Prioritized Allocation of Emergency Responders based on a Continuous-Time Incident Prediction Model

Ayan Mukhopadhyay,Yevgeniy Vorobeychik, Abhishek Dubey, Gautam Biswas
Vanderbilt University
2201 West End Ave, Nashville, TN 37235
{ayan.mukhopadhyay,yevgeniy.vorobeychik}@vanderbilt.edu

ABSTRACT

Efficient emergency response is a major concern in densely populated urban areas. Numerous techniques have been proposed to allocate emergency responders to optimize response times, coverage, and incident prevention. Effective response depends, in turn, on effective prediction of incidents occurring in space and time, a problem which has also received considerable prior attention. We formulate a non-linear mathematical program maximizing expected incident coverage, and propose a novel algorithmic framework for solving this problem. In order to aid the optimization problem, we propose a novel incident prediction mechanism. Prior art in incident prediction does not generally consider incident priorities which are crucial in optimal dispatch, and spatial modeling either considers each discretized area independently, or learns a homogeneous model. We bridge these gaps by learning a joint distribution of both incident arrival time and severity, with spatial heterogeneity captured using a hierarchical clustering approach. Moreover, our decomposition of the joint arrival and severity distributions allows us to independently learn the continuous-time arrival model, and subsequently use a multinomial logistic regression to capture severity, conditional on incident time. We use real traffic accident and response data from the urban area around Nashville, USA, to evaluate the proposed approach, showing that it significantly outperforms prior art as well as the real dispatch method currently in use.

Keywords

Incident Response; Greedy Adaptive Search; Survival Analysis; Queuing Theory

1. INTRODUCTION

Increasing urban population density has led to a number of major challenges, such as pollution, congestion, accidents, and crime. To manage incidents, including fire and crime, cities resort to diverse groups of emergency responders, including fire and police departments. From the perspective of a responder, two problems are pivotal: 1) how to respond to emergencies as they occur, and 2) how to deploy limited responder resources, such as fire depots and vehicles, so as to best anticipate, and respond to, potential future incidents. We focus on the second problem.

Indeed, there have been a number of prior efforts considering how to best allocate responders in anticipation of incidents (e.g., [34, 25, 32]). It is clear that to ensure effective deployment, a crucial subproblem is *incident forecasting* in both space and time, and indeed, this issue has also been extensively considered in prior art [17, 31, 35, 34, 25]. Nevertheless, there are several major gaps in the literature which limit the practical applicability of the approaches to date. First, forecasting methods tend to either learn distinct models for each spatial area, resulting in higher model variance or requiring simpler models and limiting generalizability, or learn a single homogeneous model for most of the urban area, potentially failing to account for important spatial heterogeneity. Second, many forecasting methods cannot capture the dependence of incident rates on arbitrary exogenous features, such as weather, time of day, and day of the week (except Mukhopadhyay et al. [25]; see below). Third, a crucial factor rarely considered is incident severity: clearly, responders need to prioritize their response based on urgency, and both incident forecasting methods, and responder allocation, must therefore be explicitly designed to account for this.

In this paper, we systematically address the three identified gaps in the prior literature by considering the problem of optimal location of responder stations and distribution of responders in these, the latter being the key distinction between this and the well-known facility location problem [13, 21]. We develop a novel optimization problem to maximize incident density coverage with restrictions on waiting times and considering incident priorities, drawing on results from queuing theory. Our approach builds on the optimization method by Silva and Serra [32], but they restrict the approach to a single responder per station, a major limitation in practice. Our extension entails a non-trivial technical contribution and makes the approach far more practically viable. We note that while most prior work deals with response optimization and incident prediction separately, the latter is a fundamental requirement of the former. Thus, in order to aid our optimization model and validate it, we develop a hierarchically structured probabilistic model to predict time, location, and severity of urban incidents, that can capture the effect of arbitrary covariates on incident occurrence. By decomposing the prediction into a separate component involving incident time, and a component pertaining to severity, we make use of survival analysis [9] to learn the incident arrival distribution in continuous time,

Appears in: *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*

and a multinomial logistic regression [4] to learn the distribution over severity categories. Our use of survival analysis in forecasting incident arrivals is similar to Mukhopadhyay et al. [25]. Our key advance on this approach is to learn spatial granularity of the model from data: specifically, we combine survival analysis with hierarchical clustering to balance spatial heterogeneity and model variance.

We evaluate our proposed approach using traffic accident data obtained from the fire department of Nashville, US (perhaps counterintuitively, traffic accidents comprise the most common incident type to which this fire department responds). We demonstrate that our approach is a significant improvement over the state of the art in terms of incident prediction efficacy, and our method for locating responder stations substantially improves upon the current approach actually in use in this US city.

2. RELATED WORK

The problem of optimally placing responders in space to respond to incidents has been well explored in literature. As an important first step, we recognize the fact that there are multiple measures of optimality in this context. The most natural measure of the quality of response is response time, and considerable research has been devoted to this [37, 36, 25]. Another common criterion is to maximize the coverage area of response vehicles [23]. It is also natural to combine these goals and create allocation algorithms that try to achieve both [32, 10].

We focus specifically on coverage models, that aim to maximize the reach of service providers to potential demand nodes. In particular, the Maximum Coverage Problem [7] looks at optimally locating p facilities such that the maximum number of demand nodes can be served. The problem of locating facilities on a network to minimize travel distance to potential nodes was actually formulated before this [14]. The problem of minimizing the average service time from a single server location where requests are queued in the absence of a server has also been explored before [3]. Another approach is to look at minimizing the expected response time to the furthest point in the network from a single server location [6]. A common issue with these approaches is that responder *depots* and responders are used synonymously, meaning that one responder is placed at a location. In practice, this is typically not the case, severely limiting the ability to apply such methods in the field. Responder services typically rent spaces that can house multiple responders, which calls for the dual optimization over depot locations as well as the number of responders per depot. We extend prior work concerning maximizing coverage with restrictions on waiting times for incidents [32] to tackle this dual optimization scenario.

