Law Enforcement Resource Optimization with Response Time Guarantees

Jonathan Chase, Jiali Du, Na Fu, Truc Viet Le
Hoong Chuin Lau *
School of Information Systems, Singapore Management University, Singapore

***Abstract*—In a security-conscious world, and with the rapid increase in the global urbanized population, there is a growing challenge for law enforcement agencies to efficiently respond to emergency calls. We consider the problem of spatially and temporally optimizing the allocation of law enforcement resources such that the quality of service (QoS) in terms of emergency response time can be guaranteed. To solve this problem, we provide a spatio-temporal MILP optimization model, which we learn from a real-world dataset of incidents and dispatching records, and solve by existing solvers. One key feature of our proposed model is the introduction of risk values that allow a planner to flexibly make a tradeoff between their resource budget and the targeted service quality. Experimental results on real-world incident data, and simulations run on learned synthetic data, show a significant reduction in resource requirements over current practice, with violating QoS or abusing resource utilization.**

***Index Terms*—Resource Allocation, Law Enforcement Staffing, Data-Driven**

I. Introduction

On 3rd June 2017, two concurrent terrorist attacks took London by surprise, as a van drove into pedestrians on London Bridge, and men armed with knives attacked passers-by in the Borough Market area. In the aftermath, the Metropolitan Police attracted praise for the speed of their attendance on the scene [1]. To achieve a response time of 8 minutes requires effective planning and a well-designed deployment in a dense, congested city. Planning for urban environments is vital and increasingly so, as the UN has identified an increasing trend of urbanization across the world [2]. With a number of terrorist attacks hitting Western cities in the last 3 years [3], urban law enforcement agencies are under pressure to respond to emergency incidents promptly and reliably, whilst simultaneously being expected to economize on running costs. A key area in which to lower expenses is through manpower reduction or redeployment, intelligently utilizing law enforcement agents to achieve a high degree of responsiveness with a lower staffing level. However, when reducing costs and manpower, it is easy to overtax the remaining staff, and in [4] it was found that the tiredness level of police officers influenced their ability to choose correctly when faced with the decision to open fire on a suspect. Given a number of questionable recent police shootings in the US, some of which have led to civil unrest [5], the work demands placed on staff are an essential consideration of any resource optimization process.

*Corresponding Author (hclau@smu.edu.sg)

In this work, we are concerned with spatio-temporal staffing optimization in the context of law enforcement, which is a 24/7 service where officers are rostered on rotating shifts of 8-12 hours [6]. The expected outcome is an efficient resource allocation that can reduce the number of man-hours while guaranteeing a certain quality of service (QoS), without excessively increasing utilization. Specifically, we study the problem of optimizing the staffing level of law enforcement agents (i.e., officers) across base locations and time periods throughout a day using a data-driven approach. Our goal is to design high-fidelity allocation strategies so as to meet response time requirements for incidents of different priorities that maximize resource savings over the current practice.

We thus make the following key contributions:

- Based on real-world incident data provided by a large law enforcement agency, we propose a mixed integer linear programming (MILP) model for the deterministic resource optimization problem with guaranteed response time requirements.
- We solve the deterministic optimization model with sampling approximation to accommodate the dynamics of the incidents drawn from historical data.
- We perform extensive experiments simulated from real-world data to test the robustness of our solution, and experimental results demonstrate the potential savings in the staffing level over the current practice to meet response time requirements, without significant stress on resource utilization.

II. Related Work

Manpower optimization focuses on solving the problems of staff allocation (number of agents on duty for a given time and place) and staff scheduling (timing and staffing levels for implementable shifts). [7] formulates an integer optimization problem to schedule manpower in a multiskill call center. Manpower optimization has received extensive attention in computational intelligence, as these techniques can be employed for a range of applications, in our case, to law enforcement scheduling. There is a need in law enforcement to develop a more sophisticated approach to staffing than simplistic historical data analysis that has characterized the status quo [8]. Crime prediction is a notoriously challenging problem and therefore if the allocation methodology is too simplistic, the uncertainty in incident occurrence will lead

978-1-5386-2726-6/17/$31.00 ©2017 IEEE

to both over-provisioning and under-provisioning for different times and places [9].

