TRANSPORTATION RESEARCH
RECORD

No. 1497

Planning and Administration

Artificial Intelligence and Geographical Information

A peer-reviewed publication of the Transportation Research Board

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY PRESS
WASHINGTON, D.C. 1995

Transportation Research Record 1497
ISSN 0361-1981
ISBN 0-309-06163-6
Price: $35.00

Subscriber Category
IA planning and administration

Printed in the United States of America

Sponsorship of Transportation Research Record 1497

GROUP 2—TRANSPORTATION SYSTEMS PLANNING AND ADMINISTRATION
Chairman: Thomas F. Humphrey, Massachusetts Institute of Technology

Transportation Forecasting, Data, and Economics Section
Chairman: Mary Lynn Tischer, Virginia Department of Transportation

Committee on Transportation Data and Information Systems
Chairman: Arthur B. Sosslau
Secretary: James J. McDonnell
David Preston Albright, Marsha Dale Anderson, Jack A. Butler, Patrick R. Cain, Ed J. Christopher, Gary Q. Coe, Keith A. J. Crawford, Kenneth J. Dueker, Patricia S. Hu, Martha M. Johnson, Charles A. Lave, Alan E. Pisarski, Charles L. Purvis, Phillip A. Salopek, Eddie Shafie, Howard J. Simkowitz, Ronald W. Tweedie, Xu Weici, George F. Wiggers

GROUP 5—INTERGROUP RESOURCES AND ISSUES
Chairman: Patricia F. Waller, University of Michigan

Committee on Artificial Intelligence
Chairman: Michael J. Demetsky, University of Virginia
Gerry B. Andeen, Nii O. Attoh-Okine, M. Hadi Baaj, Anselmo Osvaldo Braun, Edmond Chin-Ping Chang, Louis F. Cohn, David J. Elton, Ardeshir Faghri, Jerry J. Hajek, Chris T. Hendrickson, Safwan A. Khedr, Shinya Kikuchi, David Martinelli, Olin W. Mintzer, Prahlad D. Pant, Ajay K. Rathi, W. M. Kim Roddis, Shashi Kumar Sathisan, Rein Schandersson, Yung-Ching Shen, Gary S. Spring, James A. Wentworth, Charles A. Wright

Transportation Research Board Staff
Robert E. Spicher, Director, Technical Activities
James A. Scott, Transportation Planner
Nancy A. Ackerman, Director, Reports and Editorial Services

Sponsorship is indicated by a footnote at the end of each paper. The organizational units, officers, and members are as of December 31, 1994.

Transportation Research Record 1497

Contents

Foreword

This record contains a series of papers dealing with uncertainty related to the engineering decision, the use of fuzzy relations to manage uncertainty in civil infrastructure preservation, and rehabilitation of deteriorated pavement sections. Also, a series of papers focuses on use of neural networks for real-time arterial street incident detection, waterway lock service times, automatic bus vehicle location, traffic signal control, and pavement maintenance strategy.

Six papers focus on the application of the Geographic Information System (GIS), hazardous highway location identification, safety analysis at rail-highway grade crossings, assigning spatial properties to highway feature files, Intelligent Transportation System/Intelligent Vehicle Highway System (ITS/IVHS) networks, network travel time, solid and hazardous waste transportation planning, and modeling freight flows.

The two remaining papers focus on an approach for solving the train formation problem and selection of highway design parameters in the presence of uncertainty.

Selection of Highway Design Parameters in the Presence of Uncertainty

M. NAZRUL ISLAM AND P. N. SENEVIRATNE

AASHTO guidelines on highway design have drawn criticism for their inability to deal with the uncertainty of traffic operations, costs, and physical constraints. Some analysts believe that in light of changing economic and socio-environmental values, new procedures are needed to better address uncertainties and to justify engineering decisions. An analytical model that could be used to determine the optimal design curvature (D_d^*) for horizontal curves on two-lane highways is presented. The optimal curvature results in the minimum total cost, defined as the sum of construction, maintenance, and expected user costs. The expected user cost is the sum of expected accident, travel time, and vehicle operating costs. It is shown that: (*a*) D_d^* is highly sensitive to the skewness of the probability distribution of the required curvature; (*b*) when the mean operating speed is high, D_d^* does not change significantly with the changes of standard deviation of speeds compared with the low mean operating speed; and (*c*) when the mean operating speed is low, the polynomial model best represents the relationship between D_d^* and the standard deviation of the operating speeds. When the mean operating speed is high, the linear model best represents the relationship between D_d^* and the standard deviation of the operating speeds. The application of the model and sensitivity of the optimum to model parameters are illustrated using numerical examples.

The current policy on highway geometry, published by the American Association of State Highway and Transportation Officials (AASHTO) (*1*), seeks to promote safety through the use of the highest design standards. This traditional approach to highway design has drawn criticism recently (*2–5*). Critics argue that the higher design standards are not always justifiable when working under budget constraints and do not necessarily guarantee better safety due to numerous uncertainties. The most significant of these uncertainties are the characteristics of drivers and vehicles. These concerns, and the consensus that adding safety factors is not the most cost-effective and prudent way to treat uncertainty, have heightened the need for new approaches to roadway design. As in other disciplines of science and engineering, these approaches should strive for a balance between costs and benefits of a particular design when many factors are uncertain.

The development of an analytical model for determining the optimal degree of design curvature (D_d^*) of a horizontal curve on a two-lane highway is discussed. In many practical cases, the optimum is not always attainable. The decisions are affected by one or more constraints. Road geometry is a classic example of a constrained case, in which physical and environmental factors features limit design options. Thus, the optimization model is developed under two scenarios, constrained and unconstrained, and its application is illustrated by a numerical example.

M. N. Islam, Cambridge Systematics, Inc., 150 Cambridge Park Drive, Suite 4000, Cambridge, Mass. 02140. P. N. Seneviratne, Department of Civil and Environmental Engineering, Utah State University, Logan, Utah 84322-4110.

CURRENT PRACTICE

The degree of curvature required to permit the vehicle to negotiate a simple horizontal curve at a particular speed can be determined by the following fundamental relationship:

$$D = \frac{85{,}660\,(e + f)}{V^2} \tag{1}$$

where

D = curvature required by the individual vehicle speed (degrees per 100-ft curve length),
V = speed (mph),
e = superelevation rate in feet per feet, and
f = side friction factor at speed V.

If the curve is designed with a curvature of D_d, because V is a random quantity due to differences in vehicle and driver characteristics, the curvature required by a given vehicle may be less than, equal to, or greater than the design curvature. This phenomenon, and several other assumptions underlying the current practice of horizontal curve design, must be addressed by the new approach.

The two key aspects in need of attention stem from the following:

1. Currently, a high percentile speed is chosen as design speed (V_d) irrespective of the shape or form of the distribution of D.
2. Within a given functional class of roadway, variations in traffic volume and mix are disregarded.

Two other concerns have also emerged. The first is that although changes in operating speed due to inconsistencies in horizontal alignment have been found to be a leading cause of accidents (*6–8*), no formal mechanism exists to ensure consistency when selecting design speed. The second concern is related to cost-effectiveness. Although smaller D_ds mean higher construction costs, they also mean lower accident rates (*2,9,10,11*), operating costs, and travel times. But the trade-off between costs and savings is neither clear nor explicit in AASHTO (*1*), which makes assessing cost-effectiveness difficult, if not impossible.

In the next two sections, the components of the optimization model are presented, and the sensitivity of the optimal curvature to the various cost parameters is discussed.

OBJECTIVE FUNCTION

After the cost components are defined, the design degree of curvature (D_d) that minimizes the total cost can be sought in several ways.

The objective function is defined as follows:

$$TC_{\min} = C_a + C_{tt} + C_{op} + C_c + C_m \quad (2)$$

where

$TC_{\min}$ = minimum total cost,
C_a = expected accident cost,
C_{tt} = expected travel time cost,
C_{op} = expected vehicle operating cost,
C_c = expected construction cost,
C_m = expected maintenance cost.

The next section discusses the optimum solution when there are no constraining circumstances, and then identifies some common constraints and their impact on the solution.

USER COSTS

Expected Cost of Accidents

Although some researchers have expressed different views on the relationship between accident rate and D_d, the consensus is that (*a*) accident rates are higher on horizontal curves than on tangents, (*b*) rates increase as D_d increases, and (*c*) D_d is the most significant geometric feature contributing to accidents (*3–5*).

Considering all the variables cited in TRB (*12*) and Zegeer et al. (*13*), a generalized accident prediction model can be expressed as

$$A_m = g(D) = a\left[\frac{0.0189\,bI}{D_d} + cD_d - dS\right] \quad (3)$$

where

a, b, c, and *d* are calibration constants,
I = external angle in degrees,
A_m = number of accidents in a curve, and
S = superelevation in feet per feet.

If, as stated earlier, the safest design curvature ($D_{\min}$) is that which corresponds to the maximum operating speed, anything larger increases the geometry-related accident rate in proportion to the difference between D_d and $D_{\min}$. This demand-supply concept is explained in detail by the authors in previous articles (*14,15*), and by Newman and Glennon (*16*) in the case of stopping sight distances. Hirsch et al. (*5*) have also used the same reasoning that accidents occur when the design radius of curvature is smaller than the radius required by a vehicle traveling at a specific speed. However, if the selected design value is equal to or greater than the required value, they assumed the accidents were unrelated to radius.

Following the demand-supply concept, two regimes are defined for D in the present case; one when $D \geq D_d$, and the other when $D < D_d$. In the first regime, accidents are assumed to be unrelated to curvature; that is, the nonhuman-error and environment-related accidents that may occur at the curve are not directly influenced by the curvature. Therefore, the number of accidents in this regime is considered to be zero. In the second regime, in which demand exceeds supply, accidents are proportional to the deviation of D from D_d. The number of accidents can, therefore, be expressed as

$$A = \begin{pmatrix} N\,h(D)\,(D_d - D) & D < D_d \\ 0 & D \geq D_d \end{pmatrix} \quad (4)$$

where N is the annual traffic volume in millions of vehicles, and $h(D)$ is the rate change in accident per unit change of D per year per million vehicles.

Considering a generalized form of the probability density of D [i.e., $\phi(D)$], the exceedance probability [$P(D < D_d)$] or likelihood of the deficiency may be written as

$$P(D < D_d) = \int_0^{D_d} \phi\,(D)\,dD \quad (5)$$

Therefore, the expected costs of accidents when design curvature at a given site is D_d can be expressed as

$$C_a = \gamma_a N \int_0^{D_d} h(D)\,(D_d - D)\,\phi\,(D)\,dD \quad (6)$$

where γ_a is the weighed average cost per accident.

The cost parameter γ_a depends on the type of accident and the average cost of each type of accident. The first step is to define the type of accident and its proportion, and the second step is to estimate the average cost of each type of accident. Both types of accident and its costs vary from state to state and can be obtained from the state accident data base. Although the accident type can be a function of curvature, in the present case it is assumed to be a constant. Therefore, the weighted average accident cost (γ_a) can be estimated by using the formula:

$$\gamma_a \sum_{k=1}^{n} [C_k\,P_k] \quad (7)$$

where

C_k = average cost per type k accident,
P_k = proportion of type k accidents, and
k = accident type according to severity, 1, 2, 3, . . . , n.

Expected Vehicle Operating Cost

Drivers reduce their speeds when they approach a curve and accelerate after they enter or pass the curve (*17*). This speed-change cycle consumes fuel and engine oil, wears tires, and increases maintenance costs, which are all listed as significant in the AASHTO guidelines (*18*). When vehicle speeds are distributed over a wide range, the additional operating cost of each vehicle due to a particular curvature is a function of the difference between the operating costs at D_d and the operating cost at D. In other words, when $D \geq D_d$, excess vehicle operating cost on a curve is zero. Otherwise, it is assumed to be proportional to the difference in the operating costs at the two speeds. When D is a random quantity with a known density function, the expected operating cost (C_{op}) is expressed as

$$C_{op} = N\gamma_\beta \int_0^{D_d} (D_d - D)^m \phi(D)\,dD \quad (8)$$

where

C_{op} = expected vehicle operating cost,
γ_β = rate of change of operating cost per unit change of D_d per million vehicles, and
m = exponent.

Calculation of γ_β

AASHTO (*18*) provides tables of operating cost in dollars per 1,000 veh/mi above cost of tangent with respect to D_d and speed. It is also known from Islam and Seneviratne (*17*) that: $V_{85} = 62.4 - 1.46\,D_d + 0.018\,D_d^2$

Expressions for γ_β depend on the central angle. For example, when $I = 50°$, the regression equation in 1975 dollars can be formulated as

$$C_{75} = 5229 + 87.5\,D_d\,(R^2 = 0.87)$$

The preceding equation may be converted to 1992 dollars assuming a 7 percent discount rate as

$$C_{92} = 16524 + 276.5\,D_d$$

that is $\gamma_\beta = 276.5$

Expected Travel Time Cost

According to the TRB Special Report 214 (*12*), the cost of travel time should be a principal determinant of geometric elements, although Lin (*20*) has ignored it in his work. However, the value of travel time and the amount of travel time saved are the two key aspects of travel time. Using design speed to estimate travel time has no justification because all vehicles do not operate at design speed. Furthermore, if the operating speed is less than the design speed, the travel time saved is zero. Therefore, only those vehicles whose operating speeds are greater than design speed are considered. This delay can be expressed as

$$\text{Delay} = \left[\frac{L_d}{V_d} - \frac{L}{V}\right] \tag{9}$$

where

L = required length of the curve at speed V in miles,
V = operating speed on tangent in mph,
L_d = required length of curve at V_d in miles, and
V_d = design speed in mph.

Thus, the expected cost of delay after expressing all parameters in terms of D can be written as

$$C_{tt} = \alpha\gamma_t N \int_0^{D_d} [\omega\,(D) - \omega\,(D_d)]\phi\,(D)\,dD \tag{10}$$

where

γ_t = value of travel time ($/unit time),
$\alpha = 0.0189\,I$,
$\omega(D) = \dfrac{1}{DV}$,
$\omega(D_d) = \dfrac{1}{D_d V_d}$, and
$L = \dfrac{0.0189\,I}{D}$.

CONSTRUCTION AND MAINTENANCE COSTS

Construction Cost

Construction and maintenance costs depend on a variety of factors, including site conditions, labor and materials costs, design practice, and project scale. Pavement, shoulder, and side slope design standards vary from state to state. For example, New York State usually paves shoulders, whereas Virginia constructs gravel or turf shoulders. Labor costs in San Francisco are nearly double those in Jackson, Mississippi (*21*). Unit price for construction also depends on the size of the project.

The relationship between annual construction cost (C_c) and D_d for a specific site in a particular region may be linear, quadratic, or inverse, and the unit costs may vary from state to state or even from place to place. Three forms of the generalized expression for construction cost $P(D_d)$ are proposed in the present case. They are:

$$C_c = P(D_d) \tag{11}$$

where

$$P(D_d) = a - \gamma_b\,D_d \text{ or,} \tag{11a}$$
$$P(D_d) = c - \gamma_d\,D_d + \gamma_e\,D_d^2 \text{ or,} \tag{11b}$$
$$P(D_d) = \gamma_f/D_d. \tag{11c}$$

Maintenance Cost

The maintenance cost for projects is a function of the length of the roadway. The annual maintenance cost of a curve (C_m) can therefore be expressed in terms of its length as follows:

$$C_m = \gamma_k\,[L + L_c/2640] \tag{12}$$

where

γ_k = annualized maintenance cost per mile,
L = length of circular curve in miles, and
L_c = length of spiral in miles.

When the length of the curves is expressed in terms of the degree of the curvature, the cost of maintenance also becomes a function of D_d. For developing a model for obtaining D_d^*, it is necessary to know the cost of construction and maintenance per unit change of D_d. First derivative of Equations 11 and 12 with respect to D_d will give the construction and maintenance costs per degree change in D_d, and the tangent length will not be a factor.

OPTIMAL CURVATURE

Unconstrained Case

According to the definition of total cost, the objective function is the sum of the cost given by Equations 6, 8, 10, 11, and 12, which are expressed in terms of D_d. Therefore, the optimal value of D_d could be derived by taking the first derivative of the objective function expressed as Equation 2, and equating it to zero, as follows:

$$\frac{\partial(TC)}{\partial(D_d)} = 0 = f'(D_d) \tag{13}$$

When $m = 1$ and (C_c) is linear, there is no closed form solution to $f'(D_d)$. Thus, Equation 13 can be rearranged as follows and solved graphically or numerically to obtain the value of D_d leading to the lowest TC:

$$\int_0^{D_D} \phi(D)\, dDh(D_d) + \gamma_\beta + \gamma_t \alpha\omega'(D_d) = \frac{\gamma_b + \gamma_k [0.0189\, I]}{N} \quad (14)$$

Additionally, the sensitivity of D_d^* to the various parameters can be tested by changing one while the others are held constant.

Numerical Examples

Assuming that speeds are normally distributed with a mean of 50 mph and a standard deviation of 7 mph, random speeds were generated using the Monte Carlo method. These speed deviates were then substituted in Equation 1 to obtain the distribution of D for that speed distribution. A chi-square goodness-of-fit test performed on the distribution of D, [ϕ(D)] indicated that it was normally distributed with a mean of 9.12° and a standard deviation of 3.21°. For the cost parameters and the road alignment, the following values were used: I = 50°, γ_t = \$12, γ_b = \$1,175, γ_d = \$1,175, γ_e = \$30, γ_f = \$3,175, γ_k = \$116, and $h(D)$ = 0.0336. Most of these values were obtained from the TRB Special Report 214 (*12*).

The graphical solutions to Equation 14 (when m = 1, 2, and 3, and the construction cost function is linear, quadratic, and inverse) are illustrated in Figure 1. It shows that when $m = 1$, the construction cost is linear, the minimum cost occurs at D_d^* = 10°. The sensitivity of optimum solution to the construction cost can also be seen in Figure 1. When C_c is given by Equations 11(b) and 11(c), the optimum solutions are D_d^* = 8.75° and 5.5°, respectively.

The optimum solution is also sensitive to the accident cost function, but not to the same extent as construction cost. For example, as shown in Figure 2, when the accident cost function is linear or the rate change of accident rate with respect to the degree of curvature is constant (and construction cost is a linear function), the optimum occurs at D_d = 10°. When the accident cost function is nonlinear (the form given by Equation 3), the optimum occurs at D_d^* = 9.7°.

A comparison of the optimum values and the AASHTO (*1*) recommended values at different design speeds is shown in Table 1. For example, when $m = 1$, and C_c is linear, the optimum value from the model is 10°, but the AASHTO value at the 85th percentile speed of the same distribution used in the model is 5.4°. However, when $m = 1$ and C_c is inverse, the modeled value is closer to the AASHTO value when the 80th percentile is used as the design speed. A similar comparison is shown in Table 2 for the case when the accident prediction model is nonlinear.

Constrained Case

In most real-world engineering problems, the objective functions are subjected to several constraints. It is relatively easy to solve a simple optimization problem that is unconstrained, but if constraints are imposed on the problem, few efficient solution techniques are available. The mathematical technique of Lagrange multipliers (*22*) is one of those techniques, but it can be used only when constraints are strict equalities. However, Kuhn-Tucker (*22*) has taken the concept of Lagrange multipliers from mathematical models with active constraints and extended them to mathematical models with active and inactive constraints. In the present case, the follow-

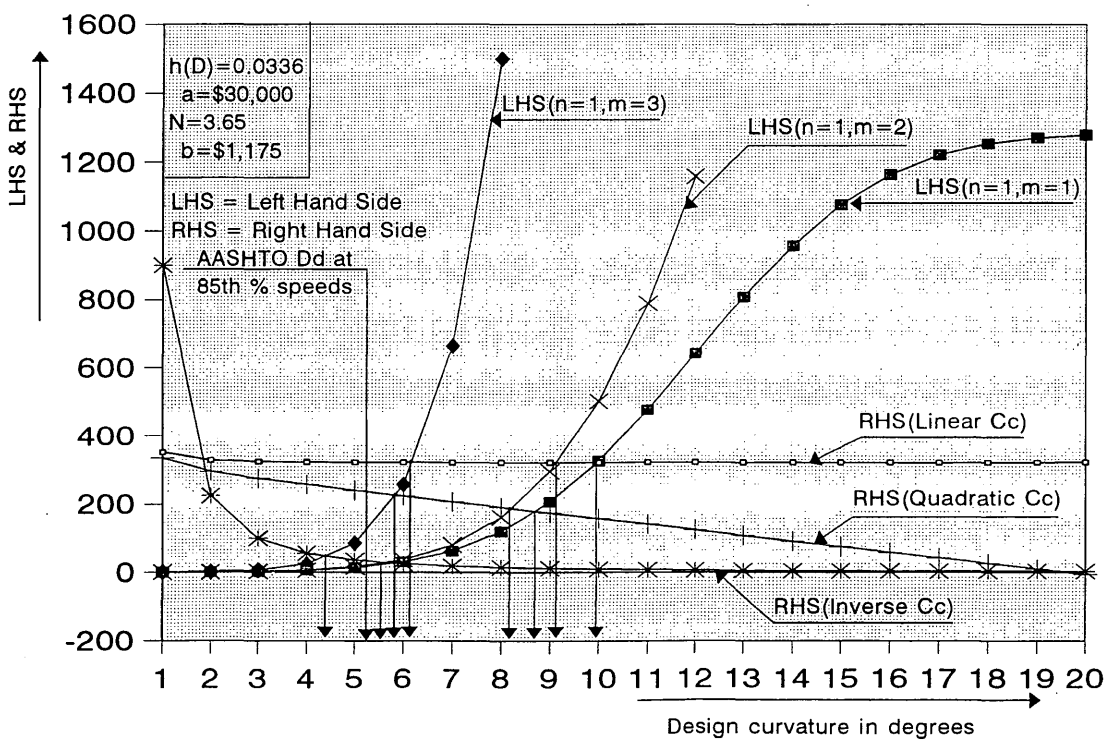


FIGURE 1 Optimum curvature under different curve density and construction cost functions.

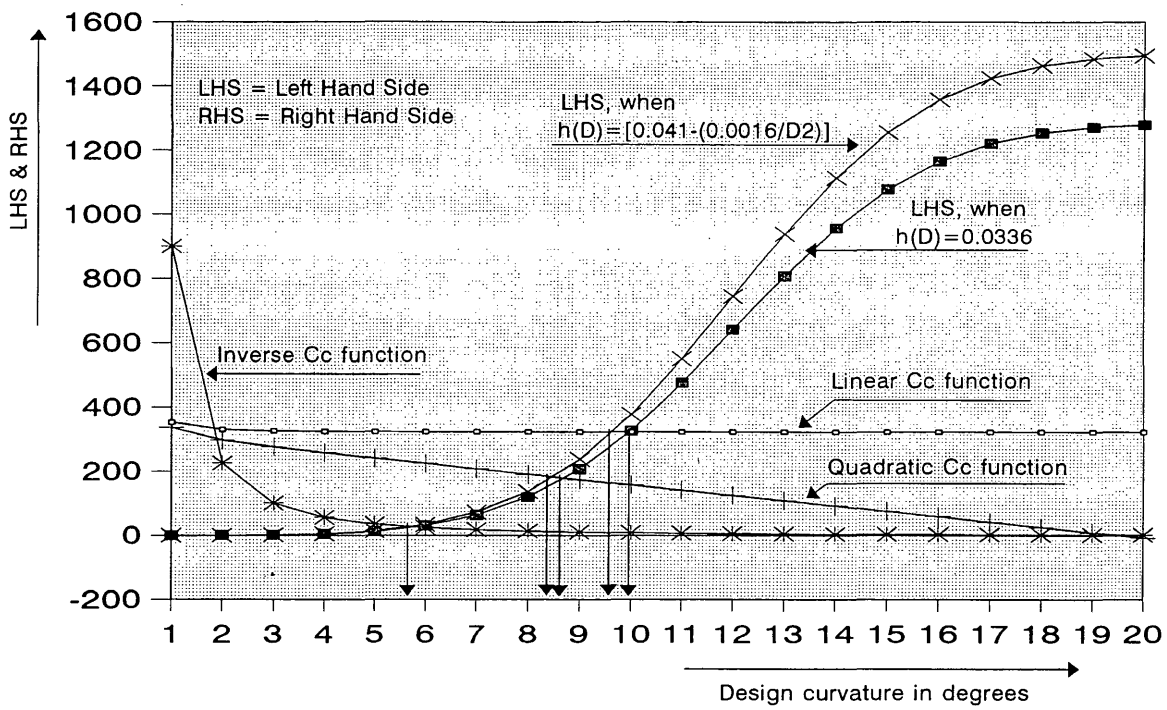


FIGURE 2 Optimum curvature under different accident prediction models.

ing two inequality constraints are considered in developing the model, and the Kuhn-Tucker technique is applied to obtain the optimum value.

Consistency Constraint

Abrupt changes in operating speeds lead to accidents on rural roads, and speed inconsistencies may be largely attributed to abrupt changes in horizontal alignment (i.e., changes in D_d). Lamm et al. (*8*), and Leisch and Leisch (*6*) studied these inconsistencies and suggested maximum allowable speed differentials between two curves and between a tangent and a curve. In the present optimization model, this condition may be expressed in terms of curvature as

$$(D_d - D_a) \leq 0 \tag{15}$$

where D_a is the maximum allowable design curvature from a consistency point of view.

This form of constraint ensures that sharp speed drops are avoided during the optimization process, and that the consistency requirements become an integral part of the analysis.

Environmental and Archaeological Constraint

Environmental and archaeological constraints are key determinants of curvature. However, as environmental and social awareness grow, roadway alignments and dimensions have to be selected in response to those needs. Therefore, a condition was included in the present model to ensure that the sight distance requirements are met under the constraints. This condition may be written in terms of the degree of curvature as follows (*19*):

$$D_s \geq D_d$$
$$\text{or } (D_s - D_d) \geq 0 \tag{16}$$

where D_s is the maximum allowable curvature when middle ordinate is fixed.

The Kuhn-Tucker function and the pertinent constraints can now be written as

$$\begin{aligned} L(D_d, \lambda_1, \mu_1) = {} & \gamma_a N h(D_d) \int_0^{D_d} [D_d - D]\, \phi(D)\, dD \\ & + \gamma_\beta N \int_0^{D_d} [[D_d - D]^m \phi(D)\, dD] \\ & + \gamma_t \alpha N \int_0^{D_d} [\omega(D) - \omega(D_d)]\, \phi(D)\, dD \\ & + C_{cm} + \lambda_1 [D_a - D_d] + \mu_1 [D_s - D_d] \end{aligned} \tag{17}$$

subject to:

$$\lambda_1 (D_a - D_d) = 0,$$
$$\mu_1 [D_s - D_d] = 0,$$
$$D_d \leq D_a,$$
$$D_d \geq D_s,$$
$$\lambda_1 \leq 0,$$
$$D_d \geq 0,$$
$$\mu_1 \geq 0,$$

where λ_1 and μ_1 are control variables associated with less-than-or-equal-to or greater-than-or-equal-to constraints.

The Kuhn-Tucker conditions give a different insight into the nature of the optimum values. For a minimization problem, the Lagrange function must be a minimum. Because it is a sum of terms, each term must be a minimum. Accordingly, in the present case, the

TABLE 1 Optimum D_d and AASHTO Recommended D_d for Different n and m Values, Construction Cost Functions, and $h(D) = 0.0336$

m	Optimum Design Curvature			AASHTO Recommended Dd			
value	Linear Cc	Quadratic Cc	Inverse Cc	@ 80th % speed	@ 85th % speed	@ 90th % speed	@ 95th % speed
1	10	8.7	5.9	5.8	5.4	4.8	4.4
2	9.3	8.2	5.8	5.8	5.4	4.8	4.4
3	6.3	5.8	4.4	5.8	5.4	4.8	4.4

term $\lambda_1(D_a - D_d)$ will be minimized when $\lambda_1 = 0$ and $(D_a - D_d) \geq 0$, or when $\lambda_1 \leq 0$ and $(D_a - D_d) = 0$. When the constraint is inactive, the Lagrange multiplier will be equal to zero. If λ_1 is equal to zero, the constrained equation will not influence the problem or its minimum value. On the other hand, when the constraint is active, $(D_a - D_d) = 0$. A similar argument can be given to justify the environmental constraint.

