Modeling Local Ambulance Resource Scheduling

Emergent Research Forum (ERF) Paper

David M. Goldberg
Virginia Tech
goldberg@vt.edu

Patricia K. Garvey
Virginia Tech
pkgarvey@vt.edu

Long Xia
Virginia Tech
longxia1@vt.edu

Sukhwa Hong
Virginia Tech
sukhwa@vt.edu

Abstract

Effective management of ambulance staffing is a complex challenge that has profound implications for public health and safety. This study analyzes the spatial and temporal distribution of 911 calls and proposes a model for staffing ambulance shifts. We demonstrate the variability in important factors across localities and the implications of these factors for response times. We also show the optimal resource scheduling proposed by our model and the variability in these schedules across localities. These results offer a useful framework from which practitioners can analyze their locality and ensure effective management of emergency medical service resources.

Keywords

Geographic information systems, simulation, emergency medical services, ambulance scheduling, decision support systems.

Introduction

Ambulances are a critical class of emergency medical services (EMS) for ensuring public health and safety. The interval of time between a 911 call and the subsequent arrival of an ambulance on scene, known as the ambulance response time, substantially impacts morbidity and mortality rates of EMS patients (O’Keeffe et al. 2011). However, despite the importance of minimizing ambulance response times, these times vary greatly across geographic areas. Furthermore, accepted standards for ambulance response times also vary geographically, and many localities lament experiencing lackluster service. Ambulance response times are increasing over time in many areas, and response times for some 911 calls have exceeded 30 minutes (Talarico et al. 2015).

Most localities do not employ all of their ambulance fleet at once, but rather organize shifts throughout the day (O’Keeffe et al. 2011). Unlike many medical goods, the nature of EMS systems is such that services only provide utility to the extent that they are utilized; that is, the services cannot be inventoried. In managing EMS systems, this facet has important implications. First, if all EMS responders on shift are currently responding to other calls, then a new call will be set into a queue, increasing the response time. Second, if not enough calls are made to occupy many of the EMS responders, then ambulance response times are essentially unaffected, but resource allocation is inefficient. Therefore, EMS managers face the difficult task of balancing intra-day and inter-day demand for EMS services with responder shift management.

Extant research on capacity management of ambulance fleets has addressed several important factors, including location of EMS resources to manage the spatial distribution of 911 calls and sizing of EMS fleets to match gross call volume (Aubin 1992; Erdoğan et al. 2010). In this work, we contribute to the literature by modeling the probabilistic rather than deterministic nature of EMS systems. 911 calls behave stochastically in that the time between two calls is unknown but follows some probability distribution.
Similarly, the distribution of drive times affects both the speeds at which ambulances reach patients and how quickly those ambulances complete those calls and become available to service additional patients. We build a model to suggest optimal ambulance resource scheduling. After incorporating these factors, we compare optimal solutions across regions to demonstrate the manner in which optimal strategies vary across geographic areas.

**Literature Review**

Emergency medical services (EMS) are often the first responders in extreme situations, offering support services and preparing patients for transportation to appropriate medical facilities (CDC 2017). The EMS umbrella includes a variety of first responders, including volunteers, emergency medical technicians, paramedics, and firefighters (CDC 2017). In 2011, almost one million individuals were credentialed to work as EMS professionals, often risking their personal safety to protect and save others (CDC 2017). Generally, response time, or the time elapsed between a 911 call and the subsequent arrival of EMS on scene, is considered the most important performance measure in generating successful outcomes for patients (McLay and Mayorga 2010). EMS response time performance is often evaluated either as an *average response time* or as a *response time threshold* (RTT), the “number (or fraction) of calls that can be reached in a fixed timeframe” (McLay and Mayorga 2010). For example, RTT may measure the fraction of calls reached within 10 minutes. However, sources do not have universally accepted “preferred” values, varying from 8 minutes to 15 minutes depending upon the type of EMS and the locality (Concannon et al. 2009). The National Highway Traffic Safety Administration (NHTSA) (2009) distributes “recommended attributes and indicators for system and service performance” but does not include quantitative benchmarks.

According to McLay and Mayorga (2010), distributing ambulances in a pattern that achieved a 9-minute RTT created a more equitable spatial distribution of successful outcomes in a rural (typically more spatially dispersed) community. Though a great deal of research investigates inequality within public health services, research focusing on the effects of social factors on EMS-specific outcomes is limited with little consensus. A study conducted by David and Harrington (2010) found that there was “no economically or statistically significant difference in mean response time [under 8 min]” for Black or White individuals. However, after controlling for the response time, Black individuals were more likely to be deceased upon arrival, illustrating the complexity of analyzing EMS and social factors (David and Harrington 2010).

**Methodology**

We utilized a dataset of 23,749 unique 911 calls placed between March of 2015 and March of 2016 in the Baltimore, Maryland area. Each 911 call record contained fields pertaining to timestamps for the call and response, the reason for the call, and the address of the call. We retained the full dataset for use in our analysis. Using this data, we first constructed spatial references for each emergency. We obtained a spatial dataset of roads in the Baltimore, Maryland area and “geocoded” all 911 calls. We then associated each with the nearest fire station, from which an ambulance would be deployed.

