Inventory Is People: How Load Affects Service Times in Emergency Response

1.1 Introduction

In many service systems, the most important inventory in the system is people. Most queueing models assume that people and objects are interchangeable. While most people recognize the significant difference, the academic literature on how that difference manifests itself is still young. We analyze a large data set of emergency medical service (EMS) responses in the city of Calgary to explore and understand how the inventory of patients in the system adapts as a function of load.

If variability is the enemy of managers, then people must be their bane. The problem with people is they are hard to predict. When your inventory consists of jelly beans (Sox et al., 1991), you can pretty much expect that the inventory will not care how many other jelly beans are in the immediate vicinity or how long they’ve been in any particular box. That’s not true with people. People react to the situation around them, and the range of reactions varies greatly. If the line is long, the reaction of the customer at the front of the line will be quite different from the server, which may be completely different from the person at the back of the line. Many people would postulate that emergency medicine is a science, and treatment should not vary based on the system load. Everyone should receive the same treatment. Many paramedics would argue that they are not scientists but caregivers. They would say good caregivers treat the entire situation and the situation includes how busy the system is.

It is this variability, complexity, and range of reactions that makes it difficult to include people in mathematical models. In 1998, Joe Thomas, working with students Ken Schultz and Dave Juran and his colleagues John Boudreau and John McClain (Schultz et al., 1998), used an experiment to show that the predictions from standard queueing models failed to predict the behavior of a simple production system. In their experiment, the difference between a high and low inventory
line changed idle time by 19%. However, output remained virtually unchanged. Following up on this analysis, Schultz [1997] Chapter 4) showed that most of
the difference between model predictions and actual performance was due to the
assumption that workers were independent of each other. The human operators
changed their behavior as a result of the people around them and changes in the
state of the system. Five years later, with the same colleagues and joined by Wally
Hopp, Dr. Thomas wrote an influential paper in the Behavioral Operations Man-
agement literature [Boudreau et al. 2003]. They argued that most mathematical
models, in order to preserve tractability, assume people are unimportant, deter-
ministic, independent, stationary, or not part of the product.

Since then, many researchers have investigated how people are important,
stochastic, interdependent, and non-stationary. In 2010 Drs. Schultz, Schoen-
herr, and Nemhbad showed that interdependent workers on a radio line changed
their work pace closer to the average pace of those around them [Schultz et al.][
2010]. In 2010, Gans et al. showed that non-stationary call center agents reduce
their average call time by 8.3% with every doubling of the cumulative number of
calls handled. Gans et al. [2010] also showed that agents were stochastic, hav-
ing widely different initial average call times, and that taking agent learning into
account is important—ignoring learning caused simulated average customer wait
to deviate by more than 20% in almost one third of their simulation experiments.
Green et al. [2013] found that interdependent nurses are more likely to be absent
when a hospital unit is understaffed, and that ignoring this interdependence in
planning staffing levels could increase costs by 2-3%.

This large collection of empirical work is analyzed by Delasay et al. [2015].
They show that there is no simple answer to the question “what is the load effect
on service times?” The relationship is dependent on the reactions of the cus-
tomers, the servers, and the networks to the system load, how long the load has
been strong, and the changeover in type of service. They show 22 different mech-
anism through which load type affects servers by changing either the speed or the
amount of work. This conceptual model, the load effect on service times (LEST)
framework, can be used to frame an investigation to predict how processing times,
and therefore inventory, will respond to changes in load.

The effects of load on service times are important. Classical queueing theory
tells us that the length of the line depends on the average and variability of the
processing times. Fig. 1.1 shows the relationship between average service times
(± 1 standard deviation) and the EMS system load, measured as the percentage of
busy ambulances, in the city of Calgary. It certainly appears that the service time
average and variability depend on load. In this paper, we explore why one EMS
crew’s service time would be non-stationary and dependent on what other crews
are doing. We have a long and rich literature exploring the effects of service time on server utilization in queueing systems. Fig. 1.1 highlights the need to explore the opposite effect—that of load on service times.

We intend to add to the emerging literature on the effects of load on service time, using a large data set of EMS calls in the city of Calgary. The effect of load is important for capacity planning: The average time to serve a patient can vary by up to 26%, depending on the system load (see Fig. 1.1). For loads below 0.75, service slows down as load increases, but for loads above 0.75, service speeds up with load. The inverted-U relationship between load and service time that we see in Fig. 1.1 is not uncommon (Tan and Netessine, 2014; Batt and Terwiesch, 2014). The causes of the inverted-U relationship depend on the context, however, and we find that the causes in the EMS context differ from those identified in other settings.

![Fig. 1.1 Load-dependent EMS service times](image)

We identify mechanisms that we believe are operative in this setting. Some of these mechanisms have been analyzed in previous empirical research. We show their effect in new settings and under different conditions. As a field, we must continue to document how, when, why, and how much these mechanisms are activated. We also identify new mechanisms as a result of applying the LEST framework to anticipate human reactions in an EMS system. In this way, we demonstrate how the LEST framework can be used to explore and understand a situation and reduce, at least by a little, the unpredictability of the human component in queueing systems.
1.2 The LEST Framework

The load effect on service times framework describes mechanisms that link load effects to service time as a path in the network shown in Fig. 1.2. Each path begins with a load characteristic (changeover, which involves the load changing from zero to a positive value; load, which refers to the instantaneous value of the load; or extended load, which incorporates the history of the load), that influences a system component (a server, the network, or a customer) in a way that impacts a service time determinant (work content or service speed). Delasay et al. (2015) identify 21 such mechanisms, which are listed in Table 1.1. As a preview of our results, Table 1.5 lists 12 mechanisms that we identify as being operative in the EMS setting. Of the 12, 11 are listed in Table 1.1 but one mechanism (task transfer) is new. Another four mechanisms (network arrangement, geographical dispersion, geographical speedup, and network chaos) were first identified by Delasay (2014). The remaining eight mechanisms have been previously documented in non-EMS settings, as discussed in Delasay et al. (2015).

