation strategy under an appropriate penalty weight and observe the trade-off between effectiveness and fairness.

Finally, we examined the optimal intervention strategies under a different resource capacity. When resource level is higher, optimal intervention strategies are both more effective and more balanced under the same allocation upper bounds. Our model, as suggested by [**?**], can be generalized to almost any population of definitive hosts with modest adjustments to the parameters and environmental factor distribution. Strategies for allocating intervention resources provided in this chapter can also be applied in the infection control of other hosts.

4.4. Limitations

In this work, we divided Chad into multiple clusters to capture the differences in the magnitude of GW infections and to examine the effective strategies of allocating tethering and abating resources among them. We assume there exists only one joined WS in each cluster, which would be far from reality. There exist numerous ponds spread in a cluster. Dividing our current clusters into even smaller clusters would create difficulties in parameter calibrations while significantly increase the workload of a simulation. Furthermore, when smaller ponds are located near each other, dogs would treat them similarly with the same traveling behavior. Thus, modeling multiple smaller ponds as an integrated larger one can be justified. The intervention strategy presented in this work is a generalization of the real-world but could still be adopted when considering different magnitudes of

clusters.

CHAPTER V

CONCLUSION

In Chapter 2, we considered resource allocation problems among multiple class customers in traffic and e-server setting. We presented Bayesian Optimization methods for optimal solution search to maximize system throughput in both settings. Our test results showed that Bayesian Optimization method is able to find solutions that are closed to true optimal while evaluating a much smaller set of alternatives comparing with exhaustive search and Cross-entropy Method. Furthermore, we are able to adopt this method in more complicated settings where system performance is stochastic, searching space is large, and secondary constraints are included.

In Chapter 3, we built an agent-based simulation with mosquito and human populations to model the spread of malaria in sub-Saharan Africa countries. Using this model, we concluded that ProCCM is very effective in reducing malaria incidence at the community level in this setting with moderate, highly seasonal transmission. Annual symptomatic malaria cases could be reduced by as much as 41.6% by conducting twice a week sweeps during the peak infection transmission season. Early start, long-term intervention, and high sweep frequency are critical factors in improving the ProCCM outcome, although these all increase costs. By varying weekly and biweekly sweeps during peak season (with higher frequency in mid-season), we proposed an alternative strategy that could reach a similar level of infection malaria incidence reduction control, while reducing the intervention cost by 33.3%.

In Chapter 4, we built a multi-water-source (MWS) agent-based simulation model to capture the Guinea Worm (GW) disease transmission among dogs in multiple clusters among Chad. We further examined strategies for allocating tethering and abating resources among three clusters in Chad. Using our MWS simulation model, we found the level of intervention coverages needed to reach nationwide eradication when interventions are performed uniformly is lower than the level suggested by [**?**] due to natural division among clusters. Next, we conducted sensitivity analysis

and found more effective allocation schemes while maintaining the same resource capacity. In particular, allocating more tethering resource in the Central cluster and more abating in the West cluster results in significantly better infection control. Finally, we presented a simulation optimization algorithm that enables the decision-maker to find the optimal intervention strategy under any resource capacity and allocation constraints. We examined the effectiveness of intervention strategies based on the percentage of dog population infected after years of intervention, and fairness of intervention strategies based on the maximum difference in intervention coverages between clusters.

APPENDIX A

MODELING THE IMPACT OF PROACTIVE COMMUNITY CASE MANAGEMENT ON REDUCING MALARIA TRANSMISSION IN SUB-SAHARAN AFRICAN COUNTRIES

A.1. Data

Table 18: Human characteristics data for the simulation model

Setting	Scenario I	Scenario II	Reference
Population	4747, within 15 villages	3762, within 14 villages	
Net Coverage	100%	98.40%	
Symptomatic Malaria Prevalence [%] (on July 8th, Sep 23rd and Nove 25th)	$[1.58, 2.97, 3.35]$	$[1.881.150.21]$	[**?**]
Treatment Seeking Rate	40%	37%	
Children under 5	15%	15%	Estimated
Immune Human	30%	30%	based on 20
Birth Rate	33.4 per 1000 ppl per year	33.4 per 1000 ppl per year	[**?**]
Death Rate	5.9 per 1000 ppl per year	5.9 per 1000 ppl per year	

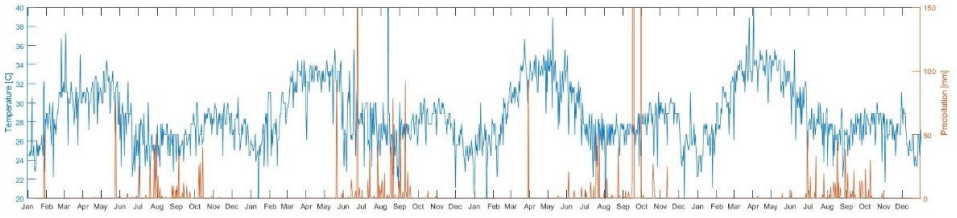

Figure 12: Temperature and precipitation data in Senegal from 2010 to 2013

A.2. Agent-based Simulation Model

A.2.1 Model of the Mosquito Population

To model the dynamics of the mosquito (Anopheles) population, we use the general framework proposed in [**?**], and [**?**]. Table **??** and Figure **??** depict the stages in the life cycle of Anopheles: aquatic stages (E, eggs; L, Larvae; P, pupae), emerging adult stage (A_{em}), nulliparous stages (A_{1h},A_{1g},A_{1o}) and parous stages (A_{2h},A_{2g},A_{2o}) [**?**]. Note that only the female mosquitoes are represented in the adult stage. The mosquito population dynamic model is based on a system of ordinary differential equations (Equation **??**). The parameters and functions are defined in Table **??** and Table **??**. The adult female mosquito population (N_v) is divided into three groups: clean mosquitos (S_v), incubating mosquitos (C_v), and infectious mosquitos (I_v), i.e., $N_v = S_v + C_v + I_v$.

Table 19: Mosquito life cycle stages

Stage	Description
E	Eggs
L	Larvae
A_{em}	Emerging adults
A_{1h}	Host-seeking nulliparous
A_{1g}	Nulliparous engorged
A_{10}	Nulliparous seeking oviposition sites
A_{2h}	Host-seeking parous
A_{2g}	Parous engorged
A_{20}	Parous seeking oviposition sites

The duration of each stage in the mosquito life cycle depends on several factors, such as temperature and water availability (precipitation). Temperature impacts the mortality and transition rates of larvae, pupae and adults, while precipitation impacts the environments carry capacity of aquatic stages, increasing the number of breeding sites available for Anopheles. For example, the climate in Senegal is tropical with high temperatures all year round and rainy season from May through November. The high transmission season starts in early July and end in late October [**?**]. We divide a year into favorable season and diapause period, where eggs stop hatching until the next favorable season when they hatch if they are immersed in water.

