Simulation Model for Evaluating the Performance of Emergency Response Fleets

K. G. Zografos, C. Douligeris, and L. Chaoxi

A simulation model for evaluating the performance of an emergency response fleet of an electric utility company is presented. The proposed model considers the spatial, temporal, and severity distribution of calls and has the capability to simulate alternative configurations of service districts and dispatching policies of the emergency response fleet. A nonstationary Poisson process is used to simulate the temporal distribution of service calls, whereas discrete simulation is employed for the spatial and severity distribution of the service calls. A mixed planar and network model is used to calculate the shortest travel time between the service calls and the location of emergency response vehicles. The model is validated on the basis of historical data. It is used to evaluate the relationship between fleet size and total incident service time and to compare alternative configurations of service districts for the same fleet size and dispatching policy.

A central problem in managing spatially distributed emergency response operations, such as police, ambulance, fire, emergency repair, and roadway assistance, is to determine the number of mobile units (fleet size) that should be available to respond to emergency calls, the service territories, and the dispatching strategies of the Emergency Response Units (ERUs) (1,2). The primary measure of effectiveness of an emergency response system is the minimization of the average response time to emergency calls (1–3). Average response time depends on the spatial, temporal, and severity distribution of the service calls and the size and deployment strategy of the emergency response fleet (1,2).

Thus, efficient deployment of an emergency response fleet requires examination of trade-offs between the cost of the emergency response fleet operations, that is, the number of ERUs available for deployment, and resulting performance, or average response time. Examination of this trade-off requires development of analytical tools that will be able to evaluate the performance of an emergency response mechanism for various levels of expected work load and manpower availability.

The objective of this paper is to present a simulation tool, developed for evaluating the performance of alternative districting patterns and fleet size of an emergency response system. The paper focuses on the development of the simulation model and its application as an evaluation tool. The work presented is motivated by the emergency repair operations of a large utility company. The proposed simulation tool is part of an integrated decision support system (4) that was developed to help the service restoration managers improve the effectiveness of the utility’s service restoration fleet.

EMERGENCY RESPONSE FLEET OPERATIONS

The need for emergency response services arises when incidents requiring prompt response and attention occur randomly in space and time. Depending on the type of incident, whether a fire, medical emergency, police emergency, emergency repair, or roadway incident, a mobile ERU should be dispatched to the scene of the incident to offer the necessary services. A crucial parameter involved in the design and evaluation of emergency response operations is the total incident service time (TIST). TIST is defined as the time elapsed between the occurrence of an incident and the completion of requested services (3,4).

TIST consists of four time intervals. The first interval, T1, the incident detection and identification time, is determined by the time elapsed between the occurrence of an incident and the arrival of a call at a dispatching center announcing the incident and requesting services. The duration of T1 depends on the technology used to detect the incident (5). For certain types of incidents, there are opportunities for automatic incident detection, depending on the capacity of the switchboard receiving the incident calls and the technology used to associate calls with the location of the incident (6,7).

The second interval, T2, or the dispatch delay, is determined by the time elapsed between the detection and identification of the incident and its assignment to the first available ERU. The magnitude of a dispatch delay depends on the ERUs’ degree of use. In the case of a congested system, when the utilization rate of the servers exceeds a threshold value, the dispatch time is the major determinant of response time (5–8).

The third component, T3, is the time interval required by an ERU to travel from its current location to the scene of the incident. The fourth interval, T4, involves the time that an ERU spends in providing the requested services. The duration of the incident service time depends mainly on the severity of the incident (1).

TIST, and consequently the performance of the emergency response system, can be enhanced substantially by reducing dispatch delay (T2) and travel time (T3). Thus, any modeling effort regarding the improvement of the deployment of an emergency response fleet should take into account the interactions between the various components of the TIST and their effect on T2 and T3.

PREVIOUS RELATED WORK

Computer simulation methods have been used extensively to study the performance of emergency response systems and are described in the literature. Simulation models offer the capability to study the trade-off between the number of servers and TIST for complex, large-scale systems that are not amenable to exact queuing theory formulations. In addition, simulation models can be used to evalu-
ate the performance of alternative districting patterns generated by
districting models.

Savas (9) used simulation as a tool to perform a cost-effective­
ess analysis of New York’s emergency ambulance services, thereby linking operations research models, decision making, and
computer techniques.