Now, D_d, which minimizes the Lagrange function, can be obtained by taking the first derivative of Equation 17 with respect to D_d and equating to zero.

Numerical Example

The impact of the constraints in the optimal solution can be best illustrated using a numerical example. Assuming the following cost parameters are used in Equation 17, the first derivative of it to D_d, when equated to zero, takes the form

$$4675 \int_0^{D_d} \phi(D)\, dD - 1175 - \frac{110}{D_d^2} - \lambda_1 - \mu_1 = 0 \qquad (18)$$

Suppose also that D_a is set equal to 15°, following Leisch and Leisch (*6*), who have suggested that the speed change between a curve and a tangent should be less than or equal to 10 mph. As for the environmental constraints, assume that the middle ordinate cannot exceed 200 ft, and the external angle I is either 50° or 20°. These two curve parameters place a lower bound of 3.2° (when $I = 50°$) or 9.3° (when $I = 20°$) on D_s.

For the Kuhn-Tucker conditions to be satisfied, either $\lambda_1 = \mu_1 = 0$; or $\lambda_1 \leq 0$ and $\mu_1 \geq 0$. If $\lambda_1 = \mu_1 = 0$, then D_d^* is not influenced by the constraints and can be derived graphically or numerically as illustrated in the previous case. Otherwise, it should lie between D_s and D_d. For example, it can be seen in Figure 3, where D_d^*s is shown under different speed distribution parameters (V_{mean} and V_{std}), that at $V_{mean} = 47$ mph and $V_{std} = 10$ mph, D_d^* is 8.0°. Therefore, it satisfies both constraints. However, if $V_{mean} = 45$ mph, then the environmental constraint becomes active in that D_d^* should be equal to D_s. Similarly, the active and inactive constraints when V_{std} varies can be seen in Figure 4.

The sensitivity of D_d^* to the cost parameters was tested by changing one parameter at a time while the others were held constant. Subsequently, regression analyses were performed to determine the extent of the sensitivity. In the case of both γ_a and γ_b, nonlinear relationships were observed:

$$D_d^* = 18.3 + 0.00177\,\gamma_a - 0.0796\,\gamma_a^{0.5} \qquad (19)$$
$$(R^2 = 0.98)$$

$$D_d^* = 5.76 + 0.005\,\gamma_b - 0.00001\,\gamma_b^2 \qquad (20)$$
$$(R^2 = 0.99)$$

The relationships are shown in Figures 5 and 6, respectively.

CONCLUSIONS

The optimization model allows uncertainties in traffic operations to be incorporated into the decision-making process by seeking the optimal design curvature under different speed distributions. Because speed distributions depend on traffic mix, terrain, and driver characteristics, if the expected distribution at a selected site can be accurately described, uncertainty stemming from the stochasticity of traffic operation can be treated effectively. In examining the sensitivity of optimal curvature to cost functions, it became evident that the form of the construction function (whether linear or nonlinear) has more of an influence than the form of the accident cost function. Likewise, the unit cost of accidents and construction have different impacts in that the rate of change in optimal curvature with respect to the costs are linear in the case of accidents and nonlinear in the case of the construction. That difference can be seen in Figures 5 and 6.

The authors believe that the most important feature of this model is its ability to make the process of selecting the design curvature a formal and integral one. The constraints and costs can be considered simultaneously instead of at different stages of the design process.

TABLE 2 Optimum Dd and AASHTO Recommended Dd for Different n and m Values, Construction Cost Functions, and $h(D) = [0.0412\text{-}(0.0016/D^2]$

m	Optimum Design Curvature			AASHTO Recommended Dd			
value	Linear Cc	Quadratic Cc	Inverse Cc	@ 80th % speed	@ 85th % speed	@ 90th % speed	@ 95th % speed
1	9.6	8.6	5.5	5.8	5.4	4.8	4.4
2	8.9	8.1	5.4	5.8	5.4	4.8	4.4
3	6.3	5.8	4.4	5.8	5.4	4.8	4.4

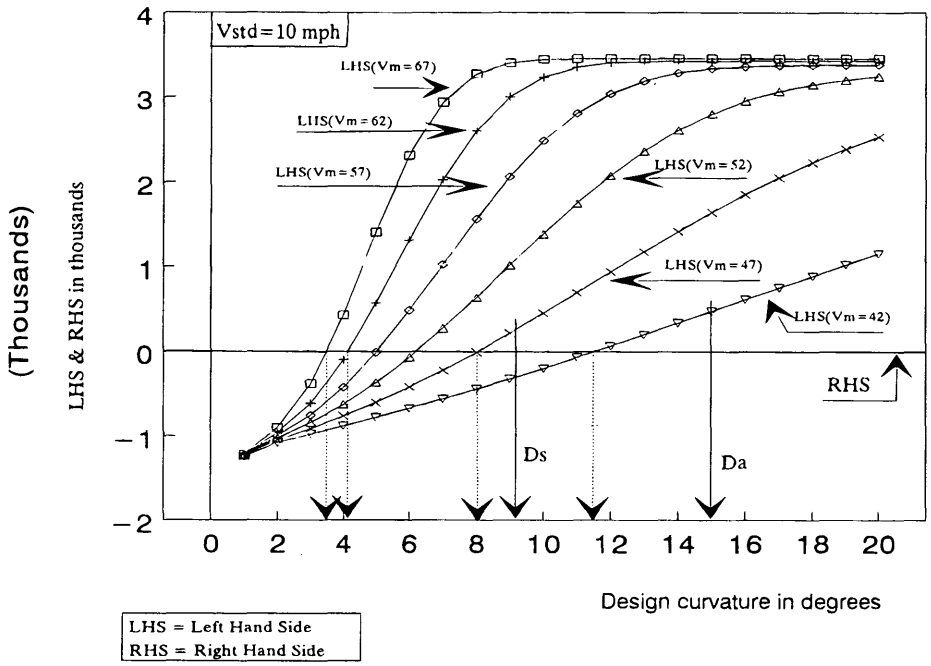


FIGURE 3 Sensitivity of optimum curvature to speed distribution parameters.

The sensitivity can be tested instantly. Moreover, the model provides the engineer with a systematic and rational basis for justifying designs under uncertainty. The ability to incorporate experience and subjective judgment into the decision-making process through the definition of exceedance probabilities and cost parameters gives the designer added flexibility and a sense of personal involvement.

The authors acknowledge that some designers may be apprehensive with this approach, particularly regarding the validity of the underlying accident prediction models and construction cost models. The authors believe, however, that this problem will be resolved as better prediction models become available. Finally, the present model is not intended to provide precise answers or to gen-

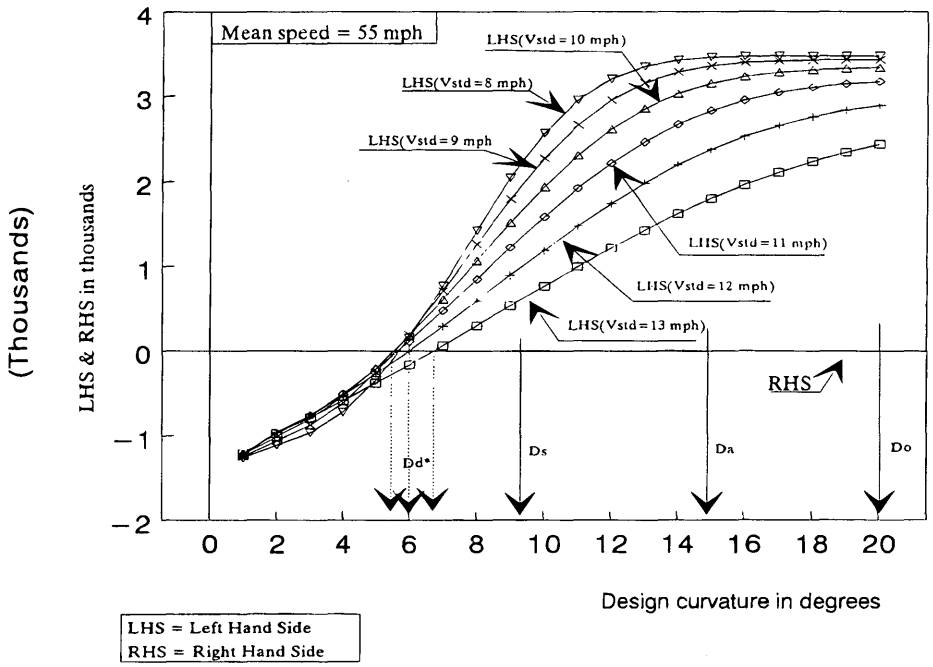


FIGURE 4 Sensitivity of optimum curvature to speed distribution parameters.

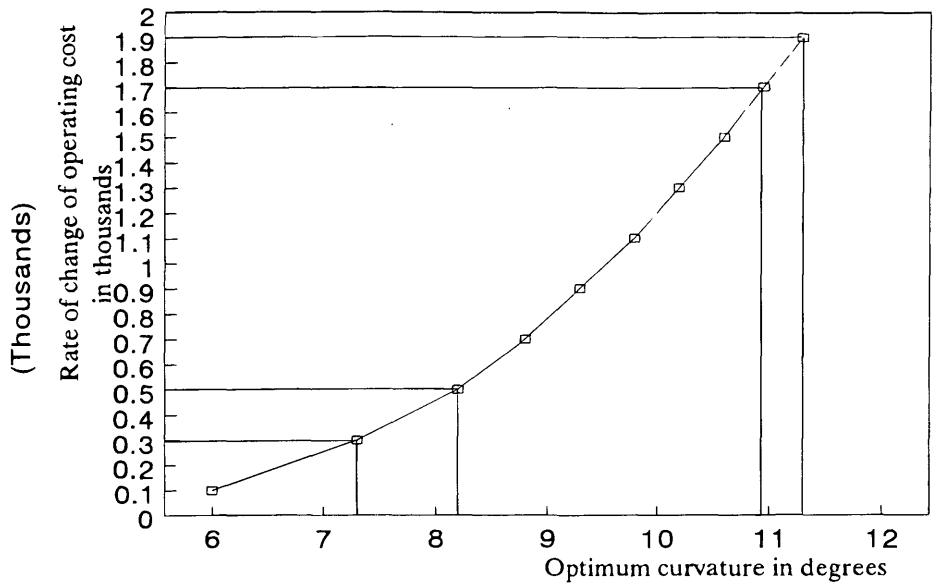


FIGURE 5 The rate of change of operating cost as a function of optimum curvature when construction cost is linear.

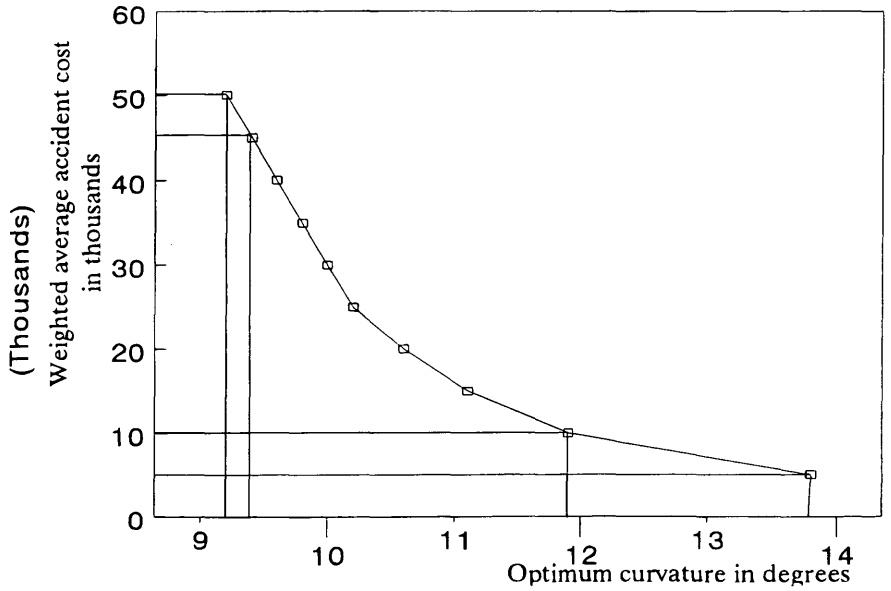


FIGURE 6 The rate of change of weighted average accident cost as a function of optimum curvature when construction cost is linear.

erate exact design values, but is meant as a tool that can be used to compare and perhaps evaluate design values in AASHTO (*1*) or similar manuals.

REFERENCES

1. *A Policy on Geometric Design of Highways and Streets.* American Association of State Highway and Transportation Officials, Washington, D.C., 1990.
2. Newman, T. R. Design Risk Analysis. Address to 71st Annual Meeting of the Transportation Research Board, National Research Council, Washington, D.C., 1992.
3. Navin, F. P. D. Safety Factors for Road Design: Can They Be Estimated? In *Transportation Research Record 1280,* TRB, National Research Council, Washington, D.C., 1990, pp. 181–189.
4. Ben-Akiva, M. Probabilistic and Economic Factors in Highway Geometric Design. *Transportation Science,* Vol. 19, No. 1, Feb. 1985, pp. 38–57.
5. Hirsch, M., J. N. Prashker, and M. Ben-Akiva. New Approach to Geometric Design of Highways. In *Transportation Research Record 1100,* TRB, National Research Council, Washington, D.C., 1986, pp. 50–57.
6. Leisch, J. E., and J. P. Leisch. New Concepts in Design-Speed Application. In *Transportation Research Record 631,* TRB, National Research Council, Washington, D.C., 1977, pp. 4–14.
7. Lamm, R., E. M. Choueiri, J. C. Haywood, and A. Paluri. Possible Design Procedure to Promote Design Consistency in Highway Geometric Design on Two-Lane Rural Roads. In *Transportation Research*

Record 1195, TRB, National Research Council, Washington, D.C., 1988, pp. 111–122.
8. Lamm, R., and E. M. Choueiri. Recommendations for Evaluating Horizontal Design Consistency Based on Investigation in the State of New York. In *Transportation Research Record 1122,* TRB, National Research Council, Washington, D.C., 1987, pp. 68–78.
9. Dart, O. K., Jr., and L. Mann, Jr. Relationship of Rural Highway Geometry to Accident Rates in Louisiana. In *Highway Research Record 312,* HRB, Washington, D.C., 1970, pp. 1–16.
10. Jorgensen, R. *Cost and Safety Effectiveness of Highway Design Elements.* Roy Jorgensen & Associates, National Cooperative Highway Research Program Report 197, TRB, National Research Council, Washington, D.C., 1978.
11. Glennon, J. C., T. R. Newman, and J. C. Jack. *Safety and Operational Considerations for Design of Rural Highway Curves.* Report FHWA/RD-86/035. FHWA, U.S. Department of Transportation, 1985.
12. Transportation Research Board. *Designing Safer Roads.* Special Report 214. TRB, National Research Council, Washington, D.C., 1987.
13. Zegeer, C. R. Stewart, E. Council, P. Reinfurt, and E. Hamilton. Safety Effects of Geometric Improvements on Horizontal Curves. Preprint No. 920893, Transportation Research Board's 71st Annual Meeting, National Research Council, Washington, D.C., 1992.
14. Seneviratne, P. N., and M. N. Islam. Treating Uncertainty in the Design of Horizontal Curves. *Proc., ASCE Conference on Infrastructure Planning and Management,* Denver, Colo., 1993, pp. 479–488.
15. Seneviratne, P. N., and M. N. Islam. Optimum Curvature for Simple Horizontal Curves. *Journal of Transportation Engineering,* ASCE, Vol. 120, No. 5, Sept./Oct. 1994, pp. 773–786.
16. Newman, T. R., and J. C. Glennon. Cost-Effectiveness of Improvements to Stopping-Sight Distance Safety Problems. In *Transportation Research Record 923,* TRB, National Research Council, Washington, D.C., 1983, pp. 26–34.
17. Islam, M. N., and P. N. Seneviratne. *Evaluation of Design Consistency of Two-Lane Rural Highways.* Institute of Transportation Engineers, Washington, D.C., Feb. 1994, pp. 28–31.
18. *A Manual on User Benefit Analysis of Highway and Bus Transit Improvements.* American Association of State Highway and Transportation Officials, Washington, D.C., 1977.
19. Islam, M. N. *Determination of Optimum Horizontal Curvature.* Ph.D. dissertation. Department of Civil and Environmental Engineering, Utah State University, Logan, Utah, 1994.
20. Lin, F. B. Flattening of Horizontal Curves on Rural Two-Lane Highways. *Journal of Transportation Engineering,* ASCE, Vol. 116, No. 2, March/April, 1990, pp. 181–196.
21. *Dodge Guide to Public Works and Heavy Construction Costs.* McGraw-Hill, New York, 1986.
22. Ravindran, A., D. Phillips, and J. Folberg. *Operations Research Principles and Practice,* 2nd ed. John Wiley and Sons, New York, 1987.

Publication of this paper sponsored by Committee on Artificial Intelligence.

Use of Fuzzy Relations To Manage Decisions in Preserving Civil Infrastructure

DIMITRI A. GRIVAS AND YUNG-CHING SHEN

The use of fuzzy relations to manage uncertain information in civil infrastructure preservation is addressed. Subjectivity and imprecision of uncertain information and ambiguity of terminological knowledge in preserving infrastructure facilities are the primary motivation for the employment of fuzzy sets and fuzzy relations. Formal decision processes use the concept of knowledge graphs that enable the establishment of general-specific or cause-effect relations between condition factors, as well as relations between condition and symptom factors. Fuzzy relations allow the characterization of the uncertainty associated with these relations. Fuzzy graphs assimilating the degree of certainty involved in decision processes are used to illustrate the connection strength of the elements in the associated fuzzy subsets. A case study in pavement preservation applied to the New York State Thruway is presented. This research makes contributions to synthesizing uncertain information involved in the decision processes of preserving civil infrastructure. In particular, the case study exemplifies the benefits of using fuzzy relations to identify feasible treatment options.

This study is concerned with the use of fuzzy relations to structure decision processes in preserving infrastructure facilities such as pavements and bridges. Feasible preservation methods aim at improving or strengthening infrastructure facilities that are in deficient condition. The complex behavior of infrastructure components renders the preservation decision into an environment of uncertainty. The uncertainty associated with the decisions is concerned mostly with personal preference and judgment, which involves graded or qualified statements that are not strictly true or false.

The qualified statements of engineering judgment are mostly expressed in the form of conditional relation between different quantities, e.g., the relation between climate and cracking of a structural component such as pavement. Frequently the relationship between two mutually dependent classes or variables is neither exact nor inexact. In other words, the linkage between objects in the two classes, or values taken by the two variables, varies gradually from a condition of weak to strong. Such situations can arise basically from two sources: (a) fuzziness in the definition of classes or variables, and (b) ambiguity of the conditional relations.

Fuzzy relations enable the characterization of the uncertainty involved in conditional statements. In civil engineering, Blockley (*1*) illustrated fuzzy relations in the uncertainty analysis of structural safety. Among other applications, this approach was exemplified in a fuzzy relation between compressive stress and longitudinal slenderness for a steel column. Brown and Yao (*2*) examined the fundamental theory of fuzzy sets and illustrated engineering decisions in estimating the strength of concrete using fuzzy conditional relations. Kikuchi and Perinchery (*3*) introduced the concept of fuzzy sets and fuzzy measure for representing two types of uncertainty: vagueness and ambiguity, in engineering planning problems.

In this paper the uses of fuzzy relations to structure decision processes concerned with condition diagnosis and treatment identification in preserving civil infrastructure are addressed. The concept of knowledge graphs (*4*) is used to identify the relations. The uncertainties involved in the relations are manipulated with fuzzy set theory and illustrated with knowledge graphs. A computational framework is formulated to calculate the strength of belief of the diagnosed conditions and identified treatments. A case study on the preservation of pavements is presented. The strengths and weaknesses of using fuzzy relations to structure decisions involving uncertainties are discussed.

INFORMATION AND UNCERTAINTY

Several types of uncertain information, defined by Klir and Folger (*5*) as the amount of uncertainty associated with the system, are present in infrastructure preservation decisions, each of which occurs under its own distinct conditions. The most significant types of uncertainties in the present study are subjectiveness, imprecision, and statistical uncertainty. Information about condition assessment of infrastructure facilities is generally presented in linguistic form, which has meaning that is inherently vague or subjective. This vagueness reflects the uncertainty represented and manipulated with fuzzy sets. The second type of uncertain information involved in preservation is a measurement or test with an instrument, or the uncertainty of imprecision. Imprecision is represented and calculated using fuzzy set theory. Also, statistical uncertainty is involved in quantitative information such as traffic volume, climate, and others. Quantitative information is represented with probability density functions that address statistical uncertainty. In this study the focus is on the uncertainties that are represented with fuzzy sets and manipulated with fuzzy calculus (*6*).

The uncertainty concerned with the reliability of descriptive information, defined by Klir and Folger (*5*) as the shortest description of the system in some standard language, is a result of ill-defined concepts (fuzziness) involved in the problem domain. Within the category of descriptive information, uncertainty (ambiguity) may occur as a result of weak implication, when an engineer is unable to establish a strong correlation between premise and conclusion of conditional statements in preservation decisions.

PRESERVATION DECISIONS

The decision-making process for preserving civil infrastructure generally consists of problem identification, determination of potential solutions, and selection of the preferred solution. In prob-

Department of Civil and Environmental Engineering, Rensselaer Polytechnic Institute, Troy, N.Y. 12180.

lem identification, existing infrastructure conditions are identified through data collection, data evaluation, and project constraints. On the basis of this information, feasible rehabilitation methods are analyzed and recommended. The preferred solution is selected by analyzing costs and by considering project constraints and nonmonetary factors. Shen and Grivas (*4*) proposed a decision framework for pavement preservation that consists of symptom observation, condition diagnosis, and treatment identification.

For infrastructure in general, symptom observation is a process of gathering data and facts required to identify existing infrastructure condition. Condition diagnosis includes the knowledge required to evaluate infrastructure condition. This part of the decision task represents actual diagnostic processes for detecting the causes of deterioration of an infrastructure component. Treatment identification recommends several potential methods that are feasible to remedy infrastructure deterioration. Condition diagnosis and treatment identification are ill-structured decision problems. There are no definitive procedures for the evaluation of infrastructure condition and the identification of treatment options

Using the concept of knowledge graphs (*4*) to formalize the decision problems allows clear identification of the relation of contexts and the structure of knowledge.

KNOWLEDGE GRAPHS

A knowledge graph is a graphical representation of a decision process that attempts to mimic the knowledge of domain experts. Figure 1 shows an example of a knowledge graph for condition diagnosis. The structure of knowledge graphs is established by formalizing the decision processes with a three-step procedure: (a) problem decomposition, (b) term interpretation, and (c) heuristics organization. Formalization of the decision process was presented by Shen and Grivas (*4*) in a study of pavement preservation.

The sample knowledge graph indicates that the terms used by maintenance engineers have different interpretations in describing structural conditions. For example, the problem of insufficient support (Node 3) is recognized by most engineers. Some emphasize that the cause of insufficient support is overloaded traffic (Node 8),

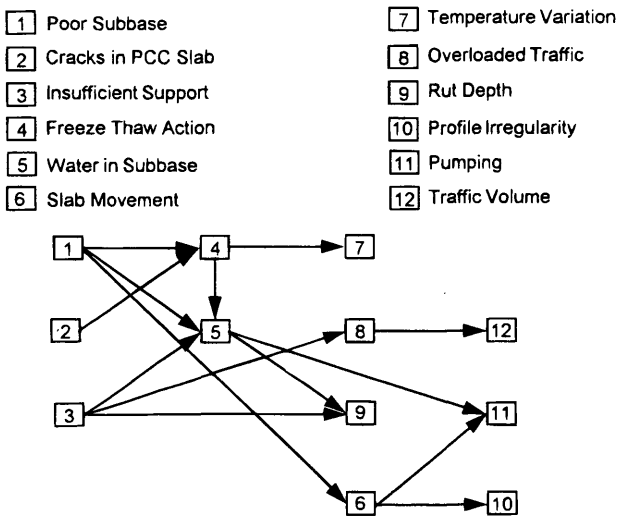


FIGURE 1 A sample of knowledge graph.

while others consider that the problem is evidenced by rut depth (Node 9), which, in turn, indicates the problem of water in the subbase (Node 5).

Understanding the precise meaning of each term was used to establish the knowledge graphs. Some terms are abstract and are defined informally and implicitly by the Thruway engineers because a standard vocabulary has not been designated. The meanings of such terms were derived mostly from engineers' experience in the maintenance actions. Therefore, the terms that experts use are generally vague because the degree of certainty (truth) of the term varies from case to case. The facts, relations, and rules of thumb (conditional statement) contained within the knowledge graphs usually manifest varying degrees of uncertainty. These uncertainties indicate either vagueness of a concept or ambiguity of a relation or both. The use of exact satisfaction of the premise of a conditional statement seems unnatural in the context of infrastructure preservation. Therefore, fuzzy relations from the derived knowledge graphs more realistically represent the decision process and, as well, more clearly account for the uncertainties involved.

FUZZY RELATIONS

The relations established in the knowledge graphs for condition diagnosis and treatment identification can be expressed as a conditional statement: If A, then B; A and B are fuzzy predicates represented by membership functions rather than the propositional variables defined in the classical propositional calculus. In essence, the conditional statement describes a fuzzy relation (*7*) between two fuzzy variables. In this study, fuzzy sets are used to account for the uncertainty (vagueness) of linguistic terms, while fuzzy relations help to clarify the confusion (ambiguity) in interpreting the terms.

Membership Matrix

Fuzzy sets help establish inexact relationships between different quantities or classes of objects presented in a membership matrix. (See Table 1.) Two classes of objects, e.g., temperature variation and slab cracking, form a cause-effect relation that describes the condition of infrastructure facilities, and can be denoted by the Cartesian product of two fuzzy sets, X and Y. In X, the elements of fuzzy sets are various degrees of temperature changes in a day that may affect deterioration (cracking) of slabs. In Y, the fuzzy elements are a six-level severity of cracking.

The Cartesian product of X and Y, $X \times Y$, forms a fuzzy relation, R, that constitutes a new universe with the ordered pairs as its elements, characterized by a membership function $\mu_R(x, y)$. A typical operation for the Cartesian product is represented as

$$\mu_R(x, y) = \mu_{X \times Y}(x, y) = \min [\mu_X(x), \mu_Y(y)] \tag{1}$$

A fuzzy relation is a subset of the Cartesian product, X × Y. An example of a fuzzy relation is established from the conditional statement: If temperature variation is high then slab cracking is moderate general. The linguistic value "high" is represented by fuzzy subset A and "moderate general" is B (*8*), as follows:

$$A = \frac{0.1}{x_1} + \frac{0.4}{x_2} + \frac{0.7}{x_3} + \frac{0.9}{x_4} + \frac{1}{x_5} \tag{2}$$

TABLE 1 Typical Two-Dimensional Membership Matrix of $X \times Y$.