![Fig. 1.2 The LEST framework](image)

1.3 EMS Services

An EMS response to a medical emergency begins when a patient or bystander calls 911. An emergency medical dispatcher (EMD) answers and triages the call through a systematic medical interrogation to determine the patient’s condition acuity. After gathering the required information, including the call address and the type of required equipment, the EMD dispatches an appropriately equipped ambulance to the incident scene.

The EMS ambulance service time begins when the ambulance receives the
Table 1.1 Mechanisms

<table>
<thead>
<tr>
<th>Load characteristic</th>
<th>Server</th>
<th>Network</th>
<th>Customer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Changeover</td>
<td>Work content Physical setup ↑ Forgetting ↑</td>
<td>Work content Network arrangement ↓</td>
<td>Work content Customer early task initiation ↓</td>
</tr>
<tr>
<td>Load</td>
<td>Work content Task reduction ↓ Engagement ↑ Server early task initiation ↓ Cognitive sharing ↑ Workload smoothing ↓</td>
<td>Work content Downstream system congestion ↑ Resource sharing ↑ Geographical dispersion ↑</td>
<td>Work content Bounce back ↑</td>
</tr>
<tr>
<td>Extended Load</td>
<td>Work content Service cancelation ↓ Service speed Social speedup pressure ↓ Social loafing ↑</td>
<td>Work content Network chaos ↑</td>
<td>Work content Deterioration ↑</td>
</tr>
</tbody>
</table>

dispatch notification and it includes the five time intervals in Fig. 1.3:

- Chute time \(T_{\text{Chute}}\): The preparation and boarding time for the ambulance crew after receiving the dispatch notification.
- Travel time \(T_{\text{Travel}}\): From the dispatch location to the incident scene.
- Scene time \(T_{\text{Scene}}\): The time that ambulance crew are on scene providing medical care to a patient.
- Transport time \(T_{\text{Transport}}\): From the scene to a hospital, if the patient requires hospital transportation.
- Hospital time \(T_{\text{Hospital}}\): The offload time to transfer the patient to the emergency department (ED) after arriving to the hospital.

Ingolfsson (2013) reviews the literature on EMS planning and management, including empirical work investigating the five time intervals. We know of no previous work that focuses on mechanisms through which load impacts EMS service times, but Alanis et al. (2013) includes graphs that illustrate the association
between EMS load and the five time intervals.

1.3.1 Data, Explanatory Variables, and Descriptive Statistics

Our analyses are based on a dataset of 108,423 call records for the EMS system of the city of Calgary, Canada, in 2009. We focus on 92,893 calls for which an ambulance was dispatched. The information for a call includes: (1) time stamps for the events in Fig. 1.3 generated by the EMD and paramedics, (2) coordinates for the ambulance dispatch location, call address, and hospital location, (3) call priority numbers assigned by the EMD (a number from 1 to 7 with 1 being assigned to Delta/Echo or the most critical priority calls), and (4) the number of busy and scheduled ambulances at the moment of call arrival. We use this information to extract the following variables for each call:

- The length of the EMS time intervals: $T_{\text{Chute}}$, $T_{\text{Travel}}$, $T_{\text{Scene}}$, $T_{\text{Transport}}$, and $T_{\text{Hospital}}$
- Response and service times for no-transport ($T_{\text{Service, No transport}}$) and transport ($T_{\text{Service, Transport}}$) calls:
  
  \[
  T_{\text{Response}} = T_{\text{Chute}} + T_{\text{Travel}} \\
  T_{\text{Service, No transport}} = T_{\text{Chute}} + T_{\text{Travel}} + T_{\text{Scene}} \\
  T_{\text{Service, Transport}} = T_{\text{Chute}} + T_{\text{Travel}} + T_{\text{Scene}} + T_{\text{Transport}} + T_{\text{Hospital}}
  \]
- Travel ($D_{\text{Travel}}$) and transportation ($D_{\text{Transport}}$) distances, calculated based on shortest paths in the Calgary road network
- Average travel speed ($S_{\text{Travel}}$) and average transportation speed ($S_{\text{Transport}}$), calculated as $S_{\text{Travel}} = D_{\text{Travel}} / T_{\text{Travel}}$ and $S_{\text{Transport}} = D_{\text{Transport}} / T_{\text{Transport}}$
- Number of busy ambulances at dispatch ($N_{\text{BDispatch}}$), scene arrival ($N_{\text{BArrival}}$), scene departure ($N_{\text{BDeparture}}$), and hospital arrival ($N_{\text{BHospital}}$). We also compute the average number of busy ambulances $\overline{NB}$ at the four instants.
- Indicator $I_{\text{Transport}}$ for whether the patient is transported to hospital
- Indicator $I_{\text{Urgent}}$ for life-threatening calls (Delta/Echo priority)
- Indicator $I_{\text{Siren}}$ for use of lights and sirens during travel to scene (corresponds to priority levels 1-4)
- Indicator $I_{\text{Changeover}}$ for an ambulance that responds from standby mode, that is, a regular service. We identify a service as regular (extended) if the time between the start of the service and the finish of the previous service by the same crew is more (less) than 10 minutes
- EMS load at dispatch ($L_{\text{Dispatch}}$), scene arrival ($L_{\text{Arrival}}$), scene departure ($L_{\text{Departure}}$), and hospital arrival ($L_{\text{Hospital}}$), calculated as the proportion of busy ambulances:

$$L_i = \frac{NB_i}{NS}, i = \text{Dispatch, Arrival, Departure, Hospital},$$

where $NS$ is the number of scheduled ambulances. We also compute the average load $\overline{L}$ at the four instants.
- Extended load at the server level, $EL_{\text{Server}}$, quantified as the amount of time the responding ambulance has been continuously busy since their last changeover. In calculating $EL_{\text{Server}}$, we consider that an ambulance is continuously busy from one call to the next if the new dispatch notification is received in less than 10 minutes of the finish of the previous service.

Tables 1.2-1.4 provide descriptive statistics for these variables. The average durations for the time intervals range from 0.96 minutes (chute time) to 69.38 minutes (hospital time). The average total service time is 44.49 minutes for no-transport calls and 117.50 minutes for transport calls. The proportion of transport calls is 58%, so the average service time for all calls is 86.84 minutes.