To increase the sophistication of law enforcement allocation, [10] applies staff scheduling to rail security patrol scheduling. Staff are allocated to patrol stations spatio-temporally, comparing three mathematical models and evaluating on a real test case. Since police dispatch data is kept electronically, there are extensive data available for historical incidents, including location, time of occurrence, and officer engagement and travel times. [11] uses historical data to devise a visual analytics-based approach to prediction as a tool for resource allocation. Seasonal Trend decomposition is used to model crime patterns and predict the spatio-temporal distribution of future incidents, although no staff allocation is proposed to accompany this analysis. [12] justifies the use of crime data in law enforcement deployment because low response times, targeted patrolling, and a sense of police 'omnipresence' all contribute to reducing the number of crime incidents in a city. The report also found that the level of crime is inversely proportional to the number of police officers, a result that we aim to achieve through intelligent resource deployment rather than manpower increases. [13] uses these insights to propose an evolutionary patrol design algorithm based on a multiobjective optimization model, to enhance security in Northern Seattle. Similarly, we adopt an optimization model in this paper, but do not rely on the assumption that the incident data follows a Poisson distribution. Instead, we optimize directly on a large set of real incident data to better account for uncertainty, and assess the performance by way of simulation, using both incident prediction and response time prediction to improve the realism of the scenario. [14] also adopts an optimization approach, but adopts an iterative Bender's decomposition method to consider the behaviour changes of criminals in response to police patrols

In addition to assessing the optimal deployment of resources to handle crime incidents, it is also necessary to schedule staff in a reasonable and humane way. [15] studies the welfare of 275 police officers working 8-hour, 10-hour, and 12-hour shifts. Quality of life is measured on a number of factors including their ability to perform, health, and off-duty employment. The study found that officers working 12-hour shifts had a lower level of well-being than those working shorter shifts, therefore in this paper we consider a number of shift patterns to see if good incident response performance can be achieved while accommodating officer welfare. In this paper we build response time guarantees into the optimization QoS, but run simulations to calculate the expected utilization rates for agents to verify that our solution provides acceptable workloads for individuals.

III. Problem Definition

We consider a law enforcement and staffing allocation problem. A set of agents are assigned to patrol predefined geographic regions, ready to respond to incidents when they occur. These incidents are received via emergency calls, and may correspond to a range of situations, from murder and bomb threats, to theft and noise disturbances. When an incident is logged, it is assigned an urgency rating, and an agent is dispatched to attend, with the aim of arriving at the incident (the response time) within the time limit defined by the QoS agreement for that incident's urgency level. Certain incidents require more than one agent to attend, in which case additional agents are also dispatched, although their response time is not factored into the QoS agreement. Agents remain at the incident until it is resolved (the engagement time), and only then are they free to attend another incident. The objective of the work outlined in this paper is to find the minimum number of agents required to be on duty, and their geographic locations, to meet the QoS for a user-defined proportion of incidents. We illustrate the scenario under consideration in Fig. 1, showing an example incident occurrence and attendance timeline.

Formally, the law enforcement staffing and allocation problem is an optimization problem that can be described by the tuple:

$$< \mathcal{R}, \mathcal{L}, \mathcal{T}, \Delta, \alpha > .$$

The goal is to decide the staff levels across different base locations $\mathcal{L}$ to meet the response time requirements, Δ, for a given set of incidents, $\mathcal{R}$, within a given risk level, α, under the travel matrix, $\mathcal{T}$. More precisely, the decision variables are the number of agents needed at different base locations at different time periods of the day in order to minimize the number of resources used while satisfying the response time requirement constraint.

Each incident, $r \in \mathcal{R}$, is a tuple, $< l, d, c, t, s >$, where l is the location of occurrence, d is the demand (number of agents needed), c is the class representing the urgency level, t is the time of occurrence, and s is the service (or engagement) time required (superscript r is omitted for simplicity). Let $T_{l,l',t}$ be the travel time (in minutes) from location l to location l' at time t. In practice, upon receiving an emergency call, the operator would first identify and assign a certain urgency level to the incident. Each urgency level has a required maximum response time, which is computed as the time between the receipt of the call and the time of arrival of the dispatched agent(s) to the location of the incident, consisting of the dispatching time and the agent's traveling time. Satisfying the response time constraint is the key metric for the QoS. Let Δ be the maximum response time vector of all urgency levels. Incidents with higher urgency should be responded to faster than those with lower urgency. That is, $\Delta_c < \Delta_{c'}$, where c is a class with higher urgency than c'.

A good resource plan is able to tradeoff between the resource cost and the service quality. Note that resource requirements are not only location-based, but are also time-varying (i.e. spatio-temporal in nature). The granularity of time, for example, can be day and night, hourly interval, peak and non-peak period, etc.