X = Temperature Variation (°F)	Y = Slab Cracking					
	y_1 (Tight Cracks)	y_2 (Tight Cracks in 3 or more slabs)	y_3 (Open Cracks)	y_4 (Open Cracks in 3 or more slabs)	y_5 (Spalled Cracks)	y_6 (Spalled Cracks in 3 or more slabs)
$x_1 = 20$	$\mu_R(x_1, y_1)$	$\mu_R(x_1, y_2)$	$\mu_R(x_1, y_3)$	$\mu_R(x_1, y_4)$	$\mu_R(x_1, y_5)$	$\mu_R(x_1, y_6)$
$x_2 = 30$	$\mu_R(x_2, y_1)$	$\mu_R(x_2, y_2)$	$\mu_R(x_2, y_3)$	$\mu_R(x_2, y_4)$	$\mu_R(x_2, y_5)$	$\mu_R(x_2, y_6)$
$x_3 = 40$	$\mu_R(x_3, y_1)$	$\mu_R(x_3, y_2)$	$\mu_R(x_3, y_3)$	$\mu_R(x_3, y_4)$	$\mu_R(x_3, y_5)$	$\mu_R(x_3, y_6)$
$x_4 = 50$	$\mu_R(x_4, y_1)$	$\mu_R(x_4, y_2)$	$\mu_R(x_4, y_3)$	$\mu_R(x_4, y_4)$	$\mu_R(x_4, y_5)$	$\mu_R(x_4, y_6)$
$x_5 = 60$	$\mu_R(x_5, y_1)$	$\mu_R(x_5, y_2)$	$\mu_R(x_5, y_3)$	$\mu_R(x_5, y_4)$	$\mu_R(x_5, y_5)$	$\mu_R(x_5, y_6)$

$$B = \frac{0.2}{y_1} + \frac{0.5}{y_2} + \frac{0.9}{y_3} + \frac{1}{y_4} + \frac{0.6}{y_5} + \frac{0.1}{y_6} \tag{3}$$

Specifically, A is a fuzzy subset of the universe of discourse, X, and B is a fuzzy subset of Y. The fuzzy relation $R(A, B)$ is characterized by a membership function $\mu_R(x, y)$ and is expressed:

$$R = A \times B = \{\mu_R(x, y)/(x, y) \mid x \in A \text{ and } y \in B \tag{4}$$

This expression is a special fuzzy relation of $X \times Y$, and the relation combines all $x \in A$ and $y \in B$ in the form of ordered pairs represented as a membership matrix:

$$A \times B = \begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.2 & 0.4 & 0.4 & 0.4 & 0.4 & 0.1 \\ 0.2 & 0.5 & 0.7 & 0.7 & 0.6 & 0.1 \\ 0.2 & 0.5 & 0.9 & 0.9 & 0.6 & 0.1 \\ 0.2 & 0.5 & 0.9 & 1 & 0.6 & 0.1 \end{bmatrix} \tag{5}$$

The membership matrix represented for a fuzzy relation can be derived from the maintenance engineers about their strength of confidence on related fuzzy elements. In the case of preserving civil infrastructure, interviews with engineering experts would be necessary to identify membership values. Typical questionnaires for the interview are:

1. Does symptom i of concrete pavements *always* indicate a problem of condition j?
2. Is symptom p of a steel-girder bridge *often* caused by the condition q?
3. Is condition r of an overlaid pavement very likely a specific case of condition s?

The relations between condition factors as well as the relations between symptoms and conditions presented in the questionnaire are associated with implicit uncertainty. A membership grade [0,1] is assigned for the linguistic terms, always, often, may-not likely, very likely, and others. For example, the connection strength between symptom i and condition j is 0.9, which represents the uncertainty value "always." Furthermore, concentration and dilation operation are modeled with the linguistic modifiers: $\mu_{\text{very A}}(x) = \mu_A^{1/2}(x)$ and $\mu_{\text{may-not A}}(x) = \mu_A^2(x)$, respectively. The created membership matrix for a fuzzy relation can also be interpreted using a fuzzy graph (*9*).

Fuzzy Graph

Elements of the matrix with nonzero membership grades are represented in the diagram by lines connecting the respective nodes. Figure 2 presents the fuzzy graph of the conditional relation "If temperature variation is high then slab cracking is moderate general." The nodes of the fuzzy graph are considered as the elements of the

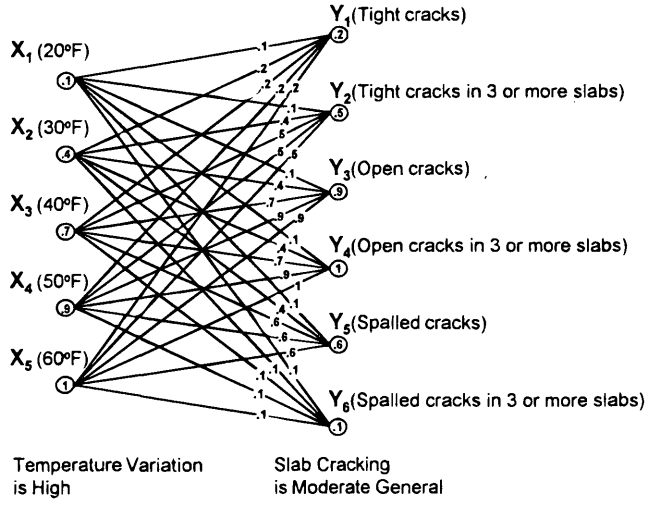


FIGURE 2 The fuzzy graph of the conditional relation "If temperature variation high *Then* cracking is moderate general."

fuzzy relation. These nodes are labeled with the membership grade of related fuzzy subsets. For a fuzzy relation, the connection strength of a path is defined as the minimum strength of related elements. As shown in Figure 2, temperature variation of 60°F (15.5°C) has the greatest influence (membership grades are 0.9 and 1.0) to open cracks on slabs. However, the 60°F (15.5°C) temperature variation does not cause a major problem to spalled cracks, where the membership grades of the relations are 0.6 and 0.1. In other words, spalled cracks are a phenomenon that temperature changes would not strongly affect.

Composition

Two fuzzy relations can be composed to a new relation. For example, a new relation can be established from a composition between symptom-condition relation and condition-treatment relation. A typical example of composition can be established from the following two conditional relations:

1. If temperature variation is high (A) then slab cracking is moderate general (B).
2. If slab cracking is moderate general (B) then seal the cracks (C).

In C, the fuzzy elements are the four generic types of treatments: preventive maintenance, minor rehabilitation, major rehabilitation, and reconstruction. The membership function of seal the cracks in this condition relation is defined as

$$C = \frac{0.8}{z_1} + \frac{1}{z_2} + \frac{0.5}{z_3} + \frac{0.1}{z_4} \tag{6}$$

Suppose that R is a relation on $A \times B$ (A and B are represented in Equation 2 and 3, respectively) and S is a relation on $B \times C$ (B and C are represented in Equation 3 and 6, respectively). One might want to know the fuzzy relation from A to C. An operator can be defined to establish the relation between A and C via B. This study follows max-min composition, denoted by $\circ$, that is:

$$R \circ S = \bigvee_{y \in Y} (\mu_R(x,y) \wedge \mu_s(y,z)). \tag{7}$$

The composition is formulated as follows:

$$R \circ S = \begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.2 & 0.4 & 0.4 & 0.4 & 0.4 & 0.1 \\ 0.2 & 0.5 & 0.7 & 0.7 & 0.6 & 0.1 \\ 0.2 & 0.5 & 0.9 & 0.9 & 0.6 & 0.1 \\ 0.2 & 0.5 & 0.9 & 1 & 0.6 & 0.1 \end{bmatrix} \circ$$

$$\begin{bmatrix} 0.2 & 0.2 & 0.2 & 0.1 \\ 0.5 & 0.5 & 0.5 & 0.1 \\ 0.8 & 0.9 & 0.5 & 0.1 \\ 0.8 & 1 & 0.5 & 0.1 \\ 0.6 & 0.6 & 0.5 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.1 \\ 0.4 & 0.4 & 0.4 & 0.1 \\ 0.7 & 0.7 & 0.5 & 0.1 \\ 0.8 & 0.9 & 0.5 & 0.1 \\ 0.8 & 1 & 0.5 & 0.1 \end{bmatrix}$$

Figure 3 shows the fuzzy graph of the composition. The connection strength between x_5 and z_2 via y_4 is the strongest among all the paths between them. Minor rehabilitation is recommended, which would require more work than just filling the sealer in the cracks. On the other hand, if temperature variation is at 50°F (10.0°C) and the cracks appear to be open on three or more slabs, then apply the treatment, and seal the cracks, as preventive maintenance.

Fuzzy Mapping

Mapping is a fuzzy transformation when uncertainties are involved in a system. When information is passed through a fuzzy system, an extra fuzziness will be added because of the fuzziness of the system

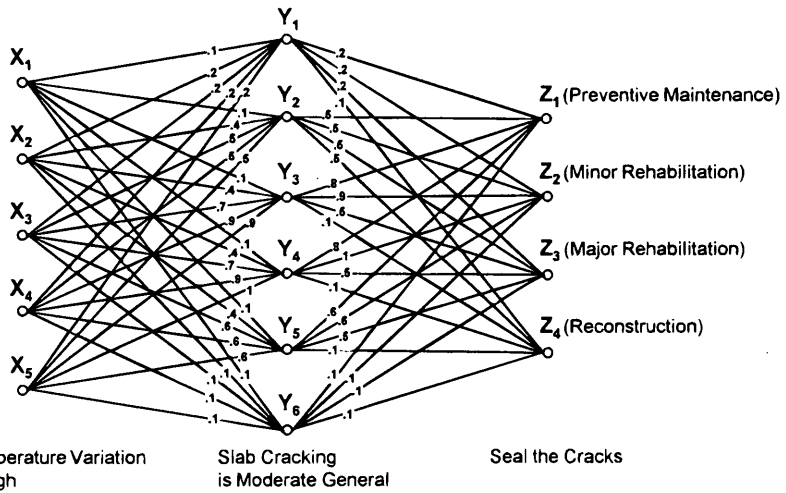


FIGURE 3 The fuzzy graph of the composition between temperature variation high and seal the cracks.

itself. Even if the input is crisp, the output will be fuzzy; if the input is fuzzy, the output will be fuzzier.

For a fuzzy transformation, a relation R between two fuzzy sets, X and Y is expressed by the membership function, $\mu_R(x,y)$. The image of A on X under this transformation, using a matrix expression, is given by

$$B = A \circ R \tag{9}$$

with the membership function

$$\mu_B(y) = \bigvee_{x \in X} (\mu_A(x) \wedge \mu_R(x, y)) \tag{10}$$

This transformation enables identification of treatments from the observed symptom once the fuzzy relations between symptoms and treatments are established.

CASE STUDY

The area covered in this study is a section of the New York State Thruway that was originally constructed in 1955. The 9-in. (22.5-cm) reinforced cement concrete pavement section was constructed over a 12-in. (30-cm) granular subbase. Later there were a 2½-in. (6.5-cm) asphalt concrete overlay, a 1-in. (2.5-cm) asphalt concrete overlay, and a 3-in. (7.5-cm) asphalt concrete overlay until 1993. The occurrence of cracking, rutting, spalling, and other conditions has affected the load-carrying of the pavement structure. This case study places emphasis on the identification of treatment options based on rut depth data collected from a roughness survey and engineering judgment about the preservation of pavement structures.

The decision process of treatment identification follows the symptom and/or data collected from the field. In accordance with the established knowledge graph (as shown in Figure 1), rut depth (Node 9) may be the effect of water in subbase (Node 5), which, in turn, indicates poor subbase (Node 1), or rut depth may be an indication of insufficient support (Node 3) of the pavement structure to carry overloaded traffic (Node 8). However, the rut depth may also be simply treated without rectifying any structural problem of the pavement.

In this study fuzzy relations are used to pursue three different decision processes: direct symptom-treatment relation explored in Method A, symptom-condition-treatment relation in Method B, and symptom-condition-condition-treatment relation in Method C. A fuzzy relation between symptom and treatment will be composed from all the parameters involved in each decision process. Fuzzy ordering enables a comparison of the strength of belief among the treatment options derived from each method. Thus, the three methods aim at establishing fuzzy relations for the conditional statement: If rut depth is large general then what kind of treatment is considered to be the most appropriate one.

Method A

The first method applies a direct symptom-treatment relation that enables identifying the strength of belief to the treatment, milling wheel rut, based on the severity of rut depth. In X (the symptom) the elements of fuzzy sets are defined as four levels of rut depth, and the five elements of fuzzy sets for Z (the treatment) are "do nothing" and four generic treatments.

X = *5 Rut Depth*	Z = *Treatments*
$x_1 <$ 1cm	z_1 = Do Nothing
2cm $> x_2 \geq$ 1cm	z_2 = Preventive Maintenance
3cm $> x_3 \geq$ 2cm	z_3 = Minor Rehabilitation
$x_4 \geq$ 3cm	z_4 = Major Rehabilitation
	z_5 = Reconstruction

The membership matrix of the relation between rut depth and treatments is established from interviewing experts about their judgments on determining potential treatments for a certain range of rut depth, and represented as

$$R = X \times Z = \begin{bmatrix} 0.8 & 0.6 & 0.2 & 0 & 0 \\ 0.4 & 0.8 & 0.3 & 0 & 0 \\ 0.2 & 1 & 0.6 & 0.1 & 0 \\ 0.1 & 0.7 & 1 & 0.5 & 0.2 \end{bmatrix}$$

A fuzzy relation is applied to the conditional statement: If rut depth is large general (A), then treatment is mill wheel ruts (C) to derive the fuzzy subset of the treatment. In the premise of the conditional statement, linguistic value "large general" is represented by fuzzy subset A as

$$A = \frac{0.1}{x_1} + \frac{0.5}{x_2} + \frac{1}{x_3} + \frac{0.8}{x_4} \tag{12}$$

The membership function (C) is obtained from mapping the "rut depth is large general" to M_A, $C = A \circ M_A$:

$$C = \frac{0.4}{z_1} + \frac{1}{z_2} + \frac{0.8}{z_3} + \frac{0.5}{z_4} + \frac{0.2}{z_5} \tag{13}$$

The membership grade for the treatment, mill wheel ruts (C), is expressed in terms of the generic treatment, Z_1. The operation of mapping the symptom to treatment is illustrated using the fuzzy graph shown in Figure 4. Membership values of the fuzzy relation are marked on the lines connecting the symptom and the treatment. Among them, the connection strength between x_3 and z_2 is the highest.

Method B

The decision path of the second method, according to the knowledge graph in Figure 1, is from symptom rut depth to a structural

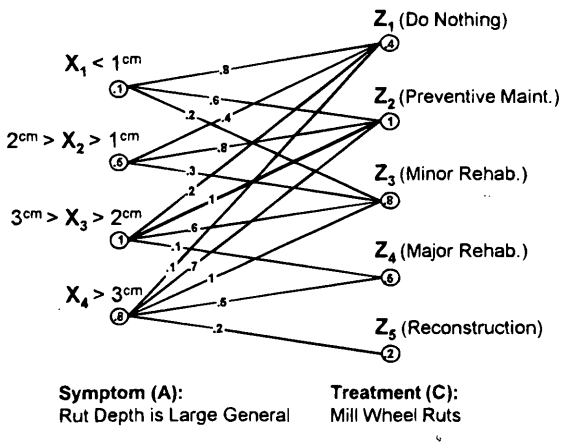


FIGURE 4 A fuzzy graph for Method A.

condition, insufficient support. The latter can be rectified by extended maintenance overlay (*4*). Thus, the conditional statements for the second method include:

1. If rut depth is large general (A), then insufficient support (B).
2. If insufficient support (B), then extended maintenance overlay (C).

The level of severity of insufficient support can be estimated from a falling weight deflectometer test. The deflection of pavement structure obtained from the test may be classified into none, small, moderate, and large, which become the elements of the fuzzy sets of defining the structural condition.

X = Rut Depth	*Y = Insufficient Support*	*Z = Treatments*
$x_1 <$ 1cm	y_1 = None	z_1 = Do Nothing
2cm $> x_2 \geq$ 1cm	y_2 = Small	z_2 = Preventive Maintenance
3cm $> x_3 \geq$ 2cm	y_3 = Moderate	z_3 = Minor Rehabilitation
$x_4 \geq$ 3cm	y_4 = Large	z_4 = Major Rehabilitation
		z_5 = Reconstruction

The fuzzy relations of symptom (X) and condition (Y), R, are established from interviewing maintenance engineers by identifying their confidence on a severity of insufficient support that causes a range of rut depth; and the fuzzy relations of condition (Y) and treatment (Z), S, are the confidence on the type of treatment for a severity of insufficient support. The relations, R and S, are represented as following:

$$R = \begin{bmatrix} 0.5 & 0.8 & 0.2 & 0 \\ 0.3 & 0.7 & 0.5 & 0.2 \\ 0.1 & 0.4 & 0.9 & 0.5 \\ 0.0 & 0.3 & 1 & 0.8 \end{bmatrix}$$

The composition of R and S forms a relation between symptom and treatment for the Method B.

$$S = \begin{bmatrix} 1 & 0.6 & 0.2 & 0 \\ 0.8 & 0.8 & 0.3 & 0.1 \\ 0.7 & 0.9 & 0.8 & 0.4 \\ 0.4 & 0.7 & 1 & 0.8 \end{bmatrix}$$

A fuzzy set representation of structural condition is obtained by mapping the membership function of "rut depth is large general" to R. Similarly, a fuzzy set representation of treatment (C) is a mapping to M_B. For the case of rut depth is large general a fuzzy subset describing the degree of truth of insufficient support (B) is

$$B = \frac{0.3}{y_1} + \frac{0.5}{y_2} + \frac{0.9}{y_3} + \frac{0.8}{y_4} \tag{17}$$

In addition, the extended maintenance overlay is given a fuzzy set representation as

$$C = \frac{0.7}{z_1} + \frac{0.9}{z_2} + \frac{0.9}{z_3} + \frac{0.8}{z_4} + \frac{0.8}{z_5} \tag{18}$$

Figure 5 presents a fuzzy graph showing the mapping from the symptom, rut depth, to the condition, insufficient support, and from the condition to the treatment. The highest connection strength among the decision paths is $X_3 - Y_3 - Z_2$ or $X_3 - Y_3 - Z_3$. This method indicates a lower certainty value on the suggested treatment, extended maintenance overlay. In addition, it shows greater uncertainty on selecting a type of treatment option.

Method C

The decision process follows a different path of the knowledge graph in Figure 1. The conditional statements for the third method include:

1. If rut depth is large general (A), then water in subbase (B^1),
2. If water in subbase (B^1), then poor subbase (B^2),
3. If poor subbase (B^2), then maintenance overlay (C).

The level of severity of water in subbase is classified into none, low, medium, and high, four levels. The certainty values for the fuzzy predicate poor subbase are defined as: (a) impossible, (b) very_low_chance, (c) it_may, (d) most_likely, and (e) certain. The elements of the fuzzy sets of each parameter used in Method C are given as follows:

X = Ruth Depth	*Y^1 = Water in Subbase*	*Y^2 = Poor Subbase*	*Z = Treatments*
$x_1 <$ 1 cm	y_1 = None	y_1 = Impossible	Z_1 = Do Nothing
2cm $> x_2 \geq$ 1cm	y_2 = Low	y_2 = Very_low_chance	z_2 = Preventive Maintenance
3cm $> x_3 \geq$ 2cm	y_3 = Medium	y_3 = It_may	z_3 = Minor Rehabilitation
$x_4 \geq$ 3cm	y_4 = High	y_4 = Most_likely	Z_4 = Major Rehabilitation
		y_5 = Certain	z_5 = Reconstruction

Applying the same interviewing procedures, fuzzy relations of X and Y^1, Y^1 and Y^2, and Y^2, and Z are shown in the following matrices, R, S^1, and S^2, respectively:

$$R = \begin{bmatrix} 0.9 & 0.3 & 0.2 & 0 \\ 0.8 & 0.4 & 0.2 & 0 \\ 0.7 & 1 & 0.6 & 0.2 \\ 0.4 & 0.8 & 0.9 & 0.5 \end{bmatrix} \tag{19}$$

$$S^1 = \begin{bmatrix} 1 & 0.9 & 0.2 & 0 \\ 0.9 & 1 & 0.3 & 0.1 \\ 0.8 & 0.9 & 0.4 & 0.2 \\ 0.6 & 0.9 & 0.7 & 0.5 \end{bmatrix} \tag{20}$$

$$S^2 = \begin{bmatrix} 1 & 0.5 & 0.2 & 0 & 0 \\ 0.8 & 1 & 0.7 & 0.4 & 0.1 \\ 0.6 & 0.9 & 0.8 & 0.6 & 0.2 \\ 0.5 & 0.8 & 0.9 & 0.7 & 0.4 \\ 0.3 & 0.7 & 0.9 & 1 & 0.8 \end{bmatrix} \tag{21}$$

The composition of the three relations in equations 19, 20 and 21 is as follows:

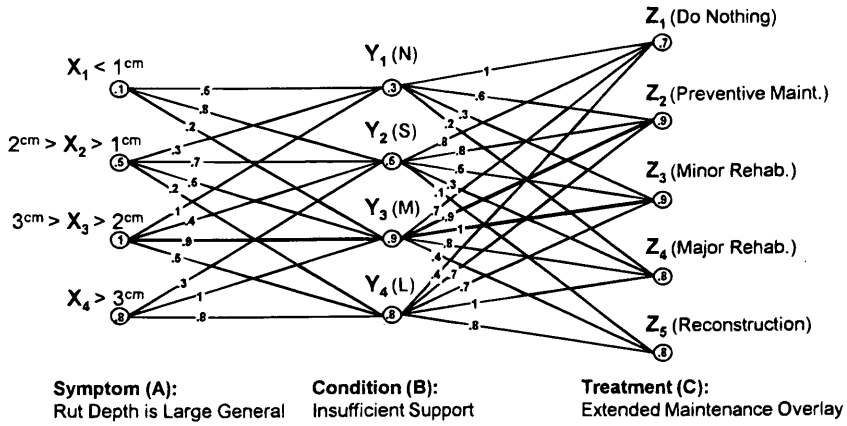


FIGURE 5 A fuzzy graph for Method B.

$$M_C = \begin{bmatrix} 0.9 & 0.9 & 0.7 & 0.6 & 0.3 \\ 0.8 & 0.8 & 0.7 & 0.6 & 0.3 \\ 0.9 & 1 & 0.7 & 0.6 & 0.4 \\ 0.8 & 0.9 & 0.8 & 0.6 & 0.5 \end{bmatrix} \tag{22}$$

A fuzzy set representation of water in subbase and poor subbase can be obtained by mapping the membership function of "rut depth is large general" to R and S^1. Similarly, a fuzzy set representation of treatment (C) is a mapping to M_C. For the case of rut depth is large general a fuzzy subset describing the degree of truth of maintenance overlay (C) is derived as

$$C = \frac{0.9}{z_1} + \frac{1}{z_2} + \frac{0.8}{z_3} + \frac{0.6}{z_4} + \frac{0.5}{z_5} \tag{23}$$

Figure 6 is a fuzzy graph showing a transformation of a fuzzy subset from symptom to treatment. The membership values obtained from maintenance expertise and represented in the fuzzy relation are marked on the lines connecting symptom, condition, and treatment. The highest strength of decision path in this graph is $X_3 - (Y^1)_2 - (Y^2)_2 - Z_2$. Although the identified treatment option is preventive maintenance, maintenance overlay, which is suggested in this case, is in the category of minor rehabilitation.

Fuzzy Ordering

Ranking the identified treatments is a major decision-making issue in preserving civil infrastructure. It is usually involved with uncertainties and fuzziness. Ordering the types of treatment is based on the strength of belief calculated from each method, and ordering the suggested treatments is based on the degree of certainty of the derived fuzzy subsets. The symptom-treatment relations established in the three methods are applied to identify a treatment for a symptom such as rut depth is large general.

Mapping the symptom to the fuzzy relations of Methods A, B, and C generates the fuzzy subsets of suggested treatments shown in equations 13, 18, and 23, respectively. The order of treatment types for Method A is $z_2 > z_3 > z_4 > z_1 > z_5$ in which preventive maintenance, mill wheel ruts, is recommended. Similarly, the order of treatment types for Methods B and C are $z_2 = z_3 > z_4 = z_5 > z_1$ and $z_2 > z_1 > z_3 > z_4 > z_5$ respectively. Preventive maintenance is also the identified treatment type for both Methods B and C. Although, extended maintenance overlay (a major rehabilitation) is recommended for Method B and maintenance overlay (a minor rehabilitation) for Method C. It appears that the associated treatments are neither totally committed for Method A nor for Method C, in accordance with the computed membership grades.

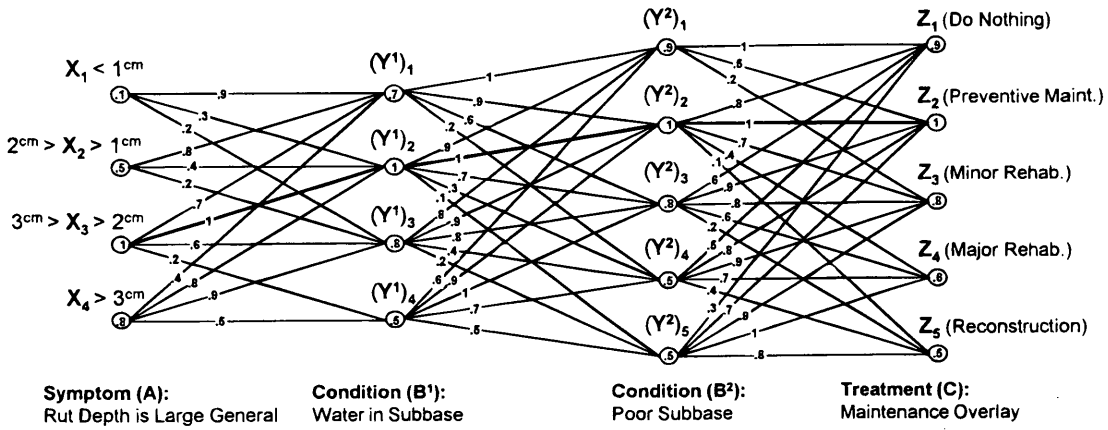


FIGURE 6 A fuzzy graph for Method C.

Ranking the suggested treatments (mill wheel ruts, maintenance overlay, and extended maintenance overlay) can be achieved from a fuzzy measure to estimate the degree of certainty.

DISCUSSION

The three different decision methods of identifying treatment options for a given rut depth information show different results. In this section, two interesting points are discussed that contribute to the investigation of this study.

First, the fuzzy relation matrices shown in Equations 11, 16, and 22 can be compared using an α-cut (*10*). α-cut is a method of converting a fuzzy set into a crisp set which provides a criterion of measuring the fuzziness of a set. A membership value of one-half has the highest degree of difficulty of deciding whether it is a member of the set or not. Membership grades close to one are closer to being in the set, membership grades close to zero are closer to being out of the set. In the present study, applying an α value to the three membership matrices, M_A, M_B, and M_C is to clarify the ambiguity of selecting a treatment among the three methods.

Setting $\alpha = 0.9$, the symptom-treatment relation obtained from Method B identifies preventive maintenance and minor rehabilitation as the potential treatment options.

$$(M_B)_{0.9} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \tag{24}$$

However, as the α value decreases to 0.7, this method essentially has no way of identifying a preferred treatment. This is evident from Equation 18 where it shows almost the same strength of belief of applying different maintenance strategies for the given symptom.

After decreasing the α-cut value to 0.6, Method C shows a higher strength of belief on the four treatment options.

$$(M_C)_{0.6} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \tag{25}$$

In comparison with Method B, Method C is less fuzzy in identifying treatment for a given symptom. On the contrary, Method C is fuzzier than Method A, because the membership matrix, M_A, shows a higher focus on preventive maintenance and minor rehabilitation.

$$(M_A)_{0.6} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \tag{26}$$

This result is consistent with the specific case study obtained from Equation 12. This higher focus means that the result obtained from Method A is less fuzzy than the results of Methods B and C.

Second, the condition fuzzy variables, insufficient support in Method B, and water in subbase and poor subbase in Method C, indicate some influence on the decisions of identifying treatment options because of the uncertainty factors of these variables involved in the processes. However, the results of Methods B and C provide more information about potential treatment availability and the degree of certainty assigned to each treatment.

The decision processes can be refined by interviewing maintenance experts who can present a more reasonable preference function (membership function) in describing the uncertain information. The membership function resulting from interviewing processes is a quantification of design preference that the engineers have the greatest confidence in using, or desire to use with respect to treatment options.