The problem of incident forecasting has also been extensively explored. This includes prediction regarding freeway accidents [33, 1], traffic congestion prediction [2], crime prediction [17, 31, 34, 25], fire accident predictions [35], and many more. It has been noted in the literature that freeway accidents are generally difficult to predict, due to an inherent random nature of these accidents and spatially varying factors [27]. Recently, freeway accidents have been predicted using panel data analysis approach that predicts incidents based on both time-varying and site-specific factors [27]. An extensive survey of the literature on crash prediction models is presented in [20], which highlights the prevalence of Poisson distribution based models [5, 22, 29], and multiple linear regression approaches [11, 28]. Further, there are incident prediction approaches that are flexible enough to fit into this domain [25]. These approaches treat accidents as homogeneous, meaning that the severity of accidents is not taken into account, which is a crucial factor in practice. We consider the problem of incident prediction as forecasting both time and place of occurrence, as well as incident severity, bridging a major gap in prior art.

3. OPTIMIZING RESPONDER PLACEMENT

A fundamental problem faced by emergency responding agencies is to optimally allocate depots in space, allocate response vehicles in depots, and assign vehicles to incidents. Commonly, depots refer to fixed responder stations, such as police and fire stations, which are pre-determined. In other settings, depots can be periodically reallocated. For example, in our target domain of fire department emergency response (see Section 5), the emergency vehicles are typically stationed in rented parking lots, which can therefore be reassigned if a need arises. Nevertheless, the time scales of depot and response vehicle allocations are typically different. Our approach can be modified directly if we wish to reassign vehicles to a fixed set of depots at an hourly or daily time scale, with the full problem (including depot allocation) solved at longer (say, monthly) time scales.

While the notion of optimality in emergency response problems can vary, we focus on the problem of maximizing coverage, one of the most common variants in this domain. To start, we discretize space into areas (henceforth referred as a set of grids G). Suppose that for each grid g_i, incidents of severity k arrive at a (predicted) rate λ_i^k. We assume that a predictive model for the concerned incident type is available. We describe one such model in Section 4. Given the arrival rates, for each grid g_i and incident severity k, jointly referred to as a pair (i, k) henceforth, we aim to allocate a depot d to respond to the associated incidents. We refer to a successful assignment of (i, k) to a responder depot d as *covering* the pair. We note that a measure of importance of such a pair is its predicted arrival rate λ_i^k. We try to maximize the total arrival rate that the model covers by optimally placing depots in a subset of the available grids, where each depot can hold a collection of responder vehicles. Thus, given a set G of discrete grid locations, p different responders (emergency vehicles) and a budget to allocate b different depots, we want to find the optimal location of the depots and the distribution of vehicles in such depots.

We now describe the formal structure of this optimization problem. For simplicity, we index depots by their grid numbers, which means that a depot located in grid j is referred to as depot j. Moreover, when there is no responder available in a depot to serve an incident that is assigned to it, we assume that the incident enters a waiting queue. We assume that each depot maintains its own queue which is ordered according to incident priorities but is non-preemptive at service time, which means that an incident already getting responded to is never left midway to attend to an incident of higher priority. In our model, lower values of k correspond to higher priorities. A similar approach has been studied previously with the aim of maximizing the total population covered [32]. However, we look at a generalized problem structure where more than one responder can be placed at a location, which significantly complicates the queuing model in

consideration by changing a single-responder priority queue model to a multi-responder priority queue model.

Formally, we consider the following optimization problem.

$$\max_{x,y,d} Z = \sum_k \sum_j \sum_i \lambda_i^k x_{ij}^k \tag{1a}$$

$$s.t. : x_{ij}^k \leq d_j \quad \forall i,j \in I, \forall k \tag{1b}$$

$$x_{ij}^k \leq y_j \quad \forall i,j \in I, \forall k \tag{1c}$$

$$\sum_{j \in I} x_{ij}^k \leq 1 \quad \forall i \in I, \forall k \tag{1d}$$

$$y_j \leq y_j d_j \quad \forall j \in I \tag{1e}$$

$$\sum_{j \in I} y_j \leq p \tag{1f}$$

$$\sum_{j \in I} d_j \leq b \tag{1g}$$

$$w_j^k \leq \tau^k \quad \forall j, \forall k \tag{1h}$$

$$y_j \in [1..p] \quad \forall j \in I \tag{1i}$$

$$x_j, d_j \in \{0,1\} \quad \forall j \in I \tag{1j}$$

where I is the set that indexes over all grid numbers, d_j is a binary decision variable which is 1 if there is a depot located at grid g_j, y_j is a decision variable that indicates how many responders are placed at depot j and x_{ij}^k is a binary decision variable which is 1 if depot j is assigned to respond to the pair (i,k) and 0 otherwise. We ensure that service standards are met by enforcing constraints that the mean waiting time (denoted by w_j^k where j and k correspond to the depot number and the priority respectively) at all depots is less than a pre-specified time limit τ^k. We assume that this information is user-specified, depending on the type of incident and the service quality required. The objective (1a) aims to maximize the total coverage by the responders. Constraint (1b) ensures that calls are assigned to locations that have depots, constraint (1c) forces that such depots have at least one responder assigned and constraint (1e) ensures that responders are placed in locations which are depots. Further, constraint (1d) ensures that each pair (i,k) is assigned at most once. Constraints (1f) and (1g) are budget restrictions on responders and depots respectively and finally, constraint (1h) ensures that the mean waiting time for incidents is within a pre-specified tolerance.