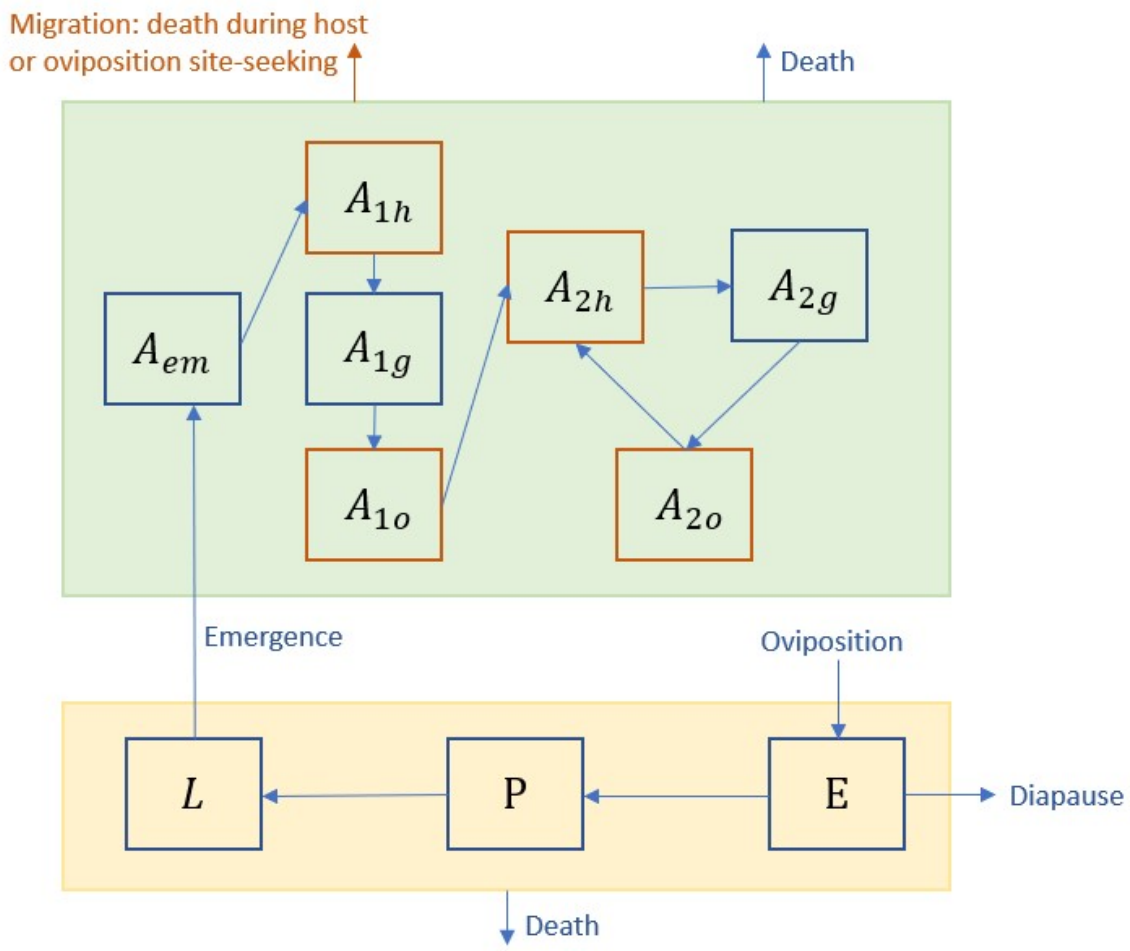

Figure 13: Mosquito population dynamic model

$$
\begin{aligned}
\dot{E} &= \gamma_{AO}(\beta_1 A_{1o} + \beta_2 A_{2o}) - (\mu_E + z f_E)E \\
\dot{L} &= (z f_E)E - [m_L(1 + \tfrac{L}{K_L}) + f_L]L \\
\dot{P} &= f_L L - [m_P + f_P]P \\
\dot{A_{em}} &= f_p P \sigma \exp[-\mu_e m(1 + \tfrac{P}{k_P})] - [m_A + \gamma_{A_{em}}]A_{em} \\
\dot{A_{1h}} &= \gamma_{A_{em}} A_{em} - (m_A + \mu_r + \gamma_{A_h})A_1 h \\
\dot{A_{1g}} &= \gamma_{A_h} A_{1h} - (m_A + f_{A_g})A_1 g \\
\dot{A_{1o}} &= \gamma_{A_g} A_{1g} - (m_A + \mu_r + \gamma_{A_o})A_1 o \\
\dot{A_{2h}} &= \gamma_{A_o}(A_{1o} + A_{2o}) - (m_A + \mu_r + \gamma_{A_h})A_2 h \\
\dot{A_{2g}} &= \gamma_{A_h} A_{2h} - (m_A + f_{A_g})A_2 g \\
\dot{A_{2o}} &= \gamma_{A_g} A_{2g} - (m_A + \mu_r + \gamma_{A_o})A_2 o
\end{aligned}
\tag{7}
$$

where $z = 0$ during diapause and 1 otherwise.

$T(t)$ and $P(t)$ represent the daily mean temperature in Celsius and precipitation in millimeters, respectively, on day t. $P_{norm}(t)$ is defined as the rainfall amount summed over a two weeks period and normalized afterwards.

Table 20: Parameters for mosquito population dynamic model

Parameters	Definition	Value	Reference
β_1	Number of eggs laid by ovipositing nulliparous females (per female)	100	
β_3	Number of eggs laid by ovipositing parous females (per female)	150	
κ_L	Standard environment carrying capacity for larvae (larvae ha^{-1})		
κ_L	Standard environment carrying capacity for purpae (purpae ha^{-1})		
σ	Sex-ratio at emergence	0.5	
μ_E	Egg mortality rate (per day)	0.05	
μ_L	Minimum larva mortality rate (per day)	0.08	
μ_P	Minimum pupa mortality rate (per day)	0.03	
μ_{em}	Mortality rate during adult emergence (per day)	0.1	[**?**], [**?**]
μ_A	Minimum adult mortality rate per day	0.02	
T_E	Minimal temperature needed for egg development (Celsius)	10.4	
TDD_E	Total number of degree-day necessary for egg development (Celsius)	110	
T_{A_g}	Minimal temperature needed for egg maturation (Celsius)	10	
TDD_{A_g}	Total number of degree-day necessary for egg maturation (Celsius)	77	
$\gamma_{A_{em}}$	Development rate of emerging adults (per day)	0.4	
γ_{A_h}	Transition rate from host-seeking to engorges adults s (per day)	0.2	
γ_{A_o}	Transition rate from oviposition site-seeking to host-seeking adults (per day)	0.2	
t_{start}	Start of favorable season	April 10th	Estimated
t_{end}	Start of favorable season	Oct 30th	Estimated

Table 21: Functions for mosquito population dynamic model

Function	Definition	Expression	Reference
f_E	Transition from egg to larva	$f_E(t) = \frac{T(t)-T_E}{TDD_E}$, if $T(t) > T_E$ $f_E(t) = 0$, if $T(t) \leq T_E$	
f_L	Transition from larva to pupa	$f_L(t) = -0.0007T(t)^2 + 0.00392T(t) - 0.3911$	
f_P	Transition from pupa to emerging adult	$f_P(t) = 0.0008T(t)^2 - 0.0051T(t) + 0.0319$	
f_{A_g}	Transition from emerging to host-seeking adult	$f_{A_g}(t) = \frac{T(t)-T_{A_g}}{TDD_{A_g}}$, if $T(t) > T_{A_g}$ $f_{A_g}(t) = 0$, if $T(t) \leq T_{A_g}$	[?], [?]
m_L	Larva mortality rate (per day)	$m_L(t) = \exp(-\frac{T(t)}{2}) + \mu_L$	
m_P	Pupa mortality rate (per day)	$m_P(t) = \exp(-\frac{T(t)}{2}) + \mu_P$	
m_A	Adult mortality rate (per day)	$m_A(t) = \max(u_A, 0.0441 + 0.00217T(t))$	
k_L	Environment carrying capacity of larvae (ha$^-$1)	$k_L(t) = \kappa_L(P_{norm}(t) + 1)$	
k_P	Environment carrying capacity of pupae (ha$^-$1)	$k_P(t) = \kappa_P(P_{norm}(t) + 1)$	
$B(T(t))$	Mosquito biting rate (per day)	$0.000203T(t)(T(t) - 11.7)\sqrt{42.3 - T(t)}$	

Based on the model proposed, we generate an adult host seeking mosquito $(A_{1h} + A_{2h})$ population for 4 years, shown in Figure **??**. The initial mosquito population contains $1,000,000$ eggs. The simulation includes a warm-up period of one year, which is not included in the plot.

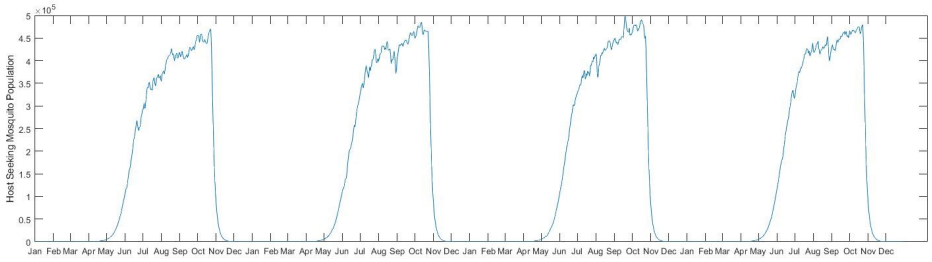

Figure 14: Host-seeking mosquito population in Saraya from 2010 to 2013 using simulation

A.2.2 Model Parameters for Human Infection

We first divide the human population into two groups: children under 5 (C) and adults (A). The transition processes between disease states follows the same patterns for both groups, but with different parameters. At a given time, a human is in one of the following states based on his/her infectious condition: clean non-immune (S_h), incubating non-immune (C_h), infectious non-immune (I_h), clean immune (SR_h), incubating immune (CR_h), infectious immune (IR_h), for $h \in \{A, C\}$. A human infection begins with a successful contact by an infectious mosquito. In an infected human body, the malaria parasite undergoes a pre-erythrocytic liver stage first, which typically lasts for one to two weeks, before the onset of the blood stage. During the blood stage, the sexual form of the malaria parasite, the gametocyte, is produced and thus the infected human becomes infectious to mosquitoes [**?**]. The duration from a human being infected by a mosquito to him/her being infectious to mosquitoes is defined as the incubation period.