The fuzzy sets representation for general-specific or cause-effect relations has no distinguished differences. The cause-effect relation between water in subbase and rut depth is part of the decision path in Method C. The membership matrix of equation 19 establishes a fuzzy relation that allows evaluation of the cause (water in subbase) of the problem using fuzzy mapping for a given symptom (rut depth). Both water in subbase and poor subbase are the causes of rutting. However, there is a general-specific relation between poor subbase and water in subbase. Structuring expertise with the approach of problem decomposition is well suited to infrastructure condition reasoning that is involved in several levels of abstraction. Higher-level knowledge sources are employed to deal with more general concepts, while lower-level knowledge sources are used to deal with much more detailed operations applied to more specific domains.

In the case study the three decision paths are represented with three different membership matrices for a given symptom. Each membership matrix is established for a relation between the symptom, rut depth, and the treatments. The matrices allow us to explore the degree of confidence on the identified treatment type, which, in turn, clarify the ambiguity of the terms interpreted in the knowledge graphs. Uncertainties involved in the knowledge graph can be assimilated using fuzzy graphs. Fuzzy graphs provide a graphical presentation for the process of fuzzy mapping and the composition of fuzzy relations. The graphs allow the connection strength between the related fuzzy elements to be identified.

SUMMARY AND CONCLUSIONS

In this study the use of fuzzy relations to manage uncertain information in civil infrastructure preservation was investigated. Emphasis was placed on structuring decision processes concerned with condition diagnosis and treatment identification in preserving infrastructure. The concept of knowledge graphs was employed to identify the relations. The uncertainties involved in the relations are manipulated with fuzzy set theory and illustrated with fuzzy graphs. A computational framework was formulated to calculate the strength of belief of the diagnosed conditions and identified treatments.

Three different paths of a decision process were examined based on the fuzzy relations for the conditional statement: If rut depth is large general, then what kind of treatment is most appropriate? Following a series of compositions of fuzzy relations in association with the observed symptoms, the degree of confidence (certainty) in the identified treatments can be obtained. The composition of membership matrices was further illustrated using fuzzy graphs for the decision processes involved in the case study. On the basis of the findings from the use of fuzzy relations in the present study, the following conclusions may be drawn:

1. Fuzzy relations allow us to synthesize the uncertain information involved in civil infrastructure preservation decision processes.

2. Fuzzy graphs provide an efficient tool to explore the strength of confidence of the related parameters represented in fuzzy sets.

3. Fuzzy set representation has the advantages of ordering decision parameters in prioritizing identified treatments.

ACKNOWLEDGMENTS

The study presented in this paper was sponsored by the New York State Thruway Authority. It is part of a broad research effort to develop a pavement management system. The input and feedback provided by Thruway Maintenance personnel during system development is gratefully acknowledged.

The authors thank Lillith Stoessel for her assistance in drafting and editing the present work. Special thanks to Dr. Shinya Kikuchi and the reviewers for providing comments and critique to improve the quality of the paper.

REFERENCES

1. Blockley, D. I. The Role of Fuzzy Sets in Civil Engineering. *Fuzzy Sets and Systems*, Vol. 2, pp. 267–278, 1979.
2. Brown, C. B., and J. T. P. Yao. Fuzzy Sets and Structural Engineering. *Journal of Structural Engineering*, ASCE, Vol. 109, No. 5, 1983, pp. 1211–1225.
3. S. Kikuchi and V. Perincherry. Use of Fuzzy Theory for Dealing with Uncertainty: Evaluating Alternatives Under Uncertainty. Proceedings of Infrastructure Planning and Management (J. L. Gifford, D. R. Uzarski, and S. McNeil eds.), 1993, pp. 573–582.
4. Shen, Y. C., and D. A. Grivas. Use of Knowledge Graphs to Formalize Decisions in Preserving Pavements. *Journal of Computing in Civil Engineering*, ASCE, Vol. 7, No. 4, 1993, pp. 475–494.
5. Klir, G. J., and T. A. Folger. *Fuzzy Sets, Uncertainty, and Information*. Prentice Hall, Englewood Cliffs, N. J., 1988.
6. Kaufmann, A. and M. M. Gupta. *Introduction to Fuzzy Arithmetic*, Van Nostrand Reinhold Co., New York, N.Y., 1985.
7. Zadeh, L. A. Outline of A New Approach to the Analysis of Complex System and Decision Processes. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-3, No. 1, 1973.
8. Grivas, D. A., and Y. C. Shen. A Fuzzy Set Approach for Pavement Damage Assessment. *Civil Engineering Systems*, Vol. 8, No. 1, 1991, pp. 37–47.
9. Kaufmann, A. *Introduction to the Theory of Fuzzy Subsets*. Vol. 1, Academic Press, New York, N.Y., 1975.
10. Zimmermann, H.-J. *Fuzzy Set Theory and Its Applications*, 2nd ed., Kluwer Academic Publishers, Boston, Mass., 1991.

Views and opinions expressed herein do not necessarily reflect those of the New York State Thruway Authority.

Publication of this paper sponsored by Committee on Artificial Intelligence.

Using a Knowledge-Based Expert System and Fuzzy Logic for Minor Rehabilitation Projects in Ohio

SAKCHAI PRECHAVERAKUL AND FABIAN C. HADIPRIONO

In the selection of a proper treatment for the rehabilitation of a deteriorated pavement section, engineers may encounter a situation in which factors besides distress conditions also contribute to the decision-making process. These factors are, among others, the expected structural integrity, functional adequacy, and performance life of a pavement section. In general, engineers make their selections based on their experience, judgment, and the use of past maintenance data, if available. For young engineers, such a selection process may lead to a poor decision. Even experienced engineers may still reach erroneous results. This study presents a methodology to overcome such problems by employing a knowledge-based expert system (KBES) and fuzzy logic. A KBES serves as a preliminary selection in which a set of alternative treatments is chosen based on pavement distress conditions and other related factors. An ordinal multiobjective decision-making model using fuzzy logic is then used to recommend the proper treatment. A computer program was written to implement such a methodology.

As required by the 1991 Intermodal Surface Transportation Efficient Act (ISTEA), the Ohio Department of Transportation (ODOT) developed a pavement management system, PMS III (*1*), to manage its highway at the network level. This system is currently being implemented. To enhance the PMS III, the development of a project-level PMS is essential. One of the objectives of the project level PMS would be to aid engineers in the selection of proper maintenance and rehabilitation (M&R) treatments. M&R treatments in Ohio are classified into three categories based upon pavement condition: major rehabilitation, minor rehabilitation, and maintenance. Pavement condition is assessed using a pavement condition rating method which provides an overall condition of a pavement section through a pavement condition rating (PCR) index (ranging from 0 to 100; the higher the number, the better the condition), and the structural condition through a structural deduct (STD) index (ranging from 0 to 65; the lower the number, the better the condition). Table 1 lists the conditions used to categorize M&R treatments (*2*).

Basically, this classification serves as an initial screen for the management of deteriorated pavement sections in a systematic fashion. Major rehabilitation projects range from structural overlay to reconstruction. On the other hand, minor rehabilitation and maintenance projects are used to restore or maintain the functionality and structural integrity of pavements. Thus, they range from crack and surface treatments to nonstructural overlay. This study focuses on the selection of proper treatments in minor rehabilitation projects. Because the selection process is usually based on experience and judgment of engineers, we propose a methodology that can be used to computerize such a process by employing a knowledge-based expert system (KBES) and fuzzy logic.

S. Prechaverakul, Department of Civil Engineering, Prince of Songkla University, Hatyai, Songkla, Thailand 90112. F. C. Hadipriono, Department of Civil Engineering, The Ohio State University, Columbus, Ohio 43210.

MINOR REHABILITATION TREATMENT SELECTION STRATEGY

In general, a minor rehabilitation treatment is selected based on primary and secondary factors. Primary factors can be defined as those directly affecting the improvement in pavement performance, such as distress condition and traffic volume. On the other hand, the secondary factors are not directly affected but more concerned with the degree to which a treatment is able to rehabilitate a pavement, and such other factors as time or budget constraints. With this in mind, we propose that the selection strategy should consist of two steps: preliminary selection and final selection. The preliminary selection involves choosing treatments by considering only the primary factors. If more than one treatment is possible then a proper treatment is selected based on the secondary factors in the final selection step.

The selection process described above is a decision-making process in which the experience and judgment of engineers play an important part. Hence, to computerize such a process, two techniques that have been proven as efficient tools to simulate the human thinking process are employed: a KBES and fuzzy logic. The KBES is used in the preliminary selection phase. In the final selection phase, an ordinal multiobjective decision-making model using fuzzy logic proposed by Yager (*3*) is employed.

KBES FOR PRELIMINARY SELECTION

The first step in the selection of a minor rehabilitation treatment is to assess pavement distress conditions. In Ohio, distress conditions are measured in linguistic terms for their severity (as low, medium, or high) and extent (as occasional, frequent, or extensive). This assessment is performed following the guidelines provided in the *Pavement Condition Rating Manual* (*4*). This information together with other factors, such as traffic volume and/or the location of the pavement section, are then used as the basic criteria to select rehabilitation treatments. This selection process may look simple when performed by a human. On the other hand, to encode the knowledge and simulate the human thinking process in a computer is not an easy task. Recently, a KBES, which was developed from the field of artificial intelligence, has proven to be an efficient tool in performing such a task. A comprehensive survey of KBESs in transportation is summarized and discussed by Cohn and Harris (*5*).

TABLE 1 Classification of M&R Treatments (*2*)

M&R Treatment	PCR and STD
Major Rehabilitation	PCR < 50 **OR** STD > 25
Minor Rehabilitation	PCR > 50 **AND** STD < 25
Maintenance	PCR > 50 **OR** STD < 25

Basically, the development of a KBES involves five steps: problem identification, knowledge acquisition, knowledge representation, implementation, and validation and extension. The problem identification phase identifies what the problem is and ensures that a KBES is more suitable than a traditional computer program in solving it. The second phase is the acquisition of knowledge from experts. This is usually done by interviews. The knowledge gained is then represented using an appropriate knowledge representation scheme. The most common scheme is the production rule system, which is also used in our study. The implementation phase encodes the knowledge in the form of production rules into a computer program. Many software packages for developing a KBES are commercially available and thus make it less difficult to program. These software packages are known as expert system shells. Once a prototype is completed, it will be tested and the validation can begin. The validation is performed by both participating and independent experts. Modification or extension can also be done, if necessary.

Based on the knowledge gained from experts, we have classified rehabilitation treatments for flexible pavements into three main categories according to the type of the problem to be corrected: cracking, surface defect problems, and structural problems. These problems can be treated using crack treatment, surface treatment, and nonstructural overlay (one- and two-course overlay), respectively. The following rules exemplify general knowledge of the experts, more refined rules have been incorporated into the knowledge base of the system.

Rule 1:
IF (Longitudinal Joint Cracking Severity is *medium* OR Longitudinal Joint Cracking Severity is *high*)
AND (Longitudinal Joint Cracking Extent is *frequent* OR Longitudinal Joint Cracking Extent is *extensive*)
THEN Treatment is Crack Treatment

Rule 2:
IF (Bleeding Extent is *frequent* OR Bleeding Extent is *extensive*)
AND (location is *intersection* OR location is *curve*)
THEN Treatment is Surface Treatment

Rule 3:
IF (Potholes Severity is *medium* OR Potholes Severity is *high*)
AND (Potholes Extent is *frequent* R Potholes Extent is *extensive*)
THEN Treatment is Overlay

Rule 4:
IF Treatment is *Overlay*
AND Traffic volume is *medium*
AND (Wheel Track Cracking Severity is *high* OR Wheel Track Cracking Extent is *extensive*)
THEN Treatment is One-Course Overlay

Rule 5:
IF Treatment is *Overlay*
AND Traffic volume is *heavy*
AND (Potholes Severity is *high* OR Potholes Extent is *extensive*)
THEN Treatment is Two-Course Overlay

Rule 6:
IF Treatment is *Overlay*
AND (Traffic Volume is *medium* or Traffic Volume is *heavy*)
AND Structural Deduct value is greater than 15
THEN Treatment is Two-Course Overlay

To illustrate how the KBES reaches the conclusion, let us consider a flexible pavement section subjected to distress conditions described in Table 2.

In addition, suppose that the traffic volume on this pavement section is medium. In this case, using the *Pavement Condition Rating*

TABLE 2 Example of Flexible Pavement Condition

Distress	Severity	Extent
Longitudinal Joint Cracking	There is multiple cracking or wide single crack greater than 1/4 inch with some spalling.	More than fifty percent of the joint length has center line cracking.
Potholes	Average depth of potholes greater than six inches in diameter is between one to two inches.	Potholes occur along ten to fifty percent of the area.
Wheel Track Cracking	There is single or intermittent multiple cracking with average crack width less than 1/8 inch or barely noticeable.	More than fifty percent of the wheel track length is within the section which exhibits cracking.

Manual (*5*), an engineer would assess the severity and extent of longitudinal joint cracking as *medium* and *extensive*, that of potholes as *medium* and *frequent*, and that of wheel track cracking as *low* and *extensive*. When this information is sent to the KBES, Rules 1, 3, and 4 are fired (using forward chaining), resulting in selecting one-course overlay. If more than one type of one-course overlay are applicable, then the final selection is performed to recommend the proper treatment. In other words, the secondary factors are taken into account along with their relative importance.

ORDINAL MULTIOBJECTIVE DECISION-MAKING FOR FINAL SELECTION

In the decision-making process, decision makers (DMs) often encounter the situation where they must select only one alternative from a set of alternatives subjected to a set of criteria or objectives to be satisfied. This type of problem is known as multiobjective decision-making. There exist many mathematical models that can be used to attack such problems, for example, mathematical programming techniques, which offer an acceptable solution when the assessments are made in a numerical fashion. An example problem would be, how to select the members of a structure that must result in a minimum weight structure while, to a certain extent, also satisfying strength, stiffness, and stability criteria. The assessment of alternatives with respect to these criteria could be done by carrying out a structural analysis. However, in many cases, such as the case of pavement treatment selection, the assessment of alternatives must be made by a DM. The DM, an engineer in this case, has to choose a treatment from a set of alternatives subjected to some criteria, such as how well the treatment would satisfy the functional and structural adequacy of a pavement.

Because humans frequently make their assessments subjectively, it may not be suitable to attempt to obtain this subjective information in a more precise way. Bellman and Zadeh (*6*) introduced an approach to tackle such decision-making problems in a fuzzy environment. Since then the use of fuzzy sets in this type of problem has been developed and has gained more and more popularity. Recently, a methodology for ordinal multiobjective decision-making based on fuzzy sets was proposed by Yager (*3*). Because of its suitability to the problem being studied, it has been chosen as a decision-making tool in the selection of minor rehabilitation treatments.

Based on the Bellman-Zadeh approach, Yager (*3*) developed a methodology to solve a special type of multiobjective decision-making problem in which the preference information about alternatives, criteria, and the relative importance of each criterion can be measured on the same ordinal scale. To illustrate Yager's model, the following notations are used:

$\{S\}$ is the finite set of elements used to indicate the preference information.

$\{X\}$ is the set of alternatives.

$Y = \{A_1, A_2, \ldots, A_P\}$ is the set of objectives (criteria) to be satisfied.

$A_i(x) \in S$ indicates the degree to which x satisfies the criterion specified by A_i.

G is a fuzzy subset of Y in which $G(A_i) \in S$ indicates the importance of the objective A_i. For the sake of simplicity, let $G(A_i) = b_i$.

$D(x)$ is the decision function from which the best alternative is to be selected.

$\cup$ is the disjunction (OR) set operator (which is equivalent to $\vee$ or a Max operator when elements are considered).

$\cap$ is the conjunction (AND) set operator (which is equivalent to $\wedge$ or a Min operator when elements are considered).

Yager proposed a general form for this type of decision function which includes the relative importance of each criterion as

$$D(x) = M(A_1(x), b_1)\ AND\ M(A_2(x), b_2) \ldots AND\ M(A_p(x), b_p) \quad (1)$$

where $M(A_i(x), b_i)$ indicates the objective A_i evaluated at alternative x, modified by its importance b_i. Yager proposed to use the following implication operation to compute $M(A_i(x), b_i)$ if S is a finite linearly ordered set:

$$M(A_1(x), b_1) = b_1' \vee A_i(x) \quad (2)$$

where b_i' is the negation of b_i. In this model, because $b_i \in S$, which is the finite linearly ordered set, the negation is defined as follows:

Let $\{S\} = \{s_0, s_1, s_2, \ldots, s_n\}$ where $i > j$ implies $s_i > s_j$. Then

$$s_i' = s_{n-i} \quad (3)$$

Hence, the decision set is

$$D = (b_1' \cup A_1) \cap (b_2' \cup A_2) \cap \ldots \cap (b_p' \cup A_p)$$
$$D = \bigcap_{i=1}^{p} (b_1' \cup A_1) = \bigcap_{i=1}^{p} C_1 = C_1 \cap C_2 \cap C_3 \cap \ldots \cap C_p \quad (4)$$

where

$$C_i(x) = b_i' \vee A_i(x) \quad (5)$$

and

$$D(x) = \text{Min}\ [C_1(x), C_2(x), \ldots, C_p(x)] = \min_i [C_i(x)] \quad (6)$$

Hence, the best alternative is the $x \in X$ that maximizes D, that is,

$$D(x^*) = \max_{x \in X} D(x) \quad (7)$$

In the application to pavement problems, suppose that after the preliminary selection, the KBES suggests three possible alternative treatments that an engineer can select to rehabilitate a pavement section. An example would be three different types of one-course overlay that differ in material types and/or thickness. In order to select the best alternative, the engineer uses the following additional criteria: functional adequacy, structural adequacy, and expected performance life. In addition, the relative importance of each criterion can also be specified to satisfy his/her requirements. The preference information set, S, can be defined as

$S = \{\text{high, medium, low}\}$.

Note that Yager's model does not require membership functions for elements in the preference information set because the preference information set must be a finite linearly ordered set. The alternative set, X, is

$X = \{\text{Treatment 1, Treatment 2, Treatment 3}\}$.

The set of criteria, Y, is

$Y = \{\text{functional adequacy, structural adequacy, expected performance life}\}$.

TABLE 3 Degree of Satisfaction of Each Criterion

Treatment x	Functional Adequacy $A_1(x)$	Structural Adequacy $A_2(x)$	Expected Performance Life $A_3(x)$
1	*high*	*low*	*high*
2	*medium*	*medium*	*high*
3	*low*	*high*	*medium*

The degree of satisfaction of each criterion, $A_i(x)$, is indicated in Table 3. Note that $A_i(x)$ must be assigned using the grades from the preference information set, S. In addition, the degree of satisfaction of each treatment subjected to each criterion must be rated relatively to other treatments and no correlation is considered among the criteria. For example, the degree to which Treatment 1 satisfies structural adequacy, $A_1(x)$, is low implies that it is low in comparison with medium and high of Treatment 2 and Treatment 3, respectively. However, low structural adequacy does not indicate that the functional adequacy of Treatment 1 must be rated in the same sense as structural adequacy. In fact, it must be rated relative to other treatments.

The relative importance of each criterion, b_i, is

$$b_i = \{\text{high, medium, medium}\}$$

Using Equation 3, the negation of b_i is obtained as

$$b_i' = \{\text{low, medium, medium}\}$$

Equation 5 yields the following:

$$C_1 = \text{low} \vee \{\text{high, medium, low}\} = \{\text{high, medium, low}\}$$
$$C_2 = \text{medium} \vee \{\text{low, medium, high}\} = \{\text{medium, medium, high}\}$$
$$C_3 = \text{medium} \vee \{\text{high, high, medium}\} = \{\text{high, high, medium}\}$$

Hence, the decision function, D, is obtained by using Equation 6.

$$D \text{ (Treatment 1)} = \text{Min } \{\text{high,medium, high}\} = \text{medium}$$
$$D \text{ (Treatment 2)} = \text{Min } \{\text{medium, medium, high}\} = \text{medium}$$
$$D \text{ (Treatment 3)} = \text{Min } \{\text{low, high, medium}\} = \text{low}$$
$$D = \{\text{medium, medium, low}\}$$

The final solution is therefore obtained from Equation 7. In this case, we have a tie, that is, Treatment 1 and Treatment 2. In the case of a tie, the engineer has three options: select one treatment from the alternatives that have tied, refine the scale, or use the following procedure, which was proposed by Yager (*3*) as well.

If there are two alternatives, x and y, which yield the same decision, then $D(x) = D(y) = \max_{z \in X} D(z)$. Because $D(x) = \min_i [C_i(x)]$, there exists some k such that $C_k(x) = D(x)$. Similarly, there exists some g such that $C_g(y) = D(y)$. Let $D'(x) = \min_i [C_i(x)]$, $i \neq k$ and $D'(y) = \min_i [C_i(y)]$, $i \neq g$. If $D'(x) > D'(y)$ then x can be selected as the solution. In the case that we have additional ties $D'(x) = D'(y)$, then the preceding procedure can be repeated until the solution is found or all the criteria are exhausted. In the latter case, the final decision will have to be made by the engineer. In sum, the alternatives that generate the same decision are progressively eliminated from the decision set until a solution (a distinct alternative) is found.

In the above example, we have

$$D' \text{ (Treatment 1)} = \text{Min } \{\text{high, } \sout{\text{medium}}\text{,high}\} = \text{high}$$
$$D' \text{ (Treatment 2)} = \text{Min } \{\sout{\text{medium}}\text{, medium, high}\} = \text{medium}$$
$$D = \{\text{high, medium}\}$$

Therefore, the final solution is Treatment 1.

COMPUTER PROGRAM

A computer program was written to implement the proposed methodology. Figure 1 shows the structure of the program which consists of four main modules: the Input Module, the Knowledge-based Module, the Multiobjective Decision-making Module, and the Output Module. The function of the Input Module is to obtain all the data needed for the Knowledge-based Module. Once the data

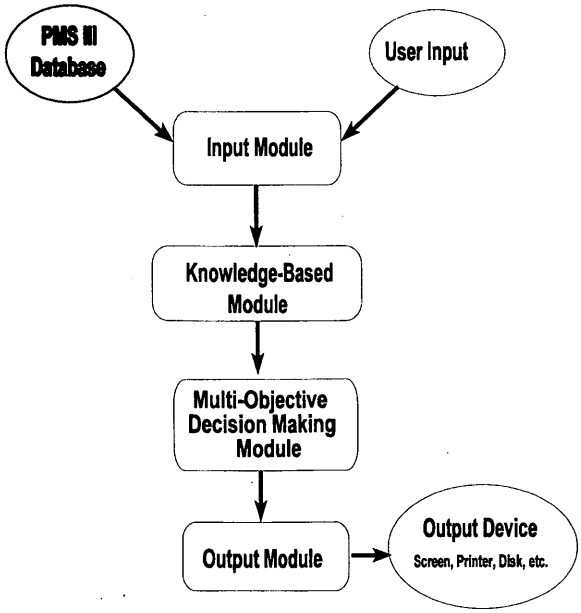


FIGURE 1 Structure of the program.

is obtained, the Knowledge-based Module proceeds with the selection of possible treatments. If there is more than one possible treatment, then the Multiobjective Decision-making Module is invoked to select and recommend the proper treatment. The solution is then reported to the user using the Output Module. Note that the Input Module also retrieves the past maintenance data from the PMS III maintenance database, which is a collection of rehabilitation project data in Ohio since 1985. The data consists of all the project records that have the same location as the new project and are presented in both graphical and text forms.

The program has been implemented using several software packages. Microsoft Visual Basic (VB) Version 3.0 (*7*) was used for the Input, Multiobjective Decision Making, and Output Modules, and Knowledge Pro Gold for Windows (KPWIN) Version 2.35 (*8*) was used for the Knowledge-based Module.

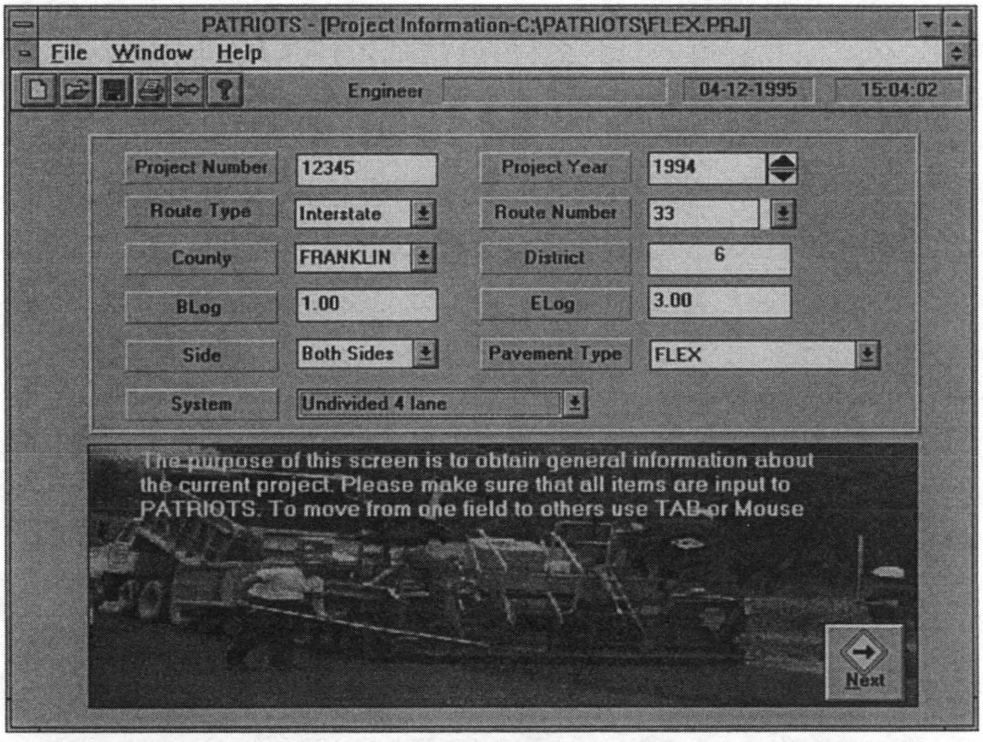
Figures 2 through 6 illustrate the above modules.

CONCLUSION

The methodology proposed in this study can be used to model the minor rehabilitation treatment selection process in Ohio. The KBES encodes the knowledge of experts at ODOT and serves as the preliminary selection tool in which a treatment or a set of alternatives are to be chosen. The ordinal multiobjective decision-making model using fuzzy logic can then be used to recommend a proper treatment by considering secondary factors along with their relative importance. Initial evaluation by the knowledge engineers and experts at ODOT indicates its feasibility and potential for use by ODOT maintenance engineers. More will be reported as the research progresses.

ACKNOWLEDGMENTS

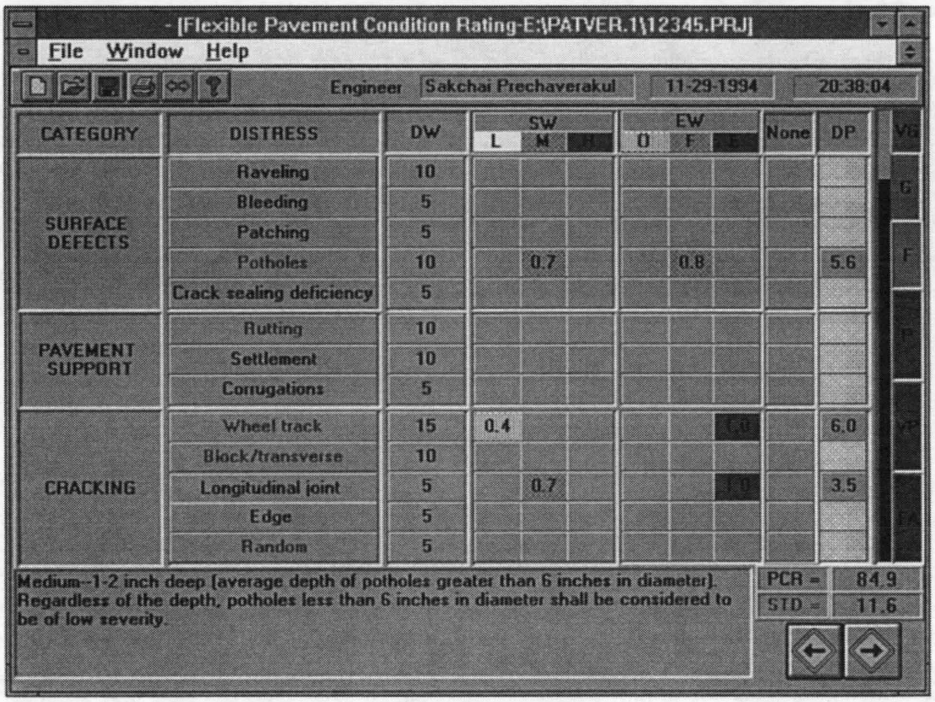
The authors would like to thank the Royal Thai Government through its Ministry of University Affairs which has sponsored the first author to study for his degree at The Ohio State University. They also would like to thank John A. Ray, Roger L. Green, and Richard D. Boyle for their valuable input to this study. The work was performed in the Construction Laboratory for Automation and System Simulation (CLASS) in the Department of Civil Engineering at The Ohio State University.