Larson (10) used the hypercube queuing model, which incorpo­
rates theoretical queuing theory results and simulation as a tool
in dispatching of police patrol cars. Brandeau and Larson (11)
extended the use of the hypercube model to the deployment of
districting module. The proposed simulation model uses historical data
describing the spatial, temporal, and priority distribution of emer­
gency repair calls and simulates alternative dispatching strategies
for the emergency repair fleet. The output of the simulation module
provides statistical information related to the performance of the
service restoration units within each truck-area. The performance of
the service restoration mechanism is evaluated in terms of the aver­
age dispatch, travel, and repair (DTR) time (i.e., the time interval
between the identification of the service call and the completion
of service). \( T_1 \) is not included in the performance analysis because it is
not affected by fleet size or the shape of the districts.

**SPATIAL, TEMPORAL, AND SEVERITY CHARACTERISTICS OF SERVICE REPAIR CALLS**

An essential step in developing the proposed simulation model was
to understand the operations and the behavior of the service restora­
tion mechanism. That was done through an analysis of historical data
describing the demand for repair services and the performance of the
service restoration mechanism. The data base used for the analysis
of the service restoration operations was provided by a major utility
company. The data base included data covering a period of 1 year
and each record of the data base corresponded to a call for an emer­
gency repair. Each record contained information regarding the time
that the service call arrived at the dispatching center, the time that a
work order (ticket) was generated for the call, the time that the ticket
was assigned to a field repair unit, the time that the field unit arrived
at the scene of the incident, and the time that the repair was com­
pleted. In addition, each record contained information describing the location of the call in terms of the \( X \) and \( Y \) coordinates of a major and
a minor reference grid used by the company to identify its facilities
and customers in the two-dimensional space. The size of the major
grid was \( 1 \text{ mi}^2 \), whereas the size of the minor grid was \( 2,500 \text{ ft}^2 \).
Finally, each record included information describing the severity of
the service call in terms of the type of the failure and the type of the
equipment that failed. A separate data base providing information on
the number of field units available on a shift basis per day was also
provided by the same utility.

Analysis of the data indicated statistically significant differences
in terms of the spatial, temporal, and severity distribution of the ser­
vice restoration calls. Important information regarding service
restoration operations was obtained by analyzing the duration of the
repair for work orders of different priorities. In this case, it was
found that the duration of the repair time was related to the severity
of the incident (4).

Information obtained from analysis of existing data was used to
develop probability density functions describing (a) the temporal
distribution of the service calls, (b) the spatial distribution of the
service calls, (c) the priority distribution of service calls, and (d) the
distribution of the average repair time for service calls having dif­
f erent priorities. In addition, the average travel time between the
centroids of the major grid atoms was calibrated using travel time
information provided by the data base.

**STRUCTURE OF PROPOSED SIMULATION
MODEL**

The simulation module input consists of the geographic definition
of the truck areas generated by the districting module and the travel
speed information for the links of the underlying transportation network. The input information is provided for all three shifts and for the entire area under consideration. The output of the simulation program provides the statistics describing the performance of the service restoration mechanism (i.e., the statistics of the total incident response time and its components). Figure 1 shows the relationship between the various components of the simulation module at a macroscopic level.

The proposed simulation program offers the opportunity to simulate a wide range of operational characteristics of the service restoration mechanism. The alternative simulations can be run by varying a set of control data in a control input file. On a more detailed level, the simulation module proceeds as follows: from the given data base, the priority distribution of service calls per shift is calculated for the whole area under consideration. Furthermore, the priority and type of service call distributions are calculated for each shift. On the basis of the calculated distributions and information regarding geographic and operational characteristics of the subarea, which is provided in control parameter input files, the following are calculated:

1. Repair time distribution for a given subarea, shift, priority, and type of service call;
2. Spatial distribution of service calls for a given shift, priority, and type;
3. Temporal distribution of service calls; and
4. Travel time matrix for the three shifts.

Once all the necessary distributions have been defined and calculated, the program proceeds with the generation of the following:

1. Calls on a shift basis;
2. Calls by priority;
3. Calls by type;
4. Call locations; and
5. Exact service call coordinates within the corresponding atom, using a uniform distribution.