After a successful infection by a mosquito, the infected human also goes through (some of) the following phases, based on the severity of their symptoms [**?**]:

$$\begin{aligned} \text{Infection} \rightarrow \text{Asymptomatic Parasitemia}(A_h) \quad &\rightarrow \text{Uncomplicated Illness}(M_h^m) \\ &\rightarrow \text{Severe Malaria}(M_h^S) \rightarrow \text{Death}. \end{aligned} \tag{8}$$

An infected human first goes through asymptomatic parasitemia phase (A_h), which lasts between

6 to 14 days [**?**], in which there are no symptoms. After this, he/she begins to show mild (uncomplicated) illness symptoms, including fever, chills, headaches, diaphoresis, etc. If not actively seeking treatment, a human in the uncomplicated illness phase could go into the severe malaria phase after 5 to 7 days for adults and 1 to 2 days for children under 5, as the parasite density accumulates within his/her body [**?**]. Since the duration of the incubation state doesnt necessarily coincide with the asymptomatic parasitemia phase, an infected human can become infectious either at the end of the asymptomatic parasitemia phase or at the beginning of the uncomplicated illness phase. We further differentiate the level of human infectiousness depending on where a human is among the phases of the infected state. As gametocytes in P. falciparum infections arise from asexual parasites (i.e., merozoites), there could be a positive correlation between the density of asexual and sexual parasite (gametocytes) [**?**]. In nonimmune individuals, hyperparasitemia ($> 5\%$ parasitemia or > 250000 parasites $/ul$) is generally associated with severe disease [**?**]. Research also suggests that the level of infectiousness of a human is a concave increasing function of female gametocyte density [**?**]. Thus, we assume that the human level of infectiousness, i_H, linearly increases from 0 to 1 during the uncomplicated illness phase, i.e.,

$$i_H = \frac{t}{T_{in}}$$

, where t is how long an infected human has been in the uncomplicated illness phase, and T_{in} is the total duration of the uncomplicated illness phase. The humans infectiousness level is highest, i.e., $i_H = 1$, during the severe malaria phase.

Table 22: State parameters for human infection dynamic model

State	Definition	Duration[day]	Reference
T_1	Infected non-immune human $\rightarrow$ Non-immune human with mild symptoms	$6-14$	[?]
T_2	Non-immune human with mild symptoms $\rightarrow$ Non-immune human with severe symptoms	$A: 5-7$ $C: 1-2$	[?]
T_3	Non-immune human with mild symptoms $\rightarrow$ (Recovery) Clean non-immune/immune human	Mean:6.6	[?]
T_4	Non-immune human with severe symptoms $\rightarrow$ (Recovery) Clean non-immune/immune human	$7-35$	[?], [?]
T_5	Infected immune human $\rightarrow$ Immune human with mild symptoms	$7-30$	
T_6	Infected immune human $\rightarrow$ (Recovery) Clean non-immune/immune human	$14-120$	
T_7	Immune human with mild symptoms $\rightarrow$ (Recovery) Clean non-immune/immune human	$7-90$	[?]
T_i	Infected human $\rightarrow$ Infectious human	$7-15$	[?]

Table 23: Transition parameters for human infection dynamic model

Arc	Definition	Parameters
$S_h \rightarrow A_h$	Non-immune human infected	$c_v^n, c_h^n, B(T)$
$SR_h AR_h$	Immune human infected	$c_v^i, c_h^i, B(T)$
$A_h \rightarrow M_h^s$	Non-immune human showing symptoms after incubation	T_1
$AR_h \rightarrow MR_h$	Immune human showing symptoms after incubation	T_5
$M_h^s \rightarrow M_h^m$	Non-immune individual with mild malaria progress to severe malaria	T_3
$M_h^s \rightarrow D_h$	Non-immune individual with mild malaria deceased	P_d^m
$M_h^m \rightarrow D_h$	Non-immune individual with severe malaria deceased	P_d^S
$AR_h \rightarrow SR_h$	Immune human recovery without showing symptoms	T_7, P_{sr}^c
$M_h^s \rightarrow S_h$	Non-immune human recovery with mild condition without gaining immunity	T_4, P_{sc}, P_{st}^m
$M_h^m \rightarrow S_h$	Non-immune human recovery with severe condition without gaining immunity	T_5, P_{sc}, P_{st}^s
$M_h^s \rightarrow SR_h$	Non-immune human recovery with mild condition while gaining immunity	$T_4, P_{sc}, P_{st}^m, P_{gi}$
$M_h^m \rightarrow SR_h$	Non-immune human recovery with severe condition while gaining immunity	$T_5, P_{sc}, P_{st}^s, P_{gi}$
$MR_h \rightarrow SR_h$	Immune human with mild condition recovery without losing immunity	T_6, P_s, P_{li}
$MR_h \rightarrow S_h$	Immune human with mild condition recovery while losing immunity	T_6, P_s, P_{li}
$_h$	Natural birth rate	$33.4/1000$
f_h	Death rate of natural causes (uniform in all stages)	$5.9/1000$

Table 24: Transition parameters for human infection dynamic model

Parameters	Definition	Value	Reference
c_h^n	Infected non-immune human to mosquito transmission efficiency	0.5	[**?**]
c_h^i	Infected immune human to mosquito transmission efficiency	0.25	Estimated
c_v^n	Infected mosquito to clean non-immune human transmission efficiency	0.5	[**?**]
c_v^i	Infected mosquito to clean immune human transmission efficiency	0.25	Estimated
P_r^m	Probability of self-recovery from mild symptom [adult, child]	$[0.2, 0.1]$	[**?**]
P_r^S	Probability of self-recovery from severe symptom [adult, child]	$[0.01, 0]$	[**?**]
P_{st}^m	Probability of seeking treatment at mild condition	0.50	Estimated based on [**?**]
P_{st}^s	probability of seeking treatment at severe condition	0.60	[**?**]
P_d^m	Mortality at mild infectious stage [adult, child]	$[0.01, 0.04]$	[**?**]
P_d	Base mortality at severe infectious stage [adult, child]	$[0.10, 0.20]$	
P_d^S	Mortality at severe infectious stage with time	linearly increase from P_d to 1	[**?**], [**?**]
P_{gi}	Probability of gaining immunity after recovery	0.3	
P_{li}	Probability of losing immunity after recovery	0.05	
P_{tf}	Probability of treatment failure	0.05	[**?**]
P_{sc}	ProCCM sweep coverage	$0.80 - 1.00$	Estimated base on [**?**]
P_s	Probability of immune infectious human showing symptoms	0.1	[**?**]

A.2.3 Interaction between Human and Mosquito Populations

A clean non-immune human is susceptible to malaria and could be infected (with probability I_v) after a bite from an infectious mosquito. We assume that every human has equal chance of being contacted by any mosquito [**?**] and children under 5 years old do not have immunity and thus, always show symptoms if infected ($P_{gi} = 0$) [**?**]. We adopt the idea of effective contact ratio that determines daily numbers of newly infected humans (NI_h) and newly infected mosquitos (NI_v) [**?**] and incorporate the individual infectiousness level, $i_H(t) \in [0,1]$. The adjusted effective contact ratio equations are:

$$\begin{aligned} NI_h &= (\tfrac{C_v B(T) I_v}{N_h})(P_{S_h} + P_{SR_h}) \\ NI_v &= \textstyle\sum_{inv \in P_{I_h}} (\tfrac{C_h B(T) i_{inv}}{N_h}) S_v + \sum_{inv \in P_{IR_h}} (\tfrac{C_h B(T) i_{inv}}{N_h}) S_v \end{aligned} \tag{9}$$

where c_v,c_h are the successful biting rates, $B(T)$ is the mosquito biting rate determined by mean daily temperature, N_h is the total number of humans in certain age group, i_{inv} is the individual infectiousness level and P_x is the total number of humans in a certain infectious condition, e.g. P_{I_A} is the total number of infectious non-immune adults.