Before we attempt to solve this problem, we present a method to calculate the waiting time for a given depot. Recall that arrival rates are available from the incident prediction model that we describe later in Section 4. We model the inter-arrival time of incidents as exponential, and consequently the arrivals are Poisson distributed. We make the standard assumption that the service times are exponential as well, giving us a queuing model with memoryless arrivals, memoryless service times and multiple servers, commonly represented as a $m/m/c$ priority-queue model using the Kendall's notation [19]. Such a model is difficult to analyze when multiple priority events are present and each follows its own service time distribution [16]. We make a simplifying assumption that although different severities follow different arrival distributions and have different service time constraints in the optimization model, they follow the same service distribution. Thus, in our formulation, we assume that all priorities are served with a common exponential distribution with mean μ. This is an assumption that is realistic in many real-life applications as severities often represent the urgency with which an incident needs to be responded to and is not an indicator of actual service time. Moreover, analysis on our dataset revealed that learning the same distribution across event severities appears to be nearly as good as learning heterogeneous distributions (see Section 5, Table 2). We present a sketch of the derivation for the waiting time of our queuing model here. The full derivation can be found in prior literature in queuing theory [8].

3.1 Calculating Waiting Time

Consider that an incident of priority k happens and has to enter the waiting queue for depot j with y_j responders because n_0 incidents of higher or equal priority are already waiting to be serviced. Also, let Λ denote aggregate arrival rates, such that $\Lambda_j^k = \sum_{i \in I} x_{ij}^k \lambda_i^k$,and let $\Lambda_j = \sum_k \Lambda_j^k$. Λ_j^k thus measures the rate of arrival of incidents of priority k for all grids that depot j serves. Let us assume that it takes t_0 time to service n_0 incidents. However, in time t_0, all arrivals of higher priority will supersede our event in the queue. Let there be n_1 such events which can be served in time t_1. Again, there can be further arrivals in t_1, and so on. Our incident, therefore, must wait for time $\sum_{l=1}^{\infty} t_l$ before it is serviced. Since we want the expected waiting time, we want to calculate $\mathbb{E}(\sum_{l=1}^{\infty} t_l)$. This can be calculated by looking at the conditional waiting time $\mathbb{E}(t_{l+1}|t_l)$, which is given by $\frac{1}{y_j \mu} \sum_{q=1}^{k-1} \Lambda_j^q t_l$, where μ is the mean service time distribution. Now, for any h, $\mathbb{E}(\sum_{l=0}^{h+1} t_l)$ is given by $\mathbb{E}(\sum_{l=0}^{h} t_l + \mathbb{E}(t_{h+1}|t_h))$. By induction and considering $h+1 \to \infty$, we get an expression for the average waiting time for an incident with priority k as

$$w_j^k = \frac{\frac{\pi}{y_j \mu}}{(1 - \frac{1}{y_j \mu} \sum_{q=1}^{k-1} \Lambda_j^q)(1 - \frac{1}{y_j \mu} \sum_{q=1}^{k} \Lambda_j^q)}$$

where

$$\pi = \frac{\left(\frac{\Lambda_j}{\mu}\right)^{y_j}}{y_j! \left(1 - \frac{\Lambda_j}{y_j \mu}\right) \left[\sum_{r=0}^{y_j - 1} \frac{\left(\frac{\Lambda_j}{\mu}\right)^r}{r!} + \sum_{r=y_j}^{\infty} \frac{\left(\frac{\Lambda_j}{\mu}\right)^r}{y_j! y_j^{r-y_j}} \right]}$$

We note that the queuing model assumes that the service time distribution is memoryless. This is a concern as the time taken by a responder to travel to an incident is not distributed exponentially. To tackle this, we assume that a depot can only respond to an incident if it is located within a small distance s of the incident, which in practice is sufficiently small that it can be treated as constant with respect to the overall service time.

3.2 Adaptive Random Search for Responder Optimization

The main challenge in solving mathematical program (1) is the fact that Constraints (1h) are non-linear and non-convex. We tackle this problem using greedy random adaptive search (GRASP). Such a procedure has been previously used in coverage maximization [32], but this previous approach cannot be directly applied when depots can have multiple responders as the search space becomes significantly more complex. We therefore propose a novel algorithm,

Algorithm 1 Restricted Candidate List Construction

INPUT: λ; p, K
OUTPUT: RCL : Restricted Candidate List
Initialize $S : \phi$
Create sorted list SN of candidate sites with respect to population/demand rate.
for $g = 1$ to b **do** ,
  Assign p servers to g grids according to λ.
  while $|S| \neq p$ **do**
    for $j \in SN$ **do**
      for $k = 1$ to K **do**
        if NoDemands(j) **then**
          Distribute$_b(j, \bar{j})$
        else
          for $i \in D_j$ **do**
            if $w_j^k < \tau^k$ **then**
              Set $x_{ij}^{[k]} = 1$
            end if
          end for
        end if
        Calculate $\Lambda_j^k = \sum_{i \in D_j} x_{ij}^k \lambda_i^k$,
      end for
    end for
    Construct RCL
    select $j^* =$ RandomSelect(RCL)
    Update $S := S \cup j^*$
    Remove demand nodes assigned to j^*
  end while
  Calculate Z_g
end for
Return Allocation S with highest Z.

Heuristic based **R**esponse **O**ptimization of **C**overage with **Q**ueuing (*HROCQ*). We break up the algorithm into two parts and describe the construction of the Restricted Candidate List in Algorithm 1 and the Local Search Phase in Algorithm 2 in sequence.

We first describe the construction of the Restricted Candidate List (*RCL*; Algorithm 1). We use D_j to denote the set of all nodes within a distance s of node j and D_{ij} to denote the i^{th} element of this set. We use $\bar{j}$ to denote all grids that have never been assigned a responder in the course of our iterative algorithm, Distribute$_b^i(a, h)$ as a method to distribute responders from grid a to i grids in set h in proportion to their demand rates (absence of i means that the distribution is done to as many nodes as possible) with a limit of b on the total number of grids that have responders, and NoDemands(a) as a method that returns *True* if no valid assignment can be made to node a and returns *False* otherwise. We use $|S|$ to denote the number of responders in our solution set and $||S||$ to denote the number of grids in S, that we iteratively build. Also, consistent with the optimization problem formulation, at any point in the algorithm, the number of responders assigned to grid j is denoted by y_j.