### 1.4 Hypotheses and Results for EMS Service Time Intervals

In this section, we employ the LEST framework ([Delasay et al., 2015]) to systematically investigate the mechanisms that cause each EMS service time interval to depend on EMS load. We consider each cell of the LEST framework and we identify how the corresponding system component and load characteristic are manifested for EMS service time intervals. Armed with this understanding, we propose mechanisms (previewed in Table 1.5) that relate the system component and the load characteristic to a service time interval. For each time interval, we aggregate the proposed mechanisms to generate hypotheses. We use multiple linear regression models to test our hypotheses.
Table 1.2  Descriptive statistics for service time intervals (min.)

<table>
<thead>
<tr>
<th>Measure</th>
<th>$T_{	ext{Chute}}$</th>
<th>$T_{	ext{Travel}}$</th>
<th>$T_{	ext{Scene}}$</th>
<th>$T_{	ext{Transport}}$</th>
<th>$T_{	ext{Hospital}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.96</td>
<td>7.13</td>
<td>28.59</td>
<td>16.57</td>
<td>69.38</td>
</tr>
<tr>
<td>Median</td>
<td>0.73</td>
<td>5.78</td>
<td>23.35</td>
<td>14.57</td>
<td>56.97</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>3.84</td>
<td>6.87</td>
<td>30.76</td>
<td>11.01</td>
<td>51.04</td>
</tr>
<tr>
<td>Coefficient of variation</td>
<td>4.00</td>
<td>0.96</td>
<td>1.08</td>
<td>0.66</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 1.3  Descriptive statistics for service time (min.)

<table>
<thead>
<tr>
<th>Measure</th>
<th>$T_{	ext{Service, No Transport}}$</th>
<th>$T_{	ext{Service, Transport}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>44.49</td>
<td>117.50</td>
</tr>
<tr>
<td>Median</td>
<td>34.58</td>
<td>105.9</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>44.57</td>
<td>55.50</td>
</tr>
<tr>
<td>Coefficient of variation</td>
<td>1.06</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Table 1.4  Descriptive statistics for explanatory variables

<table>
<thead>
<tr>
<th>Measure</th>
<th>$N_S$</th>
<th>$N_B$</th>
<th>$L$</th>
<th>$D_{	ext{Travel}}$ (km)</th>
<th>$D_{	ext{Transport}}$ (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>42.78</td>
<td>18.84</td>
<td>44%</td>
<td>3.89</td>
<td>13.27</td>
</tr>
<tr>
<td>Median</td>
<td>43.00</td>
<td>19.00</td>
<td>43%</td>
<td>3.11</td>
<td>12.47</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>7.28</td>
<td>6.97</td>
<td>13%</td>
<td>3.07</td>
<td>6.94</td>
</tr>
<tr>
<td>Coefficient of variation</td>
<td>0.37</td>
<td>0.17</td>
<td>29%</td>
<td>0.79</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Table 1.5  Mechanisms for the effect of EMS load on service time

<table>
<thead>
<tr>
<th>Changeover</th>
<th>System components</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Server</strong></td>
<td></td>
</tr>
<tr>
<td>Work content</td>
<td></td>
</tr>
<tr>
<td>Physical Setup</td>
<td></td>
</tr>
<tr>
<td><strong>Network</strong></td>
<td></td>
</tr>
<tr>
<td>Work content</td>
<td></td>
</tr>
<tr>
<td>Network arrangement</td>
<td></td>
</tr>
<tr>
<td><strong>Customer</strong></td>
<td></td>
</tr>
<tr>
<td>Service speed</td>
<td></td>
</tr>
<tr>
<td>Social speedup pressure</td>
<td></td>
</tr>
<tr>
<td>Downstream system congestion</td>
<td></td>
</tr>
<tr>
<td>Geographical dispersion</td>
<td></td>
</tr>
<tr>
<td>Extended load</td>
<td></td>
</tr>
<tr>
<td>Work content</td>
<td></td>
</tr>
<tr>
<td>Fatigue</td>
<td></td>
</tr>
<tr>
<td>Network chaos</td>
<td></td>
</tr>
<tr>
<td>Work content</td>
<td></td>
</tr>
<tr>
<td>Network chaos</td>
<td></td>
</tr>
<tr>
<td>Deterioration</td>
<td></td>
</tr>
</tbody>
</table>

1.4.1 Chute Time

Chute time is the first time interval and can be seen as the setup time for the EMS service.

**Chute Time Mechanisms.** Given that chute time can be seen as a setup time, we pay special attention to the server–changeover mechanisms listed in Table 1.1.
Table 1.6 Chute time mechanisms

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Physical setup</strong></td>
<td>Regular service requires a longer setup than extended service.</td>
</tr>
<tr>
<td><strong>Server early task initiation</strong></td>
<td>During periods of high load, ambulance crews in standby mode initiate preparation tasks before receiving a dispatch notification.</td>
</tr>
</tbody>
</table>

Physical setup could be at work for chute time. It is reasonable to believe that a regular service (where the ambulance is dispatched from standby mode, corresponding to \( I_{\text{Changeover}} = 1 \)) will have a longer chute time than an extended service. Indeed, according to Aehlert (2011, p. 654), chute time for extended services is zero in most cases because the ambulance crew is already in the vehicle. Once identified, this mechanism seems obvious. An advantage of systematic use of the LEST framework is that such obvious mechanisms are not ignored.

Turning to server–load mechanisms, paramedics might engage in server early task initiation. Information about load may lead a crew to form an expectation about the likelihood of being dispatched in the near future. When the load is high, a crew may respond by initiating preparation tasks before receiving a dispatch notification.

Table 1.6 lists our proposed mechanisms for the chute time.