Note

Blog: Beginning Log
Elog: Ending Log

FIGURE 2 Project information (input module). Note: Blog, beginning log; Elog, ending log.



Note

DW:	Distress Weight	SW:	Severity Weight
EW:	Extent Weight	DP:	Deduction Point
VG:	Very Good	G:	Good
F:	Fair	P:	Poor
VP:	Very Poor	FA:	Fail
PCR:	Pavement Condition Rating		
STD:	Structural Deduct		

FIGURE 3 Pavement condition (input module). Note: DW, distress weight; SW, severity weight; EW, extent weight; DP, deduction point; VG, very good; G, good; F, fair; P, poor; VP, very poor; FA, fail; PCR, pavement condition rating; STD, structural defect.

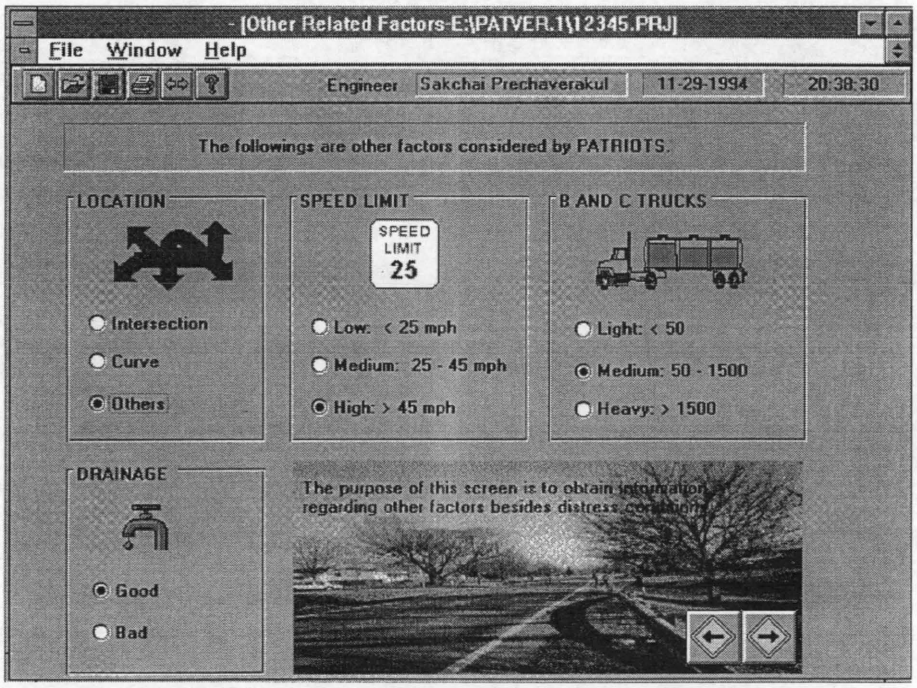


FIGURE 4 Other related factors (input module).

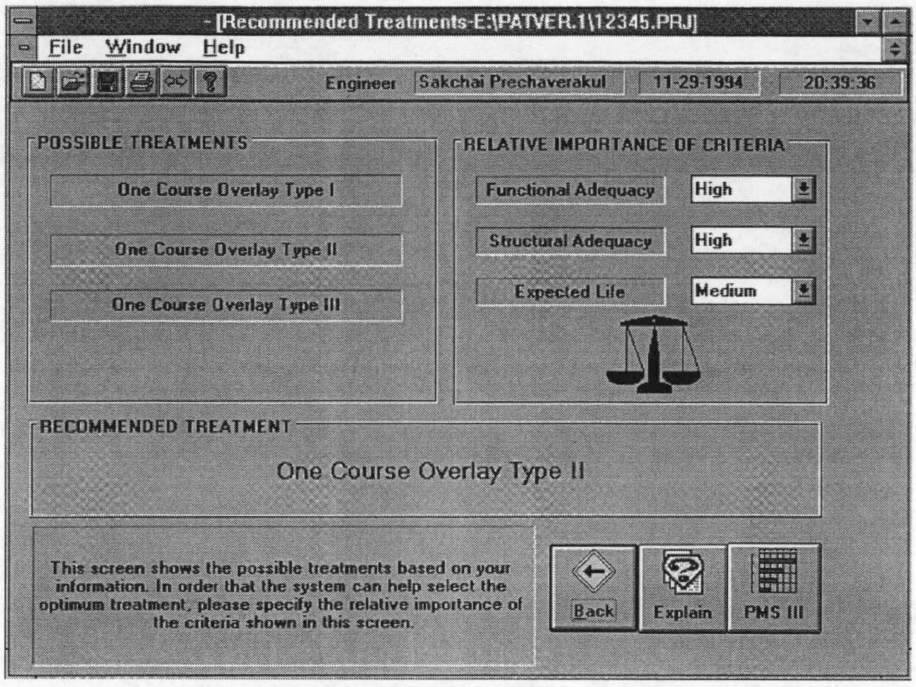


FIGURE 5 Multiobjective decision-making module.

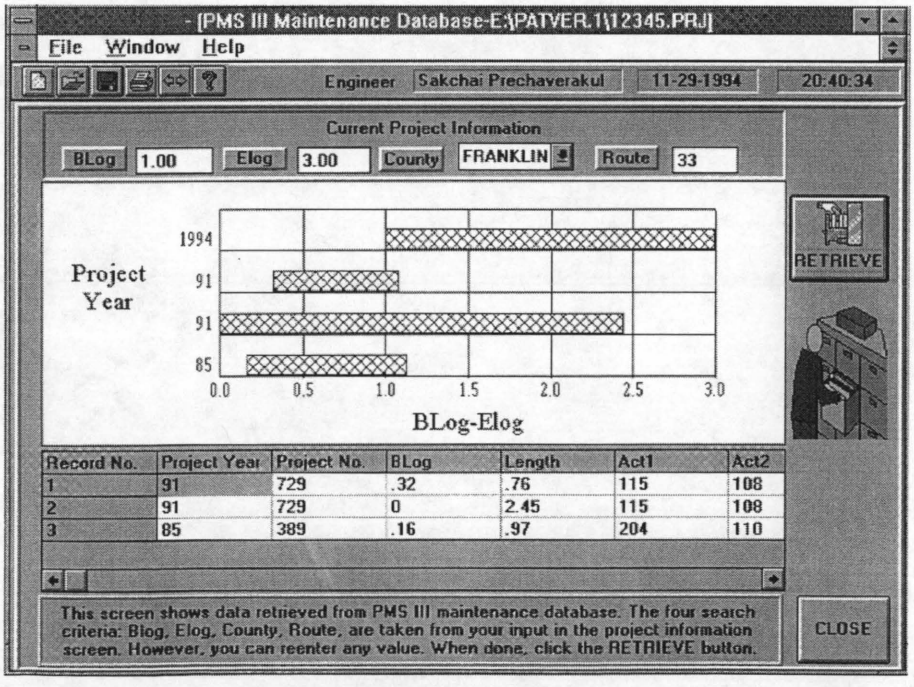


FIGURE 6 PMS III maintenance database.

REFERENCES

1. Majidzadeh, K., G. J. Ilves, J. C. Kennedy, and C. Saraf. *Implementation of Ohio Pavement Management System*, Vol. 1. Final Report Prepared for the Ohio Department of Transportation, Resource International Inc., Columbus, Ohio, 1990.
2. *Location and Design Manual*, Vol. 1, Ohio Department of Transportation, Columbus, Ohio, 1992.
3. Yager, R. R. A New Methodology for Ordinal Multiobjective Decisions Based on Fuzzy Sets. In *Readings in Fuzzy Sets for Intelligent Systems* (D. Dubois, H. Prade, and R. R. Yager, eds.), Morgan Kauffmann Publishers, Inc., 1993, pp. 751–756.
4. Majidzadeh, K., and A. Abdulshafi. *Implementation and Revision of Developed Concepts for ODOT Pavement Management System.* ODOR, Final Report, Vol. 2: Pavement Condition Rating Manual, The Ohio Department of Transportation, Columbus, Ohio, 1987.
5. Cohn, L. F., and R. A. Harris. Knowledge Based Expert Systems in Transportation. *NCHRP Synthesis 183*, TRB, National Research Council, Washington, D.C., 1992.
6. Bellman, R. E., and L. A. Zadeh. Decision Making in a Fuzzy Environment. *Management Science*, Vol. 17, 1970, pp. 141–164.
7. *Microsoft Visual Basic Version 3.0 Professional Edition*, Microsoft Corporation, 1993.
8. *KnowledgePro Gold for Windows Version 2.35*, Knowledge Garden, Inc., 1993.

Publication of this paper sponsored by Committee on Artificial Intelligence, Subcommittee on Fuzzy Systems and Uncertainty.

Real-Time Data Fusion for Arterial Street Incident Detection Using Neural Networks

JOHN N. IVAN, JOSEPH L. SCHOFER, FRANK S. KOPPELMAN, AND LINA L. E. MASSONE

This research contributes to the development of an automatic incident detection system for detecting traffic-delaying events on arterial street networks for an Advanced Traveler Information System demonstration called ADVANCE. Data describing current traffic conditions will be gathered in real-time from two distinct sources: inductive loop detectors and specially equipped vehicles that measure and report their travel times on roadway links. Two approaches are considered for data fusion, the combination of information from these sources to produce a single decision about the presence or absence of incidents on each link. In the integrated fusion approach, observed traffic data are combined directly using a neural network. In the algorithm output fusion approach, separate incident-detection algorithms individually preprocess data from each source, reporting outputs that are combined using a neural network. Data for calibrating these system components were generated using computer simulation. The algorithm output fusion network performed better than the other approach, detecting over 80 percent of the incidents with almost no false alarms. Fusing algorithm outputs using neural networks was thus found to improve the capability provided by separate source incident detection algorithms operating alone. The importance of validating these results through calibration and testing with field data, as well as improving performance through introduction of an additional data source is discussed.

Highway facilities are designed to operate acceptably within some range of demand. When an incident occurs, such as a traffic accident, a load spill, or a vehicle breakdown in the roadway, there is a sudden, temporary decrease in the capacity of a particular section of the facility. When demand exceeds this temporarily reduced capacity, queues, delays or perhaps more accidents result, as well as increased difficulty in clearing the scene (*1*).

Identifying incidents quickly is important for reducing their impacts. Researchers over the years have developed Automatic Incident Detection (AID) systems for freeways that monitor traffic flow information from a highway facility and automatically detect such incidents (*2*) to prevent secondary accidents from occurring and to dispatch emergency or cleanup crews promptly. A few researchers have studied incident detection on arterial streets (*3,4,5*), using techniques similar to those used on freeways. However, the arterial street environment is much more challenging because traffic flow discontinuities are introduced by traffic signals, traffic entering and leaving side streets and driveways, and variation in signal timing and geometric characteristics. Therefore, the effect of the same type of incident varies for different sets of arterial conditions.

J. N. Ivan, University of Connecticut, Transportation Institute U-37, Storrs, Conn. 06269-2037. J. L. Schofer and F. S. Koppelman, Northwestern University Transportation Center, Evanston Ill. 60208. L. L. E. Massone, Northwestern University, Departments of Electrical Engineering and Computer Science, Department of Biomedical Engineering,. Evanston, Ill. 60208

Recently, there has been a great deal of interest in increasing the efficiency of existing highways through the deployment of Advanced Traveler Information Systems (ATIS), a type of Intelligent Transportation System (ITS). An ATIS provides current traffic information to system users to help them reduce their travel times. The Illinois University Transportation Research Consortium (IUTRC), Motorola, and the Illinois Department of Transportation (IDOT), with funding from the Federal Highway Administration (FHWA), are preparing to launch a demonstration of such a system in the Chicago suburbs, called ADVANCE, or Advanced Driver and Vehicle Advisory Navigation Concept (*6*). ADVANCE will provide between 3,000 and 5,000 demonstration participants with shortest path routings to specific destinations using up-to-the-minute travel times on arterial and freeway links in an approximately 600 sq km service area. Advising drivers of traffic operational problems on the highway network will be an important function of the demonstration, so an AID system will be important for enhancing the value of information provided by the system and promoting constructive driver response (*7*). Detailed information about current traffic conditions and causes will be particularly valuable for predicting travel times.

METHODOLOGY

Data Sources

AID systems described in the literature have operated almost exclusively using data from fixed detectors, sensor systems that measure traffic characteristics at a fixed location, such as inductive loop detectors (ILDs) or video cameras (*2*). Fixed detectors measure the following traffic flow quantities:

1. Volume, the arrival rate of vehicles passing the detector during the measurement period;
2. Occupancy, the percentage of time that the space above the detector is occupied by a vehicle during the measurement period; and
3. Speed, the rate of motion of a vehicle as it passes the detector (the detector may provide individual vehicle speeds or average speed over the measurement period).

Traffic signals on primary arterial routes (on which traffic flow is most vulnerable to unusual congestion) are often connected in closed loop systems, which coordinate signal timings and facilitate collection of data in real-time over data communication lines. Not all of the roadway links in the ADVANCE street network will be instrumented with these systems, so additional data sources are required to cover other network sections.

One potential alternative source is observed travel times collected in real-time from probe vehicles traveling the street network, a common feature of many ITS implementations, including ADVANCE. Vehicles participating in ADVANCE will automatically report observed link travel times to a Traffic Information Center (TIC). Probe vehicles can help locate congestion on any roadway segment in the area served by the vehicle communications medium, without the spatial limitations of fixed detectors, although network coverage is limited by the market penetration rate.

ADVANCE will collect data from both of these sources, which are considerably different from each other both qualitatively and temporally. It is thus desirable to use separate procedures, or incident detection algorithms, for each data source to determine whether or not there is an incident (or the likelihood of there being an incident) on each link. A data fusion process would then solve the more clearly defined task of combining the incident decisions made by the two procedures.

Data Fusion Approaches

The focus of this research is the development of this data fusion process. These data sources are inherently imperfect and not entirely reliable, so this procedure must be able to:

1. Identify incident conditions under a variety of input data patterns,
2. Integrate inferences from input data with varying degrees of certainty; and
3. Account for complex relationships among input sources.

A number of information processing techniques have proven useful for decision-making and combining uncertain information in a variety of contexts. Following is an evaluation of the potential effectiveness of several such techniques for solving the data fusion problem posed here.

Decision support systems interpret surveillance information and recommend a course of action for the system operator, who must then make a decision. Prosser and Ritchie (*8*) describe such a system for incident management. This system really filters available information, conserving the operator's attention for confirming computer generated results. While it might be necessary for an operator to monitor incident detection system operation to avoid broadcasting false alarms, particularly in the early stages of implementation, it is more desirable for the ADVANCE incident detection system to operate without requiring regular operator response.

The best score approach (*9*) simply chooses the information source that is considered most valid (according to a predetermined quality score and an aging or decay rate) and uses it alone. This procedure may be better described as a "winner-take-all" strategy rather than data fusion. Discarding other information ignores possible interactions among the various data sources that contribute to system performance.

Virtual sampling (*10*) regards several estimates of an unknown quantity as unique random samples of observations drawn from a given population whose mean value is the unknown quantity. Estimates with low standard errors are considered to represent larger samples. For combining data sources that represent independent observations of the same phenomenon, this approach seems quite attractive, since it places more emphasis on values that are more certain and requires no weights to be calibrated. However, it does require that standard errors be provided with each estimate.

Artificial Neural Systems (ANS), also known as neural networks, are information processing structures that attempt to replicate the process of learning and decision-making observed in the operation of the human brain (*11,12*). The nature of neural networks makes them appropriate for solving many complex problems that have proven to be quite cumbersome for conventional processing systems. For example, problems in which the precise interrelationships among elements are not well understood, such as continuous speech processing and pattern recognition, are good applications for ANS. An ANS is best implemented as a partner to a traditional system, with the traditional system (for example, our incident detection algorithms) performing precise, specific computations and the ANS analyzing less precisely defined tasks (such as data fusion).

Approach Concepts

Two data fusion approaches using neural networks are considered here.

Algorithm Output Fusion

In this approach, depicted in Figure 1, two algorithms, each uniquely developed for one of the two data sources (fixed detectors and probe vehicles), determine the likelihood that an incident is occurring at particular locations on the street network. A separate data fusion process using a neural network then combines the output from these algorithms.

Integrated Fusion

Here the functions of the single source incident detection algorithms and the fusion process are combined in a single neural network. In the first approach, the algorithms effectively censor the unprocessed input data, translating them into a single output value, preventing unprocessed input data from one source from helping to interpret data from the other source. This network will read input directly from the data sources, then fuse data and detect incidents simultaneously.

NETWORK TRAINING

Training Data

Humans learn by repeatedly observing the outcomes of their responses to external stimuli. In the same way, network training in

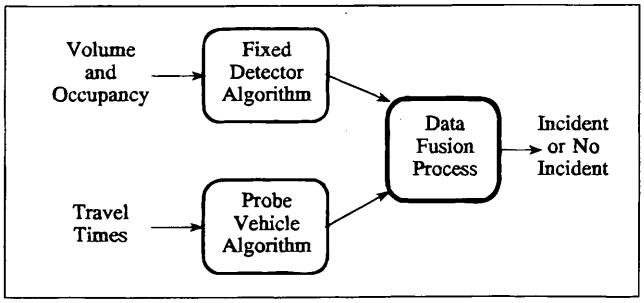


FIGURE 1 Algorithm output fusion concept.

this application requires a set of data that represent the variety of traffic conditions under which the incident detection system must operate, along with the outcomes it should return.

Most incident detection systems are calibrated for specific road sections: a different set of parameter values—for example tolerance thresholds—is used for each pair of detectors, or road section (*2*). For an arterial street network with many road segments, such a calibration would be exceedingly tedious. Instead, for this incident detection system, it was desired to calibrate more general algorithms that would apply to any of the arterial segments in the service area.

To do this, the training data must incorporate the variation that arises naturally in traffic flow on urban arterial streets. This implies an enormous number of combinations of street, traffic, driver, and incident patterns, of which the data set must be a suitable sample. Much of this variability is derived from human behavior (driving patterns, the occurrence of incidents), and real traffic conditions are the best source for unbiased observation of these phenomena. However, real-world traffic data involving incidents are considerably difficult to obtain (particularly for the sources involved here), so the training data used here were generated through traffic simulation using INTRAS, a microscopic freeway corridor traffic simulation model (*13*).

The arterial street network simulated for data collection is a representation of an approximately 5-km section of major arterial streets in the ADVANCE network, as depicted in Figure 2. This road section has 39 loop (fixed) detectors at eight intersections that are located to collect data for signal controllers. The simulation model includes these detectors along with additional detectors such that detector stations are located at each eastbound intersection approach. Signal timing was varied at several intersections to study the effect of incidents under different congestion (volume to capacity ratio) conditions. Similarly, incidents were placed at different locations on links to study the effect of the distance from the signal on traffic operation.

More than 100 simulation runs were performed to generate data for a variety of incident and corresponding nonincident conditions. A data aggregation interval of 7 min was selected so that no cycles would run between two intervals (all signals have cycle lengths of 140 sec); flow variations through the cycle thus do not taint the traffic measurements, but the interval is short enough to permit timely traffic condition updates. For each incident simulation, a number of nonincident simulations with identical control variables were also performed. Incidents were simulated on six different links, at three or four locations on each link, for durations of from 5 to 10 aggregation intervals, and with up to three different signal timing patterns at selected signals. All of the incidents were simulated in the eastbound direction, so only data from the eastbound links are used in the analysis.

Training data were prepared by extracting aggregated occupancy, volume, and travel time reports from INTRAS output corresponding to each simulated incident. For the Algorithm Output Fusion Network, input vectors were generated by processing traffic data with the calibrated single source incident detection algorithms (*14*). Each input vector corresponds to the conditions on one link during a particular incident simulation time interval, and consists of the fixed detector and probe vehicle algorithm scores scaled to keep their values between -1.0 and $+1.0$ (negative values indicate no incident, positive values indicate an incident) and the target output equal to 1.0 if there was an incident on the link during the time interval and 0.0 if not.

A similar process was used to organize the simulated traffic surveillance data into training vectors for the integrated fusion approach, but the following input values replace the algorithm scores on each vector:

1. Volume ratio, equal to the traffic flow at the detector station during the time interval divided by the average total traffic flow at the station under nonincident conditions;
2. Occupancy ratio, equal to the average traffic occupancy measured at the station on the analysis link during the time interval divided by the average traffic occupancy for the same station under nonincident conditions; and
3. Travel Time Ratio, equal to the average of the travel times observed on the analysis link during the time interval divided by the average travel time for that time period on the link under nonincident conditions.

All of these values were also scaled to keep their values between 0 and 1.

Training Procedure

Both data fusion approaches were developed with feed forward networks trained using error back propagation (*12*). Training prepares a network for application and involves presenting a series of input arrays, or vectors, to the network one at a time along with their corresponding target output values. The network adjusts connection weights over a series of many epochs, or iterations, so that it can reproduce the desired output values.

The network structure for the two approaches are depicted in Figures 3 and 4. Both use a single hidden layer of five units and a single output unit. All hidden layer units and the output unit add a bias threshold to their net inputs. For each input vector presentation, the output value is calculated using Equation 1

$$y = f\left(v_0 + \sum_{j=1}^{m} v_j f\left(w_{0j} + \sum_{i=1}^{n} w_{ij} x_i\right)\right) \tag{1}$$

where

y = resulting network output;
f = activation function;
v_0 = bias threshold on the output unit;
m = number of hidden units;
v_j = weight on the connection from hidden unit j to output unit;
w_{0j} = bias threshold on the hidden unit j;
n = number of input units;
x_i = input signal from input i; and
w_{ij} = connection weight from input i to hidden unit j.

The logistic, or sigmoid function, is used as the activation function on hidden and output units, because its output closely resembles a threshold-step function and is differentiable; it is depicted in Equation 2

$$f(z) = \frac{1}{1 + \exp(-z)} \tag{2}$$

Next, the output calculated for each input pattern is compared to the corresponding target output, and gradient steepest descent, using the square of the difference between the target and observed output as an error function, is used to adjust the connection weights and bias thresholds so that the network output will be closer to the target output value the next time it is presented the input pattern. Because the sigmoid function only asymptotically approaches 0 or 1, (non-incident or incident), target values of 0.1 and 0.9 are used instead. Partial derivatives of the error function are taken with

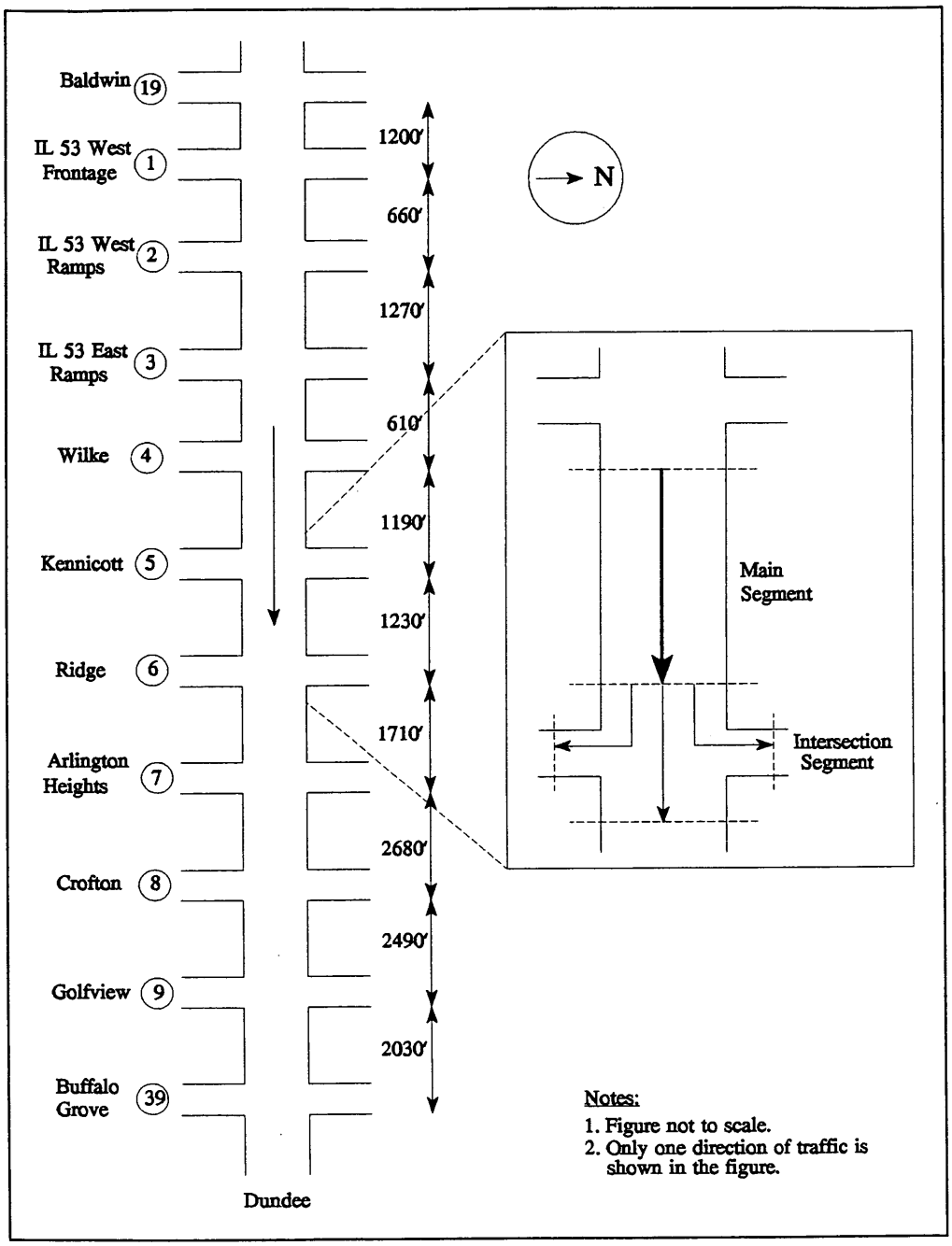


FIGURE 2 Simulation network.

respect to each connection weight and are then used to adjust the weights as expressed in Equation 3

$$\Delta w_{ij} = \eta \delta_j x_i \tag{3}$$

where

Δw_{ij} = change computed for the connection weight from unit i to unit j;
η = learning rate (controls the rate at which the network makes adjustments);
δ_j = propagated backward through unit j, and
x_i = activation of unit i.