In order to decide the number of servers to be placed in a depot, we first sort the depots according to their event demand rate. We refer to this sorted node list as SN. Then, for the g^{th} run of the construction phase, we greedily assign p responders (our budget) to the first g depot locations in the SN, in proportion to their demand rates. Then, iteratively, for each node j that has been assigned a responder, we inspect D_j. For each node i in D_j and priority k (starting from the highest priority), we assign depot j to respond to the pair (i, k). After making each assignment, we ensure that no waiting time constraint is violated. We stop assigning calls to a depot when its waiting time constraint is violated and move to the next depot location in SN. After a phase of assignments, we look at the total demand rate Λ_j. We identify the highest serving depot as $j = \text{argmax}_i \Lambda_i$ and its corresponding service rate as $\Lambda_j^{'}$. Then, we select all nodes in our RCL that have a service rate of at least $\gamma \Lambda_j^{'}$. Finally, to finish one run of an assignment, we randomly select a node from the RCL and permanently fix its assignments and remove the pairs assigned to it from being considered in the future.

We stop when all depots that were assigned responders have been assigned a pair of calls, or when there are no more pairs to assign. This entire process is run b times, and we get a feasible solution from each such run, which we then carry forward to the local search phase, described next.

In the local search phase, described in Algorithm 2, we iteratively look at each depot from the current solution and deallocate all pairs assigned to it. We also deallocate the responders that were assigned to this depot. We distribute these responders iteratively to other potential depots not in the current solution set in proportion to their demand rates. We then calculate the updated objective value by replacing the unassigned node with the newly assigned set of nodes (referred to as S^i), where i is the number of grids that have received the freed responders. Finally, if any such assignment improves the objective value, we accept the updated solution. This method, repeated iteratively, performs a local search both with respect to the depots and the number of servers per depot.

Algorithm 2 Local Search Phase

INPUT: Restricted Candidate List
OUTPUT: Updated Solution
$Z^* := Z^S$, objective with current solution set S
for j in S **do**
  $\bar{S} := G \backslash S$ Find all nodes that are not in the Solution
  $S := S \setminus j$ Deallocate assignments
  Deallocate y_j responders
  for $i = 1$ to $||\bar{S}||$ **do**
    Distribute$_b^i(j, \bar{S})$
    Calculate Z^{S^i}
  end for
  $i^* := \text{argmax}_i \, Z^{S^i}$
  if $Z^{S^{i^*}} > Z^*$ **then**
    $S = S^{i^*}$
    $Z^* = Z^{S^{i^*}}$
  else
    Revert Deallocations.
  end if
end for
Return S

4. INCIDENT FORECASTING MODEL

4.1 Predicting Incident Arrival Time

Having looked at a model that can optimally allocate responders given predicted arrival rates, we now describe our

incident prediction model. In predicting incidents, we aim to use data to learn a continuous-time model $f(t|w)$ of incident arrival given an arbitrary set of features w. A natural fit for this problem is survival analysis, which has recently been used to predict urban crime incidents [25]. We first provide a brief overview of such survival models, and then present our specific contribution. Survival Analysis is a broad class of methods that model the distribution of time to an incident arrival. Our specific approach uses an accelerated time effect (AFT) model in which covariates increase or decrease the expected time to next incident [24]. Formally, a survival model is $f_t(t|\gamma(w))$, where f_t is a probability distribution for a continuous random variable T representing the inter-arrival time, which typically depends on covariates w as $\log(\gamma(w)) = \rho_0 + \sum_i \rho_i w_i$. The survival function is defined as $S(t) = 1 - F_t(t)$, where $F_t(t)$ is the cumulative distribution function of T. In order to model and learn $f(t)$ and consequently $S(t)$, we chose the exponential distribution, which has been widely used to model inter-arrival time and has recently been used to predict urban incidents [25]. Since we model time in an accelerated failure setting, $S(t|\gamma(w)) = S(\gamma(w)\, t)$.

Armed with the basic machinery of survival modeling, we now address the specific question of interest: how to use it to model spatio-temporal distribution of incident arrival. A natural way to capture the incidents in space is to first discretize space into areas (corresponding to the set G in our responder optimization model) and then learn survival models independently for each grid. The main concern with this approach is overfitting: each grid induces relatively little data, and there are surely considerable structural similarities of the incident process across multiple grids that we can leverage. On the other hand, learning a single "universal" model for all grids may fail to capture all of the existing heterogeneity not explicitly modeled in the feature space w. We present a principled way of tackling this problem by using a Hierarchical Clustering approach [18].

We first introduce some notation. For an incident, let the feature set w be divided into two parts w_s and w_d, where w_s represents a set of *static* features, such as population density in a grid, which remain relatively stable, while w_d will denote *dynamic features*, such as the amount of rainfall in a day or day of the week. We hypothesize that the set w_s can be used to identify similarity between distinct spatial grids. To operationalize this hypothesis, we propose a hierarchical clustering algorithm, shown in Algorithm 3 which we now describe at a high level. We start by treating each grid as a distinct cluster. Iteratively, we merge two grids that are most similar, with similarity between grid i and grid j measured as the distance between associated w_s^i and w_s^j. At each step, we check whether the updated set of clusters decreases the predicted likelihood computed on the training data set compared with the previous iteration by more than a predefined limit, and stop as soon as marginal improvement in likelihood is below this limit. We maintain a high likelihood difference tolerance level initially to promote exploration of the solution space, and lower it as the algorithm progresses.

