An Efficient Data-driven Approach for Emergency Medical Services

Lavanya Marla
University of Illinois at Urbana-Champaign

Collaborators: Prof. Ramayya Krishnan (CMU), Dr. Yisong Yue (Caltech)
Talk outline

• Ground Realities for EMS in Emerging Economies
• Data-driven Simulation
• Mathematical Formulations
• Results
• Ongoing and Future Work
EMS in Emerging Economies: Ground Realities

- Highly resource constrained
  - 75M people, 750 ambulance bases (AP)
- Large-scale
- Prior to this operator, no central ambulance provider
  - Hospital ambulances, taxis
- Public-private partnership
  - No fees charged for service (paid by state)
- Cell-phone-based communications
- DATA COLLECTION
Service Area, Bases and Calls

Data-driven Models for Ambulance Location and Redeployment
Inefficiencies in spite of sophisticated models

Existing literature: medium-scale

Multiple heterogeneous resources – ALS, BLS

‘Discrete’

Non-linearities between survival and service time

Network effect – Propagation effect of ambulances in use

Source:
Challenges in EMS in Emerging Economies

- Traffic congestion
  - Public acceptability
    - Clear traffic for ambulance
- Competition with ad-hoc networks
  - Decreases utilization of ambulances
- No real-time position availability
- New cities
  - New traffic patterns
  - New modes of transport
Key Questions of Interest

• How can performance be improved using existing resources (e.g., ambulances)?
  – Static allocation?
  – Dynamic redeployment?
  – Change dispatch policy?

• How to characterize the state of the system?
  – Metrics

• How to model how the system is affected by current allocation and dispatching policy?

• Can a decision support tool be developed to answer these questions?
Key concepts

- Network consists of ambulances located at bases
- Each base’s coverage area is approximately a set of grids around it
- Each call has a priority queue of bases
  - Best served by first base in queue
- A served call consists of:
  - ambulance arriving from its base to the scene
  - taking the patient to a hospital
  - returning to (same/another) base
Design Principles

• Do not add extra bases or ambulances than those determined by the operator
  – Logistical challenges

• Consistency with current dispatching model
  – Calls served FCFS
  – Assign nearest free ambulance available
  – Priority queue for ambulances: learn from data logs (congestion implicit)

• Derive congestion information from data logs
Contributions

**Models**
- Problem-driven, data-driven models
- Problem structure, solution quality, tractability

**Algorithms**
- Static allocation of ambulances
- Dynamic redeployment of ambulances

**Applications**
- Emergency Medical Systems
- Disaster response, humanitarian logistics
- Facility location
Our approach

- Use data collected by the operator (call logs)
  - Capture time-dependent travel times
  - Optimize for metrics like preparedness, survival probabilities
  - Scalability

- Learn from the system data

- Build a solution that is faithful to the data (call logs)

Goal 1: Efficient and robust ambulance allocation

Goal 2: Dynamic repositioning policy
Solution Approach Summary

Data-driven Models for Ambulance Location and Redeployment
Talk outline

• Ground Realities for EMS in Emerging Economies
• Data-driven Simulation
• Mathematical Formulations
• Results
• Ongoing and Future Work
Modeling Concept: Chain Formation

Dependency Graph
Modeling Concept: Chain Formation
Modeling Concept: Chain Formation

25 min

18 min
(12 minutes saved)

15 min
(30 minutes saved)

25 min
(no vehicle assigned previously)

Changed set of dependencies for new allocation
Modeling Concept: Dependency Chains

*Given:* One ambulance each at b1 – b5; dispatch policy; request (call) set

- Call 1: \{B1, B2, ..\}
- Call 2: \{B2, B3, ..\}
- Call 3: \{B1, B2, B3, ..\}
- Call 4: \{B1, B2, B3, B4\}
- Call 5: \{B3, B5\}
- Call 6: \{B3, B4, B5, B1\}
- Call 7: \{B1, B5, B4\}
Simulation Framework to Compute Allocation Cost

Simulation approach to evaluate ambulance-to-base allocations

- Simulate Dispatch Officer assigning ambulances to calls
- Simulate response times and outcomes
- Data-driven approach (based on actual call logs)

\[ L_{\downarrow R}(A) = \sum_{r \in R^{\uparrow \downarrow}} L_{\downarrow r}(y_{\downarrow r}, o_{\downarrow r}); \{y_{\downarrow r} = \text{ambulance allocation at } o_{\downarrow r}\} \]