When j is the output unit, the propagated error is given by Equation 4,

$$\delta_y = (y^* - y) f'(v_y + \sum_{j=1}^{m} v_j x_j) \tag{4}$$

where

δ_y = delta value for the output unit;

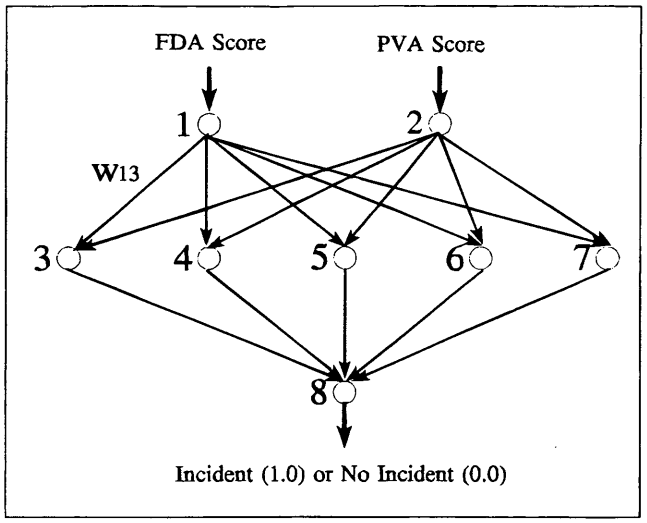


FIGURE 3 Algorithm output fusion network structure.

y^* = desired target output value;
$f'(z)$ = first order derivative of the sigmoid function;
x_j = activation of the jth hidden units; and remaining symbols are as previously defined.

For hidden units, Equation 5 is used

$$\delta_j = \delta_y w_{jy} f'(w_{0j} + \sum_{i=1}^{n} w_{if} x_i) \quad (5)$$

where δ_j is the delta value for hidden unit j.

The learning rate determines how much to change each weight value after each input vector presentation. A better performance of the learning algorithm can be achieved by incorporating a momentum term, as shown in Equation 6

$$\Delta w_{ij}(t) = \eta(\delta_j x_i)_t + \alpha \Delta w_{ij}(t-1) \quad (6)$$

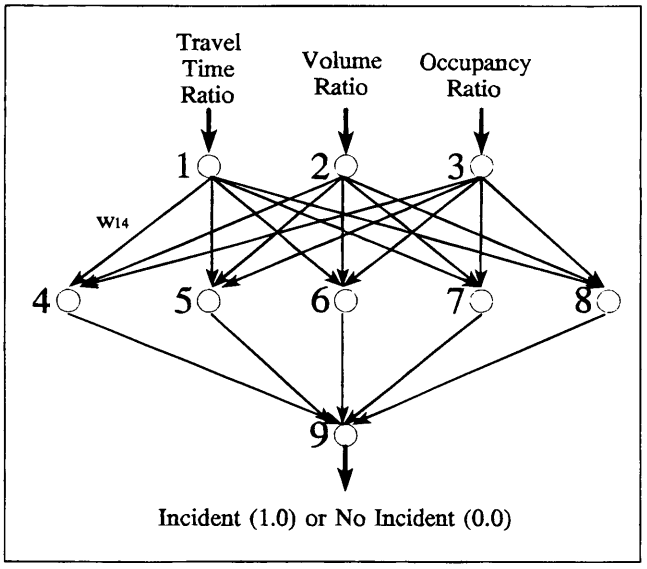


FIGURE 4 Integrated fusion network structure.

where $\Delta w_{ij}(t)$ is the weight change for the connection from unit i to unit j for presentation t, the values subscripted with t are determined for presentation t, and α is the momentum rate. The momentum term causes weight changes to move steadily in the same average direction and prevent it from settling in a local minimum. For both of these networks, a learning rate of 0.2 and a momentum rate of 0.8 were used.

The following procedure was followed for each network:

1. Ten percent of the training data files were selected randomly and set aside for testing.
2. Network training began with asymmetric connection weights randomly assigned values between −0.5 and + 0.5 (*15*); the remaining data files were divided into seven groups and added to the network training set one at a time every 50 epochs.
3. Starting at 500 epochs, network performance was tested on both the training and reserved data sets at regular 250-epoch intervals; training was discontinued when root mean square error stopped decreasing on both training sets or began to increase on the reserved set (indicating overfitting of the training data) (*11*).

RESEARCH FINDINGS

The Algorithm Output Fusion Network was trained using the parameters and procedure just described. Figure 5 shows the relationship between the root mean square error (RMSE) of the network output (from the target values) and epoch (or iteration) as the network trained. The wide oscillations early on result from the incremental introduction of data files into the training vectors. Note that the error does not change much after 1000 epochs of training. Figure 6 shows the same plot for the Integrated Fusion Network. RMSE again drops quickly by the thousandth epoch, but instead of holding at a constant value, it continues to decrease at an extremely slow rate. However, detection performance through this period does not appreciably improve, so training was terminated.

The performance of both networks on both the reserved test data and the training data is depicted in Table 1. The following performance measures are shown:

1. RMSE over all vectors in each data set;
2. Detection rate, the proportion of known incident observations (each individual period that an incident occurred) correctly classified; and
3. The false alarm rate, the proportion of nonincident observations incorrectly classified as incidents.

The Algorithm Output Fusion Network detects all of the incident observations in the reserved data set with no false alarms, but detects only 81 percent and misclassifies 0.11 percent of the nonincident observations in the training data set. The Integrated Fusion Network did not train as well. RMSE is 0.0901 for the training data and 0.1051 for the reserved data. Although RMSE at this stage continues to decrease with the training data, detection rate on both data sets stabilized at 66 percent on the training data and 70 percent on the reserved data, so further training would not yield better results. This network resulted in an unacceptably high false alarm rate for both data sets.

It is also worth noting that the networks perform much better on the reserved test data than on the training data with which they learned. This is an unexpected result, as it is analogous to a student scoring better on questions she had not seen before than on the ones

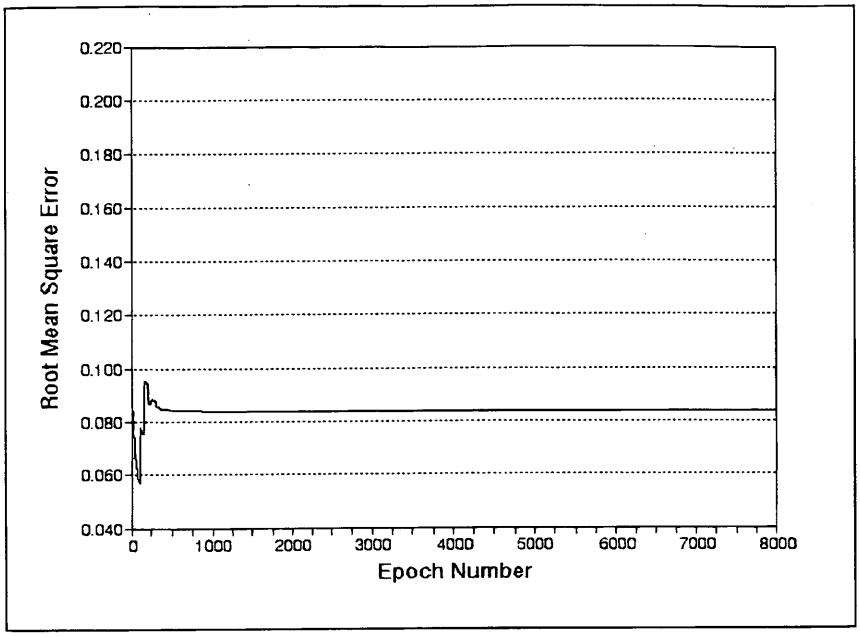


FIGURE 5 Algorithm output fusion network error by epoch.

she rehearsed prior to the examination. It turns out that the incidents in the reserved data set (though selected randomly) caused more extreme traffic conditions on average than did the incidents in the training data set. This does not implicitly invalidate this partitioning of the data files; it is simply necessary to test the network with both data sets to understand its true performance.

Table 1 also lists performance measures for the incident detection algorithms for comparison. The Integrated Fusion Network does not perform much better than the algorithms, but the Algorithm Output Fusion Network dominates all other processes with its much greater detection rates. Data fusion process and algorithm performance can be compared more directly by plotting adjusted algorithm output scores against each other on a grid for each incident, marking each observation according to whether or not the incident was detected. These plots are shown in Figure 7 for the Algorithm Output Fusion Network and in Figure 8 for the Integrated Fusion Network.

The X and Y axes divide each plot into four quadrants. The lower left quadrant contains known incident observations which both algorithms fail to detect, and the upper right quadrant, those which

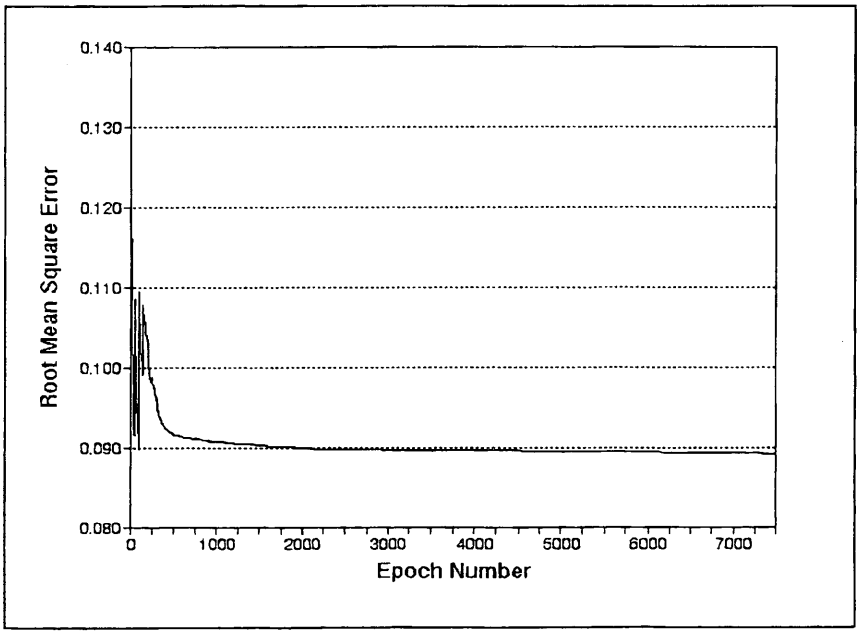


FIGURE 6 Integrated fusion network error by epoch.

TABLE 1 Neural Network Performance Summary

	RMS Error	Detection Rate	False Alarm Rate
Algorithm Output Fusion			
Training Data	0.0838	81.5%	0.11%
Reserved Data	0.0288	100.0%	0.00%
Integrated Fusion			
Training Data	0.0901	65.7%	0.54%
Reserved Data	0.1051	70.0%	0.96%
Fixed Detector Algorithm	—	65.9%*	0.00%*
Probe Vehicle Algorithm	—	53.7%*	0.00%*

— indicates value not available
* Source: *(14)*

both algorithms do detect. Observations in the upper left quadrant are missed by the fixed detector algorithm but detected by the probe vehicle algorithm, and those in the lower right quadrant are detected by the fixed detector algorithm but missed by the probe vehicle algorithm. This interpretation helps identify improvements offered by each data fusion process over the algorithms.

The Integrated Fusion Network offers no appreciable improvement over the incident detection algorithms, as it misses many incidents that both algorithms detected, and detects few that were missed by either algorithm. The Algorithm Output Fusion Network does much better, detecting all but two incidents that were detected by at least one of the algorithms. Even though it misses two incidents detected by the fixed detector algorithm, it detects five that the algorithm missed.

SUMMARY AND CONCLUSIONS

The results show that neural networks can be trained to detect incidents in arterial street settings at least as well as many conventional algorithms do in the less challenging freeway setting. The Algorithm Output Fusion Network detected well over 85 percent of the incidents in the data sets with no false alarms. A performance evaluation of a number of prominent conventional freeway incident detection algorithms (2) found false alarm rates over 0.5 percent associated with detection rates this high. Note, however, that a greater variety of incident types is included in these other studies and that the networks considered here were all trained with data collected from a traffic simulation rather than with field data as the conventional algorithms were. To the extent that the traffic simulation program used to generate the training data was calibrated to replicate the operation of a real street, it may well be reasonable to compare performance with the other algorithms directly. Nevertheless, confidence in this result would increase if similar results were obtained from a network tested (or trained) with field data.

It has also been shown that incident detection system performance can be improved by combining information from different data sources, in this case fixed detectors and probe vehicles. This idea is partially supported by the plots of algorithm output scores for incident records classified by whether each incident was detected or missed, but more positively by the detection rates reported on Table 1. Since the algorithms were calibrated to report no false alarms, they miss many marginal incidents which the networks are able to detect by combining the algorithm reports. This is good news for ITS demonstrations such as ADVANCE, which use information from a variety of sources. The bad news is that data from fixed detectors appear to be more reliable than data collected by probe vehicles, as evidenced by the superior performance of the fixed detector algorithm (*14*), and the availability of fixed detector data will be limited in ADVANCE. However, the probe vehicle incident detection algorithm was able to detect as many as 61 percent of the incidents alone with no false alarms (*14*), so a reason-

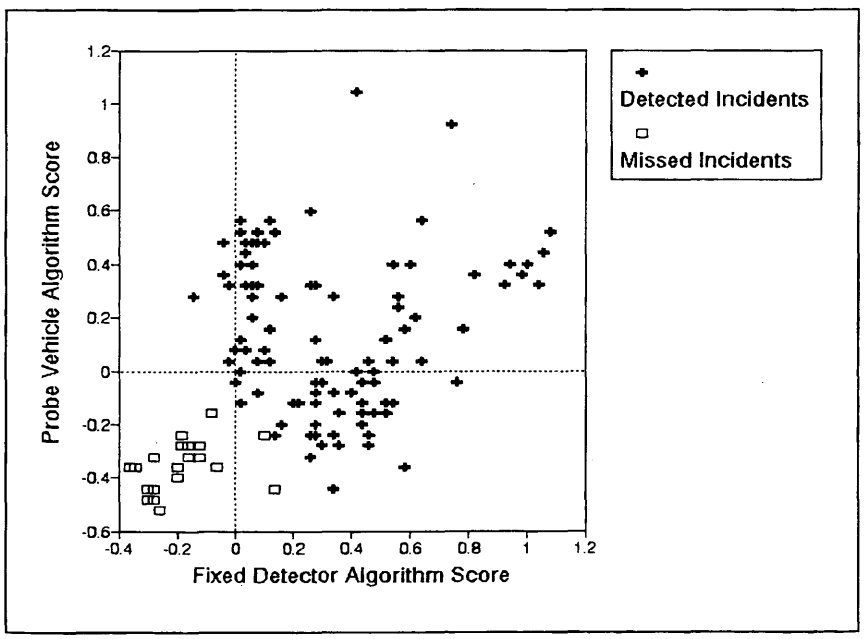


FIGURE 7 Algorithm scores for all incident observations—algorithm output fusion network.

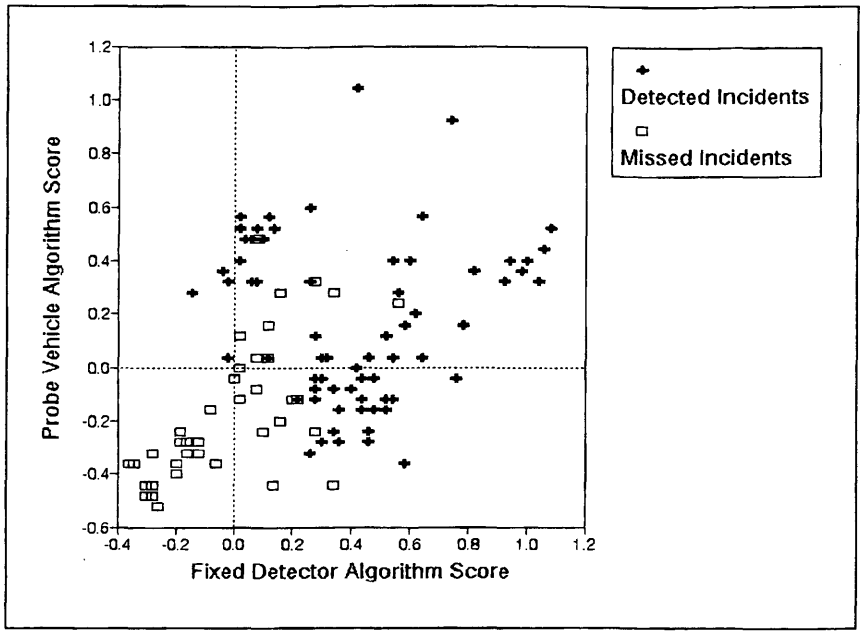


FIGURE 8 Algorithm scores for all incident observations—integrated fusion network.

able incident detection system is attainable even without fixed detector data; these data simply permit the system to detect a few more less-severe incidents.

FUTURE DIRECTIONS

A desirable objective of this (or any) incident detection system is to be an "off-the-shelf" algorithm which does not require recalibration for each implementation site. While the network performance observed suggests that it was not necessary to learn different parameters for each highway link, this total portability feature will become much more reliable as the variety of traffic and street characteristics in the training data, and thus, the transferability of the result, increases. The ADVANCE demonstration, once it becomes operational, can provide field-collected fixed detector and probe data that implicitly include this variability. Observed travel times will be collected regularly for all links traveled by the participating probe drivers; this information could be combined with any fixed detector data available for those links.

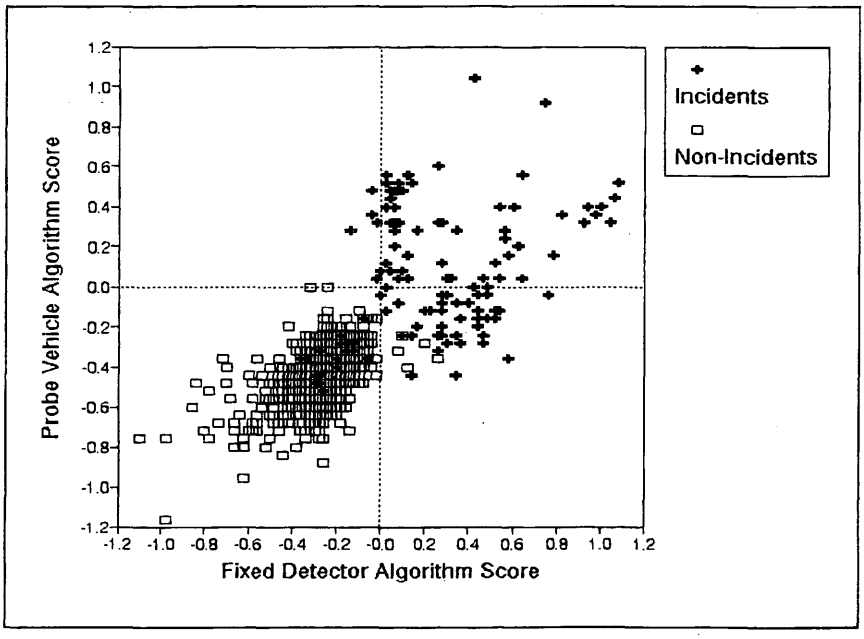


FIGURE 9 Algorithm score patterns.

A number of additional future research directions are suggested by this work. The networks should be retrained with field data, as discussed, to confirm these findings and to investigate other issues such as optimal network and input data representations and the effect of performance of using additional data sources (such as motorist calls and emergency dispatch communications).

The simulation data used for calibration of the probe vehicle incident detection algorithm and for training the neural networks considers 25 percent of the vehicle stream to be probes. The effect of much smaller probe vehicle proportions and the number of travel time reports included in each aggregation interval should be investigated.

This research is concerned with detecting incidents on arterial streets; a logical extension is development of systems that use this capability to modify traffic control parameters (e.g., signal timings) in response to observed traffic conditions. Such systems will become increasingly important as ITS implementations attempt to extract more and more capacity from existing highway networks.

ACKNOWLEDGMENTS

The information in this report was partially funded and developed for the ADVANCE Project, a joint ITS undertaking by the (FHWA) Federal Highway Administration, the Illinois Department of Transportation, the University of Illinois and Northwestern University operating under the auspices of the Illinois Universities Transportation Research Consortium, Motorola, Inc., and other participating organizations.

REFERENCES

1. Hall, R. W. Non-Recurrent Congestion: How Big is the Problem? Are Traveler Information Systems the Solution? *Transportation Research Part C,* No. 1, March 1993, pp. 89–103.
2. Stephanedes, Y. J., A. P. Chassiakos, and P. G. Michalopoulos. Comparative Performance Evaluation of Incident Detection Algorithms. *Transportation Research Record 1360,* TRB, National Research Council, 1992, pp. 50–57.
3. Han, L. D., and A. D. May. *Automatic Detection of Traffic Operational Problems on Urban Arterials.* UCB-ITS-RR-89-15, Institute of Transportation Studies, University of California at Berkeley, 1989.
4. Hounsell, N. B., M. McDonald, and C. F. S. Wong. Traffic Incidents and Route Guidance in a SCOOT Network. In *Traffic Management and Road Safety, Proceedings of Seminar B, Planning and Transport Research and Computation (International) Co. Ltd. (PTRC) Transport and Planning Summer Annual Meeting,* University of Bath, England, September 12–16, 1988.
5. Thancanamootoo, S., and M. G. H. Bell. *Automatic Detection of Traffic Incidents on a Signal-Controlled Road Network.* Transport Operations Research Group, University of Newcastle upon Tyne, Research Report No. 76, 1988.
6. Boyce, D. E., A Kirson, and J. L. Schofer. Design and Implementation of ADVANCE: The Illinois Dynamic Navigation and Route Guidance Demonstration Program. In *Vehicle Navigation and Information Systems Conference Proceedings,* Society of Automotive Engineers, Dearborn, Michigan, October 20–23, 1991, P-253 Part 1.
7. Khattak, A. J., J. L. Schofer, and F. S. Koppelman. Factors Influencing Commuters' En Route Diversion Behavior in Response to Delay. In *Transportation Research Record 1318,* TRB, National Research Council, 1991, pp. 125–136.
8. Prosser, N. A., and S. G. Ritchie. Real-Time Knowledge-Based Integration of Freeway Surveillance Data. Presented at the Transportation Research Board 70th Annual Meeting, January 13–17, 1991.
9. Sumner, R. Data Fusion in Pathfinder and TravTek. In *Vehicle Navigation and Information Systems Conference Proceedings, P-253 Part 1,* Society of Automotive Engineers, Dearborn, Michigan, October 20–23, 1991.
10. Hamburger, H. Representing, Combining and Using Uncertain Estimates. *Uncertainty in Artificial Intelligence,* (L. N. Kanal and J. F. Lemmer, eds.) Amsterdam: Elsevier Science Publishers B.V., 1986.
11. Caudill, M. Neural Networks Primer, Parts I, II and III. Reprinted from *AI Expert,* Miller Freeman Publications, San Francisco, 1990.
12. Rumelhart, D., and J. McClelland. *Parallel Distributed Processing Volume I: Foundations.* MIT Press, Cambridge, Mass., 1986.
13. Goldblatt, R. B. *Development and Testing of INTRAS, A Microscopic Freeway Simulation Model, Vol. 3, Validation and Application.* FHWA Offices of Research and Development, Traffic Systems Division, 1980.
14. Koppelman, F. S., J. L. Schofer, N. Bhandari, V. Sethi, and J. N. Ivan. *Calibration of Probe and Fixed Detector Algorithm Parameters with Simulated Data.* Evanston, Illinois: ADVANCE Project Technical Report TRF-ID-151, The Transportation Center, Northwestern University, 1994.
15. Hertz, J., A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation.* Addison Wesley, 1991.

Publication of this paper sponsored by Committee on Artificial Intelligence.

Neural Network Estimation of Waterway Lock Service Times

YEON MYUNG KIM AND PAUL SCHONFELD

Good service-time estimates at locks are essential for evaluating waterway performance, planning improvements, and controlling operations. Difficulties in estimation are due to great variations in lock characteristics, vessel characteristics, operating options, and environmental conditions. In this study several artificial neural network models for lock service-time estimation are developed and compared. Results show that simple artificial neural network models yield lower prediction errors than simple regression models, that systematic removal of outliers can reduce the number of artificial neural network prediction errors, and that combined service-time models for locks with dissimilar chambers can be obtained without unreasonably compromising accuracy.

Inland waterway transportation in the United States is used for shipping heavy or bulky commodities because it is inexpensive and energy efficient. There are 216 lock chambers at 167 lock sites operated in the United States. The lock structures (Figure 1) used to raise or lower vessels across dams constitute the major bottlenecks in the U.S. waterway network and generate extensive queues, which lead to costly delays.

Locks have one or two parallel chambers whose characteristics may differ greatly. A commercial tow typically consists of a tow boat and a number of barges. If a tow has more barges than the chamber can accommodate, it must be disassembled into several pieces (called cuts) to pass through the chamber, and must be reassembled later. The lock service time mainly depends on the chamber size and tow size. The number of barges, number of cuts, and tow direction also affect the lock service time.

PROBLEM STATEMENT

Good estimates of lock service times are essential for improving lock operations, either through long-term investments or short-term control. However, service times are quite complex and are influenced by numerous factors. Lock service time is defined as the sum of all times (approach time, entry time, chambering time, exit time, time between cuts, turn-back time, etc.) spent processing a given tow through a specific lock.

Several studies of lock service time have used traditional methods such as regression analysis (*1*) and simulation (*2*), and have obtained relatively inaccurate models. Dai and Schonfeld (*2*) had to use historical service-time distributions rather than estimated models in their simulation. In this paper, we explore the possibility of obtaining better service-time models using neural network methods. In the following sections we discuss candidate variables, neural network models, comparative regression model building, model results, and model validation.

Transportation Studies Center, University of Maryland, College Park, Md. 20742.

IDENTIFICATION OF CANDIDATE VARIABLES

For this work we used the data from the Corps of Engineers' performance monitoring system (PMS) 1988 data base, which provides comprehensive records of the arrival and processing times for all vessels using a lock. Lock 27 on the Mississippi River was selected. It has a large main chamber (33.55 m × 366 m) and a half-size auxiliary chamber (33.55 m × 183 m). From the PMS data base, 14 candidate variables were selected based on their high correlations with lock service time. From those 14 variables, 6 input variables (tow direction, index of same direction, number of cuts, number of barges, ratio between tow length and chamber length, and ratio between tow width and chamber width) and 1 output variable (service time) are defined. The service time is the sum of approach time, entry time, chambering time, and exit time. If the tow must be cut to get through the lock, the time between cuts and turn-back time are added to the service time.

Statistical Analysis of Service Time

It is difficult to define the specific distribution of service time due to its complexity and the great variation in causal factors. The service-time distribution can be checked by analyzing its statistical characteristics. During 1988, 8090 tows were observed through the main chamber and 3784 through the auxiliary chamber. Table 1 shows the summary statistics for the service times of the main and auxiliary chambers.

The mean service times for the main and auxiliary chambers are 44.218 min and 26.490 min, respectively. Histograms for both chambers in Figure 2 show that service-time distributions are skewed to the right. Very few tows have large service times. The maximum deviation of service time from the mean is 6.4 standard deviations (σ) for the main chamber and 11.2 σ for the auxiliary chamber.

Because data collection is performed by lock operating personnel, mistakes are sometimes made. Some data may be recorded incorrectly or illogically. Such flawed data compromise the accurate estimation of service time and must be removed from the input. The data collected during lock failure conditions are also excluded from the input in this study.

NEURAL NETWORK MODELS

Neural networks are biologically inspired. They are composed of elements that perform in a manner that is analogous to the most elementary functions of the biological neuron. These elements are then organized in a way that may be related to the anatomy of the brain (*3*). The neural networks are also called connective systems or par-

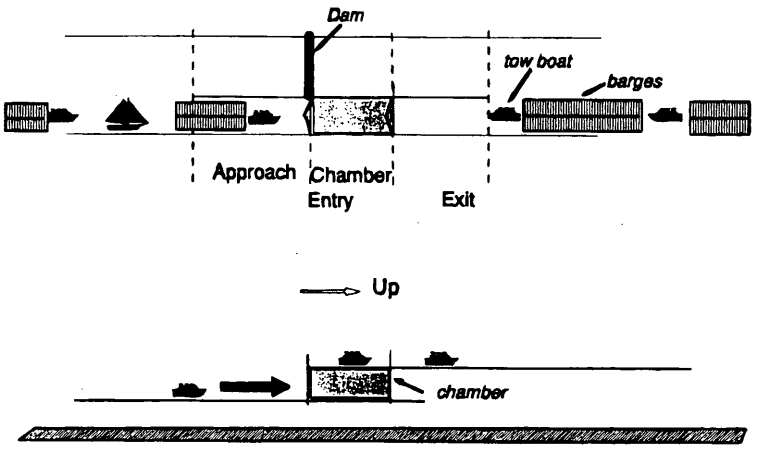


FIGURE 1 Lock system.

allel distributed processors. The many areas addressed by a neural network approach include data compression, recognition, prediction, classification, image processing, decision making, control, and optimization. For the estimation of service time, a backpropagation neural network model with three layers was constructed, as shown in Figure 3.