Algorithm 3 Hierarchical Clustering

INPUT: Grids G, Static Features W_s
OUTPUT: Clusters C, with optimal likelihood
At iteration 0,initialize each grid g_i as a cluster in list C^0.
for iteration m in max_iter **do**
    Calculate Similarity Matrix S, where $S_{i,j} = ||w_s^i - w_s^j||$
    $i, j = \text{argmin}_{ij} S$
    Merge c^i, c^j into c^i
    Update $w_s^i = \frac{(w_s^i + w_s^i)}{2}$
    Calculate Likelihood L^m
    if $L^m - L^{m-1} > \frac{\sigma}{m}$ **then**
        Return C^m
    end if
    Return C^m
end for

4.2 Predicting Incident Severity

Predicting time to arrival for incidents is crucial, but treating all incidents as identical, as is commonly done, is problematic in practice. As an example, consider two grids g_1 and g_2 with similar rates of predicted traffic incidents with one major difference: most incidents that happen in g_1 require immediate medical attention while most incidents in g_2 are minor accidents. Focusing solely on incident rates to allocate medical response vehicles would clearly be unwise from the perspective of saving lives. Consequently, it is imperative that we also predict the severity of an incident or the urgency with which it needs a response. Here, we point out that dispatching emergency response based on predictions is undesirable, as real-time information must be taken into account for accurate severity assessment. However, planning aggregate depot and responder locations necessitates predicting severities.

One way to capture incident severity is to use a distinct model for each incident type. However, past incident prediction models [12] have found that sacrificing the scale of data for achieving heterogeneity can produce noisy estimates, as it limits the data available for learning each distribution. We address this issue by learning a joint distribution over arrival time t and incident severity k, $f(t, k|w)$, where k is a discrete ordinal random variable representing the severity class of the incident from K possibilities. As a first, step, we represent $f(t, k|w) = f(t|w)f(k|t, w)$. This decomposition helps us in two ways: first, our model for predicting arrival times described in Section 4.1 can now be used as is to learn the density over arrival times, and second, we can now use the entire dataset to learn distribution over arrival times and severities, rather than fracturing it by severity category. To learn the severity distribution (conditional on incident time and the feature vector w) $f(k|t, w)$, we use the multinomial logistic regression model [4].

$$\begin{bmatrix} P(y=1|x,\theta) \\ P(y=2|x,\theta) \\ . \\ . \\ P(y=k|x,\theta) \end{bmatrix} = \begin{bmatrix} \frac{e^{\theta_1^T x}}{\sum_{j=1}^K e^{\theta_j^T x}} \\ \frac{e^{\theta_2^T x}}{\sum_{j=1}^K e^{\theta_j^T x}}) \\ . \\ . \\ \frac{e^{\theta_3^T x}}{\sum_{j=1}^K e^{\theta_j^T x}} \end{bmatrix} \quad (2)$$

Multinomial logistic regression (MLR) generalizes the standard logistic regression by extending the output variable to a general categorical variable. Formally, given a training set

$\{(w^1, t^1, y^1), (w^2, t^2, y^2), ..., (w^n, t^n, y^n), \}$, where $w^i \in \mathbb{R}^m$, $t \in \mathbb{R}$ and $y^i \in 1, 2..., K$, MLR models $P(y = k|w, t) \forall k \in 1, 2, ..., K$. The hypothesis function in this case can be represented as shown in Eq. (2) where we represent the set $\{w, t\}$ by a generic feature vector x. The cost function that we try to minimize is:

$$J(\theta) = \sum_{i=1}^{m} \sum_{k=1}^{K} \mathbb{1}\{y^i = k\} \log \frac{e^{\theta_k^T x_i}}{\sum_{j=1}^{K} e^{\theta_j^T x_i}}$$

The cost function is typically minimized by an iterative optimization algorithm such as gradient descent.

4.3 Feature Set

The next step in this model of incident prediction is to select a set of features w that can be used to learn the predictive model for traffic accidents. We describe here the features that constitute the set w.

- **Temporal Cycles** We used preliminary analysis and prior work in incident prediction to identify the different types of seasonality that affect incidents in our dataset. We use a binary feature for each season and another one for encoding weekdays and weekends. In order to look at the effect of time of day on incidents, we split each day into six zones of four hours each, and captured these by binary features.
- **Temporal and Spatial Incident Correlation** For each grid, we looked at the past incident counts in the last week and month in it as well as neighboring grids as features to capture the effect of temporal and spatial correlation among incidents. We also treated the number of past incidents in each severity category as a feature while predicting incident severity, and considered the long-term effect of temporal correlation by looking at the average number of incidents in the past year.
- **Weather** It is known that weather affects traffic incidents [33]. We included a collection of features, such as rainfall, snowfall, and mean temperature to capture this effect.
- **Transportation Features** The effect of roadway geometry on accidents has been extensively studied [26], [30]. For each grid, we used the total number of roadway and highway intersections as features.

We use average incident counts in a grid and the set of transportation features as the set of static features w_s while the other features form the set of dynamic features w_d.

5. RESULTS

We present the results in two different categories. First, we evaluate the effectiveness of our approach in predicting incidents, and second, we evaluate the performance of our responder optimization methods.

5.1 Data

Our evaluation uses traffic accident data obtained from the *fire department* in Nashville, USA, with a population of approximately 700,000. For this fire department, traffic accidents comprise a large majority of incidents it responds to (fires, in contrast, are relatively rare). We looked at data for 26 months, from 2014 - 2016, comprising of a total of 20148 traffic accidents. Each accident is accompanied by its time of occurrence, the time at which the first responding vehicle reached the scene and the time at which the last responding vehicle was back at service, which refers to completion of servicing an incident. To predict incidents, we extracted highway and street intersections from Open Street Maps [15] and weather data was collected at the county level.