By building a simulation model for each individual region, we may understand how each shift’s ambulance availability affects average response time. Figure 1 displays a schematic of our simulation model. For each of four distinct six-hour shifts during the day, we fit distributions to our dataset for 911 call inter-arrival times (IAT’s) and services times, including driving, pick up, and drop off. For each region, we find that distributions vary throughout the day. 911 call IAT’s tend to follow beta distributions, but drive times tend to follow triangular distributions. All distributions fit at the 0.05 level. Next, according to these distributions, we simulated the conditions of each shift over a 100-day period while modulating the available ambulances in each shift. We recorded average response times at each level of availability. We also modeled the capability for EMS systems to “borrow” capacity from neighboring regions at peak capacity; however, doing so is only a last resort as it increases response time.
Figure 1: EMS simulation model

Given estimated average response times obtained from our simulation model, we construct a mixed-integer programming model to optimize allocations of EMS responders to shifts. We measure the total efficacy of the EMS system in terms of the average response time yielded by the simulated allocation procedure.

We define $t_{as}$ as the average response time for shift $s$ given an allocation of $a$ ambulance teams, where $a = 1, \ldots, n$ and $s = 1, \ldots, m$. Furthermore, let $x_{as}$ be defined as a binary variable equal to 1 if $a$ ambulance teams are allocated to shift $s$ and 0 otherwise. Next, we define $u_s$ as the relative usage of ambulances (proportion of 911 calls) on shift $s$. Our objective, or the average response time of an ambulance allocation scheme as weighted by usage, is expressed in (1). Each EMS manager is constrained in that they may only allocate some number of ambulance teams in a day due to budgetary constraints, employee off-time, etc. Let $N$ equal the maximum total ambulance team allocation. Our constraint on the total number of allocated ambulance teams is expressed in (2). Finally, each shift can only receive one allocation variable (that is, 5 ambulance teams can be allocated, but not 3 ambulance teams and 2 ambulance teams) as expressed in (3).

As such, the following represents our complete mixed-integer programming model.

$$
\begin{align*}
\text{Minimize} \quad Z &= \sum_{s=1}^{m} u_s \sum_{a=1}^{n} x_{as} t_{as} \\
\text{Subject to:} \quad \sum_{s=1}^{m} \sum_{a=1}^{n} a(x_{as}) &\leq N \\
\sum_{a=1}^{n} x_{as} &= 1 \forall s
\end{align*}
$$

Results

We focused our analysis on two neighboring regions, termed ND (West) and NE (East) by the local EMS districting, with each region feeding into a different Baltimore hospital (see Figure 2). Although located adjacent, the NE region experienced a far greater density of 911 calls than the ND region. Interestingly, 911 call density appeared strongly related to socioeconomic and demographic factors. A regression model using density of crime and the percentage of households earning under $60,000 per year to predict density of EMS 911 calls generated a multiple $R^2$ of 0.93, or a highly explanatory linear fit.
Figure 2: Spatial comparison of ND (West) and NE (East) regions

For each region, we found that as we increased the value of $N$ in our model, or the total number of ambulances available in a day, the optimal average response time decreased. However, at each tested value of $N$, the NE region had a greater optimal average response time than the ND region. It also appeared that the NE region was far more sensitive to the value of $N$; that is, each additional ambulance available resulted in a greater reduction in optimal average response time in the NE region than in the ND region. We display these findings in Figure 3.

Figure 3: Ambulance availability versus optimal average response time

Figure 4: Optimal allocation of 18 ambulances to shifts by region
The results of our model also demonstrate some fundamental differences in optimal shift allocation. In Figure 4, we show an example in which each region is constrained to 18 ambulance shifts in a day. First, the needs of the different regions suggest different optimal allocations of resources. While the NE region requires an enormous allocation of resources in Shift 4, the ND region actually has the greatest need in Shift 3. Second, the optimal allocation for the ND region is more uniform throughout the day. The trend that the NE region required the most ambulances during the night shift whereas the ND region required the most ambulances during the afternoon shift was consistent across different values of $N$.

Limitations

Our study is subject to several limitations. We focused our study on the Baltimore, Maryland area, although future works may also determine whether the same conclusions hold in different localities. We focused on ambulance response times, but we acknowledge that EMS also include additional services that fell beyond the scope of our paper. Our modeling in this paper is largely based on optimizing the average ambulance response time, but this is certainly not the only metric for assessing EMS effectiveness. In future research, we intend to contrast the average response time measure with RTT and examine any potential differences when optimizing for these measures across localities. In future revisions, the model may also consider emergency room availability, as this factor could allow for rerouting in the event of overcrowding.

Conclusion

From the perspective of a city-wide EMS planner, our results indicate which regions will benefit most from additional resources. The ND region’s optimal response times outperform the NE region’s optimal response times at each allocation. Thus, even at equal allocations, the outcomes are not equitable. However, each additional ambulance allocated to the NE region can provide a more substantial benefit to that region’s response time. The two EMS regions have distinct needs as unique geographic and demographic regions. The predictability of EMS needs based on crime and demographics could allow additional regions to strategically plan resource allocation using our technique and findings. In addition to the demand for EMS at each time of day, we consider the distribution of calls and the distribution of times required to service patients. Our findings further demonstrate the spatial disparity across regions and the corresponding manner in which allocations of resources have different effects across those regions.

REFERENCES

CDC. 2017. "Emergency Medical Services Workers."
NHTSA. 2009. "Emergency Medical Services Performance Measures: Recommended Attributes and Indicators for System and Service Performance."