**Chute Time Hypotheses.** The physical setup and server early task initiation mechanisms lead directly to the following two hypotheses about chute time, which we test using Model (1.1):

**Hypothesis 1**  
Chute time increases with changeover.

**Hypothesis 2**  
Regular chute time decreases with load.

\[
T_{\text{Chute}} = \beta_0 + \beta_{I_{\text{Changeover}}} I_{\text{Changeover}} + \beta_{L_{\text{Dispatch}} \times I_{\text{Changeover}}} L_{\text{Dispatch}} \times I_{\text{Changeover}} + \beta_{L_{\text{Dispatch}} \times (1 - I_{\text{Changeover}})} L_{\text{Dispatch}} \times (1 - I_{\text{Changeover}}) + \beta_{\text{NS}} NS + \beta_{D_{\text{Travel}}} D_{\text{Travel}} + \beta_{I_{\text{Urgent}}} I_{\text{Urgent}} + \beta_{D_{\text{Travel}} \times I_{\text{Urgent}}} D_{\text{Travel}} \times I_{\text{Urgent}} + \beta_{X} X + \varepsilon. \tag{1.1}
\]

In Model (1.1) and the other models of this section, \( X \) denotes a vector of control variables including day, time, and day and time interaction variables, \( \varepsilon \) is the error term, and “\( \times \)” represents interactions.

A significant positive coefficient for \( I_{\text{Changeover}} \) supports Hypothesis 1 and provides evidence for the physical setup mechanism. A significant positive coefficient for \( L_{\text{Dispatch}} \) and a significant negative coefficient for \( L_{\text{Dispatch}} \times I_{\text{Changeover}} \) supports Hypothesis 2 and provides evidence for the early task initiation mechanism for regular services. We include \( L_{\text{Dispatch}} \times (1 - I_{\text{Changeover}}) \) to check whether load has
Table 1.7  Chute time

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>$t_{\text{Chute}}$ (min.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$-1.29(1.11)$</td>
</tr>
<tr>
<td>$\Delta_{\text{Changeover}}$</td>
<td>$0.39(0.16)^*$</td>
</tr>
<tr>
<td>$L_{\text{Dispatch}} \times \Delta_{\text{Changeover}}$</td>
<td>$-0.42(0.16)^{**}$</td>
</tr>
<tr>
<td>$L_{\text{Dispatch}} \times (1 - \Delta_{\text{Changeover}})$</td>
<td>$-0.13(0.31)$</td>
</tr>
<tr>
<td>$\mathcal{N}$</td>
<td>$0.04(0.02)$</td>
</tr>
<tr>
<td>$D_{\text{Travel}}$</td>
<td>$0.08(0.01)^{***}$</td>
</tr>
<tr>
<td>$\mathcal{I}_{\text{Urgent}}$</td>
<td>$-0.01(0.08)$</td>
</tr>
<tr>
<td>$D_{\text{Travel}} \times \mathcal{I}_{\text{Urgent}}$</td>
<td>$-0.04(0.01)^{***}$</td>
</tr>
<tr>
<td>$R^2$</td>
<td>$0.0045$</td>
</tr>
<tr>
<td>P-val.</td>
<td>$&lt; 2.2e-16$</td>
</tr>
</tbody>
</table>

an effect on extended services. We include $D_{\text{Travel}}$ in Model (11) because it may take longer for an ambulance crew to find the call address and to plan a route if the address is outside the unit’s normal coverage region.

The estimation results for Model (1.1) in Table 1.7 provide support for both hypotheses. Average chute time is estimated to be 23 seconds longer for regular calls (providing evidence for physical setup) and an increase in load from 10% to 90% is estimated to shorten average chute time by 20 seconds for regular services (providing evidence for early task initiation). In contrast, load does not have a significant impact for extended services. We also see that average chute time increases by 4.8 seconds per kilometer of travel distance, which could be interpreted as a physical setup time that increases with distance. The distance effect is smaller for urgent calls. All of these effects are statistically significant.

1.4.2  Travel Time

Travel time has natural measures for work content and service speed: the travel distance and the average driving speed. Some of our proposed travel time mechanisms are about the distance and some are about the speed.

**Travel Time Mechanisms.** Table 1.8 summarizes our proposed mechanisms for the travel time.

**Social speedup pressure:** This is a potential server–load mechanism affecting driving speed. EMS systems are under pressure to meet response time targets. Paramedics’ actions are tracked by computer-aided-dispatch systems and therefore, the paramedics are likely to feel pressure to speed up by driving faster as the EMS load increases.

---

Footnote: In Table 1.7 and other regression result tables, $\ast$, $\ast\ast$, $\ast\ast\ast$, and $\ast\ast\ast$ denote statistical significance at the 0.1%, 1%, and 5% significance levels, respectively. Standard errors are shown in parentheses.
Table 1.8  Travel time mechanisms

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social speedup pressure</td>
<td>Driving speed increases with EMS load.</td>
</tr>
<tr>
<td>Geographical dispersion</td>
<td>Travel distance increases with load.</td>
</tr>
<tr>
<td>Geographical speedup</td>
<td>Travel speed increases with distance.</td>
</tr>
<tr>
<td>Network chaos</td>
<td>Extended load increases travel times as the actual ambulance locations deviate from the planned locations.</td>
</tr>
<tr>
<td>Network arrangement</td>
<td>Travel distance is shorter for regular (not extended) services.</td>
</tr>
</tbody>
</table>

Geographical dispersion: It is natural to view an EMS system and the EDs as the two nodes of a queueing network where the EDs are the downstream nodes. We will exploit this view in Section 1.4.5 (hospital time). Here, we expand our perspective of network to include routes in the city road network. From this perspective, patients need simultaneous service from two servers (first, an ambulance and its paramedics, and second, from a route) during the travel and transport times. This perspective allows us to identify new network mechanisms. Travel time delays the start of medical care to a patient by an amount that depends on the travel distance. The travel distance depends on the dispatch policy and on the paramedics’ selection of route to the call address.

Geographical dispersion causes longer travel distance when load is high. High EMS load means fewer available ambulances to cover a fixed geographic area (a city), which results in longer average travel distance. Geographic dispersion could also be relevant for other services, including repair and tow truck services, hospital porters, taxi and delivery services, and other emergency services (fire or police).

Geographical speedup: Geographical speedup mitigates geographical dispersion by enabling the paramedics to travel faster on longer trips that involve at least some highway or main artery travel (Budge et al., 2010). Unlike geographical dispersion, geographical speedup impacts the service speed.

Network chaos: Ambulance locations are chosen to cover as many calls as possible within a target response time. If EMS load remains high for a long period, the actual locations of the available ambulances could deviate more and more from their planned positions, leading to suboptimal dispatch decisions and longer travel times. We call this the network chaos mechanism.