The neural network can also be defined as an interconnection of neurons such that neuron outputs are connected, through weights (e.g., W_{ij} and V_{ij}), to all other neurons including themselves. Both lag-free and delay connections are allowed (*4*). Figure 3 has one input layer of neurons, one output layer, and one hidden layer between the input and output layers. (This type of network may have more than one hidden layer.) Each of the neurons in a layer is connected to each of the neurons in the next layer. Table 2 defines the variables in Figure 3.

Normalization of Input Data

The weighted sums of inputs are compressed by the activation function into output values between 0 and 1. This study used the unipolar sigmoid activation function expressed in Equation 1 (*4*).

$$f(W'X) = \frac{1}{1 + \exp(-\lambda W'X)} \tag{1}$$

The normalization facilitates error convergence when the models are trained. The six input variables used here were normalized by dividing their actual values by their maximum values. Service times are normalized using the following equation:

$$Z_i' = \frac{Z_i - Z_{\min}}{Z_{\max} - Z_{\min}} \tag{2}$$

where Z_i' = normalized service time and Z_i = service time.

Training the Neural Network

A number of neural network models and training algorithms are currently available. Because of its reliability and its applicability to this study, the backpropagation algorithm that has been widely applied for prediction was chosen (*4*). The following backpropagation training procedure was used for n given training pairs:

Step 1: Select a learning constant (η) and initialize the weight vectors W and V using random numbers.

Step 2: Present the input data and compute the layers' output based on the unipolar activation function.

Step 3: Compute the error value: $E_{k+1} = (d_k - o_k)^2 + E_k$

Step 4: Compute the error signal for the output layer (δ_z) and hidden layer (δ_y):

$$\delta_z = (d_k - o_k)(1 - o_k)o_k \tag{3}$$

$$\delta_z = y_j(1 - y_j)\sum_{k=1}^{K} \delta_z w_{kj} \tag{4}$$

TABLE 1 Summary of Statistics of Tow Service Times at Mississippi Lock 27

Type of chamber	Mean	Standard deviation	Min/Max. service time	P_1†	P_{99}‡	No. of tows
Main chamber	44.218	14.857	8/140	14	82	8090
Auxiliary chamber	26.490	11.749	9/158	12	64.66	3784

† the service time which has 1% probability in the cumulative distribution.

‡ the service time which has 99 % probability in the cumulative distribution

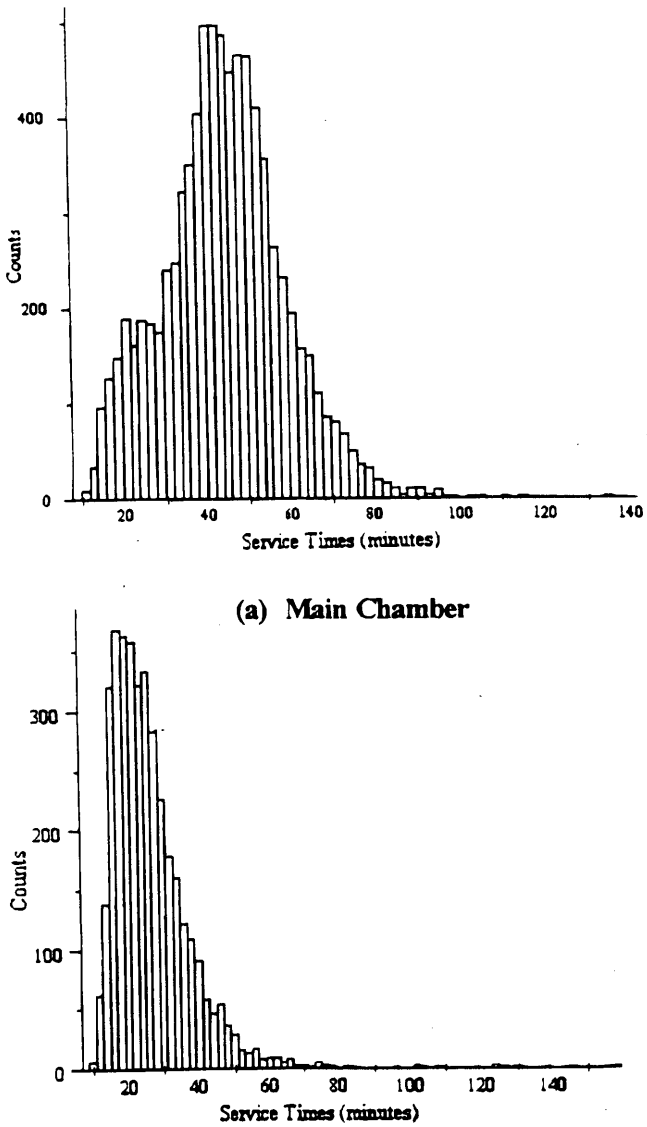


FIGURE 2 Histograms of service times before removing outliers.

Step 5: Adjust output layer weights ($W_{kj} = \eta\delta_z Y$) and hidden layer weights ($V_{ji} = \eta\delta_y Z$) to minimize the error signal.

Step 6: If n pairs of data are all trained, go to Step 7. Otherwise, go to Step 2.

Step 7: If the stopping rule is satisfied, terminate. Otherwise, go to Step 2.

In order to control the learning speed, the algorithm was run using different learning constants (η). The effectiveness and convergence of the error backpropagation learning algorithm depended on the value of learning constant η. In general, however, the optimum value of η depends on the problem being solved. The purpose of the momentum (M) method was to accelerate the convergence of the error backpropagation algorithm (*3*). For best results, different input parameter values were used to train the neural network:

- The number of hidden nodes (H): 3, 4, 5;
- The value of the learning constant (η): 0.4, 0.45, 0.5; and
- The value of momentum (M): 0.2, 0.3, 0.4.

To run the program, the input data were divided into two groups of training data and test data (4090 training data and 4000 testing data for the main chamber, and 1984 training data and 1800 testing data for the auxiliary chamber).

Performance Evaluation

The test data sets, 4000 pairs for the main chamber and 1800 pairs for the auxiliary chamber, were used to verify the trained neural network. Each test data set was evaluated after training the neural network through 50 iterations. The following three types of prediction errors were considered in assessing the neural network model performance:

- Maximum error between actual service time and estimated service time
- Average error between actual service time and estimated service time
- Mean absolute percent error (MAPE):

$$\text{MAPE} = \frac{1}{n}\left(\sum_{i=1}^{n}\frac{|A_i - E_i|}{A_i} * 100\right) \quad (5)$$

where

A_i = actual service time of testing data,
E_i = estimated service time from neural network model, and
n = number of testing data set.

Here, MAPE was mainly used to assess the model's prediction accuracy.

IMPLEMENTATION AND RESULTS

To estimate the lock service time, the neural network model developed in the previous section was applied. The backpropagation algorithm was encoded in the C language, and run on the Unix system. The PMS data were divided into two groups of training data

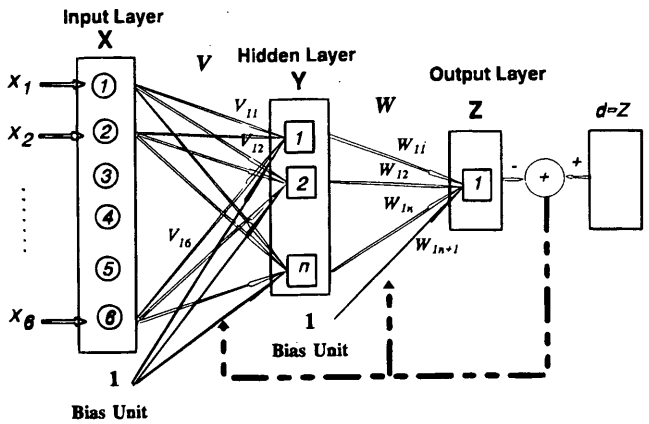


FIGURE 3 Backpropagation neural network.

TABLE 2 Input and Output Variables

Variables	Definition	Ranges
X_1	Tow direction	1 or 2
X_2	Index of tow with same direction (0) or opposite direction (1) from the previous one	0 or 1
X_3	Number of cuts in a tow	1 - 3 cut
X_4	Number of barges in a tow	0† - 23 barges
X_5	Ratio between tow length and chamber length	0† - 3‡
X_6	Ratio between tow width and chamber width	0† - 2‡
Y_i	Hidden layer output	-
V_{ij}	Weight matrix for hidden layer	-
W_{ij}	Weight matrix for output layer	-
Z	Service time, output layer	8 - 158 min

† Recreational boats have zero values.
‡ If X_5 or X_6 are greater than 1.0, tow must be divided into cuts to fit into lock chambers.

and test data. The neural networks were trained with 4090 input data for the main chamber and 1984 input data for the auxiliary chamber at each iteration. All experiments were limited to a maximum of 1000 iterations. At every 50 iterations, the service time was estimated with testing data based on the trained neural network models. The experiments were performed for every combination of parameter values, for a total of 27 experiments (3 types of hidden node × 3 learning rate values × 3 momentum values). The test solution with the best MAPE was saved from these experiments. The estimation results obtained with neural network models show that the MAPE of training data usually converges to one value. However, in some experiments the MAPE fluctuates. A possible reason is the inappropriate choice of values for such parameters as learning rate (η) or momentum (M). Table 3 shows the initial best MAPE solution in the specific experiments.

Data Manipulation

As described previously, some data might have been recorded incorrectly or illogically, since data were collected by humans. These data hinder accurate estimation of service time and should be removed if they can be properly detected. Barnett and Lewis (*5*) define an outlier as "an observation which appears to be inconsistent with the remainder of that set of data" and explain the relationships between the extreme observations, outliers, and contaminants. An outlier can also be defined as an extreme observation that has errors that are considerably larger in absolute value than the others, about 3 or 4 standard deviations from the mean (*6*). In order to detect outliers, the deviations between actual and estimated service times were computed. These deviations and the outliers with deviations beyond 3 σ (3 standard deviations) are summarized in Table 4.

The summary shows that the mean values of deviations for both chambers are negative (−2.03 and −0.56), which means that service times are slightly overestimated. Bell-shaped histograms of deviations between actual and estimated service times for both chambers are shown in Figure 4. The error analysis detected 81 outliers for the main chamber and 59 for the auxiliary chamber. Histograms of service times for both chambers after removing outliers are shown in Figure 5. The service time for the cleaned data sets was then reestimated with same procedure initially used in service-time estimation. The results are shown in Table 5.

Without outliers, the new MAPEs for the main chamber are about 16.8 percent lower and for the auxiliary chamber about 16.1 percent lower than those in Table 3. The main chamber and auxiliary chamber have their best solutions when the numbers of hidden nodes are 5 and 4, learning rates are 0.4 and 0.5, and momentum values are 0.3 and 0.4, respectively. The auxiliary chamber shows a higher MAPE, largely because the service times at the auxiliary chamber are more variable than at the main chamber.

TABLE 3 Performance Value of Neural Network Without Removing Outlier Data

	Maximum absolute error (minutes)	Average absolute or (minutes)	MAPE (%)
Main chamber	61.996	7.855	21.049
Auxiliary chamber	77.696	5.872	23.461

TABLE 4 Summary Statistics of Deviation

	Mean	Standard Deviation (SD)	Total number of data	Number of outliers
Main Chamber	-2.031	10.928	8090	81
Auxiliary chamber	-0.556	8.109	3784	59

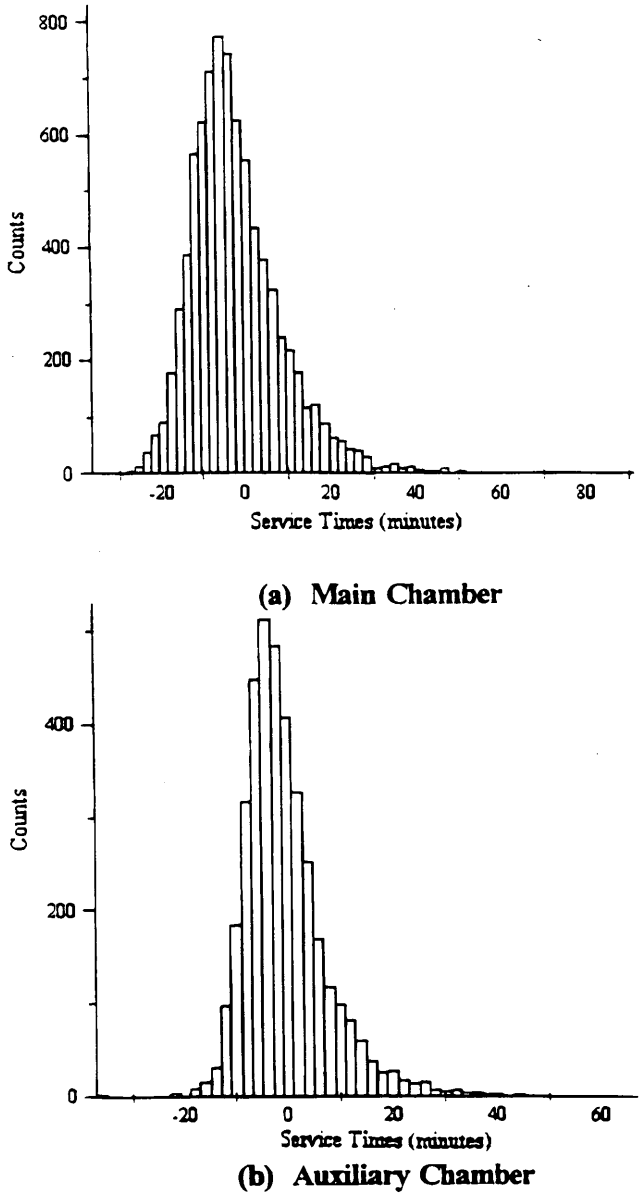


FIGURE 4 Histograms of deviations between actual and estimated service times.

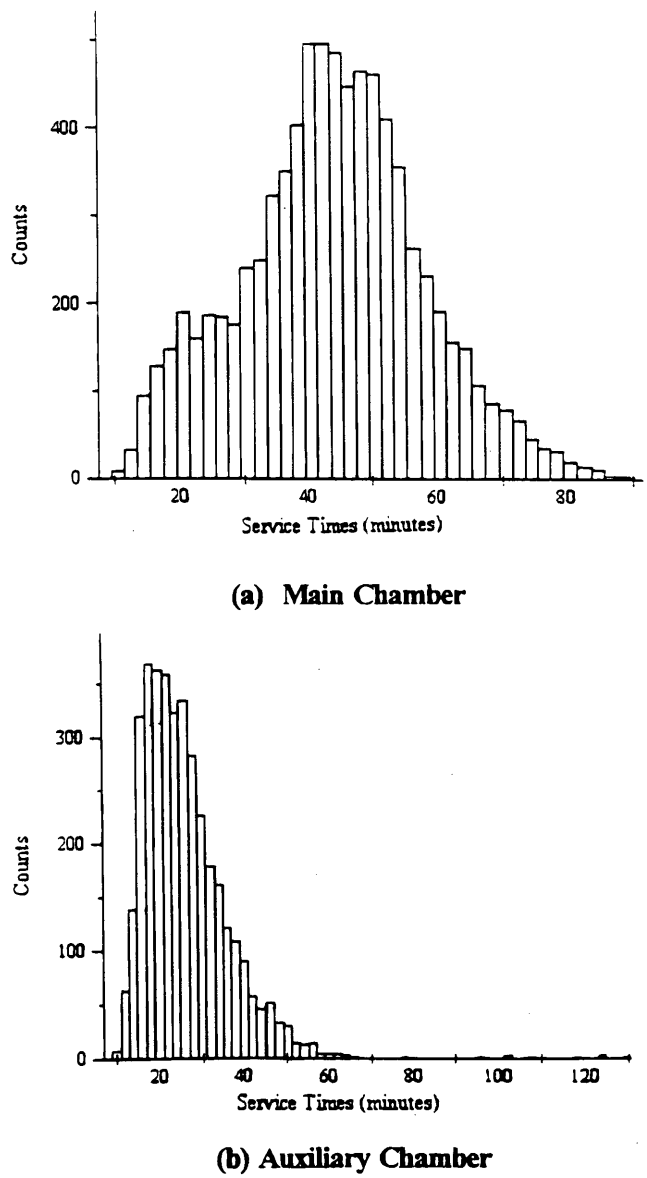


FIGURE 5 Histograms of service times after removing outliers.

TABLE 5 Performance Value of Neural Network After Removing Outlier Data

	Maximum absolute error (minutes)	Average absolute error (minutes)	MAPE (%)	% MAPE improvement (%)
Main chamber	37.059	7.846	17.516	16.8
Auxiliary chamber	27.003	5.502	19.683	16.1

Multiple Regression Model

For a comparative assessment of prediction accuracy, multiple regression models of service times were also developed. Model I is a linear function and Model II is a nonlinear function that can be transformed to linear form by taking the logarithms of both sides.

- Model I:

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6$$

- Model II:

$$Y = a^*\exp(b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6)$$

Table 6 shows the results of multiple regression analysis based on the same training data sets. Both Model I and Model II have lower *R*-squared values. Because the *R*-squared values are related to linear models, they were not used as performance measures. Instead, the MAPE was used for a comparative assessment.

Overall, MAPEs are considerably lower (by up to 30.9 percent) for the neural network models than for the regression models. A possible reason for the superior neural network performance is the ability to search for any linear or nonlinear relation without explicitly defining that relation or specifying its properties.

ESTIMATION OF COMBINED SERVICE TIME FOR TWO-CHAMBER LOCK

In earlier sections, separate neural network models were developed to separately estimate the service times for main and auxiliary chambers. There are, however, some practical applications in which it is not known in advance which tows will use which chamber. To allow such applications, a combined service-time model was developed for a two-chamber lock (Mississippi Lock 27).

Combined Input Data

Previously, six variables were used as inputs. The combined service-time estimation models used the same variables except for the ratio between tow and chamber length, which was replaced by tow length. (That ratio is not known until a chamber is selected.) Thus, the six input variables were tow direction, index of same direction, number of cuts, number of barges, tow length, and ratio between tow width and chamber width. The two separate data files for the chambers were combined into one input file with 12,160 tows based on 1988 PMS data at Mississippi Lock 27.

Figure 6 shows the cumulative distribution and histogram of actual combined service times. The mean actual combined service time is 38.93 min and the standard deviation is 16.39. The combined input data were trained using Neuroshell 2 software (*7*).

Training the Neural Network

Backpropagation networks are known for their ability to generalize well on a wide variety of prediction problems. Backpropagation networks are a supervised type of network, that is, trained with both inputs and outputs. Three different types of backpropagation networks, standard connection, jump connection, and recurrent, were used to train the input and output data. To find the best combined service-time estimation model, the following neural network models were selected for training the input and output.

- COM271: three-layer standard connection backpropagation network; that is, every layer is connected or linked to the previous layer.
- COM272: four-layer standard connection backpropagation network.
- COM273: five-layer standard connection backpropagation network.
- COM274: three-layer jump connection backpropagation network; that is, every layer is connected or linked to every previous layer.

TABLE 6 Results of Multiple Regression

		R-square	MAPE
Model I	Main chamber	0.4801	25.34 %
	Auxiliary chamber	0.4151	26.01 %
Model II	Main chamber	0.5811	24.75 %
	Auxiliary chamber	0.4071	27.86 %

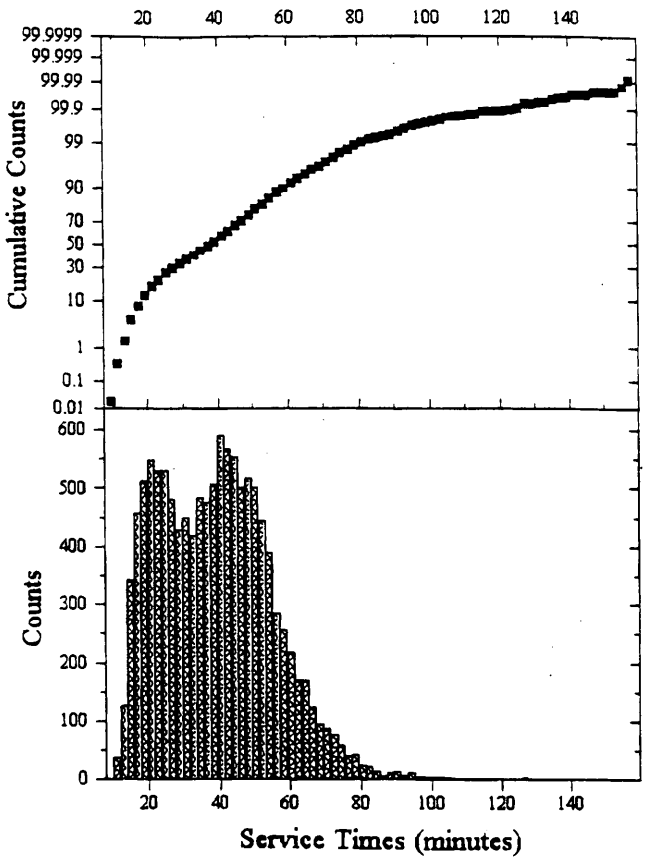


FIGURE 6 Cumulative probability distribution and histogram of actual service times for combined chambers.

- COM275: four-layer jump connection backpropagation network.
- COM276: three-layer recurrent backpropagation network with dampened feedback.
- COM277: General regression neural networks (GRNNs). There are no training parameters such as learning rate and momentum for these networks, but a smoothing factor determines how tightly the network matches its predictions to the data in the training patterns.

It should be noted that a three-layer network has one hidden layer and a four-layer network has two hidden layers.

Results for Combined Service-Time Models

Each model was trained until stopping rules were satisfied. Training stopped when either the average error was below a predefined level, or when 50,000 events occurred without improvement in the minimum average error. The values of the learning constant and momentum were updated from 0.1 to 0.5 by 0.1 increments at every iteration. The weight vectors were also updated to minimize the error between actual values and estimated values. The best test set was saved every time it reached a new minimum average error. Combined service times were estimated for each model at Mississippi Lock 27 based on the saved best test set. Table 7 shows the summary of statistics for combined estimation models.

The means and standard deviations were calculated from the best test set. As shown in the table, the COM277 (GRNN) model has the lowest MAPE. Figure 7 shows the cumulative probability distribution and histogram of service times estimated by the COM277 model, which has a tendency to estimate the service time as two values of 22 min and 48 min.

SUMMARY AND CONCLUSIONS

This study has statistically analyzed lock service times, developed neural network models for service-time estimation, and comparatively assessed neural network models and regression models. First, the statistical analysis of lock service times shows that the main chamber has a mean service time of 44.218 min and a standard deviation of 14.857 min. The auxiliary chamber has a mean service time of 26.490 min and a standard deviation of 11.749 min. Both distributions are skewed to the left. The maximum deviations of service times from the mean are 6.4 σ for the main chamber and 11.2 σ for the auxiliary chamber.

Second, neural network models for estimating service times were developed separately for the main and auxiliary chambers at Mississippi Lock 27. The estimation was performed with six input variables and one output variable based on 1988 PMS data. The MAPEs are 21.05 percent for the main chamber and 23.46 percent for the auxiliary chamber. After removing the outliers (beyond 3 σ), the MAPEs decreased by 17.52 percent for the main chamber and 19.68 percent for the auxiliary chamber. For a comparative assessment of prediction accuracy, two multiple regression models were developed and the lock service times were estimated. The MAPEs of regression models range from 24.75 percent to 27.86 percent. Comparisons between these neural network models and regression mod-

TABLE 7 Summary of Estimation Statistics

ANN network types	Mean (min)	Standard deviation	MAPE(%)
Actual service time	38.93	16.39	
COM271	38.925	12.580	21.354
COM272	44.424	21.784	30.036
COM273	38.864	12.492	21.280
COM274	39.672	12.217	22.519
COM275	39.498	17.411	26.018
COM276	39.121	11.361	22.915
COM277	38.915	12.680	21.040

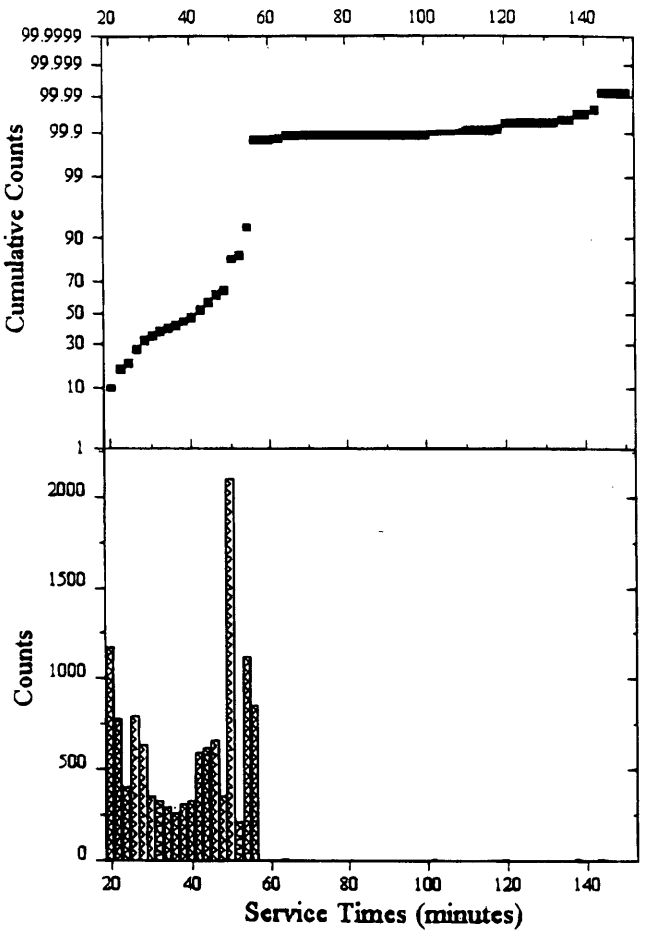


FIGURE 7 Cumulative probability distribution and histogram of estimated service times for combined chambers.

els show that the MAPEs are considerably lower (by about 24.3 percent for the main chamber and 29.2 percent for the auxiliary chamber) for the neural network models.

Third, combined service-time models for locks with dissimilar chambers were developed based on six input variables and one output variable. The results show that the combined actual service times have a mean of 38.932 min and a standard deviation of 16.389. The best combined service-time estimation model (COM277) has a mean of 38.915 min and a standard deviation of 12.680. The MAPE of the best set is 21.039 percent. This combined service-time estimation model can estimate the lock service time without unreasonably compromising accuracy, even before knowing which tows will use which chamber.

Based on these results, the prediction accuracy of neural network models is considerably better than for the regression models considered. Neural network models clearly have considerable potential for improving lock service-time estimation.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the funding from the Institute for Waterway Resources of the Corps of Engineers and the advice received from Dr. L. George Antle in support of this work.

REFERENCES

1. Chang, C. *Models for Estimating Lock Service Times at Waterway Locks.* Transportation Study Center Working Paper 92–25. University of Maryland, College Park, 1992.
2. Dai, D. M., and P. Schonfeld. Simulation of Waterway Transportation Reliability. In *Transportation Research Record 1313,* TRB, National Research Council, Washington, D.C., 1989.
3. Wasserman, P. D. *Neural Computing: Theory and Practice.* Van Nostrand Reinhold, New York, N.Y., 1989.
4. Zurada, J. M. *Introduction to Artificial Neural Systems.* West Publishing, St. Paul, Minn., 1992.
5. Barnett, V., and T. Lewis. *Outliers in Statistical Data.* John Wiley & Sons, New York, N.Y., 1984.
6. Montgomery, D. C., and E. A. Peck. *Introduction to Linear Regression Analysis.* John Wiley & Sons, New York, N.Y., 1982.
7. *NeuroShell 2 User's Manual.* Ward Systems Group, Frederick, Md., 1993.