To this end, we first propose a Mixed Integer Linear Programming (MILP) model for the problem defined above. We then solve the model with consideration of incident dynamics

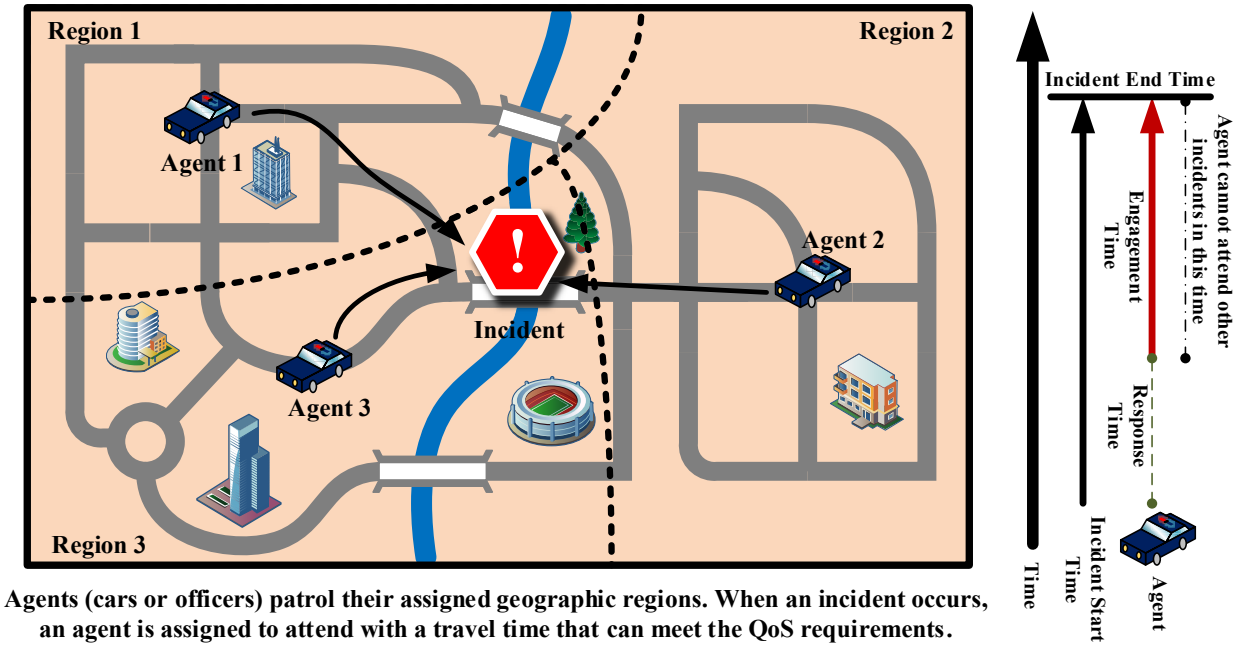

Fig. 1. Illustration of the system model under consideration. Agents attend incidents, with success defined as the first arriving agent's response time falling within the time limit defined by the QoS agreement.

from historical data using Sample Average Approximation (SAA) [16].

IV. Mathematical Model

In this section we describe the optimization model in detail. The key notations are summarized in Table I.

TABLE I
KEY NOTATIONS USED IN OPTIMIZATION MODEL.

Variable	
i	Agent index and its base location l^i
r	Incident index and its location l^r
T_{l^i,l^r}	Travel time from location of agent i to location of request r
d^r	Number of agents required to attend request r
c^r	Class (urgency level) of request r
t^r	Start time of request r
s^r	Service (engagement) time of request r
y_i	Binary variable indicating if agent i is required
y_i^r	Binary variable indicating if agent i serves incident r
$y_i^{q,r}$	Binary variable indicating if i serves r after serving q
ρ_i^r	Binary variable indicating if agent i is the first responder to request r
z^r	Binary variable indicating if the response time target was met for request r
δ_r	Calculated response time for incident i
e_i^q	Calculated ending time for agent i upon serving q

The proposed MILP model is given in (1)-(19). Given a dataset comprising a large number of historical incident records, we solve this model optimally based on the Sample Average Approximation (SAA) method proposed by [16]. More specifically, the objective is to minimize the total number of agents on duty subject to the requirement that the QoS constraint is met over the set of incidents given in the dataset.

$$\min \quad \sum_i y_i \tag{1}$$

s.t.