Based on call logs we can model:

- Call congestion patterns
- Chains and other long-range system effects
- Utilization of various base locations
Breaking dependencies improves service

1 ambulance each at B₁ – B₅
Add ambulance to B₁
Add ambulance to B₂
Modeling Abandonment

- Customer calls multiple service providers, limited patience for waiting
  - Choose the one which arrives first

Abandonment model

$$\log \frac{\text{Prob(abandoned)}}{\text{Prob(not abandoned)}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

- $x_1 = 1$ if request from a rural area
- $x_2 = \text{base-to-scene} \ast \text{if (urban, peak hour)}$
- $x_3 = \text{base-to-scene} \ast \text{if (rural, peak hour)}$
Talk outline

• Ground Realities for EMS in Emerging Economies
• Data-driven Simulation
• Mathematical Formulations
• Results
• Ongoing and Future Work
Data-driven Models for Ambulance Location and Redeployment

Mathematical Formulations

- $R = \text{Request set}$
- $G_R = \text{dependency graph}$
- $L_R(A) = \text{total cost of allocation A for request R, from evaluating } G_R$

Utility of Static Allocation: $F_{\downarrow R}(A) = L_{\downarrow R}(\emptyset) - L_{\downarrow R}(A)\uparrow$

Static Allocation Objective: $^{\wedge}A\in\mathcal{M}(A) : |A| \leq K \arg\max_{A} \downarrow F_{\downarrow R}(A)$

- $s_t = \text{state of the system at time } t$
- $W_{st} = \text{currently free allocation}$

Dynamic Redeployment Utility: $F_{\downarrow R}(\pi) = E(\downarrow s_1, \ldots, s_T) E[\sum_{t=1}^{T} \downarrow F(\pi(s_{\downarrow t}) | s_{\downarrow t})] \uparrow$

Dynamic (myopic) Redeployment: $^{\wedge}A\in\mathcal{M}(A,W\downarrow s_{\downarrow t}) : |A| \leq W \downarrow s_{\downarrow t} \arg\max_{A} \downarrow F_{\downarrow R \downarrow t}(A | s_{\downarrow t})$
Claim: F is submodular?

- \( F(A) \) is submodular \iff \n
\[ \forall A \subseteq B, \forall a, \delta \downarrow F aA \geq \delta \downarrow F aB \]

Gain of ambulance a only decreases with larger allocations

Rare case in data but happens nonetheless!
Simulation-Optimization (Greedy algorithm)

Goal: Allocate $N$ ambulances among $M$ bases

- Number ambulances added $< N$?
- Test adding an ambulance at location $b \in B$
  - Simulate calls: $F(A + a_1)$
  - Simulate calls: $F(A + a_2)$
  - Simulate calls: $F(A + a_{|B|})$
- Find location $b^*$ with most improvement in $F(A)$
- Add ambulance at $b^*$

Running time = $NB * O($Simulator$)$
Non-submodularity of $F$ (static and dynamic)

- If monotone submodular, greedy algorithm returns solutions that achieve

$$F(A) \geq (1 - 1/e) \text{OPT}$$

- Approximate monotonicity:

$$\forall A, \forall a, \delta \downarrow F aA + \epsilon \downarrow m \geq 0$$

- Approximate submodularity:

$$\forall A \subseteq B, \forall a, \delta \downarrow F aA + \epsilon \downarrow s \geq \delta \downarrow F aB$$
Theoretical Guarantees and Bounds (1)

**Theorem:** Let $F$ be approximate submodular with additive violation and approximate monotone with additive violation. Let $A_1, ..., A_k$ denote the intermediate solutions of Greedy as it optimizes on $F$ for a budget of $K$ ambulances, the greedy algorithm produces an allocation $A$ that satisfies

\[ \mathcal{F}(A) \geq \left(1 - \frac{1}{e}\right)\OPT - \sum_{l=1}^{k} \left(\epsilon_{\downarrow S} + \epsilon_{\downarrow m}\right)(A_{\downarrow l} - 1) \]