Network arrangement: As load returns to normal, or more specifically when an individual ambulance finishes service without being immediately dispatched to another call (that is, a regular service), that ambulance can return to its planned location. Network arrangement causes travel distances to be shorter for regular (not extended) services.

Travel Time Hypotheses. Based on the mechanisms in Table 1.8, we develop
three hypotheses for the travel time: one for the changeover effect, one for the load effect, and one for the extended load effect. The network arrangement mechanism directly leads to:

**Hypothesis 3**  Travel time decreases with changeover.

The social speedup pressure, geographical dispersion, and geographical speedup mechanisms all involve the effect of EMS load on the travel time. The two speedup mechanisms and the geographical dispersion mechanism predict opposite effects on the travel time. It is not clear which one—the increase in the work content or the increases in the service speed—will be the dominant effect. To clarify the interactions between these mechanisms, we use a simple model (Fig. 1.4) of the physics of ambulance travel time that was proposed by Kolesar et al. (1975) and formed the basis for a statistical model of ambulance travel times in Budge et al. (2010). The model assumes that (1) the ambulance accelerates at a constant rate $a$ at the beginning of a trip and decelerates at the same constant rate $a$ at the end of a trip and (2) if the trip is long enough, the ambulance reaches a constant cruising speed $v$ during the middle of the trip. The two graphs of Fig. 1.4 illustrate a trip that is so short that the cruising speed is never reached and a trip that is long enough to reach the cruising speed.

![Fig. 1.4 Simple model of the Physics of ambulance travel time](image)

From this model, one can derive the following Model (1.2) for the travel time $T$, the distance $D$, and the average speed $S = D/T$:

$$T = \begin{cases} \frac{2\sqrt{D}}{a} & D < \frac{v^2}{a}, \\ \frac{v + D}{a} & D \geq \frac{v^2}{a}. \end{cases}$$

(1.2)

Using the parameters of Model (1.2), we can specify the social speedup pressure, geographical dispersion, and geographical speedup mechanisms more precisely as follows:
- Social speedup pressure: Increased load increases the acceleration $a$, the cruising speed $v$, or both.
- Geographical dispersion: Increased load increases the distance $D$.
- Geographical speedup: Increased distance $D$ increases the average speed $D/T$.

The model implies that the geographical speedup effect will occur, because we can derive:

$$S = \frac{D}{T} = \begin{cases} \frac{1}{2} \sqrt{aD} & D < \frac{v^2}{a}, \\ \left( \frac{v}{aD} + \frac{1}{v} \right)^{-1} & D \geq \frac{v^2}{a}. \end{cases}$$

Inspection of the two right-hand-side expressions for $S$ shows that when $D$ increases, $S$ increases.

The remaining question is, what is the combined effect of social speedup pressure and geographical dispersion? That is, if load increases, causing $a$, $v$, and $D$ all to increase, what happens to $T$? The answer is that it can go either way, as one can verify numerically using Model (1.2). If $a$, $v$, and $D$ all increase by the same percentage, then $T$ will remain unchanged for long trips but will decrease for short trips. If $a$ and $v$ increase by a larger (smaller) percentage than $D$, then $T$ will tend to decrease (increase). Thus, it is not clear which mechanism (social speedup pressure or geographical dispersion) will dominate, and therefore we include all possibilities in the following hypothesis:

**Hypothesis 4** (a) Travel time increases with load. (b) Travel time decreases with load. (c) Travel time does not change with load.

The network chaos mechanism, which involves the effect of extended load, leads directly to:

**Hypothesis 5** Travel time increases as high load periods last longer.
We use Models (1.3)-(1.6) to test Hypotheses 3-5 and the supporting mechanisms:

$$T_{\text{Travel}} = \beta_0 + \beta_{I\text{Changeover}} I_{\text{Changeover}} + \beta_{L\text{Dispatch}} L_{\text{Dispatch}} + \beta_{\text{NSNS}} + \beta_{I\text{Urgent}} I_{\text{Urgent}} + \beta_{I\text{Siren}} I_{\text{Siren}} + \beta_X X + \epsilon,$$

(1.3)

$$T_{\text{Travel}} = \beta_0 + \beta_{I\text{Changeover}} I_{\text{Changeover}} + \beta_{L\text{Dispatch}} L_{\text{Dispatch}} + \beta_{\text{NSNS}} + \beta_{I\text{Urgent}} I_{\text{Urgent}} + \beta_{I\text{Siren}} I_{\text{Siren}} + \beta_{D\text{Travel}} D_{\text{Travel}} + \beta_X X + \epsilon.$$

(1.4)

$$D_{\text{Travel}} = \beta_0 + \beta_{I\text{Changeover}} I_{\text{Changeover}} + \beta_{L\text{Dispatch}} L_{\text{Dispatch}} + \beta_{\text{NSNS}} + \beta_{I\text{Urgent}} I_{\text{Urgent}} + \beta_{I\text{Siren}} I_{\text{Siren}} + \beta_X X + \epsilon.$$

(1.5)

$$S_{\text{Travel}} = \beta_0 + \beta_{I\text{Changeover}} I_{\text{Changeover}} + \beta_{L\text{Dispatch}} L_{\text{Dispatch}} + \beta_{\text{NSNS}} + \beta_{I\text{Urgent}} I_{\text{Urgent}} + \beta_{I\text{Siren}} I_{\text{Siren}} + \beta_{D\text{Travel}} D_{\text{Travel}} + \beta_X X + \epsilon.$$

(1.6)

In Model (1.5), a significant negative coefficient for $I_{\text{Changeover}}$ supports Hypothesis 3 and the sign and the significance of the coefficient for $L_{\text{Dispatch}}$ determines which of the alternatives in Hypothesis 4 is supported. We expect the magnitude and significance of the coefficient for $L_{\text{Dispatch}}$ to attenuate as we control for distance, in moving from Model (1.3) to Model (1.4).

In Model (1.5), a significant negative coefficient for $I_{\text{Changeover}}$ implies that average travel distance is shorter for regular services than for extended services, supporting network arrangement. A significant positive coefficient for $L_{\text{Dispatch}}$ implies longer average travel distance under high load, which supports geographical dispersion.