A.2.4 Treatment and Residual Parasitemia

A human with uncomplicated malaria or severe malaria has the possibilities of self-recovering, e.g., self-treatment, seeing a traditional healer, taking herbal medication [**?**] or simply recovering as the immune system responds, although the chance is small in the case of severe malaria. If a symptomatic human seeks treatment, at any stage after developing symptoms, we assume that he/she will receive artemisinin-based combination treatments (ACTs) for 3 days [**?**], and the recovery process starts immediately after the first day [**?**]. Thus, the human infectiousness level would stop increasing as soon as the treatment begins. Possibilities of treatment failure (P_{tf}) and mortality depends on the how long one has been infected as well as the severity of his/her symptoms. After the completion of treatment, some humans (50%) would be parasite free and the others would have residual parasitemia [**?**]. Humans with residual parasitemia have higher gametocyte density after treatments, and longer gametocyte carriage durations compared with those who are parasite free after ACTs [**?**]. In this model we assume that, after ACTs, a human preserves 50% to 75% of the infectiousness before

treatment, and this level linearly decreases until the end of gametocyte carriage duration.

A.2.5 Acquired Immunity against Malaria

After recovery, a non-immune human can gain immunity with a certain probability (P_{gi}) [**?**]. An immune human will become infected after a successful bite from an infected mosquito with a smaller chance [**?**]. After being infected, the immune human will then go through the asymptomatic parasitemia phase with a longer duration [**?**]. It is shown that with immunity, parasite density within an infected human body would be lower. Thus, an immune human with asymptomatic parasitemia will have a much smaller possibility of showing symptoms, move to the uncomplicated malaria phase, and will be extremely unlikely to eventually progress into the severe malaria phase [**?**]. The majority of the immune humans with asymptomatic parasitemia will self-recover. Some who develop symptoms after the asymptomatic parasitemia phase may recover if they seek treatments. Losing immunity after recovery (P_{ii}) is also very unlikely under consistent heavy malaria exposure [**?**]. Newborns begin their lives as clean non-immune humans. All humans are susceptible to natural death at a fixed rate. Traditionally, an infected human would only actively seek treatment (P_{st}) if he/she developed symptoms. Treatment-seeking possibilities are different between humans in the adult group and children. Those who dont seek treatment will keep infecting clean mosquitos. ProCCM could actively detect infected humans with symptoms, give them treatments and terminate the infection loop. All the states are listed in Table **??**. The duration range for each state is adopted from corresponding literature. We assume that all distributions are uniform within the given ranges. The transition diagram for our human infection model is shown in Figure **??**. The parameters used in this simulation model are listed in Table 7 and Table 8.

A.3. Model Validation and Sensitivity Analysis

In [**?**], $1,669$ symptomatic infection cases are documented in scenario II from July to November, where 89% of the weekly sweeps are conducted and the total population is $3,762$. With all sweeps conducted, the estimated range for symptomatic infection cases is $[1,669, 1,750]$. Adjusting this number to scenario I with a higher population, the estimated range for symptomatic infection cases is $[2,105, 2,208]$. Our simulation results show that about 90% of the infection cases occurred during peak season. Considering this, the estimated range for year-round infection cases is $[2338, 2,454]$.

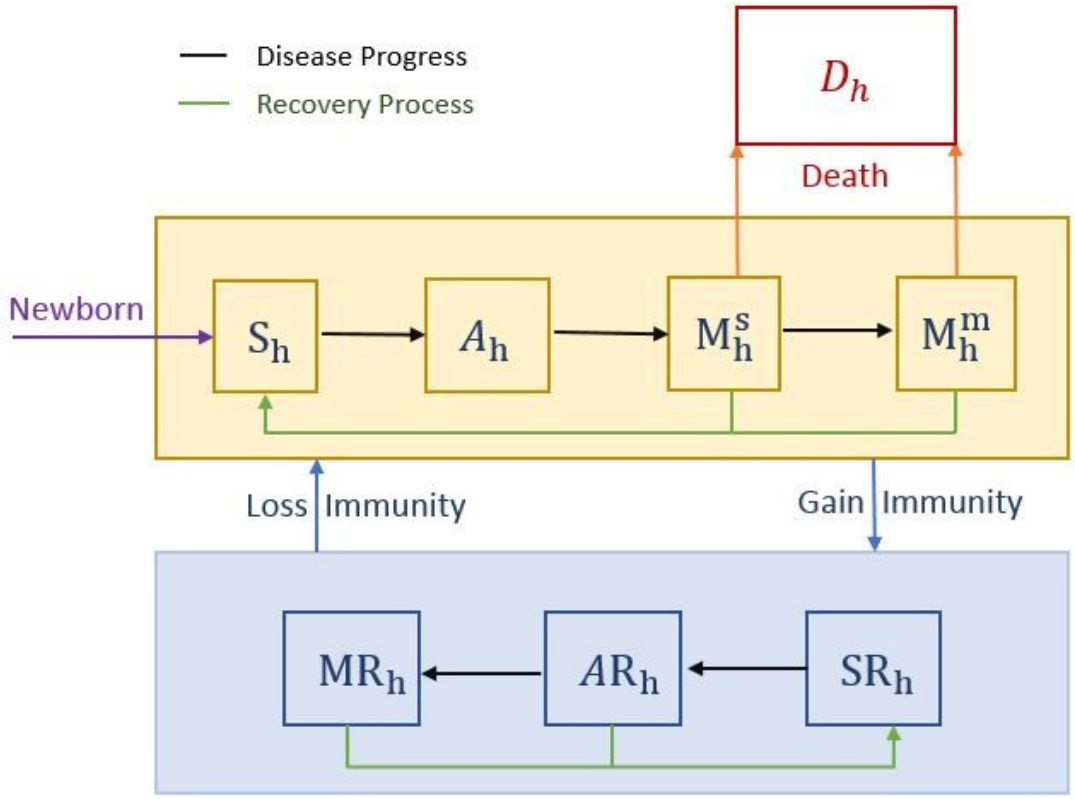

Figure 15: Human infection transition graph

The symptomatic infection cases reported in Table **??**(2,496) lies on the edge of this range. Moreover, the total number of positive RDT tests given during sweeps reported in Linns work is 647. The estimated number of positive RDP tests in scenario I after adjustment is 1005, which lies in the 95% confident interval of our results. Thus, our model is validated. To test the robustness of model results, we conducted sensitivity analysis on ProCCM sweep coverage and simulation year. We adjusted the coverage from 80% to 95% while adopting weekly sweep strategy (C) and alternative strategy (I); the results are shown in Table 10 and Table 11 in the Appendix **??**. Since the weekly sweep strategy (strategy C in Table **??**) provides higher frequency interventions, the results (number of malaria infection cases) are more robust to coverage change. Decreasing coverage from 100% to 80% would result in 1.4% more infection cases per year. When adopting a strategy with lower intervention frequency (strategy I in Table **??**), 3.1% more infection cases result per year. We repeated the simulation from 2010 to 2013 when adopting the following strategies: i) no sweeps

conducted, ii) weekly sweeps conducted during transmission season and iii) alternative strategy (I) conducted. We compared the total number of symptomatic malaria infection cases per year in Table 9 in the Appendix **??**. The results show consistency from year to year: i) the total number of symptomatic infection cases per year is around 4000 in sweep strategy A (i.e., scenario I); ii) conducting weekly sweeps could reduce 36% to 42% of year-round infection cases; iii) conducting alternative strategy (I) could reduce 32% to 36% of year-round infection cases, while reducing 33% of the implementation cost.

A.4. Experimental Results for Malaria Intervention Simulations.

Table 25: List of simulation results for all test cases in year 2013

Test Strategy [*No.* sweeps]	Total Symptomatic Infection [incidences per 1000 population]	STD	Peak Symptomatic Prevalence [max infection cases /population]	STD	RDT	STD
A. No Sweeps [0]	0.846	0.0002	0.0514	0.001	0	0
B. 3 Sweeps, 100% coverage [3]	0.738	0.0002	0.0487	0.001	262	6
C. Weekly Sweeps, 100% coverage of symptomatic individuals [21]	0.526	0.0002	0.0190	0.001	1005	31
D. Weekly Sweeps, 50% coverage [21]	0.574	0.0002	0.0249	0.001	842	22
E. Bi-weekly Sweeps, 100% coverage [11]	0.578	0.0002	0.0284	0.001	825	20
F. Twice a week, 100% coverage[41]	0.494	0.0001	0.0159	0.001	1232	37
G.1. Weekly Sweeps for first 6 weeks, 100% coverage [7]	0.665	0.0002	0.0301	0.001	461	14
H.1. Bi-Weekly Sweeps for first 12 weeks, 100% coverage [7]	0.588	0.0001	0.0289	0.001	559	17
G.7. Weekly Sweeps for 6 weeks during mid-season, 100% coverage [7]	0.696	0.0002	0.0350	0.001	493	13
I. Weekly sweep from week 5 to week 11, biweekly sweep at week 1, 3, 12, 14, 16, 18,20.[14]	0.557	0.0002	0.0270	0.001	929	23
J. Weekly sweep from week 5 to week 11, biweekly sweep at week 3, 12, 14, 16. [11]	0.579	0.0002	0.0280	0.001	889	23

Table 26: Sensitivity analysis for 21 weekly sweeps with different coverage ($C*$)

Coverage	Total Symptomatic Infection [incidences per 1000 population]	Peak Symptomatic Prevalence [max infection cases /population]	RDT
1	0.526	0.019	1005
0.95	0.527	0.020	987
0.9	0.529	0.021	969
0.85	0.530	0.023	935
0.8	0.532	0.024	920

Table 27: Sensitivity analysis for alternative sweeping strategy with different coverage ($I*$)

Coverage	Total Symptomatic Infection [incidences per 1000 population]	Peak Symptomatic Prevalence [max infection cases /population]	RDT
1	0.557	0.027	929
0.95	0.557	0.028	897
0.9	0.564	0.032	854
0.85	0.570	0.033	821
0.8	0.578	0.035	801

Table 28: Total Infections for Strategies A, C, and I in multiple years.