$$\delta^r \le T_{l^i,l^r} \cdot y_i^r + t^r + M \cdot (1 - y_i^r) \quad \forall r, i \tag{2}$$

$$\delta^r \ge W^r \quad \forall r \tag{3}$$

$$T_{l^i,l^r} \cdot y_i^r + t^r + M \cdot (1 - y_i^r) \ge W^r \quad \forall r, i \tag{4}$$

$$T_{l^i,l^r} \cdot y_i^r + t^r + M \cdot (y_i^r + \rho_i^r - 2) \le W^r \forall r, i \tag{5}$$

$$y_i^r \ge \rho_i^r \quad \forall r, i \tag{6}$$

$$\sum_i \rho_i^r = 1 \quad \forall r \tag{7}$$

$$y_i^r \le y_i \quad \forall i, r \tag{8}$$

$$\sum_i y_i^r = d^r \quad \forall r \tag{9}$$

$$y_i^{q,r} \le y_i^q \quad \forall i, q, r, q \ne r \tag{10}$$

$$y_i^{q,r} \le y_i^r \quad \forall i, q, r, q \ne r \tag{11}$$

$$\sum_q y_i^{q,r} = y_i^r \quad \forall i, q, r, q \ne r \tag{12}$$

$$y_i^{q,r} \le \frac{t^r - e_i^q}{M} + M(1 - y_i^q) + y_i^q \quad \forall i, q, r \tag{13}$$

$$e_i^r \leq t^r + T_{l^i,l^r} + s^r + M(1 - y_i^{q,r}) \quad \forall i, q, r \tag{14}$$

$$e_i^r \geq t^r + T_{l^i,l^r} + s^r + M(y_i^{q,r} - 1) \quad \forall i, q, r \tag{15}$$

$$y_i^{q,r} \leq \frac{t^r - e_i^q}{M} + 1 \quad \forall i, q, r \tag{16}$$

$$z^r \geq \frac{\delta^r - \Delta_c}{M} \quad \forall r, c^r = c \tag{17}$$

$$\frac{\sum_r z^r}{|\mathcal{R}|} \leq \alpha \tag{18}$$

$$z^r \in \{0, 1\} \quad \forall r \tag{19}$$

1) **Computing the response time**: A QoS key criterion is the response time to incidents. If more than one agent is dispatched to an incident, the response time is taken to be the first car to arrive at the scene. This is linearized in constraints (2) – (7), where δ^r denotes the response time for incident r, and T_{l^i,l^r} represents the travel time from the agent's location l^i to the incident location l^r when r happens. δ^r can be computed as the minimum travel time of all the responding agents, where the binary indicator $y_i^r = 1$ indicates agent i is dispatched to attend r. W^r denotes a lower bound on the response time δ^r, binary variable ρ_i^r serves as a location indicator that shows where the first agent comes from, and M is a large number. That is, i will be the first agent to be dispatched if $y_i^r = 1$ and $\rho_i^r = 1$. If i is not dispatched, ρ_i^r must be 0, as shown in (6).
2) **Preventing resource preemption and fulfilling demand of incidents**: When an agent is attending an incident, it must remain at the incident for the entire duration required to service the request, it cannot be pre-empted to serve another request. In addition, incidents belonging to different classes may request different numbers of agents and those demands must be fulfilled. This can be achieved by the constraints (8) – (13). Constraints (8) – (11) require that i can only be dispatched if it is available. The sum over all agents must fulfil the demand d^r, for all r, as shown in (9). Constraint (13) enforces that once an agent is occupied, it cannot be assigned to another incident until the current service has completed. That is, if i is serving q, and r occurs before q ends, then i cannot serve r and (13) forces y_i^r to be 0.
3) **Dispatching agents when incidents arrive**: Since incidents can be served by agents from different locations, the problem of planning the spatial and temporal resource supply becomes challenging. To this end, we use an intermediate variable $y_i^{q,r}$ to denote the sequence that i serves. $y_i^{q,r} = 1$ indicates the process that i serves r after q. We have constraints (14) – (16) where e_i^r is the ending time of serving incident r for agent i, i.e., the moment i becomes available after attending r. Suppose q was the last incident i attended, i.e., $y_i^{q,r} = 1$, then e_i^r consists of three parts: $r's$ starting time, t^r, the response time, T_{l^i,l^r}, of the dispatched agents and the engagement time, s^r, to attend r represented in (14) and (15). Thus, due to constraint (16), the necessary condition of $y_{q,r}^i = 1$ is $t^r \geq e_i^q$.
4) **Ensuring risk embedded QoS:** Without loss of generality, we assume a single QoS constraint that, given $0 \leq \alpha < 1$, allows at most α fraction of the incidents within a planning horizon to fail, i.e., whose response times exceed the respective maximum thresholds for the incidents given in $\mathcal{R}$. This requirement is linearized in constraints (17) – (19) where δ_r is the response time for request r, with a response time target of Δ_c for incidents of class c. If the response time for an incident, r, exceeds the target, the binary variable, $z^r = 1$, indicating a failure.

TABLE II
FEATURES USED TO LEARN A PREDICTIVE MODEL FOR THE TRAVEL TIME.