In Model (1.6), a significant positive coefficient for $L_{\text{Dispatch}}$ implies higher average speed under high load, supporting social speedup pressure. A significant positive coefficient for $D_{\text{Travel}}$ implies higher average speed for longer distances, supporting geographical speedup.

Tables 1.9-1.10 present the results for Models (1.3)-(1.6). Model (1.5) supports the network arrangement mechanism, which led us to Hypothesis 3. However, the significant positive coefficient of $I_{\text{Changeover}}$ in Model (1.3) does not support Hypothesis 3. One explanation would be that the regular service requires an acceleration phase before reaching a cruising speed phase, as in the simple model of the physics of ambulance travel time (Fig. 1.4), whereas the extended service could be initiated when the ambulance is already in the cruising speed phase. The resulting higher average speed for extended service could surpass the advantage of shorter average distance for regular service. The negative coefficient for $I_{\text{Changeover}}$ in Model (1.6) supports this argument.

Model (1.3) supports Hypothesis 4(a), indicating a 0.42-minute increase in average travel time associated with a 10% increase in load. By comparing Models (1.3)-(1.5), we see that the increase in average travel time occurs primarily through geographical dispersion: A 10% increase in load leads to a 0.44-kilometer
increase in average distance (Model (1.5)), which translates into a (0.44 × 0.97)-minute increase in average travel time (Models (1.4)-(1.5)), whereas the direct effect of a 10% increase in load on average travel time is not significant (Model (1.4)). The coefficients of determination ($R^2$) of Models (1.3) and (1.4) show that distance explains almost 20% of the variability in travel time.

Model (1.5) indicates that average distance increases with load and Model (1.6) indicates that average speed increases with distance. Taken together, these results support geographical speedup. In contrast, load does not have a significant effect on average speed in Model (1.6), thus failing to support social speedup pressure. However, we do see evidence of social speedup pressure in short trips, as shown in the last column of Table 1.10 where we re-estimate Model (1.6), including only the services with travel distances less than 5 kilometers.

### 1.4.3 Scene Time

In the field of prehospital care, there is an ongoing debate about the relative merits of the scoop and run (transport the patient to a hospital as quickly as possible) ver-
Table 1.11  Scene time mechanisms

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task transfer</td>
<td>As load increases, up to a threshold, paramedics spend more time on scene in order to stabilize the patient on scene and avoid hospital transportation.</td>
</tr>
<tr>
<td>Workload smoothing</td>
<td>As load increases beyond a threshold, paramedics spend less time on scene in anticipation of short hospital times due to hospital surge capacity protocols.</td>
</tr>
<tr>
<td>Fatigue</td>
<td>If the paramedics have been busy for an extended period, then they take longer to complete their tasks on scene.</td>
</tr>
<tr>
<td>Deterioration</td>
<td>Increased load increases response time, which has an adverse effect on the patient’s medical condition, which increase the required on-scene care time.</td>
</tr>
</tbody>
</table>

**sus stay and play** (paramedics initiate primary treatment and stabilize the patient on the scene before making the transport decision) strategies (Smith and Conn, 2014). Paramedics need to exercise discretion to decide not only on how much treatment to provide on scene but also whether to transport the patient to hospital, and in doing so they must consider such factors as the medical urgency, distance to the closest hospital, and their own qualifications and skill level. Because of the discretionary nature of the on-scene care, it is likely that EMS load also affects paramedics’ decisions.

**Scene Time Mechanisms**. Table 1.11 summarizes our proposed mechanisms for the scene time. **Task transfer** and **workload smoothing** mechanisms are related to the effect of EMS system load on paramedics’ on-scene decisions. **Fatigue** is the effect of extended load (how long the paramedics have been busy since their last service gap) on paramedics’ behavior. **Deterioration** in patient health due to longer response times caused by high EMS load can also impact the scene time.

**Task transfer** and **workload smoothing**: If paramedics are aware of the system load (through radio communications, for example), then they may form expectations about the time they would need to spend in a hospital to offload the patient, if they transport the patient to a hospital. As we discuss in Section 1.4.5, **downstream system congestion** causes hospital time to increase with load up to a threshold, and to decrease with load beyond that threshold, because of surge capacity protocols. We postulate that predictability of the hospital time, based on the current EMS load, affects paramedics’ decisions on the scene.

Specifically, when EMS load is below a critical threshold, paramedics spend more time on scene as load increases in order to stabilize the patient and to avoid hospital transportation, where they expect a long wait. We call this mechanism
On the other hand, when EMS load increases beyond the critical threshold, paramedics may prefer to shorten the scene time and instead continue the care process in the ambulance, en route to a hospital, which we call workload smoothing, in the anticipation of short hospital times due to surge capacity protocols that come into effect in hospitals when EMS load is in critical situation.

**Fatigue:** We measure the extended load at the paramedic crew level as the total amount of time that the crew has been busy without a break between calls (we referred to this as extended service in Section 1.3.1). We expect that a fatigue mechanism could increase scene time for two reasons. First, scene time involves a demanding combination of physical and mental tasks. Second, paramedics have an opportunity to slow down intentionally to take a break if needed.

**Deterioration:** Patients do not have direct information about load but they experience longer response time (chute time + travel time) when EMS system load is higher due to longer travel times, as supported in Hypothesis 4. Long response times can result in deterioration of a patient’s medical condition (Feero et al., 1995; Blackwell and Kaufman, 2002), which can increase the amount of time the paramedics need to spend on the scene.