Year	Total Symptomatic Infection (A) [incidences per 1000 population]	Total Symptomatic Infection (C) [incidences per 1000 population]	Total Symptomatic Infection (I) [incidences per 1000 population]
2010	0.830	0.520	0.549
2011	0.853	0.523	0.553
2012	0.858	0.521	0.555
2013	0.846	0.526	0.557

Table 29: Simulation results for 7 consecutive weekly sweeps with different starting dates

Strategy	Sweep Duration	Total Symptomatic Infection [incidences per 1000 population]	STD	Peak Symptomatic Prevalence [max infection cases /population]	STD	RDT	STD
$G.1$	$07/08-08/19[7]$	0.665	0.0001	0.034	0.0011	461	15
$G.2$	$07/15-08/26[7]$	0.659	0.0001	0.032	0.001	454	14
$G.3$	$07/22-09/02[7]$	0.657	0.0002	0.033	0.001	459	14
$G.4$	$07/29-09/09[7]$	0.656	0.0001	0.032	0.0009	458	13
$G.5$	$08/05-09/16[7]$	0.656	0.0002	0.031	0.0011	463	14
$G.6$	$08/12-09/23[7]$	0.669	0.0002	0.033	0.0011	470	14
$G.7$	$08/19-09/30[7]$	0.696	0.0002	0.035	0.0012	493	15
$G.8$	$08/26-10/07[7]$	0.701	0.0001	0.037	0.001	545	14
$G.9$	$09/02-10/14[7]$	0.73	0.0002	0.036	0.001	576	16
$G.10$	$09/09-10/21[7]$	0.756	0.0002	0.038	0.0011	645	18
$G.11$	$09/16-10/28[7]$	0.77	0.0002	0.042	0.0011	664	15
$G.12$	$09/23-11/04[7]$	0.776	0.0002	0.043	0.0012	683	16
$G.13$	$09/30-11/11[7]$	0.788	0.0002	0.049	0.0015	693	17

Table 30: Simulation results for 7 consecutive biweekly sweeps with different starting dates

Strategy	Sweep Duration	Total Symptomatic Infection [incidences per 1000 population]	STD	Peak Symptomatic Prevalence [max infection cases /population]	STD	RDT	STD
$H.1$	$07/08-09/30[7]$	0.588	0.0001	0.029	0.0010	559	17
$H.2$	$07/15-10/07[7]$	0.601	0.0001	0.031	0.0011	584	18
$H.3$	$07/22-10/14[7]$	0.628	0.0001	0.032	0.0008	639	19
$H.4$	$07/29-10/21[7]$	0.635	0.0001	0.032	0.0009	608	17
$H.5$	$08/05-10/28[7]$	0.650	0.0001	0.036	0.0011	637	18
$H.6$	$08/12-11/04[7]$	0.678	0.0001	0.037	0.0012	639	15
$H.7$	$08/19-11/11[7]$	0.692	0.0001	0.038	0.0010	663	17

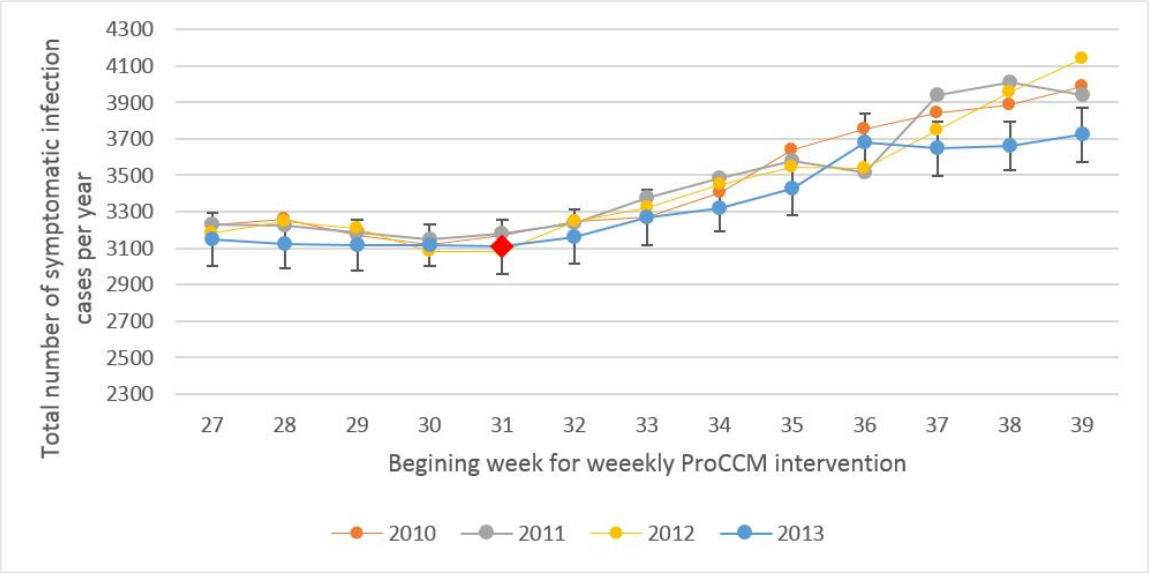

Figure 16: Sum of symptomatic infection cases per year for 7 consecutive weekly sweeps with different starting dates

ProCCM sweeps started on July 8th and ended on Nov 25th of 2013 in the pilot study. To find the optimal starting date of sweeps, we first proposed 13 sweep strategies, all of which are composed of 7 consecutive weekly strategies and each of them starting on a different week during peak season. We only selected 7 out of 21 sweeps in this comparison set to reduce the cost of intervention. By exhaustive search, we identified the optimal strategy with the current constraints: strategy $G.5$ which starts on the 31st week and lasts until the 37th week. With 29.2% reduction in infection cases, its still not comparable to having weekly sweeps, but two-thirds of the cost is reduced. We repeated the analysis on years 2010, 2011 and 2012. In all three cases, the optimal starting weeks for 7 consecutive weekly sweeps are week 30 and 31, corresponding to strategy $G.4$ and $G.5$. Other options like $G.1$ postponed the rapid growth of malaria prevalence in both populations, yet the duration of intervention is not long enough to consolidate its effect over the season. Strategies $G.7$ to $G.13$ are shown to be ineffective compared to strategy B, where only three sweeps are adopted. Since the peak of malaria prevalence is reached at the mid-late peak season, starting intervention during mid-peak season is not effective when the malaria prevalence within the mosquito population is already very high. Early intervention is needed for effective malaria control. Next, we examined the effect of having a longer intervention period by using biweekly