Feature	Description
`traffic_travel_time`	Time-dependent travel time computed by Google Maps
`mean_travel_time`	Hourly average travel time from location to location captured by data
`is_urgent`	Binary classification whether the incident is urgent or non-urgent
`is_resource_sharing`	Whether the responding car comes from a different base location
`num_cars`	Number of cars dispatched to respond to the incident
`hours`	Integer hour of the incident's occurrence time (0–23)

V. Travel Time Prediction

While historical data may provide details for the specific time and location of incidents, as well as the required engagement time, the aim of our model is to determine staffing levels that are robust against future occurrence of incidents. Hence, we propose to solve the model using historical data (viewed as training data), but evaluating the resulting solution on extensive simulated (testing) data.

One key challenge is to be able to generate realizations of incident data together with accurate response time for an agent at its location to respond to a particular incident. To that end, we apply a machine learning technique to predict response time values for an agent attending an incident. We first assume that the dispatched agent always starts from its base location. This is due to the lack of complete information about the agent's actual location as it patrols from point to point throughout the day. We propose a regression model that predicts the travel time (in minutes) from its base location to the incident location using the relevant features derived from the provided data. Table II summarizes these features.

In this regression model, the true travel time captured by the data (i.e., the duration from when the agent's car is dispatched to its arrival time at the incident location) is the response variable. The features in Table II were derived from a combination of both regression analysis and random forest feature importance. `traffic_travel_time` is the estimated travel time from the agent's base location to the incident location computed by Google Maps API at the time the agent was dispatched. `mean_travel_time` is the hourly average travel time from the agent's base location to the

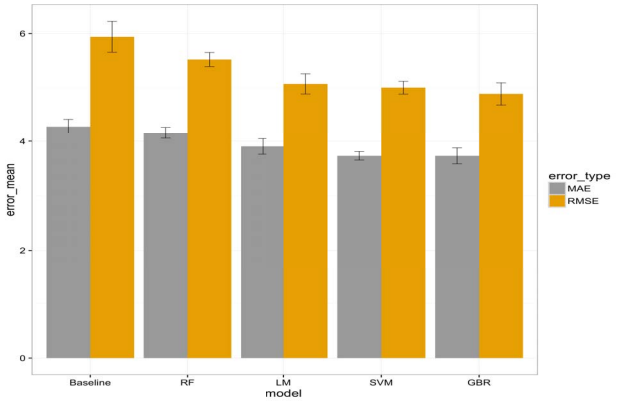

Fig. 2. Evaluation of predictive models for travel time prediction using 10-fold CV.

incident's location. `is_urgent` is a binary classification whether the incident is urgent or not. we assume the number of urgency levels to be 2 without loss of generality, i.e. incidents are classified as either urgent or non-urgent, with their respective response time requirement. Here, without loss of generality, we assume the number of urgency levels to be 2, i.e. incidents are classified as either urgent or non-urgent. `is_resource_sharing` is a binary variable indicating whether the dispatching car comes from a different base location. This typically happens when the resources near the incident's location are being deployed and unavailable, thus resources from another (neighboring) base location are called for. `num_cars` is the number of responding cars dispatched. `hours` is the integer hours of the incident's timestamp. Thus, both the spatio-temporal features of the incident and its response information are taken into consideration in the predictive model.

The following models are evaluated: random forest, linear regression, support vector machine (SVM), and gradient boosting regression (GBR). All these models use the features listed in Table II. We additionally evaluate a naive "baseline" model that uses the `mean_travel_time` feature as the predicted travel time for a given test incident. For the SVM model, the RBF (radial basis function) kernel is used. For the GBR model, we use the efficient implementation in the XGBoost package [17]. We use both MAE (mean absolute error) and RMSE (root-mean-square error) to evaluate the models. We perform 10-fold cross-validation (CV) on the provided data and take the mean errors across the folds. The results are shown in Fig. 2 with the mean error rates and the variances (error bars) over 10 folds.

Fig. 2 shows that the model that performs the best (by both MAE and RMSE) is the GBR model with average MAE well below 4 minutes. Unsurprisingly, the baseline model performs the worst (since it uses only one feature). Our experiments show that GBR is the best model overall. GBR is a powerful ensemble learning method that produces a predictive model in the form of an ensemble of regression trees. It has been shown to be robust against overfitting (hence, suitable for highly skewed and long-tailed data such as travel time) in many machine learning contests including the Netflix prize [18]. Therefore, we choose GBR as our predictive model for travel time estimation T_{l^i,l^r} in our MILP model.