**Scene Time Hypotheses.** Deterioration increases the scene time. Task transfer also increase the scene time, but only up to the critical load threshold. Workload smoothing causes the service times to decrease with load above the threshold. Taken together, we hypothesize:

**Hypothesis 6** Scene time increases with load below a critical threshold and decreases with load above the threshold.

The fatigue mechanism leads directly to:

**Hypothesis 7** Scene time increases with extended load.

We use Model (1.7) to test Hypotheses 6 and 7.

\[
T_{\text{Scene}} = \beta_0 + \beta_{L_{\text{Arrival}}} L_{\text{Arrival}} + \beta_{L_{\text{Arrival}} L_{\text{Arrival}}} L_{\text{Arrival}} + \beta_{E_{\text{Server}}} E_{\text{Server}} + \beta_{T_{\text{Response}}} T_{\text{Response}} + \\
\beta_{N_{\text{NS}}} + \beta_{I_{\text{Urgent}}} I_{\text{Urgent}} + \beta_{T_{\text{Transport}}} T_{\text{Transport}} + \beta_{L_{\text{Arrival}} I_{\text{Urgent}}} L_{\text{Arrival}}^2 I_{\text{Urgent}} + \\
\beta_{L_{\text{Arrival}} I_{\text{Urgent}}} L_{\text{Arrival}} I_{\text{Urgent}} + \beta_X X + \epsilon.
\]

A significant positive coefficient for \(L_{\text{Arrival}}\) and a significant negative coefficient for \(L_{\text{Arrival}}^2\) support the concave relation between load and scene time that is postulated in Hypothesis 6 and for the underlying task transfer and workload smoothing mechanisms. To isolate the effect of the deterioration mechanism, we control for response time in (1.7). A significant positive coefficient for \(T_{\text{Response}}\)

---

2This mechanism was referred to as engagement in Delasay (2014) and Delasay et al. (2015) but task transfer better captures the nature of this mechanism in an EMS setting.
Table 1.12 Scene time

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>$T_{scene}$ (min.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>−1.95(8.91)</td>
</tr>
<tr>
<td>$L_{Arrival}$</td>
<td>−29.47(4.79)**</td>
</tr>
<tr>
<td>$T_{Response}$</td>
<td>26.40(4.42)**</td>
</tr>
<tr>
<td>$EL_{Server}$</td>
<td>0.01(0.00)**</td>
</tr>
<tr>
<td>$T_{Response}$</td>
<td>0.50(0.01)**</td>
</tr>
<tr>
<td>$NS$</td>
<td>0.65(0.23)**</td>
</tr>
<tr>
<td>$I_{Urgent}$</td>
<td>3.93(1.83)**</td>
</tr>
<tr>
<td>$T_{Transport}$</td>
<td>−12.28(0.20)**</td>
</tr>
<tr>
<td>$L_{Arrival} \times I_{Urgent}$</td>
<td>28.37(9.24)**</td>
</tr>
<tr>
<td>$L_{Arrival} \times I_{Urgent}$</td>
<td>−24.15(8.41)**</td>
</tr>
</tbody>
</table>

$R^2$: 0.0647
P-val.: $< 2.3e−16$

Table 1.12 presents results for Model (1.7). The results support the concave relation between load and scene time (Hypothesis 6) and provide evidence for the task transfer, workload smoothing, and deterioration mechanisms. Fig. 1.5 plots estimated average scene time versus load for transported urgent and non-urgent services based on the results in Table 1.12 and with all independent variables except load, urgency, and transport set to their average values. We see that the impact of load on average scene time is indeed much more pronounced for non-urgent calls. Model (1.7) also supports the fatigue mechanism and its effect on increasing scene time due to the extended load.

1.4.4 Transport Time

Like travel time, transport distance and driving speed are the natural measures for the work content and service speed of the transport time interval. However, we do not expect the geographical dispersion and geographical speedup mechanisms to be relevant for transport time, for the following reasons. Travel time is from A (dispatch location) to B (call location), where A is impacted by load but B is not. Transport time, in contrast, is from B (call location) to C (the closest hospital to B), and we do not expect B and C to be impacted greatly by load.

In some EMS systems, an ED can request to divert incoming ambulances
Inventory Is People: How Load Affects Service Times in Emergency Response

1.4.5 Hospital Time

Hospital time is the final and the longest EMS service time interval.

**Hospital Time Mechanisms.** Hospital time is the contact point between EMS and the ED. Therefore, we pay special attention to the network mechanisms listed in Table 1.11, especially the downstream system congestion mechanism. We propose no changeover mechanisms for hospital time. Table 1.13 summarizes our proposed mechanisms for hospital time.

---

**Table 1.13  Hospital time mechanisms**

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Downstream system</td>
<td>ED congestion causes hospital time to increase with load.</td>
</tr>
<tr>
<td>congestion</td>
<td></td>
</tr>
<tr>
<td>Social speedup pressure</td>
<td>When EMS load is very high, ED staff and paramedics speed up the admission of ambulance patients.</td>
</tr>
<tr>
<td>Fatigue</td>
<td>Extended load increases the amount of overwork and slows the work of paramedics at the hospital.</td>
</tr>
</tbody>
</table>

---

to neighboring hospitals during periods of overcrowding. This phenomenon is known as “ambulance diversion” (Deo and Gurvich 2011; Gurvich et al. 2014). In such systems, one would expect the average transport distance to the closest hospital that is not on diversion to increase with load—similar to the network chaos mechanism for travel time. This does not apply directly in Calgary because the system is centrally managed and individual hospitals do not have the freedom to decide to go on diversion.

We estimated Models (1.3)-(1.6) with travel time replaced with transport time. As expected, we did not observe statistically significant impacts of load on transport time.

---

Fig. 1.5  Scene time vs. load based on Model (1.7); $NS = 42.79$, $\gamma_{Response} = 8.09$ min., and $EL_{Server} = 6.75$ min.
Downstream system congestion: Congestion in the ED is likely to impact the hospital time interval through the downstream system congestion mechanism (Forster et al., 2003; Asaro et al., 2007; Hillier et al., 2009). EMS load is positively correlated with the ED load. EMS patients add workload to an ED and a congested ED causes ambulances to back up to offload patients, resulting in longer hospital time.

Social speedup pressure: We expect the effect of downstream system congestion to be mitigated at high loads. Paramedics and ED staff are likely to feel pressure to shorten hospital time when the EMS load is close to its limit. Some jurisdictions have formal protocols that prescribe actions that must be taken to expedite patients when EDs and EMS systems are under high load (Handel et al., 2010; Villa-Roel et al., 2012). For example in Edmonton and Calgary, Alberta, Canada, “ED surge capacity protocols” require EDs to accelerate admission of ambulance patients when fewer than seven ambulances are available for service in the city (Alberta Health Services, 2010).