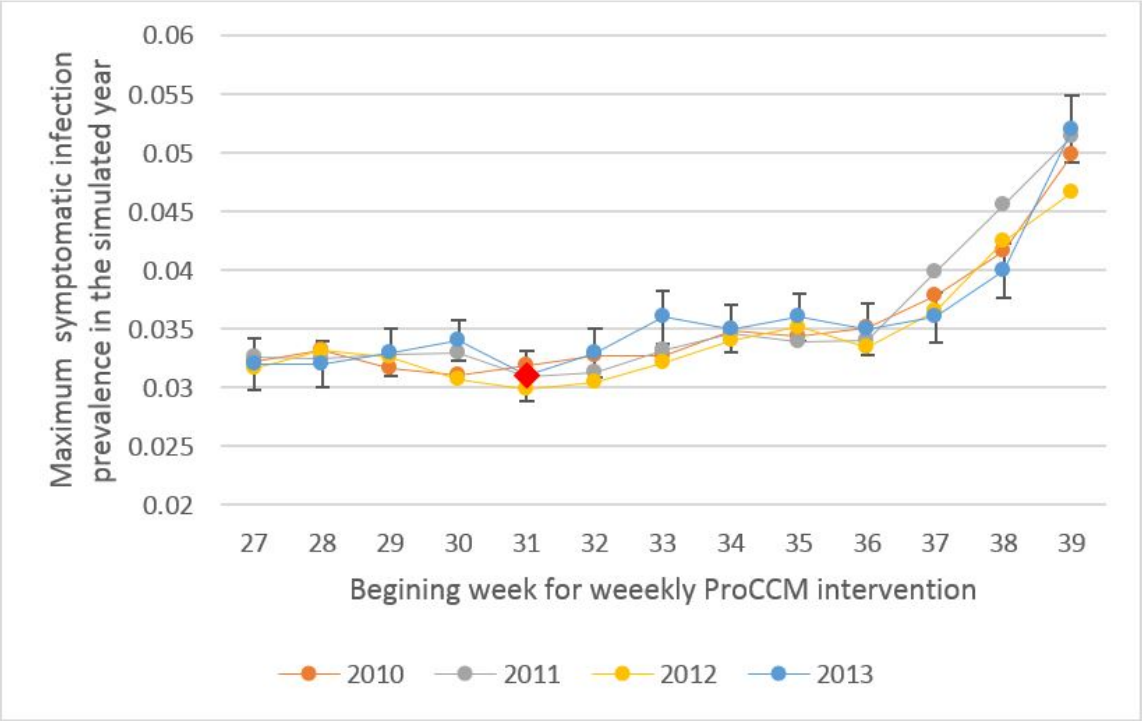

Figure 17: Maximum infection prevalence in 2013 for 7 consecutive weekly sweeps with different starting dates

sweeps. Strategies E conducted 11 biweekly sweeps converging the same 21-week period starting on July 7th. The biweekly strategy could reduce the cost by half while still follow the requirement of an early start and long duration. Alternatively, Strategy *D* conducted 21 weekly sweeps but each time only having 50% of the population coverage instead of 100%. Reduction in malaria infections in both cases are shown to be similar, with strategy D slightly better. In the later setting, there are infectious humans who got detected earlier than a regular two-week period due to randomness. Theoretically having 50% coverage each sweep could also reduce the cost by half, yet the idea of randomly selecting 50% of the symptomatic humans to test is hard to implement. Covering half of the villages on odd weeks and the other half on even weeks in more likely to be carried out in practice. Without the randomness in the selection, the outcome is going to be different. For most of the sub-Sahara African countries, funding for malaria intervention is a major barrier. Thus, we aim at exploring a better option thats more cost-efficient while providing promising results in infection control. Assuming each sweep has the same coverage, with the same amount of homecare providers, the average number of malaria infection cases identified per sweep would be an essential indicator

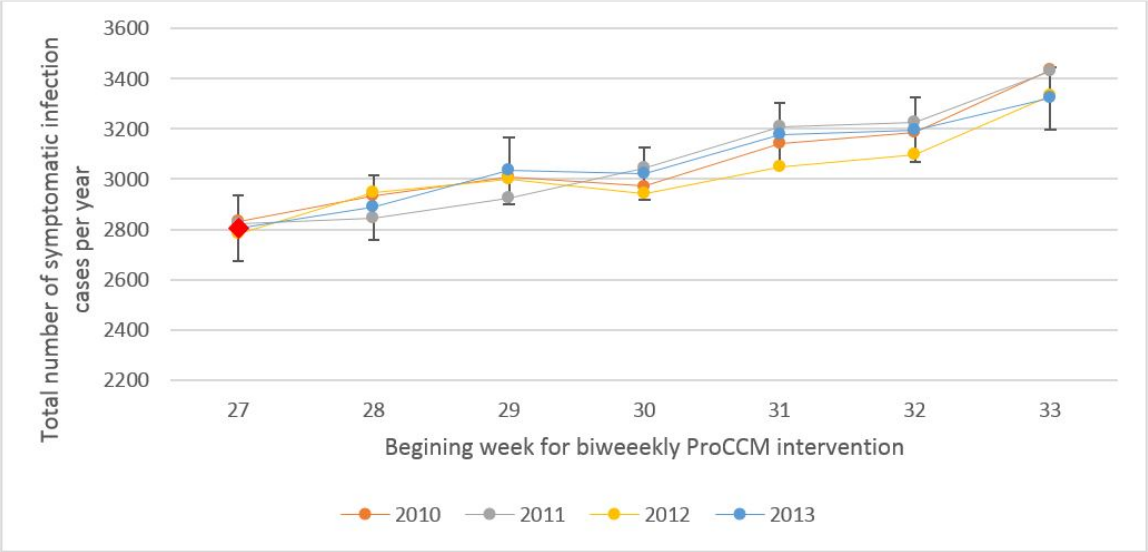

Figure 18: Sum of symptomatic infection cases per year for 7 consecutive bi-weekly sweeps with different starting dates

of the cost efficiency for ProCCM strategies. Obviously, to not implement ProCCM at all would not induce any cost. Within the low-cost setting, we examined the optimal starting date for 7 consecutive biweekly sweeps on years $2010, 2011, 2012$ and 2013. In all four cases, the optimal starting week for 7 consecutive biweekly sweeps is at the beginning of peak season on week 27, corresponding to strategy $H.1$. While having the same amount of sweeps, the outcome of the optimal biweekly strategy, $H.1$, is better than the optimal weekly strategy, $G.5$. Less frequent interventions over a longer duration is preferred over higher frequency over a shorter duration.

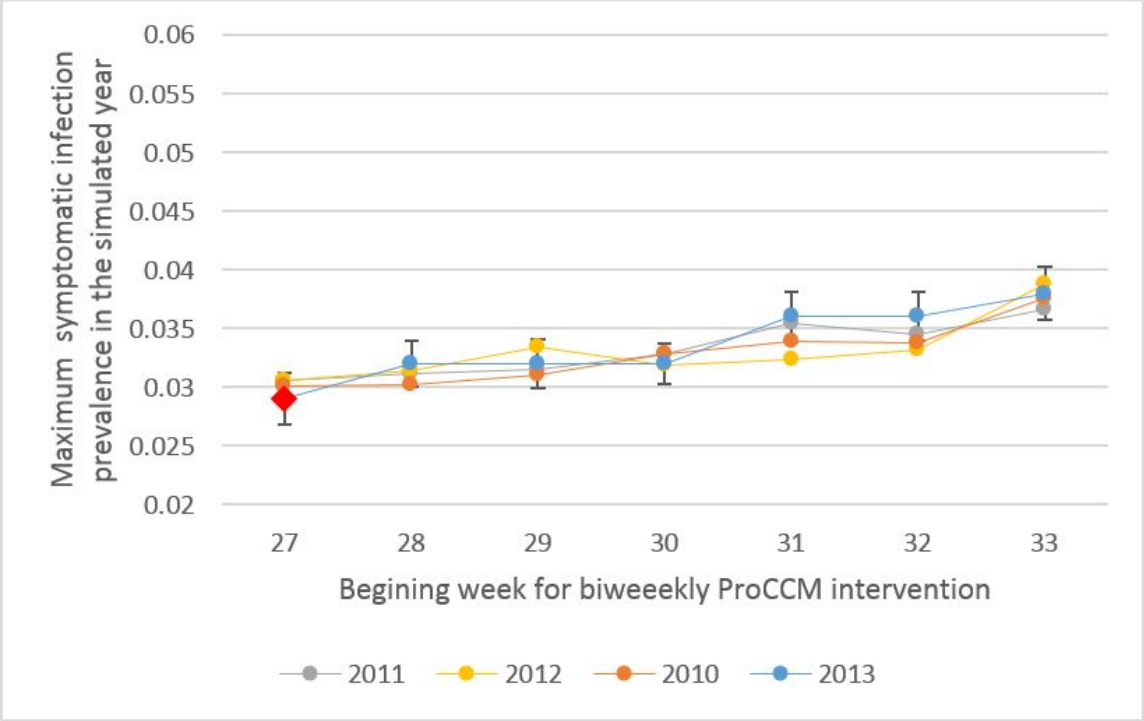

Figure 19: Maximum infection prevalence in 2013 for 7 consecutive bi-weekly sweeps with different starting dates

APPENDIX B

MODELING THE TRANSMISSION OF GUINEA WORM DISEASE IN DOGS AMONG MULTIPLE WATER SOURCES AND EVALUATING THE EFFECTIVENESS OF INTERVENTION METHODS IN CHAD.