VI. Experimental Evaluation

Our experimental evaluation consists of two parts. First, we solve the optimization model using real-world incident data provided by a national law enforcement agency (details are omitted for the purposes of national security). The dataset spans a one-year period and contains more than 200,000 incidents in total that require resource deployment. For each incident, the data records the detailed information of the location (latitude and longitude), timestamp, type and urgency as well as dispatch information such as travel time and service duration. In other words, the data tells us where, when, and what happened, and how the incidents were responded to. The optimization model was implemented using CPLEX [19] as the main solver and run on a cluster with 2 Intel Xeon E5-2665 2.90GHz processors (with a total of 24 threads) and 256GB of RAM. Our model provides time-varying resource plans at each base location for a 24-hour planning horizon. Solving the model for a year provides 365 individual solutions, each satisfying a risk level of 0.10, with a single resource allocation plan chosen conservatively to satisfy demand on all days with the exception of occasional outliers. This single allocation plan is then used as the input for the second part of the experimental evaluation. Using a set of synthetic incidents generated by applying machine learning techniques to the provided historical data, we implement a simulated agent dispatch system in Python [20]. The simulator reads the set of incidents in chronological order, and assigns agents to attend incidents, recording the observed risk and utilization for each day.

A. Optimization Model: Computational Performance

It is well established that integer problems suffer from computational tractability problems, with the traditional branch-and-bound algorithm having exponential complexity [21]. For our dataset, each day has an average of 550 incidents, which is too large to reach an optimal solution quickly. To improve computation time, we solve the optimization model for twelve 2-hour, four 6-hour and two 12-hour periods, computed in parallel, and compare the computation time and memory requirements to solve for a typical day. These results are given in Table III. For 2-hour and 6-hour periods we set a 1-hour run time cutoff, and for 12-hour set a 2-hour cutoff, recording the duality gap to show how close to optimality the solution came within the allotted time. This result shows clearly that executing 12 parallel optimization problems running on a 2-hour time window offers large performance gains over four 6-hour optimization runs, and particularly over two 12-hour optimization runs.

TABLE III
IMPACT OF TEMPORAL DECOMPOSITION ON COMPUTATIONAL DEMANDS

Time Period (hrs)	Execution Time (s)	Memory (GB)	Average Duality Gap
2	33.5	3.1	0.9%
6	3650.4	11.4	6.5%
12	7316.7	23.4	33.9%

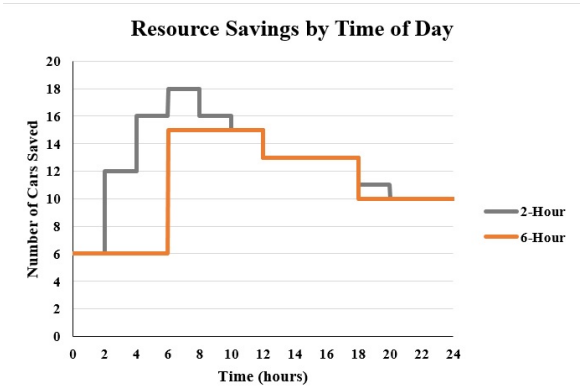

Fig. 3. The average savings (in number of agents) at 2-hour vs. 6-hour intervals.

B. Optimization Model: Resource Savings from Shorter Time Windows

In addition to a faster computation time, optimizing on a shorter time period allows a more nuanced allocation result, as the quantity and spatial distribution of demand can vary significantly with the time of day. Optimizing over a 2-hour period allows us to examine a greater range of shift handover times and shift lengths without needing to rerun the optimization, which, for a whole year, is a time-consuming process. Whilst the 12-hour solution takes prohibitively long to reach a sub-optimal solution, the 6-hour option is more reasonable. However, it lacks the potential for flexible planning that the 2-hour option allows, and overestimates the number of agents required for a given time period, as illustrated in Fig. 3, where we show a comparison of the two approaches for one major geographic area from the historical data. The difference between the two methods is particularly highlighted in the early morning time period, where a peak in demand is closely followed by a trough, which the 6-hour time window does not adequately capture, making it imprecise when designing shift patterns.

C. Optimization Model: Shift Design Savings

A key goal of law enforcement resource optimization centers around the goal of reducing the manpower required to maintain good response times. We calculated three alternative shift systems and evaluate the savings offered by the optimization model over current practice, as shown in Fig. 4. Shift Model 1 follows the current practice of 4 teams rotating in 12-hour shifts, with shift handover taking place at 8am and 8pm. Agents follow a rotational pattern of a day shift, then a night shift, followed by two rest days. Shift Model 2 considers