The simulation program continues with the assignment of a service call to a particular service unit according to the defined dispatching policy. The activity of the ERU is followed throughout its shift. Any unfinished work load is transferred to the next shift. Performance and utilization statistics are gathered throughout the process. Figure 2 shows the detailed structure of the proposed simulation model. A 10-min delay is assigned to each ticket to compensate for $T_0$. Tickets are served according to their priority. Unserved tickets are transferred to the next shift.

**Temporal Generation of Calls**

Analysis of historical data describing the arrival of the service calls to the switchboard of the emergency repair system revealed that arrival of service calls can be described by a nonstationary Poisson process.

Therefore, a nonstationary Poisson process with rate $\lambda(t)$ was used to fit the service call arrival rate data (16). The rate $\lambda(t)$ was expressed by a polynomial function of the form

$$\lambda(t) = \sum_{i=0}^{n} a_it^i$$

where the constants $a_i$ and the order $n$ of the polynomial were calculated on the basis of the best fit to the historical data.

The generation of the temporal distribution of the service calls requires the solution of the following problem:

Given the instant $t_i$ of the arrival of the $i$th call, find $t_{i+1} = t_i + \Delta t$ (i.e., find the time of the next service call arrival).

From the nonstationary Poisson process, the distribution of an interval $(t_i, t_{i+1})$ is given by

$$F(t_i, \Delta t) = 1 - \exp \left[- \int_{t_i}^{t_{i+1}} \lambda(\tau) d\tau \right]$$

From the time interval distribution we get

\[ F(t_i, \Delta t) = 1 - \exp \left[- \int_{t_i}^{t_{i+1}} \lambda(\tau) d\tau \right] \]
The following computational procedure was used to determine the interval $\Delta t$:

1. Generate a uniform random number $\xi$.
2. Calculate $z_1 = \ln(1 - \xi)$, and let $z_2 = 0$, $t_0 = t$. These are the initialization conditions for dummy variables $z_2$ and $t_0$, which are used subsequently for the determination of the stopping criterion and the updating procedure of the algorithm. Let $\Delta t$ be a chosen small increment.
3. On the basis of the chosen $\Delta t$ and Simpson's rule, calculate

$$\Delta z_2 = \frac{\Delta t}{3} [\lambda(t_0 + \Delta t) + 4\lambda(t_0 + 2\Delta t)]$$

Let $z_2 = z_2 + \Delta z_2$ and $t_0 = t_0 + 2\Delta t$.
4. If $z_2 \geq z_1$, there is a need to get a smaller increase by interpolation, then
\[ t_{n+1} = t_0 - (z - z_i) \frac{2\Delta t}{\Delta z_2} \]

else go to Step 3.

5. End.

The procedure described above provides a stable solution to the temporal generation of calls.

Spatial Generation of Calls

A two-step procedure was used for the spatial generation of the service calls. First, the atom to which a call belongs was specified, and then the exact location within a specific atom was determined. To generate service calls in atoms, a discrete probability distribution was used.

The next step proceeds with the uniform generation of the service call within the specific atom of area \( A \), as determined in the previous step. A common procedure to generate a point uniformly in an atom of area \( A \) requires first the generation of a point within a rectangle that covers the given region and then a check to determine whether the point is located within the specific atom. This two-step procedure is repeated until a candidate point is finally acceptable (i.e., located within the atom).

For this paper, a new procedure is used, which assigns two random numbers to a corresponding point in the given atom of area \( A \) without the need for an iteration and checking procedure.

Because most of the atoms met in practical problems can be described or closely approximated as polygons, the problem is formulated for a polygon, which is described by the \( N \) counterclockwise-ordered vertices \((x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_N, y_N)\). We assume that a point \((x_0, y_0)\) can be selected within the polygon so that the segment line \((x_0, y_0) - (x_i, y_i)\) is located within the polygon for \( i = 1, 2, \ldots, N \) (Figure 3a). A ray from \((x_0, y_0)\) intersects the polygon at only one point. Let us denote the following:

- \( A_i = \) area defined by the triangle with vertices at \((x_0, y_0), (x_i, y_i)\), and \((x_{i+1}, y_{i+1})\) (Figure 3b);
- \( A = \sum_{i=1}^{N} A_i \) the total area of the polygon;
- \( B_i = \left( \sum_{j=i}^{N} A_j \right) / A \), the fraction of the polygon area in the first \( i \) triangles; and
- \( B_0 = 0 \).