B.1. Regional Clustering

Our data documented GW infections within 5 regions, 19 districts, 88 zones, and 1674 villages of Chad. We considered district as the smallest units for clustering to balance the level of complexity and data quality (of GW infection in each unit). We used the relative worm emergence per month, latitude, longitude, position along river and elevation of each district center as inputs for K-means to generate clustering result. We compare the results from K-means ($k = 4$) with the 5 administrative regions in Chad (Figure **??**).

Since the distances from SLM region to any other regions, and to the river, is more than 40 km, we consider the districts within SLM region to be isolated from the others. Additionally, we removed the third cluster, shown in Figure 1, due to insufficient data. Our data documented GW infections within 5 regions, 19 districts, 88 zones, and 1674 villages of Chad. We considered district as the smallest units for clustering to balance the level of complexity and data quality (of GW infection in each unit). We used the relative worm emergence per month, latitude, longitude, position along river and elevation of each district center as inputs for K-means to generate clustering result. We compare the results from K-means ($k = 4$) with the 5 administrative regions in Chad (Figure **??**).

Since the distances from SLM region to any other regions, and to the river, is more than 40 km, we consider the districts within SLM region to be isolated from the others. The number of exuding dogs from 2016 to 2018 within each cluster are plotted in Figure **??**. Mois Cluster differs from the other three in both infection seasonality and infection magnitude. Additionally, there are 13 months within the three-year-period where no exuding dogs are documented in this cluster. Mois cluster is too small to be considered as an individual cluster and its model is hard to validate due to insufficient

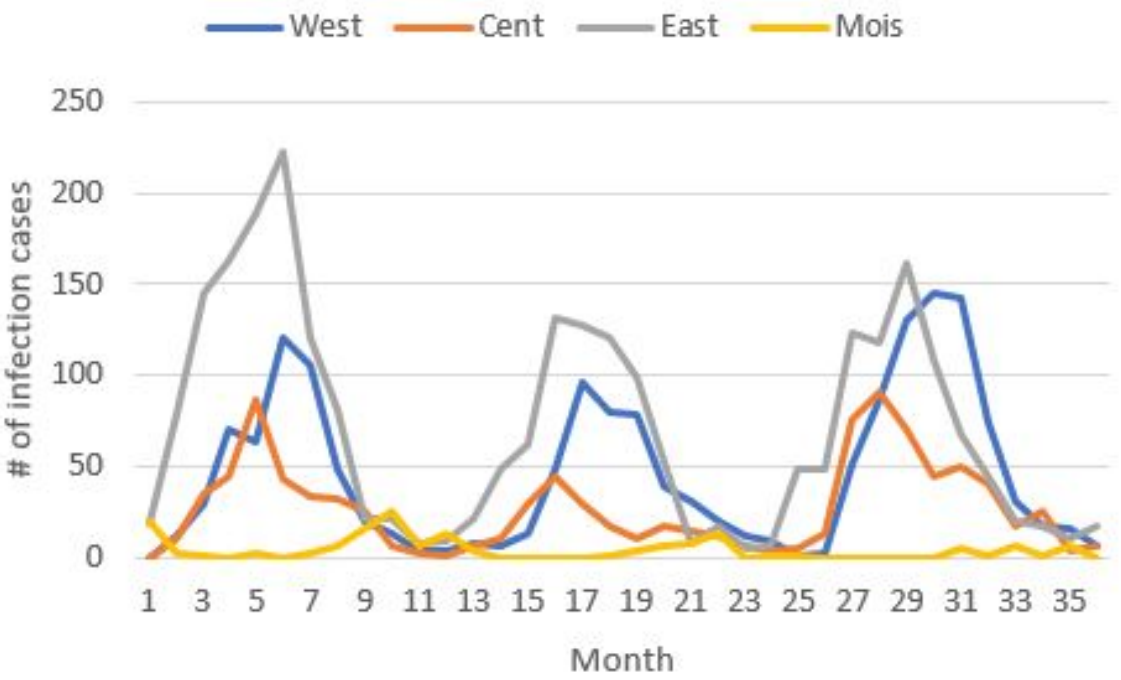

Figure 20: Regional infection data.

data. Thus, we removed this cluster from our model.

B.2. Parameter Calibration

We first integrate the number of worms emerging from dogs per month and the number of dogs with an emerging worm per month for the years 2016-2018 for each cluster (Table **??**). Then, we collected the percentage of tethering conducted for each region per year (Table **??**). Note that starting from 2017, tethering status for a certain number of infected dogs are not documented, thus we calculated ranges for the actual tethering coverage.

Table 31: Tethering coverage for each cluster in Chad

Cluster (Figure**??**)	Tethering Coverage [2016]	Tethering Coverage [2017]	Tethering Coverage [2018]
1 West Chari	0.63	$[0.43, 0.83]$	$[0.43, 0.88]$
2 East Chari	0.72	$[0.40, 0.86]$	$[0.32, 0.95]$
4 Central Chari	0.74	$[0.62, 0.90]$	$[0.49, 0.88]$

We first calibrate parameters for a shared infectivity curve among all clusters. Even with the

best set of parameters, the simulation model failed to capture the difference in GW infection cases between clusters. Thus, for each cluster, we calibrate parameters for infectivity curve separately. We calibrate the model parameters by minimizing mean square error (MSE) of the exuding dogs. While empirical data is more likely to be underreported than overreported for both the number of exuding dogs and GWs, the former one is more accurate. We came up with three scenarios for each cluster, where the tethering coverage of 2017 is the minimum, mean and the maximum of the documented range. For each scenario, we calibrate parameters for infectivity curves. The simulation is initialized with data from 2016 (number of infected dogs and number of emerging worms), and the remaining data from 2017-2018 are used for calibration. We take 20 replications for each simulation scenarios. Parameter searches are done using Cross Entropy method. Final calibration results are shown from Figure **??** to Figure fig:GW8.

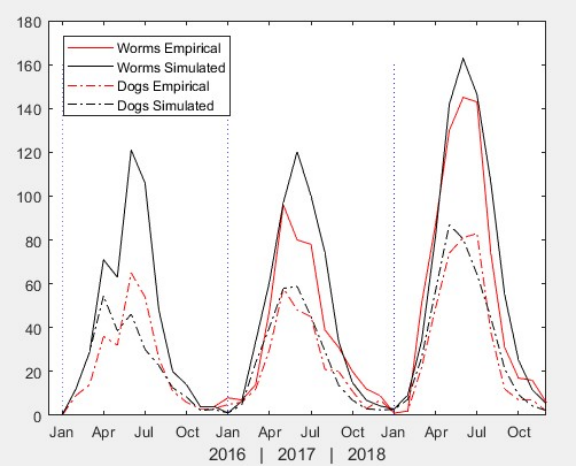

Figure 21: GW infection simulation for the West Cluster.

B.3. Dogs' Traveling Behavior

McDonalds research on the ecology of Guinea Worm infection in dogs suggested that 80% of dogs visit ponds that are within the 100-meter-range from the dog owners house [**?**]. Additionally, a dogs the mean travel range is 4.4 square kilometers in Chad, much higher in dry season than in wet season [**?**]. Based on the geographical data, we assigned groups to dogs within each cluster. The interactions between clusters are:

1 Dogs do not use WS from the other clusters during the wet season (from June to October).

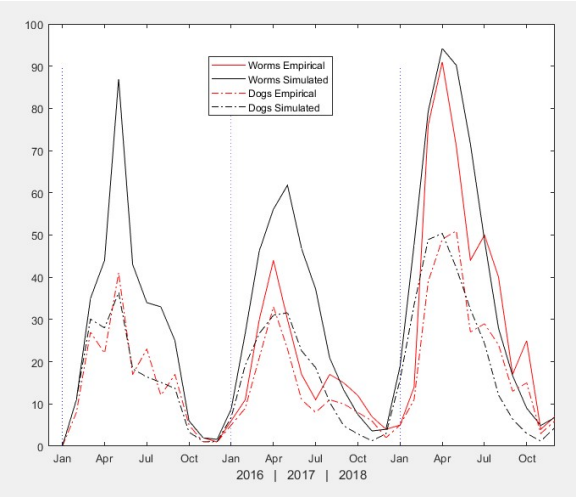

Figure 22: GW infection simulation for the Central Cluster.