Before presenting the results, we highlight the three central problems faced by the fire department that are addressed by our framework: 1) optimal choice of locations for the fire depots, 2) optimal decision about which vehicles should reside in which depots, and 3) optimal decision about which depots are assigned to respond to which traffic accidents, with particular emphasis on minimizing cross-depot dispatch due to considerations such as vehicle-maintenance. The third of these is a particularly acute concern, as they view their current policy of assigning vehicles to accidents to be inefficient, unnecessarily increasing response times as well as mileage on the fire vehicle fleet (the latter resulting in more frequent and costly repairs).

5.2 Incident Prediction

We evaluate the performance of our incident prediction model by comparing it to a recent state-of-the-art prediction approach that uses survival analysis with a coarse categorization of spatial grids [25]. To evaluate incident prediction performance, we split our data into three overlapping data sets of 22 months each. For each such set, we use 80% of the data as our training set and 20% as our test set. For learning the hierarchical model, we clustered grids based on past accident counts and the number of road intersections. A comparison of likelihoods is presented in Fig. 1. We show that a hierarchical clustering approach improves (log)-likelihood significantly in all test sets, with an average improvement of about 13%.

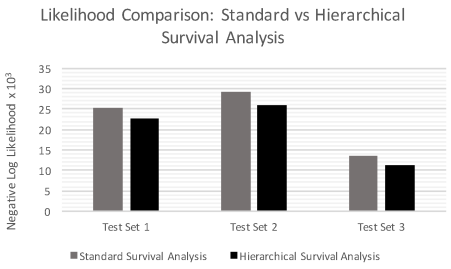

Figure 1: Likelihood Comparison (Lower is better)

Next we evaluate the quality of severity prediction, using accuracy as a metric. The results are shown in Table 1, where 3 severity classes were used. We find an average accuracy of about 66%, a reasonable performance on a 3-class classification problem. To delve more deeply, note that in emergency response settings, incorrectly predicting high severity incidents is more costly than overestimating severity. To evaluate this aspect, we also considered the fraction of times severity was underestimated (termed False Negatives) and overestimated (False Positives). We can see that the model rarely underestimates severity.

Note that the performance of both the arrival prediction methodology as well as severity prediction can also be measured indirectly by the performance of the response opti-

Table 1: Severity Prediction Accuracy

Test Set	Percent Accuracy	False Negative	False Positive
Set 1	64.7%	3.4%	31.9%
Set 2	65.4%	1.6%	33%
Set 3	67.4%	1.8%	30.8%

mization framework that uses predicted data to learn arrival rates. We address this next.

5.3 Response Optimization

Having looked at the performance of our predictive model, we now evaluate the performance of the response optimization model. To do this, we show several measures of performance of our model in a way that highlights its performance with respect to actual patrol policies and sheds some light on the structure and the components of the model. To achieve this, we use the same data sets for validating the response optimization mechanism as the training mechanism. First, we show how the components of *HROCQ* aid its performance. The two main stages of the algorithm involve *a)* greedily building a restricted candidate set for each and then *b)* choosing a candidate depot location from the set to form a partial solution to the problem. We evaluate the importance of the local search phase that works over the greedy-random candidate list phase. We present results by varying the number of responders and the parameter γ and check how the two stages of GRASP work. The results are presented in Fig. 2. For different problem configurations, we show how the objective function varies between the two phases. We note that the local search phase usually improves the existing solution. Although the improvement is typically small, it is crucial in emergency responder placement.

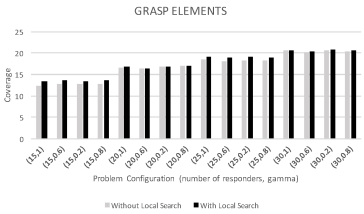

Figure 2: Phases of GRASP: Objective Value Comparison

Having seen the importance of the local search phase, we turn to tuning the parameter γ in our model. For a run of the algorithm, γ is fixed *a priori*. The parameter can be interpreted by understanding that $\gamma = 1$ corresponds to a completely greedy construction and $\gamma = 0$ corresponds to a construction that recognizes all temporary depot locations as candidate solutions in the RCL phase. In order to determine the parameter, we perform a discrete search over its domain, by varying it from 0 to 1 in steps of 0.1. We looked at cumulative results over 20 runs to determine the average objective value attained by each value of γ. We present the results for one training set in Fig. 3. Observe that usually a γ of 0.7 produces the best placement of responders and depots.

Finally, we compare our model to the existing responder

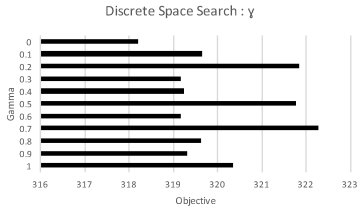

Figure 3: Discrete Search over γ's Space

deployment strategy in Nashville, TN. The accident data that we have reveals 3 types of priorities among incidents, namely, *A,B* and *D*, in order of increasing priority. First, we validate our assumption that learning a common service time distribution across event priorities is sufficiently accurate. We compare the *AIC* scores of two models: *a)* Learning a separate service time distribution for each of the priorities and *b)* Learning a common distribution across event priorities. The AIC Score is defined as: $\text{AIC}_m = 2k - 2\ln(L)$ where k is the total number of parameters estimated and L is the likelihood of the data under model m. We present the findings in Table 2, which shows that there is little loss in assuming a common service time distribution across event priorities.