- Need to compute $\epsilon_{\downarrow S}$ and $\epsilon_{\downarrow m}$
- Integer program written based on dependency chain model
Utility of an *Omniscient dispatcher* \((G)\):

\[
s.t. \quad = 1
\]

**Theorem:** The objective, \(G\), as measured by simulating an omniscient dispatcher, is monotone submodular. Furthermore, for and \(A\) and \(R\), we have \(G\). Also, for any \(A\) with \(|A|=K\),
Talk outline

• Ground Realities for EMS in Emerging Economies
• Data-driven Simulation
• Mathematical Formulations
• Results
• Ongoing and Future Work
Cost Function F

• \( L(y) = \begin{cases} 0 & \text{if service time} \leq 15 \text{ min} \\ 1 & \text{if service time} \leq 30 \text{ min} \\ 2 & \text{if service time} \leq 60 \text{ min} \end{cases} \) otherwise
Data-driven Models for Ambulance Location and Redeployment

**Metrics and Static allocation**

\[
L_{\downarrow r}(y) = \begin{cases} 
0 & \text{if service time} \leq 15 \text{ min} \\
1 & \text{if service time} \leq 30 \text{ min} \\
2 & \text{if service time} \leq 60 \text{ min} \\
\text{otherwise} & 
\end{cases}
\]

*Result:* Greedy solution improves upon baseline allocation of operators.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Improvement over baseline allocation</th>
</tr>
</thead>
<tbody>
<tr>
<td># Calls w/ Base-to-scene &lt; 15 min</td>
<td>6.1% (increase)</td>
</tr>
<tr>
<td># Calls w/ Base-to-scene &lt; 30 min</td>
<td>3.4% (increase)</td>
</tr>
<tr>
<td># Ambulances Busy</td>
<td>42.7% (decrease)</td>
</tr>
<tr>
<td># Calls serviced by primary base</td>
<td>9.4%</td>
</tr>
</tbody>
</table>
Bounds

**Result:** Greedy solution close to bound from optimal dispatch allocation => ‘close’ to optimal
Dynamic repositioning

- Under high demand regions
  - ‘System stress’
- Re-position ambulances in real-time
  - Move free ambulances from ‘home’ base to nearby bases
  - Waiting on street corners
Dynamic repositioning vs. static allocation

Result 1: More redeployment produces better service

Result 2: Most impacted metric = number of calls served

![Graphs showing fraction with base-to-scene <= 15 min and 30 min, 45 min, 60 min, 75 min redeployment intervals.]

![Graphs showing fraction requests not served with static and 30 min, 45 min, 60 min, 75 min redeployment intervals.]

Data-driven Models for Ambulance Location and Redeployment
Value of Dynamic Repositioning

- Value in dynamic repositioning compared to static

<table>
<thead>
<tr>
<th>Look-ahead = 45 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calls served with base-scene &lt;15 min</td>
</tr>
<tr>
<td>Calls with base-scene &lt;30 min</td>
</tr>
<tr>
<td>Calls served by primary base</td>
</tr>
<tr>
<td>Calls not served (vehicles busy)</td>
</tr>
</tbody>
</table>

- Most impacted metric: calls served
- Value higher when greater flexibility in repositioning – example: more often, more ambulances allowed to be repositioned
Robustness under congestion fluctuations

**Result:** Even under variability in demands and travel times, the *Greedy* solution shows improvement over default.

<table>
<thead>
<tr>
<th></th>
<th>0% increase in demand</th>
<th>10% increase in demand</th>
<th>15% increase in demand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base-to-scene &lt;15 min</td>
<td>6.1%</td>
<td>5.7%</td>
<td>5.0%</td>
</tr>
<tr>
<td>Base-to-scene &lt;30 min</td>
<td>3.4%</td>
<td>3.5%</td>
<td>3.8%</td>
</tr>
<tr>
<td>Served by primary base</td>
<td>9.4%</td>
<td>10.1%</td>
<td>10.3%</td>
</tr>
<tr>
<td>Calls not served</td>
<td>-42.7%</td>
<td>-36.2%</td>
<td>-33.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0% increase in travel time</th>
<th>10% increase in travel time</th>
<th>15% increase in travel time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base-to-scene &lt;15 min</td>
<td>6.1%</td>
<td>5.7%</td>
<td>4.9%</td>
</tr>
<tr>
<td>Base-to-scene &lt;30 min</td>
<td>3.4%</td>
<td>3.9%</td>
<td>3.7%</td>
</tr>
<tr>
<td>Served by primary base</td>
<td>9.4%</td>
<td>10.4%</td>
<td>10.5%</td>
</tr>
<tr>
<td>Calls not served</td>
<td>-42.7%</td>
<td>-35.0%</td>
<td>-31.5%</td>
</tr>
</tbody>
</table>