Fatigue: We expect that the fatigue mechanism increases hospital time for the same reasons that we mentioned for scene time. First, the interval involves a demanding combination of physical and mental tasks. Second, paramedics have an opportunity to slow down intentionally to take a break if needed.

Hospital Time Hypotheses. Based on the downstream system congestion and social speedup pressure mechanisms, we hypothesize:

Hypothesis 8 Hospital time has a concave relationship with load.

The fatigue mechanism leads directly to:

Hypothesis 9 Hospital time increases with extended load.

We use Model (1.8) to test Hypotheses 8 and 9. A significant negative coefficient for $L_{Hospital}^2$ supports the concave relation between the hospital time and load as proposed in Hypothesis 8. A significant positive coefficient for $EL_{Server}$ supports Hypothesis 9 and the fatigue mechanism.

$$T_{Hospital} = \beta_0 + \beta_{L_{Hospital}^2} L_{Hospital}^2 + \beta_{L_{Hospital}} L_{Hospital} + \beta_{EL_{Server} EL_{Server}} + \beta_{NS NS}$$

$$+ \beta_{I_{Urgent} I_{Urgent}} I_{Urgent} + \beta_{L_{Hospital} I_{Urgent}} L_{Hospital} \times I_{Urgent}$$

$$+ \beta_{L_{Hospital} \times I_{Urgent} L_{Hospital} \times I_{Urgent}} + \beta_X X + \epsilon. \quad (1.8)$$

The results of Model (1.8), as presented in Table 1.14 support the concave relation between load and hospital time (Hypothesis 8). Fig. 1.6 shows the estimated relationship between load and average hospital time, for urgent and non-urgent calls, with all other independent values set to their average values. As shown,
the effect of social speedup pressure is more pronounced for non-urgent patients, although the differences are not statistically significant (see the coefficients of the two interaction variables in Table 1.14). The results in Table 1.14 do not support Hypothesis 9 and the fatigue mechanism about the effect of extended load on increasing the hospital time.

### 1.5 Conclusion and Future Directions

Inventory is people. People react to the situation around them in complicated ways. We use the LEST framework to understand some of those complicated interactions to help predict how service times react to changes in load. A better understanding of variations in service time is important to capacity and scheduling of service operations and queues. Academics already have a fairly good understanding of how service times affect utilization. We inform our understanding of
how utilization affects the service time.

In Fig. 1.1 we showed how EMS service times are concave with load. In this paper, we have shown how that concavity develops from a complicated mix of multiple facets of care delivery with both behavioral and technical responses to changes in load. We used the LEST framework to explore these complexities. We have found changeover, load, and extended load responses. We have seen responses in servers, networks, and one in the customers. We have seen changes in both speed and work content. Some of the mechanisms in operation here are, once identified, rather obvious. Others are less so. We use the nature of the service encounter to break down the service into component parts and we use the LEST framework to guide our exploration into potential mechanisms.

We have identified 12 different mechanisms applying to EMS response times that are both logical and supported by our data set. Eight mechanisms were previously identified as shown in [Delasay et al. (2015)]. Four mechanisms have not been previously examined.

1. **Physical setup** increases the work content under changeover when paramedics respond after having had a break between missions.
2. **Network arrangement** reflects the work reduction on changeover corresponding to ambulances being well dispersed.
3. **Server early task initiation** reflects paramedics being more prepared to respond when they know the system is under load.
4. **Task transfer** is the tendency of paramedics to do more tasks on scene and delay transport to the hospital (stay and play) as load increases up to a threshold.
5. **Workload smoothing** comes into play above a threshold as paramedics do fewer tasks on the scene as load extends beyond a threshold in order to begin transfer to the hospital.
6. **Fatigue** occurs when servers under load for an extended period of time slow down.
7. **Deterioration** is the possibility that delays in response time will accompany additional work requirements upon arrival on the scene.
8. **Geographical dispersion** acknowledges that workload increases under load as fewer available ambulances are more widely dispersed.
9. **Social speedup pressure** is the response felt by servers to respond quickly as travel distances increase with load.
10. **Geographical speedup** reflects server speed increasing as longer distances allow increased use of faster roads when load is high.
11. **Network chaos** is the tendency of the system to move towards a worse response position when load is high for an extended period.
(12) *Downstream system congestion* occurs when hospital congestion, correlated with high ambulance load, delays transfer at the hospital.

These mechanisms affect processing times in almost all stages of the response process. Average chute time decreases by 23 seconds if no changeover is required and, as load increases, chute times decrease. Effects of load on travel time are more complicated. While average distance (work) increases with load, average speed increases with distance. These effects combine so that a 10% increase in load is accompanied by a 0.42-minute increase in average travel time. The relationship between load and scene time is concave and more significant for non-urgent calls. For non-urgent calls, scene time increases from 17 minutes to 22 minutes as load increase up to 55% and then, it drops to 14 minutes as load gets close to 100%. We find no effects on Transport time. Hospital time is also concave with load for non-urgent care with a tipping point around a load of 80%. The average hospital time increases from 33 minutes to 65 minutes when load increases from 10% to 80% due to downstream system (ED) congestion.

These findings will be of interest to EMS managers in identifying and understanding how their systems respond to load. The concave nature of the response curve suggests an identifiable tipping point beyond which system response begins to deteriorate. Managers should be able to find the tipping point using local historical data and consider policies to mitigate the negative effects.

This work also helps us to gain a better understanding of these 12 mechanisms and how they affect response times. If we wish to generalize the common effects of load on service time, much more empirical exploration of the mechanisms involved is required. We, as a field, need to continue to record empirical evidence of different mechanisms before we can begin the process of generalization.

A third contribution of this work is demonstrating how the LEST framework can be used to understand the relationships of load and time. The framework allows us to take a very complicated problem and break it down into its component parts. We can help both researchers and managers explore the relationship of load and time by giving them the framework under which the larger topic can be split and questions directed at the component parts. Providing structured questions is one of the strongest contributions any theory can make.
Bibliography


Gans, N., N. Liu, A. Mandelbaum, H. Shen, H. Ye. 2010. Service times in call centers: Agent heterogeneity and learning with some operational consequences. *Institute of