Table 2: AIC Score Comparison : Service Distribution Models

Model	AIC Score
Separate Servicing Models for each Priority	326724.55
Common Servicing Model	326721.16

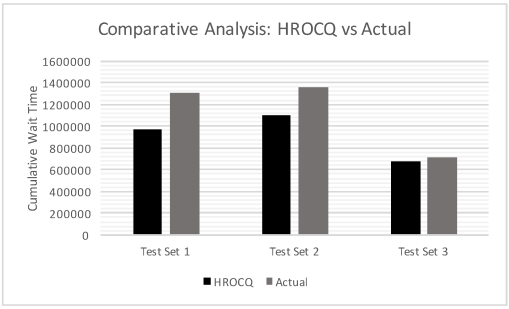

Figure 4: Waiting Time Comparison : *HROCQ* vs Actual

Next, we evaluate the overall performance of *HROCQ*. To compare the waiting time to serve incidents, we calculated the performance of *HROCQ* on our traffic accident data. The total number of responders in the model is taken as 25, which is the number of fire department vehicles available. We fix the waiting time upper bounds as 4 minutes, 8 minutes and 12 minutes for categories *D, B and A* respectively. Also, we set the value of s (max travel distance) as 3 miles, whose travel time is small with respect to service times (an assumption made in the model). Before presenting the results, we point out that in our dataset, each incident is marked with the time when the first responder arrives at the scene as well as the time when the last responder returns after servicing an incident. We assume that these are

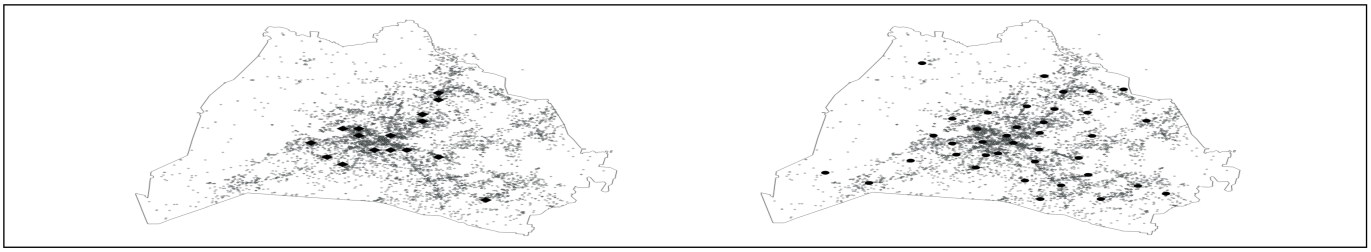

Figure 5: Location of Fire Stations: Optimized (Left) and Actual (Right)

the exact emergency responders needed, which is not always the case; consequently, it offers a *conservative* evaluation of the relative performance of our approach compared to actual response times. As an example, a common scenario for an accident is that the nearest police vehicle visits it first and the actual medical response vehicle arrives later. Similarly, while emergency responders return when their task is finished in the real world scenario, our model's validation must assume (due to unavailability of data) that the last vehicle to get back to service was the medical response vehicle, increasing wait times.

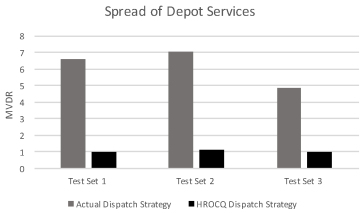

Figure 6: Spread of depots responding to a location

The mean travel time to incidents in our dataset is about 2.14 minutes. Although this is small in comparison to the service time, and is not explicitly contained in the model, we take into account the travel time in our validation to draw a fair comparison to the existing strategy used by the Fire Department. For each test dataset, we make a comparison based on total waiting time across all incidents and present the results in Fig. 4. We note that in all cases, the proposed formulation reduces the total service time, with an average improvement of about 16%, with the comparison heavily weighted against our approach. We highlight that this is a remarkable improvement, both in terms of performance and due to the importance of the incidents served, which often involve life-and-death scenarios. We show in Fig. 5 the difference in placement of actual responder placement stations and the station placed by our model. We observe that *HROCQ* focuses on areas with high density of incidents by placing more than one responder in these depots rather than distributing responders across the urban area. The low density areas are then covered by stations in high density areas as and when required.

It was observed in the urban area of concern that the incidents happening in the same location are serviced by responders from many distinct depots, which is likely a consequence of greedily assigning the closest responder to each incident. This naive approach makes responders from depots with high densities extremely busy and unavailable when accidents occur in their designated areas, and is viewed as a major concern by the fire department. We now consider how *HROCQ* addresses this issue compared to the actual responder policy. We point out the *HROCQ* implicitly tackles this problem by assigning grids to specific depots. However, not all potential demand locations are covered, and in such scenarios, other depots are called into action. To compare the results, we calculate for each grid i the average number of depots that respond to it; we denote this quantity by r^i. We show how structured the dispatch mechanism is by calculating $(\sum_i r_i)/(\sum_i i)$, which we refer to as the mean variation in depot response ($MVDR$). We see how $MVDR$ varies in practice versus *HROCQ* in Fig. 6. We can observe that the proposed approach provides a far more structured response mechanism.

6. CONCLUSION

We proposed a novel optimization problem that maximizes coverage of locations that need response while maintaining service time requirements by finding optimal location of response depots as well as the distribution of responders in them. In order to solve the optimization problem, we then extended prior work by proposing a novel greedy random adaptive algorithm that can accommodate the presence of multiple responders per demand node.

In order to predict incidents to aid the optimization model, we proposed learning a joint probability distribution over incident arrival and severity of incidents to tackle incident predictions heterogeneously based on priorities. We decomposed this distribution into a distribution over arrival times and a conditional distribution over incident severity given arrival times. To learn the former, we proposed a novel hierarchical clustering approach to extend the use of survival analysis to predict incidents by learning from data the spatial granularity of the model. We used a Multinomial Logistic Regression model to learn the distribution over incident severity. We showed from real traffic accident data from Nashville, TN, USA, that our algorithm results in significant reduction in waiting time for incidents and provides a structured and systematic dispatch policy. We also showed that our prediction model outperforms a state-of-the art incident prediction technique.

7. ACKNOWLEDGEMENT

This research was partially supported by the National Science Foundation (CNS-1640624, IIS-1526860, CNS-1238959), Office of Naval Research (N00014-15-1-2621), Army Research Office (W911NF-16-1-0069), and Air Force Research Laboratory (FA 8750-14-2-0180).