The procedure is described as follows: First, generate a pair of uniformly distributed random numbers \((R_1, R_2)\) in \((0, 1)\). Second, find the location of \( R_1 \) among the \( B_i \). If \( B_{i-1} \leq R_1 < B_i \), calculate \( \alpha, x', y' \) such that

\[ \alpha = \frac{R_1 - B_{i-1}}{B_i - B_{i-1}} \]

\[ x' = x_i + \alpha(x_{i+1} - x_i) \]

\[ y' = y_i + \alpha(y_{i+1} - y_i) \]

The desired point \((x, y)\) is given by

\[ x = x_0 + \sqrt{R_2} (x' - x_0), \ y = y_0 + \sqrt{R_2} (y' - y_0) \]

This procedure directly generates points within the polygon without the need to check for acceptance or rejection of a certain point and the subsequently required iteration.

Priority, Type, and Repair Time Generation of Calls

A discrete distribution based on historical data is used to generate the priority of the service calls, the type of service calls of a certain priority, and the repair time of the service calls for a given shift, priority, and type.

Travel Time Estimation

The travel time estimation involves two steps. First, the travel time between the centroid of the atom where an ERU is located and the centroid of the atom where the service call originated is calculated. Second, the travel time from the actual location of the ERU and the service call to the corresponding centroid is calculated. The calculation of the travel time between the centroids of all atoms results in a travel time matrix whose elements are stored in a working file.
The matrix is calculated once for each shift. When a call is generated, the program identifies the atoms where the call and the ERUs are located and reads the corresponding travel time from the travel time matrix.

For the estimation of the shortest travel time matrix, all atoms are represented by the centroid \((x^*, y^*)\) of the atom.

Because of the nature of the underlying transportation network (i.e., regular grid pattern), the Manhattan metric was selected to represent the distance for the surface street system of the network (i.e., local streets and main arteries). When the assumption of the regular grid pattern is violated because of travel on freeways, the existence of barriers to movement, a rural area’s sparse road network, or for other reasons, special links are defined, and the network that was created by the special links is superimposed on the plane. Special links for freeways were defined by nodes representing entrance points to the freeway and the centroids of the neighboring atoms of the freeway. Special links around barriers were generated by a procedure described by Larson and Li (18).

Travel speed data for local streets, arterials, and freeways were collected for different hours of the day. For the surface street system, average travel speeds for each shift along the two major directions in the area \((v_x, v_y)\) (the X axis corresponds to an east-west movement and the Y axis to a north-south movement) were obtained from the utility’s personnel.

Arterial street speeds first were translated into equivalent local street speeds.

For freeway links, an average travel speed, \(v\), for each shift was used; whereas for the special links, an average speed, \(v_s\), for each shift was used. An algorithm is used to find the shortest travel time between all pairs of nodes and to determine the travel time matrix.

Given the atom center coordinate \((x_i, y_i)\), the average travel speed \((v_{x_i}, v_{y_i})\) within the atom, the coordinates of freeway points, and the allowable travel directions on it (entrance, exit, or both), the average travel speeds between two freeway points, special link information, and the travel speeds on special links, the following algorithm determines the travel time matrix.

1. Initialize travel matrix: \(T_{ij} = \infty\).
2. Calculate the travel time for each link.

For link \(i-j\) connecting the centroids of atoms \(i\) and \(j\), we use the weighted travel speeds:

\[
T_{ij} = \frac{|x_i - x_j|}{\alpha_i v_{x_i} + \alpha_j v_{x_j}} + \frac{|y_i - y_j|}{\alpha_i v_{y_i} + \alpha_j v_{y_j}}
\]

where \(\alpha_i\) is the percentage of the distance between the two atom centroids that belongs to atom \(i\) (2).