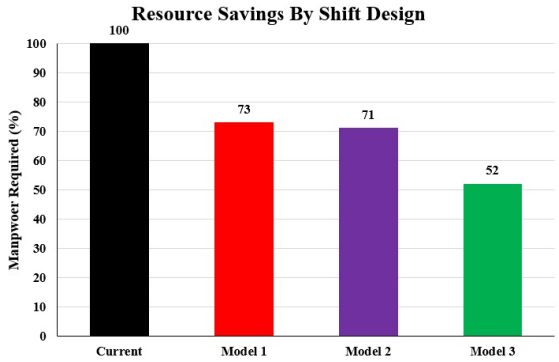

Fig. 4. Resource savings offered by each shift pattern.

alternative handover times, in this case 6am and 6pm, but maintains the 4 team rotational pattern used by Model 1. Shift Model 3 reorganizes the 4 team rotation into 3 teams that remain on duty for a 24-hour period, with members of the team alternating between rest and active duty, in 8-hour periods from midnight to 8am, 8am to 4pm and 4pm to midnight. This model employs a different rotational pattern, with a team working for a whole day, followed by two rest days. The transformation of 4 teams into 3 permits a larger saving of overall resources.

D. Simulation: Effect of Savings on Utilization

In addition to the goal of saving resources through optimization, it is important to ensure that individual agents do not become overworked. To determine the real-world viability of our proposed solution, we test each shift pattern against a year of synthetic incidents by simulating a chronological dispatching process. An optimization problem can essentially 'see the future', as agents are allocated to incidents with full knowledge of subsequent incidents. However, in the real world, an emergency dispatcher would not have this knowledge, thus this simulation verifies that our solution is able to achieve the targeted risk level in practice, without the advantage of full knowledge about the states of the system. Given the savings offered by each shift pattern, as observed in Fig. 4, we plot the observed utilization rates from the simulation, and provide the probability density function and cumulative distribution function for the observed data in Fig. 5. Each day rarely exceeds the target utilization rate of 40%, showing that the proposed savings are still reasonable in practice, without overtaxing agents. Whilst this target rate may seem low, there are a number of overheads to allow for that are not captured by the engagement time of an incident, such as the time spent doing paperwork, and scheduled training, that may also take place during working hours. Interestingly, these results show that a 20% difference in resource savings between Model 3 and Models 1 and 2 does not equate to such a significant difference in utilization rates. This shows that the intelligent design of a shift system can permit greater resource savings without significant sacrifices in QoS.

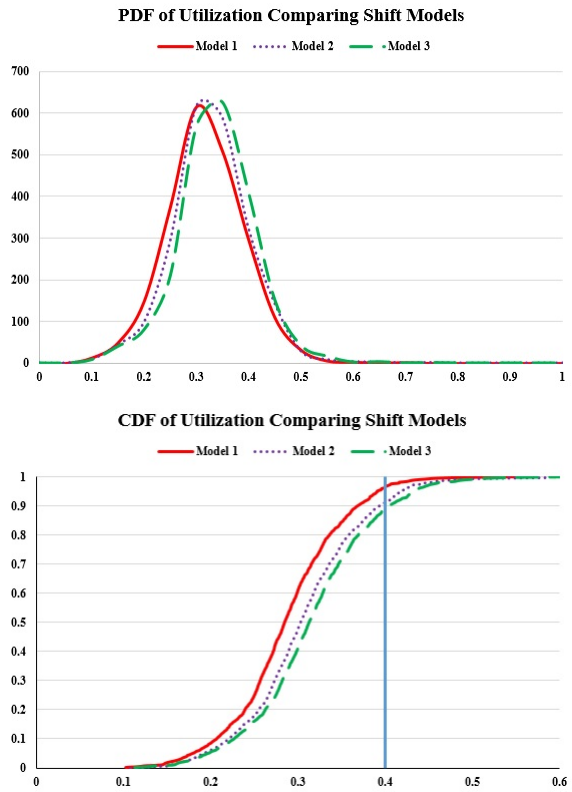

Fig. 5. Effect of different shift patterns on utilization.

VII. Conclusions and Further Work

In this work, we considered the problem of law enforcement manpower allocation and proposed a deterministic resource optimization model that is evaluated using real world emergency request data, with response time prediction based on machine learning techniques. Our solution framework embeds decision makers' risk attitudes, which allows the planners to flexibly choose their own tradeoffs between the resource cost and the risk they are willing to take. Our experimental results indicate significant resource savings over the current practice, as well as simulating a real world dispatch process to demonstrate reasonable utilization rates. We believe that this work can provide a practical solution for law enforcement agencies to efficiently and effectively respond to crimes and incidents. Future work should introduce stochasticity allowing us to account for greater seasonality in demand, permitting for the dynamic selection of an allocation solution given changing season and environmental features. The assumption that agents start from a fixed location should also be relaxed, incorporating patrol design and more realistic travel time prediction. The incorporation of these features should provide a powerful and flexible tool for law enforcement agencies to handle the demands of a security-conscious world.