REFERENCES

[1] W. Ackaah and M. Salifu. Crash prediction model for two-lane rural highways in the ashanti region of ghana. *IATSS research*, 35(1):34–40, 2011.

[2] K. Balke, N. Chaudhary, C.-L. Chu, S. Kuchangi, P. Nelson, P. Songchitruksa, D. Swaroop, and V. Tyagi. Dynamic traffic flow modeling for incident detection and short-term congestion prediction: Year 1 progress report. Technical report, 2005.

[3] O. Berman, R. C. Larson, and S. S. Chiu. Optimal server location on a network operating as an m/g/1 queue. *Operations research*, 33(4):746–771, 1985.

[4] D. Böhning. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44(1):197–200, 1992.

[5] J. A. Bonneson and P. T. McCoy. Estimation of safety at two-way stop-controlled intersections on rural highways. *Transportation Research Record*, (1401), 1993.

[6] M. L. Brandeau and S. S. Chiu. A center location problem with congestion. *Annals of operations research*, 40(1):17–32, 1992.

[7] R. Church and C. R. Velle. The maximal covering location problem. *Papers in regional science*, 32(1):101–118, 1974.

[8] A. Cobham. Priority assignment in waiting line problems. *Journal of the Operations Research Society of America*, 2(1):70–76, 1954.

[9] D. R. Cox and D. Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.

[10] D. J. Eaton, M. S. Daskin, D. Simmons, B. Bulloch, and G. Jansma. Determining emergency medical service vehicle deployment in austin, texas. *Interfaces*, 15(1):96–108, 1985.

[11] J. Frantzeskakis, V. Assimakopoulos, and G. Kindinis. Interurban accident prediction by administrative area application in greece. *ITE journal*, 64(1):35–42, 1994.

[12] W. Gorr, A. Olligschlaeger, and Y. Thompson. Short-term forecasting of crime. *International Journal of Forecasting*, 19(4):579–594, 2003.

[13] S. Guha and S. Khuller. Greedy strikes back: Improved facility location algorithm. *Journal of Algorithms*, 31(1):228–248, 1999.

[14] S. L. Hakimi. Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations research*, 12(3):450–459, 1964.

[15] M. M. Haklay and P. Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, Oct. 2008.

[16] M. Harchol-Balter, T. Osogami, A. Scheller-Wolf, and A. Wierman. Multi-server queueing systems with multiple priority classes. *Queueing Systems*, 51(3-4):331–360, 2005.

[17] C. Ivaha, H. Al-Madfai, G. Higgs, and J. A. Ware. The dynamic spatial disaggregation approach: A spatio-temporal modelling of crime. In *World Congress on Engineering*, pages 961–966, 2007.

[18] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[19] D. G. Kendall. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *The Annals of Mathematical Statistics*, pages 338–354, 1953.

[20] V. Kiattikomol. Freeway crash prediction models for long-range urban transportation planning. 2005.

[21] S. Li. A 1.488 approximation algorithm for the uncapacitated facility location problem. *Automata, Languages and Programming*, 6756:77–88, 2011.

[22] M. J. Maher and I. Summersgill. A comprehensive methodology for the fitting of predictive accident models. *Accident Analysis & Prevention*, 28(3):281–296, 1996.

[23] F. Majzoubi. *A Mathematical Programming Approach for Dispatching and Relocating EMS Vehicles*. PhD thesis, University of Louisville, 2014.

[24] R. G. Miller Jr. *Survival analysis*, volume 66. John Wiley & Sons, 2011.

[25] 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 *Conference on Decision and Game Theory for Security*, 2016.

[26] M. Poch and F. Mannering. Negative binomial analysis of intersection-accident frequencies. *Journal of Transportation Engineering*, 122(2):105–113, 1996.

[27] Y. Qi, B. L. Smith, and J. Guo. Freeway accident likelihood prediction using a panel data analysis approach. *Journal of transportation engineering*, 133(3):149–156, 2007.

[28] P. T. V. Resende and R. F. Benekohal. Development of volume-to-capacity based accident prediction models. In *Traffic Congestion and Traffic Safety in the 21st Century: Challenges, Innovations, and Opportunities*, 1997.

[29] T. Sayed and F. Rodriguez. Accident prediction models for urban unsignalized intersections in british columbia. *Transportation Research Record: Journal of the Transportation Research Board*, (1665):93–99, 1999.

[30] V. Shankar, F. Mannering, and W. Barfield. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis & Prevention*, 27(3):371–389, 1995.

[31] M. B. Short, M. R. D'Orsogna, V. B. Pasour, G. E. Tita, P. J. Brantingham, A. L. Bertozzi, and L. B. Chayes. A statistical model of criminal behavior. *Mathematical Models and Methods in Applied Sciences*, 18(supp01):1249–1267, 2008.

[32] F. Silva and D. Serra. Locating emergency services with different priorities: the priority queuing covering location problem. *Journal of the Operational Research Society*, 59(9):1229–1238, 2008.

[33] P. Songchitruksa and K. Balke. Assessing weather, environment, and loop data for real-time freeway incident prediction. *Transportation Research Record: Journal of the Transportation Research Board*, (1959):105–113, 2006.

[34] C. Zhang, A. Sinha, and M. Tambe. Keeping pace with criminals: Designing patrol allocation against adaptive opportunistic criminals. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 1351–1359, 2015.

[35] M. Zhanli, Z. Yi, Y. Bozhong, and Z. Lei. Forecast of fire accidents based on grey-markov model [j]. *Engineering Sciences*, 1:019, 2010.

[36] K. G. Zografos, K. N. Androutsopoulos, and G. M. Vasilakis. A real-time decision support system for roadway network incident response logistics. *Transportation Research Part C: Emerging Technologies*, 10(1):1–18, 2002.

[37] K. G. Zografos, T. Nathanail, and P. Michalopoulos. Analytical framework for minimizing freeway-incident response time. *Journal of Transportation Engineering*, 119(4):535–549, 1993.