*Measured using simulation on independent data, for a period of one month*
Result: Optimality gap remains similar in the case of abandonment => ‘close’ to optimal

Bounds with abandonment

Result: Optimality gap remains similar in the case of abandonment => ‘close’ to optimal
Solutions with abandonment

**Result:** Improvements with respect to all metrics

![Bar chart showing improvements](chart.png)
Opportunity cost of abandonment

• Abandoned calls add inefficiency to the system
• Ambulance could have served another customer (with a better service level)
• How much is lost due to abandoned calls?
  – Find optimal allocation when abandoned calls existed
  – Remove abandoned calls and measure impact of optimal allocation
• 12% calls abandoned in data set
  – ~6% improvement when abandoned calls ignored
  – Remaining 6% of calls do not reduce service level
Opportunity Cost of Abandonment

<table>
<thead>
<tr>
<th>Default</th>
<th>Static</th>
<th>Dynamic 30 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.33</td>
<td>0.32</td>
<td>0.33</td>
</tr>
<tr>
<td>0.325</td>
<td>0.32</td>
<td>0.325</td>
</tr>
<tr>
<td>0.32</td>
<td>0.32</td>
<td>0.325</td>
</tr>
<tr>
<td>0.315</td>
<td>0.315</td>
<td>0.325</td>
</tr>
<tr>
<td>0.31</td>
<td>0.315</td>
<td>0.325</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Default</th>
<th>Static</th>
<th>Dynamic 30 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.71</td>
<td>0.72</td>
<td>0.73</td>
</tr>
<tr>
<td>0.72</td>
<td>0.73</td>
<td>0.74</td>
</tr>
<tr>
<td>0.73</td>
<td>0.74</td>
<td>0.75</td>
</tr>
<tr>
<td>0.74</td>
<td>0.75</td>
<td>0.76</td>
</tr>
<tr>
<td>0.75</td>
<td>0.76</td>
<td>0.77</td>
</tr>
<tr>
<td>0.76</td>
<td>0.77</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Default Static Dynamic 30 min

<table>
<thead>
<tr>
<th>Default</th>
<th>Static</th>
<th>Dynamic 30 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.93</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>0.98</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>0.99</td>
<td>0.99</td>
<td></td>
</tr>
</tbody>
</table>

- Removing abandoned calls
- Including abandoned calls

Data-driven Models for Ambulance Location and Redeployment
Takeaways

• Static allocation provides good results compared to baseline operations.
• More repositioning makes more ambulances available where needed; covers requests better
  – Reposition often if idle travel cost is low
• Greedy algorithm is quick, particularly for dynamic redeployment (<~10s)
• Solutions from our algorithm are robust
• Opportunity cost of abandonment is about 50% that of fraction of abandoned calls
Talk outline

- Ground Realities for EMS in Emerging Economies
- Data-driven Simulation
- Mathematical Formulations
- Results
- Ongoing and Future Work
New/needed technology: Traffic models

- Existing routes
  - Currently use data-driven models for traffic congestion capture
  - Allows to extrapolate data for routes taken in past

- New routes?
  - Crowdsource/obtain traffic information from other ambulances
  - Communication between ambulances to share traffic data
New/needed technology: Human behavior models

- ‘Conflict’ between existing ad-hoc networks and the operator’s network
- Customer calls multiple service providers
  - Choose the one which arrives first
- Modeled higher abandonment in select urban areas
- How to improve ambulance utilization?
  - Better dispatching models?
- What system can lead to improved social welfare?
Robust and Dynamic Approaches for Evolving Infrastructure

Robust Planning

Dynamic Reconfiguration

Repair/Recovery

Optimization

Simulation

Prediction

Real-time Information

Environmental Impact

Data-driven Models for Ambulance Location and Redeployment
THANK YOU!