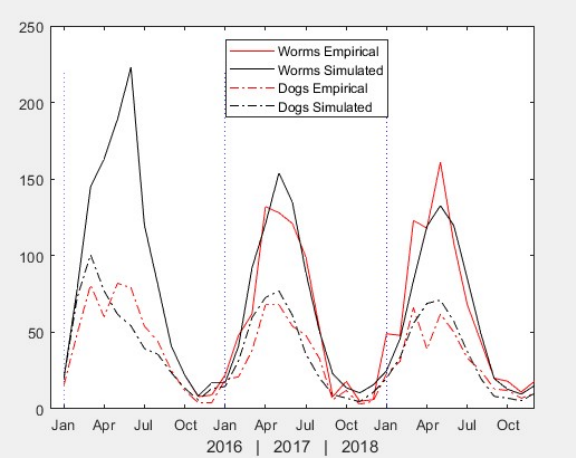

Figure 23: GW infection simulation for the East Cluster.

2 Dog group I (40%) in the West Cluster uses the Central WS as their alternative with a probability of 20% during the dry season, group II (60%) uses the West WS with 100% probability.

3 Dog group I (10%) in the East Cluster use the Central WS as their alternative with a probability of 20% during the dry season, group II (90%) uses the East WS with 100% probability.

4 Dog group I (40%) in the Central Cluster uses the West WS as their alternative with a probability of 20% during the dry season. Dog group III (10%) of dogs uses the East WS as their alternative with a probability of 20% during the dry season. Dog group II (50%) uses the Central WS with 100% probability.

5 Dogs do not switch groups.

B.4. Simulation Optimization Model

Table 32: Notations for simulation optimization model

Notation	Meaning
$T_w \in [0,1]$	**[Decision Variable]** Tethering coverage in the West cluster.
$T_c \in [0,1]$	**[Decision Variable]** Tethering coverage in the Central cluster.
$T_e \in [0,1]$	**[Decision Variable]** Tethering coverage in the East cluster.
$A_w \in [0,1]$	**[Decision Variable]** Abating coverage in the West cluster.
$A_c \in [0,1]$	**[Decision Variable]** Abating coverage in the Central cluster.
$A_e \in [0,1]$	**[Decision Variable]** Abating coverage in the East cluster.
$\bar{T}$	Tethering strategy $\bar{T} = [T_w, T_c, T_e]$.
$\bar{A}$	Abating strategy $\bar{A} = [A_w, A_c, A_e]$.
$\bar{T_U}$	Uniform tethering strategy.
$\bar{A_U}$	Uniform abating strategy.
$Inf_i(\bar{T}, \bar{A})$	Total number of dog infections in cluster i, with intervention strategy $\bar{T}, \bar{A}$.
P_i	Dog population in cluster i.
T_{max},A_{max}	Allocation upper bound for tethering/abating.
C_t, C_A	Resource capacity for tethering/abating.
M	Penalty weight.
D	Maximum difference between $[T_w, T_c, T_e]$.

$$
\begin{aligned}
\min \quad & \frac{\sum_{i\in W,C,E} Inf_i(\bar{T},\bar{A})}{\sum_{i\in W,C,E} P_i} + MD \\
\text{subject to} \quad & T_w W_w + T_c W_c + T_e W_e = C_T; \\
& \tfrac{A_w+A_c+A_e}{3} = C_A; \\
& \tfrac{Inf_i(\bar{T},\bar{A})}{P_i} \leq \tfrac{Inf_i(\bar{T_U},\bar{A_U})}{P_i}, \forall i; \\
& T_w, T_c, T_e < T_{\max}; \\
& A_w, A_c, A_e < A_{\max}.
\end{aligned}
\tag{10}
$$

B.5. Results and Discussion

Next, we considered allocation strategies where tethering or abating coverages are not uniform among clusters while maintaining the same resource level. Without loss of generality, we set the resource level of tethering and abating to be the national average between 2016 and 2018. We examined varies options of allocating one resource while keeping the other one uniform. Tethering coverages are weighted based on the number of infected dogs in each cluster, but abating coverages are not. For each set of experiments, we move on intervention resource from one or two clusters to another one, while keeping the other one constant. We also included strategy As, in the opposite direction in comparison with uniform allocation strategy (B) to verify that intervention outcomes are worse-off otherwise.

Table 33: Sensitivity analysis on tethering resource allocation set I

Strategy (Abating 21% for all)	Tethering coverage in West	Tethering coverage in Central	Tethering coverage in East	Percentage of dogs infected
A_1	71%	68%	80%	3.69%
B	71%	71%	71%	3.23%
C_1	71%	74%	62%	3.04%
D_1	71%	77%	53%	2.76%
E_1	71%	80%	44%	2.01%
F_1	71%	83%	35%	1.79%
G_1	71%	86%	26%	1.72%
H_1	71%	89%	17%	1.62%
Base	0%	0%	0%	8.67%

Table 34: Sensitivity analysis on tethering resource allocation set II

Strategy (Abating 21% for all)	Tethering coverage in West	Tethering coverage in Central	Tethering coverage in East	Percentage of dogs infected
A_2	74%	68%	74%	3.84%
B	71%	71%	71%	3.23%
C_2	68%	74%	68%	2.97%
D_2	65%	77%	65%	2.45%
E_2	62%	80%	62%	2.22%
F_2	59%	83%	59%	2.09%
G_2	56%	86%	56%	2.04%
H_2	53%	89%	53%	1.96%
Base	0%	0%	0%	8.67%

Table 35: Sensitivity analysis on tethering resource allocation set III

Strategy (Abating 21% for all)	Tethering coverage in West	Tethering coverage in Central	Tethering coverage in East	Percentage of dogs infected
A_3	68%	71%	77%	3.41%
B	71%	71%	71%	3.23%
C_3	74%	71%	65%	3.16%
D_3	77%	71%	59%	3.10%
E_3	80%	71%	53%	2.96%
F_3	83%	71%	47%	2.85%
G_3	86%	71%	41%	2.93%
H_3	89%	71%	35%	3.08%
Base	0%	0%	0%	8.67%

Finally, to verify that allocation 0% abating is indeed the optimal intervention strategy, we search for optimal solutions when abating is fixed from 10% to 50% and compare the intervention outcome with the true optimal solution.

Table 36: Sensitivity analysis on abating resource allocation set I

Strategy (Tethering 71% for all)	Abating coverage in West	Abating coverage in Central	Abating coverage in East	Percentage of dogs infected
A_4	21%	19%	23%	3.48%
B	21%	21%	21%	3.23%
C_4	21%	23%	19%	3.13%
D_4	21%	25%	17%	3.00%
E_4	21%	27%	15%	2.88%
F_4	21%	29%	13%	2.69%
G_4	21%	31%	11%	2.61%
H_4	21%	33%	9%	2.58%
Base	0%	0%	0%	8.67%

Table 37: Sensitivity analysis on abating resource allocation set II

Strategy (Tethering 71% for all)	Abating coverage in West	Abating coverage in Central	Abating coverage in East	Percentage of dogs infected
A_5	24%	16%	24%	3.26%
B	21%	21%	21%	3.23%
C_5	19%	26%	19%	2.80%
D_5	16%	31%	16%	2.51%
E_5	14%	36%	14%	2.40%
F_5	11%	41%	11%	2.29%
G_5	9%	46%	9%	2.19%
H_5	6%	51%	6%	2.10%
Base	0%	0%	0%	8.67%

Table 38: Sensitivity analysis on abating resource allocation set III

Strategy (Tethering 71% for all)	Abating coverage in West	Abating coverage in Central	Abating coverage in East	Percentage of dogs infected
A_4	19%	21%	23%	3.38%
B	21%	21%	21%	3.23%
C_4	23%	21%	19%	3.17%
D_4	25%	21%	17%	3.12%
E_4	27%	21%	15%	3.04%
F_4	29%	21%	13%	2.94%
G_4	31%	21%	11%	2.88%
H_4	33%	21%	9%	2.80%
Base	0%	0%	0%	8.67%

Table 39: Optimal intervention strategies with fixed abating coverage in the Central clutser.

		Optimal Intervention Strategy						
		Tethering Coverage			Aabting Coverage			Percentage of infected dogs
		West (Tw)	Central (Tc)	East (Te)	West (Aw)	Central (Ac)	East (Ae)	
Fxied Abating Coverage at the Central Cluster	10%	77%	92%	0%	42%	10%	9%	1.40%
	20%	85%	86%	2%	32%	20%	8%	1.48%
	30%	85%	86%	2%	29%	30%	1%	1.52%
	40%	89%	80%	14%	19%	40%	1%	1.56%
	50%	95%	78%	8%	10%	50%	0%	1.57%
	(none)	60%	95%	25%	50%	0%	10%	1.31%