Acknowledgment

This research is partially funded by the National Research Foundation Singapore under its Corp Lab@University scheme.

References

[1] S. Hewitt, "London bridge attack: Eight-minute police response the result of years of training," http://www.abc.net.au/news/2017-06-05/london-bridge-police-response-the-result-of-years-of-training/8589472 accessed 3 August 2017, 2017.

[2] G. K. Heilig, "World urbanization prospects: The 2011 revision," *United Nations, Department of Economic and Social Affairs (DESA), Population Division, Population Estimates and Projections Section, New York*, 2012.

[3] B. Singman, "Timeline of recent terror attacks against the west," http://www.foxnews.com/world/2017/06/19/timeline-recent-terror-attacks-against-west.html accessed 3 August 2017, 2017.

[4] J. Barton, A. Vrij, and R. Bull, "Shift patterns and hardiness: Police use of lethal force during simulated incidents," *Journal of Police and Criminal Psychology*, vol. 19, no. 1, pp. 82–89, Mar 2004. [Online]. Available: https://doi.org/10.1007/BF02802577

[5] M. Park, "Three trials, no convictions in fatal police shootings," http://www.cnn.com/2017/06/25/us/police-shooting-trials/index.html accessed 3 August 2017, 2017.

[6] Accenture, "Study of police resource management and rostering arrangements," http://www.intellicate.com/whitepapers/HO_police_rostering.pdf accessed 15 September 2016, 2004.

[7] M. T. Cezik and P. L'Ecuyer, "Staffing multiskill call centers via linear programming and simulation," *Management Science*, vol. 54, no. 2, pp. 310–323, 2008.

[8] J. M. Wilson and A. Weiss, "A performance-based approach to police staffing and allocation," *US Department of Justice Office of Community Oriented Policing Services*, 2012.

[9] N. Malleson and M. A. Andresen, "Spatio-temporal crime hotspots and the ambient population," *Crime Science*, vol. 4, no. 1, pp. 1–8, 2015.

[10] H. C. Lau, Z. Yuan, and A. Gunawan, "Patrol scheduling in urban rail network," *Annals of Operations Research*, vol. 239, no. 1, pp. 317–342, 2016.

[11] A. Malik, R. Maciejewski, S. Towers, S. McCullough, and D. S. Ebert, "Proactive spatiotemporal resource allocation and predictive visual analytics for community policing and law enforcement," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1863–1872, Dec 2014.

[12] L. Sherman, D. C. Gottfredson, D. L. MacKenzie, J. Eck, P. Reuter, and S. Bushway, "Preventing crime: What works, what doesn't, what's promising. research in brief. national institute of justice," 01 1998.

[13] M. Muaafa and J. E. Ramirez-Marquez, "Engineering management models for urban security," *IEEE Transactions on Engineering Management*, vol. 64, no. 1, pp. 29–41, Feb 2017.

[14] A. Mukhopadhyay, C. Zhang, Y. Vorobeychik, M. Tambe, K. Pence, and P. Speer, "Optimal allocation of police patrol resources using a continuous-time crime model," in *Decision and Game Theory for Security - 7th International Conference, GameSec 2016*, 2016, pp. 139–158.

[15] K. L. Amendola, D. Weisburd, E. E. Hamilton, G. Jones, and M. Slipka, "An experimental study of compressed work schedules in policing: advantages and disadvantages of various shift lengths," *Journal of Experimental Criminology*, vol. 7, no. 4, pp. 407–442, 2011.

[16] B. Pagnoncelli, S. Ahmed, and A. Shapiro, "Sample average approximation method for chance constrained programming: theory and applications," *Journal of optimization theory and applications*, vol. 142, no. 2, pp. 399–416, 2009.

[17] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 785–794.

[18] R. M. Bell and Y. Koren, "Lessons from the Netflix prize challenge," *ACM SIGKDD Explorations Newsletter*, vol. 9, no. 2, pp. 75–79, 2007.

[19] IBM, "Cplex optimizer," https://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/ accessed 7 August 2017, 2017.

[20] P. S. Foundation, "python," https://www.python.org/ accessed 7 August 2017, 2017.

[21] C. P. Gomes, "Structure, duality, and randomization: Common themes in AI and OR." in *AAAI/IAAI*, H. A. Kautz and B. W. Porter, Eds. AAAI Press / The MIT Press, 2000, pp. 1152–1158. [Online]. Available: http://dblp.uni-trier.de/db/conf/aaai/aaai2000.html#Gomes00