For atom \(i\) with centroid coordinates \((x_i, y_i)\) to freeway point link \(j\) with coordinates \((x_j, y_j)\) (note that freeway points are treated as centroids of atoms), use the average speed \((v_{x_j}, v_{y_j})\) within the atom:

\[
T_{ij} = \frac{|x_i - x_j|}{v_{x_j}} + \frac{|y_i - y_j|}{v_{y_j}}
\]

For a freeway point link connecting points \((x_i, y_i)\) to \((x_j, y_j)\), use given speed \(v\):

\[
T_{ij} = \frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{v}
\]

For special links connecting points \((x_i, y_i)\) to \((x_j, y_j)\), use given average speed \(v_{a_i}\):

\[
T_{ij} = \frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{v_{a_i}}
\]

3. Find the shortest way between all pairs of nodes using Floyd’s algorithm (a node is the centroid of an atom or a freeway point).

**CASE STUDY**

The simulation model described in the previous section was used to examine the trade-off between the number of available ERUs and the service restoration time and to evaluate the effectiveness of alternative dispatching and districting patterns produced by the solution of the districting problem. Initially, the simulation model was validated in terms of its capability to reproduce the inputs of the simulation process (i.e., temporal, spatial, and priority distribution of calls and distribution of the duration of the repair time). Statistical significance tests were performed to compare the means of the simulated and observed data for various shifts and districts of the study area. The results of the tests suggest that there is no statistically significant difference between the observed and simulated data.

The model is applicable to any type of area, transportation network, and population density. The calibration of the specific parameters is subject, however, to a case-by-case validation.

The validated model was used to study the relationship between the number of available ERUs (the fleet size) and the performance of the service restoration mechanism, as it is manifested by the duration of DTR time and its components \(T_2\), \(T_3\), and \(T_4\). First-come-first-served (FCFS) and nearest neighbor (NN) policies were used for the evaluation. Two distinct districting procedures were evaluated: constrained districting and unconstrained districting. According to constrained districting, the whole districting was divided by the utility personnel in three subdistricts, which satisfied their previous organizational and managerial structure, and then the model was used independently in each subdistrict. In unconstrained districting, the whole district was divided optimally into truck areas according to our model.

Figures 4 through 7 summarize the results of the evaluation for a particular district, for various numbers of ERUs, and compare them with results for the current districting pattern. The results presented involve Shift 1 (7 a.m. – 3 p.m.) and are run for \(N = 11–17\). The current districting pattern has \(N = 14\). A first result emerging from this analysis is that the average total DTR time decreases as the number of servers increases. The reduction is attributed mainly to the reduction of the dispatch delay (\(T_2\)) component. The equalization of work load achieved by the districting module has a very clear impact on \(T_2\); reductions in travel time are not that significant. The differences in average time shown in these figures were found to be statistically significant at the \(\alpha = 0.05\) level.

The results of the comparisons suggest that the new districting patterns lead to a more efficient performance of the service restoration mechanism, which is manifested through a statistically significant (\(\alpha = 0.05\)) reduction of the average DTR time for the same number of servers. In most cases, as many as two ERUs can be removed before any change in quality of service, as it is manifested by the value of DTR. Given the cost of operating each ERU, one can easily determine the cost savings from observed reductions in fleet size.
FIGURE 4  Relationship between DTR and number of ERUs (unconstrained districting, FCFS, Shift 1).

FIGURE 5  Relationship between DTR and number of ERUs (unconstrained districting, NN, Shift 1).

FIGURE 6  Relationship between DTR and number of ERUs (constrained districting, FCFS, Shift 1).
A comparison of the results of constrained and unconstrained districting indicates that unconstrained districting (Figures 4 and 5) provides better performance than constrained districting (Figures 6 and 7).

A comparison of the results of FCFS and NN policies indicates that NN gives consistently better results than FCFS, especially when the work load becomes very high (as with a small number of trucks).

CONCLUDING REMARKS

A simulation model for evaluating operations of the emergency repair fleet of an electric utility company was presented. The proposed simulation model has the capability to simulate alternative dispatching strategies (i.e., FCFS, NN, and their combinations). It is able to evaluate the performance of the emergency repair fleet in terms of the duration of DTR time. The model provides utility companies with an effective analysis and decision-making tool. Sample applications of the simulation model include an examination of the trade-offs between fleet size and duration of DTR, evaluation of alternative designs of service restoration districts, and examination of the effectiveness of alternative dispatching policies under various conditions of temporal, spatial, and severity distributions of service calls.

ACKNOWLEDGMENT

This work was supported by Florida Power & Light Company.

REFERENCES