Publication of this paper sponsored by Committee on Artificial Intelligence.

Modeling Schedule Deviations of Buses Using Automatic Vehicle-Location Data and Artificial Neural Networks

RAVI KALAPUTAPU AND MICHAEL J. DEMETSKY

The establishment of the Advanced Public Transportation Systems program has encouraged bus transit operators to experiment with implementing automatic vehicle-location systems for real-time monitoring and supervision of operations. While the focus has primarily been on the implementation of technologies, such as automatic vehicle-location systems, it is necessary to experiment and develop advanced performance analysis and evaluation procedures that can assist in schedule planning and real-time service-control tasks. One potentially useful and effective approach to these tasks is system behavior modeling. In this study this method is used to model schedule behavior of buses on a route using schedule-deviation information. The primary objective of this study is to investigate the application of artificial neural networks, which have been shown to hold promise when applied to nonlinear dynamic system-modeling problems, for developing schedule behavior models. Models are developed using the schedule-deviation information obtained from Tidewater Regional Transit's automatic vehicle-location system. The time-series analysis approach is adopted for the development of schedule behavior models at the route level. The results of a case study are encouraging and demonstrate the usefulness of artificial neural network techniques, especially the Jordan networks and the Elman networks, for modeling schedule deviations of buses on a route.

In recent years, bus transit operators have been testing and implementing automatic vehicle location (AVL) systems for real-time monitoring and supervision of operations. However, implementation of technologies such as AVL systems needs to be complemented by the development of advanced performance analysis and evaluation procedures for assisting in operational planning, management, and real-time service-control tasks. While real-time monitoring provides useful information on bus transit operations, advanced analysis and evaluation procedures such as system behavior models can be useful for schedule planning and design of real-time service-control strategies.

The central idea of this study is that a schedule behavior model can provide an understanding of the past system behavior of buses on a route. Such a model has several potential uses. It can be used for prediction of schedule deviations at a downstream stop based on current and past schedule deviations of buses at timepoints in the upstream section. The predictive model can assist in the design, development, and real-time implementation of service-control strategies. Also, the schedule behavior models can be used for updating and modifying schedule plans. The models can be used to speed up and automate performance analysis and evaluation of service-control strategies. Currently there are no automated procedures available to evaluate the effect of implementing service-control strategies. The models can be integrated into an automated decision-support system to assist dispatchers and supervisors in real-time decision making on schedule and headway adjustments to improve service reliability.

In real-time operating conditions, the time required to make decisions based on graphical display codes and other features of a number of buses is not sufficient to make reasonable decisions regarding schedule or headway changes. Dispatchers and supervisors have to make quick decisions based on the information presented graphically on the screen. However, they must monitor several buses simultaneously, leading to information overload. Hence there is a need to develop computer-based analysis and decision-support tools to model the system's performance and use it to predict the future schedule behavior of a bus on a route. The key purpose of this study is to introduce the concept of schedule behavior modeling as a performance analysis tool for bus transit operations. The primary objective is to investigate the development of schedule behavior models using historical AVL data and artificial neural network (ANN) techniques. In this paper, system behavior is referred to as schedule behavior of buses and is used to denote the key performance indicator, schedule deviation, which is the difference between the actual arrival times computed from the location information and the scheduled arrival times of a bus at a timepoint on a specific route.

ANN modeling techniques have been of great interest to many researchers. These techniques have certain advantages such as not requiring to assume a priori the nature of the relationship between the dependent and independent variables. The modeling approach using neural networks performs two important tasks. First, the model learns the system performance using past and current AVL data. Secondly, the ANN models can be used for predicting the behavior of the buses. Such a system behavioral modeling approach has been successfully used in other dynamic-system performance analysis and control problems (*1,2,3*). The literature reviewed indicated that ANNs have the potential to capture the dynamic and interactive effects of schedule deviations of buses on a route network. In addition, they are able to capture the trend in a time series, especially when the relationship is nonlinear.

The basic approach adopted for ANN modeling of the performance of a bus transit system was to develop separate models for the different routes instead of one complete model for the entire transit route network. By using this approach the ANN modeling process becomes simpler and the training process is perhaps faster because of its reduced complexity: there is a smaller domain space to learn for one route, compared to learning all the routes in the transit network. In addition, such a modeling approach is appropriate and justified by the different physical, traffic, and environmental characteristics of the various routes. Modeling at the route level can

Department of Civil Engineering and Applied Mechanics, Thornton Hall, University of Virginia, Charlottesville, Va. 22903.

help reduce the complexity of the modeling process and simplify and ease the implementation of the service-control strategies. In addition, such a modeling approach can help reduce the time required for system identification, and subsequent selection and implementation of a service-restoration strategy.

A number of ANN architectures and learning algorithms have been proposed and investigated for various problems. Since the primary objective of this study is to illustrate the applicability of the ANN approach for the problem of bus schedule behavior modeling, we discuss only the advantages of ANNs and the applicability of alternative strategies to developing ANN models for this problem. The fundamental concepts of ANNs are discussed in detail in the vast collection of relevant literature (*4,5,6,7,8,9*).

ARTIFICIAL NEURAL NETWORKS

ANNs are a type of learning system that has gained some prominence in the last decade because they can be trained to identify, classify, and predict nonlinear patterns and can solve complex problems much faster than traditional techniques. ANNs are a paradigm for intelligent processing of information for some specific objective such as classification, pattern recognition, decision-making, system behavior identification, and prediction. ANNs have a highly distributed parallel structure and when combined with powerful digital hardware technology can make model simulations economically and with relative ease. ANNs mimic human learning processes and therefore hold great potential as adaptive learning systems. ANNs can handle complex and nonlinear relationships that are common to dynamic systems like bus transit operations. In the case of nonlinear systems, ANNs have the distinct advantage over a standard regression method of not having to know the form of the function a priori. Unlike other mathematical techniques, ANN models' learning can be continuous, so that they can automatically adapt to the changing characteristics of the operating environment of buses. What this implies is that a base ANN model can be developed using historical AVL data and this base model can be updated and modified using new online data. The potential advantage of an ANN learning method is that, unlike mathematical simulation models, ANNs can be trained using observed data only, without requiring any knowledge of the internal structure of the system or of modeling techniques (*10*). This ability to approximate unknown functions through the presentation of past states of a system makes ANNs a useful modeling tool in engineering applications, such as bus transit schedule behavior modeling.

Lapedes and Farber (*1*) reported that simple neural networks can outperform conventional methods. Sharda and Patil (*11*) concluded from their work on 75 different time series that the simple neural network model could forecast about as well as the Box-Jenkins forecasting technique. Tang et al. (*12*) in their comparative study of the performance of ANNs and conventional statistical techniques concluded that for short-term memory series, ANNs appear to be superior to the Box-Jenkins model. A review of relevant literature indicated that each of the methods performed better than the other about half of the time.

In this study the focus is on using three different ANN architectures, namely feedforward networks with input windows, Jordan nets, and Elman nets. Jordan and Elman nets are two types of partial recurrent neural networks. These three network architectures have been feasible for modeling a number of engineering problems, such as system behavior identification and prediction and time-series modeling, among others (*1,2,3,12,13*). Hence our initial efforts are aimed at investigating these three ANN architectures.

The main distinction between the feedforward and the partial recurrent nets is in the network topology. The two types of partial recurrent nets (Jordan and Elman nets) have memory layers in addition to the basic architecture of a feedforward network. These network architectures are discussed briefly in the next two sections.

Feedforward Networks

Feedforward networks are the most commonly used network architectures for neural network modeling. Depending on the representation scheme, feedforward networks can be different types. Figure 1 illustrates the schematic architecture of a feedforward network with an input window. The most basic approach for handling time series is using an input window that holds a restricted part of the time series. This type of feedforward network seems appropriate for our modeling problem. A feedforward network with an input window has been shown to be superior to a simple feedforward network (*2,13,14*). The input window provides the network with information on previous states in the form of units in the input layer. This allows it to incorporate knowledge about previous states or past values of a time series. Therefore such an architecture is suitable for modeling spatiotemporal sequencing problems such as bus schedule behavior.

Partial Recurrent Neural Networks

A second way that a neural network can model and predict a time series is to incorporate an internal state that enables it to learn the relationship of an indefinitely large set of past inputs to future states. This is achieved via recurrent connections, and such a network is known as a recurrent network. If the recurrent networks are updated like feedforward networks (with a single update per time step) they are known as partial recurrent networks (*5*).

Partial recurrent networks have been suggested and proven to be applicable by many researchers (*7,8*) for dynamic problems involving temporal sequencing. The problem of bus schedule behavior prediction can be considered a spatiotemporal problem. The sched-

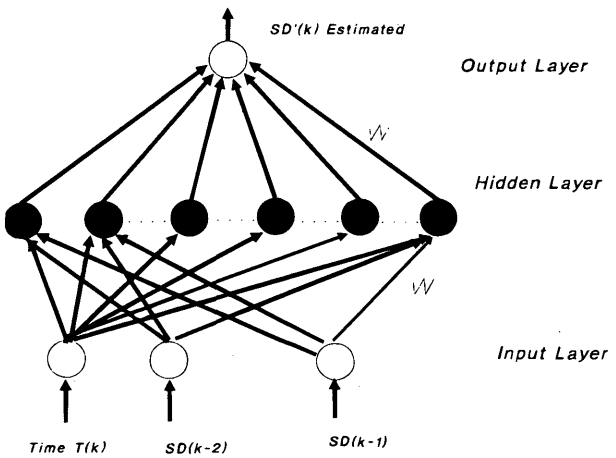


FIGURE 1 Architecture for a feedforward network with input windows.

ule deviation at a point in time is affected by the schedule deviation at previous timepoint(s). The spatiotemporal sequencing of the schedule-deviation information can be modeled and investigated for the purpose of predicting the schedule deviations at a timepoint downstream in the route network. This sequential information, regarded as short-term memory of the system's performance, can be an effective approach for developing an intelligent model of the bus transit schedule behavior. Partial recurrent networks, through their architecture, have the ability to store and use information about the previous state, and therefore are appropriate for the problem of bus schedule behavior modeling.

Jordan Networks

Jordan networks (*7*) are a type of partial recurrent neural network. Figure 2 illustrates the basic architecture for a Jordan network. The network has the following features:

- The input layer is fully connected to the hidden layer, and the hidden layer is fully connected to the output layer.
- Output units are connected to context units by recurrent one-to-one connections. Every context unit is connected to itself and also to every hidden layer unit.
- The number of context units is equal to the number of output units.

A partial recurrent network has an input consisting of two components. The first component is the pattern vector, which is also the only input to the partial recurrent network. The second component, the state vector, is given through the next-state function in every step. In this manner the behavior of a partial recurrent network can be simulated with a feedforward network that receives the state not implicitly through recurrent links, but as an explicit part of the input vector (*7*). These networks are regarded as having memory, as the recurrent connections allow the network's hidden units to see its own previous output. Therefore, behavior can be shaped by previous responses. This network memory concept can be used to model the schedule behavior of buses. The knowledge of schedule deviation of a bus at the previous timepoint or stop can be useful for developing a model of the system for eventual use as a prediction tool. The adoption of such a structure to the ANN model is appropriate for the bus schedule behavior problem because the schedule deviation at a timepoint has a strong relationship to the schedule deviations at the previous timepoints. The extent of previous timepoints that should be considered is yet to be researched. The approach of this study is to take advantage of these features of Jordan nets and investigate their applicability to schedule behavior modeling.

Elman Recursive Networks

An Elman recursive network is a type of partial recurrent network that is also commonly used for learning to recognize and generate sequences of inputs. The Elman net, in addition to the basic topology of a single-hidden-layer feedforward network, has a set of additional units at the input level that are referred to as context units. These context units are responsible for the dynamic behavior of the network. A typical architecture of an Elman recursive network is illustrated in Figure 3. The number of context units is equal to the number of hidden units. After each time step, the output values of the hidden units are copied to the context units. The context units thus provide the network with memory of the previous state through implicit representation in the internal state of the network (*8*). The important distinction between Jordan nets and Elman nets has to do with where the context units are present. The two networks are both essentially memory models, but they differ in whether they have the previous state's inputs or outputs in the memory.

SCHEDULE BEHAVIOR MODELS USING ANNS

Modeling Approach

In prediction modeling there are two basic approaches that have gained prominence and are often adopted. With the fundamental approach, it is believed that the forecasting process should at least approximately model the mechanisms that underlie the determination of the key variable being predicted (*13*). The key factors that affect schedule behavior and cause schedule deviations are:

$$SD\,(R_i, j, k, T) = \phi\,(\text{Traffic, Driver, Vehicle, Environment, Loading and Unloading}) \quad (1)$$

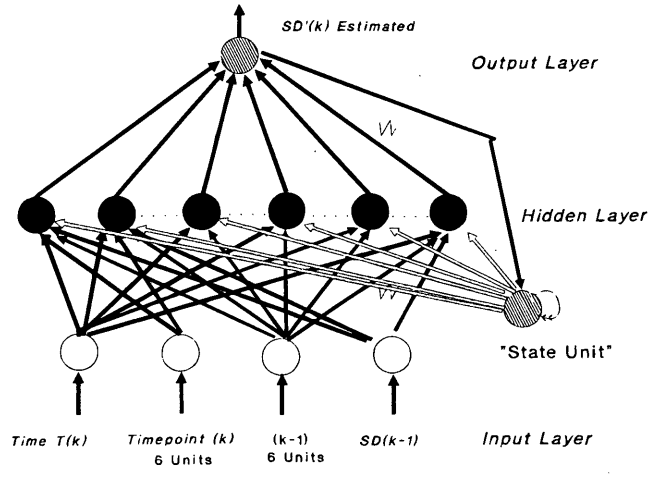


FIGURE 2 Architecture of a Jordan network.

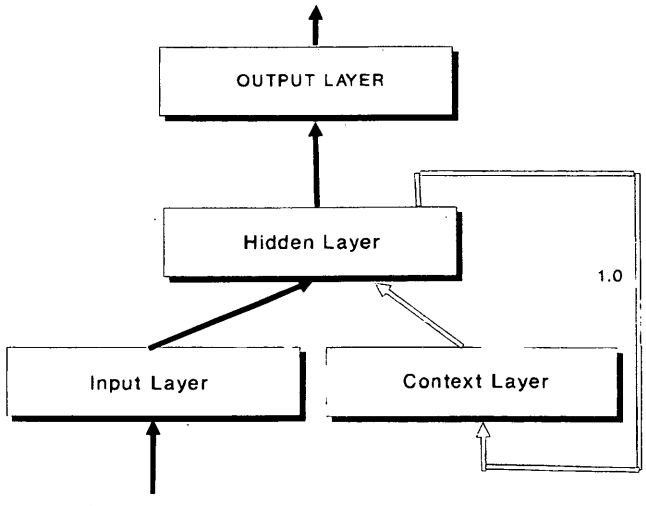


FIGURE 3 Architecture of an Elman network.

where

SD = Schedule Deviation,
R_i = Route i,
j = Direction,
k = Timepoint,
T = Scheduled Arrival Time, and
ϕ represents an unknown function that the network would try to ascertain during the training process.

System behavior modeling on this approach is currently not feasible due to lack of adequate information in the data set on many of the above factors that affect schedule behavior. For example, no information is collected on loading and unloading characteristics at each timepoint on a given route. The second modeling approach is to assume that all the available information (on key factors affecting schedule behavior) has already been represented by the values of the key variable being predicted (*13*). For example, with schedule deviation prediction, the values that indicate "early or late" have been influenced by the various factors that affect it, namely traffic conditions, driver characteristics, passenger loading and unloading characteristics, and vehicle condition. Therefore nothing else is considered while trying to predict the future of the system behavior except the past states of the key prediction variable, the schedule deviation. Hence a time-series approach is adopted that is mathematically represented as follows:

$$SD(k) = \phi\,[SD(k-1), SD(k-2), \ldots, SD(k-n)] \qquad (2)$$

where $SD(k)$ denotes the schedule deviation at timepoint k on a specific route and in a specific direction of travel. The term n represents the length of the input time series, or in other words, the short-term memory about the schedule deviations of a bus at timepoints in the upstream part of a route (k − 1, k − 2, . . . , etc.).

The focus of this study is on developing ANN models for one particular scenario, that is, given a particular route and direction of travel. Two different ANN model sets, depending on the length of short-term memory about the time series ($n = 1$ and $n = 2$) provided, are investigated. For Model Set I, the schedule deviation at the previous timepoint [$SD(k-1)$] is provided, while in the Model Set II two previous schedule deviation values [$SD(k-2)$, $SD(k-1)$] are provided. For each of these model sets, two different cases (Case A and Case B) are examined. The difference between the two cases is that in case B, the spatial information about the timepoints (k, $k-1$, $k-2$) are also provided to the network as inputs. The distinction between the two cases is illustrated in the input layer of Figures 1 and 2. This was done in order to investigate the effect of providing information about the spatial location of the buses on the route to the ANNs. The two sets of models are briefly described in the following section.

Model Set I: Using Short Input Series of Length n = 1

Case A

Input units: Schedule Arrival Time $T(k)$, Schedule Deviation $SD\,(k-1)$;
Output unit: Schedule Deviation $SD(k)$.
$SD(k) = \phi\,[SD(k-1)]$

Case B

Input units: Scheduled Arrival Time $T(k)$, Timepoint k, Timepoint $k-1$, Schedule Deviation $SD(k-1)$;
Output unit: Schedule Deviation $SD(k)$.
$SD(TP_j) = \phi\,[SD(k-1), k, k-1]$

Model Set II: Using Short Input Series of Length n = 2

Case A

Input units: Scheduled Arrival Time $T(k)$ Schedule Deviation $SD(k-2)$, Schedule Deviation $SD(k-1)$
Output unit: Schedule Deviation $SD(k)$.
$SD(k) = \phi\,[SD(k-1)]$

Case B

Input units: Scheduled Arrival Time $T(k)$, Timepoint k, Timepoint $k-1$, Timepoint $k-2$ Schedule Deviation $SD(k-2)$, Schedule Deviation $SD(k-1)$;
Output unit: Schedule Deviation $SD(k)$.
$SD(k) = \phi\,[SD(k-1), k, k-1]$

CASE STUDY: A SAMPLE ROUTE FROM TIDEWATER REGIONAL TRANSIT

Data Collection

In order to examine the concept of system behavior models and to investigate the application of neural networks to their development, real data from Tidewater Regional Transit's (TRT's) AVL system was obtained. A sample route (Rt. 23) was chosen for this study. The raw data stored in the form of binary files in the TRT's VAX system was converted into ASCII format. The history data files were preprocessed to extract only the desired information for developing ANN models. AVL information comprising 26 weekday (Monday through Friday) data was considered for modeling purposes. The focus is limited to weekday operations since insufficient weekend (Saturday and Sunday) data was collected.

Modeling Process

The ANN models were developed using the following procedure.

Step 1: Data preprocessing
Step 2: Network selection
Step 3: Learning algorithm and update function selection
Step 4: Weights initialization
Step 5: Network training
Step 6: Network testing and performance evaluation

These steps are discussed in detail in the following sections.

Data Preprocessing

Data preprocessing is the critical step in ANN modeling. In this case it covered about half of the modeling process. Data preprocessing

involved two important steps: elimination of outliers or noise, and data scaling. Noise elimination involved removing "outliers" or absurd values of schedule deviation at that specific timepoint and replacing them with the value of schedule deviation from the timepoint immediately preceding it. The data was normalized using minimum and maximum values of the variables over the entire data set. The scaling of these two variables was accomplished using the following expression:

$$Xnorm = \left(2.0^* \frac{X}{MAX - MIN}\right) + \left[\left(-2.0^* \frac{MIN}{MAX - MIN}\right) - 1.0\right] \quad (3)$$

where X is the variable to be normalized, and MAX and MIN denote the maximum and minimum values of variable X in the data set.

In this study, for the scheduled arrival time (T) variable, MAX = 1440 min and MIN = 300 min. The scaling using the above expression, converts the data into the [−1, 1] interval. It is important to set the scaling so that the units do not affect the net's output (that is, the inputs should be either unitless ratios or else chosen so that percentage changes are the same across monotonic transformations of input values). Having most or all inputs scaled identically to the output function can speed convergence. Normalization of the output data to the [−1, 1] region prevents the propagation of large error signals during training, which could force the middle-layer nodes to saturate and become insensitive to training. The output variable, schedule deviation, was also normalized using the expression given in Equation 3 and the corresponding schedule-deviation values. The timepoint data was also transformed into a binary vector. There were six timepoints located on the route being studied. Therefore, a vector of length 6 was considered and the timepoints were transformed. For example, timepoint k = 1 was binarized as [1 0 0 0 0 0]. The data set consisting of 26 weekday AVL data was divided into three sets: one a training set consisting of 24 days of data, and two test sets consisting of one day's data each.

Network Architectures

As discussed earlier, three basic neural network architectures were examined in this study: feedforward networks with an input window Elman recurrent networks, and Jordan recurrent networks. All three types of networks had one hidden layer. The network features used in this study are given below.

Model Set I

Case A: The networks had two inputs. The input consisted of the scheduled arrival time $T(k)$ and an input window representing schedule deviation SD at the timepoint $k - 1$ immediately preceding the current timepoint k location. All the networks for this case had five hidden units and one output unit.

Case B: The networks, in addition to the inputs discussed in Case A, had 12 units representing the current timepoint location k and the previous timepoint location $k - 1$. The networks had a total of 14 input units. All the networks had 20 hidden units and 1 output unit.

Model Set II

Case A: Input Units: 3, Hidden Units: 6

Case B: Input Units: 21. Eighteen units correspond to timepoint location, 1 unit to scheduled arrival time $T(k)$, and 2 units to input windows for schedule deviations.

Hidden Units: 21.

ANN architectures are denoted as $I \times H \times O$, where I, H, and O represent number of input, hidden, and output units, respectively.

Learning Algorithm and Update Functions

Since both the input (time and location, etc.) and output (schedule deviation) variables were known quantities, the schedule behavior modeling using ANNs constituted a supervised learning problem; hence supervised learning algorithms such as Quickprop were useful. QuickProp, which was developed by Fahlman (*9*), is a faster and more efficient version of the standard backpropagation algorithm.

Weight Initialization

The weights were initialized depending on the type of network architecture selected. The weights for the connections were randomly chosen between −0.001 and +0.001 for a feedforward network.

Network Training

The networks were trained with the QuickProp learning algorithm until there was no substantial decrease in the mean square error (MSE) for every 1000 iterations. The TanH (hyperbolic tangent) activation function was used for the hidden units. Both the MSE and sum of square errors (SSE) were computed for each iteration of the training process. MSE was used as a stopping criterion during the training phase.

Network Testing and Performance Evaluation

The networks were tested on the two test data sets, and the MSE and SSE were computed. The network performance was evaluated using average percentage error (PE_{avg}) to check the accuracy of the trained ANN models on the test data sets. The percentage error PE_{avg} was calculated for each point in the test data set (having n patterns) using the following expression:

$$PE_{avg} = \sum_{i=1}^{n} \frac{(SDact_i - SDpred_i)}{SDact_i} \times 100 \quad (4)$$

where

$SDact$ is the actual schedule deviation;

$SDpred$ is the network predicted schedule deviation; and

PE_{avg} is used to justify the accuracy and validity of the ANN models.

DISCUSSION OF RESULTS

The performance results of various ANNs are summarized in Table 1. The results indicate that for Case A, the average percentage error PE_{avg} was 3.5 to 6.30 points lower for Model Set II than for Model

TABLE 1 Comparison of Predictive Performance of Various Neural Network Models

ANN NETWORKS	Mean Square Error, MSE	Average % Error (PE_{avg})
	TEST Data	TEST Data
MODEL SET I		
Case A : 2x5x1 NETS		
Feedforward Net	0.00279	28.19
Elman Net	0.00412	24.95
Jordan Net	0.00219	27.65
Case B : 14x20x1 NETS		
Feedforward Net	0.00421	26.55
Elman Net	0.00720	26.97
Jordan Net	0.00403	28.38
MODEL SET II		
Case A: 3x6x1 NETS		
Feedforward Net	0.00232	24.25
Elman Net	0.00416	21.27
Jordan Net	0.00203	21.38
Case B: 21x21x1 NETS		
Feedforward Net	0.00479	24.57
Elman Net	0.00663	25.60
Jordan Net	0.00476	25.00

Set I. This leads to the conclusion that increasing the input series from $n = 1$ to $n = 2$ results in lower error values and more accurate models. Thus, providing the networks with longer input time series (for the example route, $SD(k - 6)$, $SD(k - 5)$, $SD(k - 4)$, . . . , $SD(k - 1)$) leads to improved results. More inputs will provide more information, and are thus likely to provide more accurate results. It is interesting to note that providing additional information on the spatial location (timepoints) did not improve the accuracy. In the case of Model Set I, the PE_{avg} for Case B was higher than for Case A, for both the Elman and Jordan nets. This can be attributed to the increase in number of inputs that resulted in an increase in the complexity of the network and causes a higher MSE for the same number of training iterations. The same error behavior was also observed for Model Set II. In addition, the training times for Case B models were significantly higher (nearly 1.5 to 2 times) than those of Case A models, especially for Model Set II. Therefore it was concluded that there is no distinct advantage in including the spatial location information for schedule behavior modeling. The overall accuracy of the models ranged from 71 to 78 percent. Since no previous work on schedule behavior modeling has been reported in the published literature, no comparative study of the results could be made. The lower accuracy in network performance can be attributed to the following reasons: inadequate training data set, nonoptimal training of networks or shorter input time series ($n = 1, n = 2$). The training data set consisted of only 24 days of AVL information. It is believed that a larger data set, consisting of at least 6 months of AVL data, will improve the accuracy of the various neural network models. This is because ANNs are data-driven models and using a larger data set would result in a much better generalization of the schedule deviations.

The actual versus network-predicted schedule deviations are illustrated in Figures 4a, 5a, 6a for Model Set I, Case A, and in Figures 4b, 5b, 6b for Model Set II, Case A. The figures show that the ANN models performed well in capturing the trend in the schedule deviations at different times of day. The three networks learned the decreasing (or larger values of schedule delays) trends very well for Model Set II. While the best results were obtained for the Elman and Jordan networks, there were no significant differences between the networks. Hence, no definitive conclusion can be made on the superiority of one architecture over the other. The partial recurrent net architectures incorporate knowledge about the past states internally, and therefore seem more suitable for our schedule behavior modeling problem.

The schedule behavior models can be used for predicting the schedule deviations at timepoints $(k + 1)$, $(k + 2)$, and so forth, if the schedule deviation at the current timepoint k is known from the

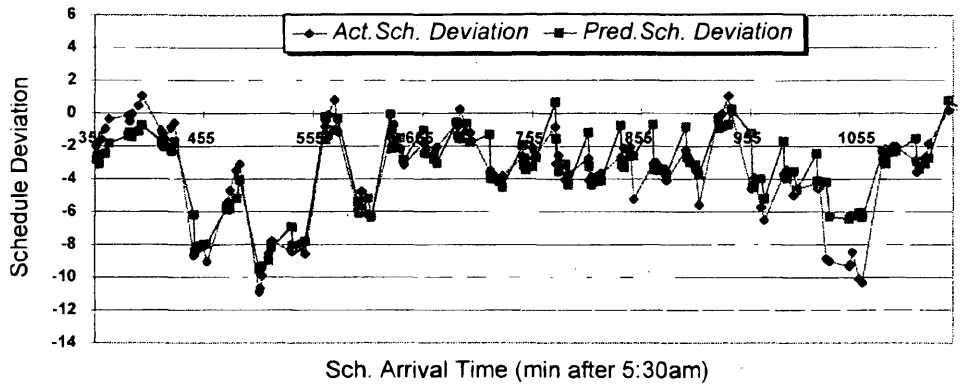


FIGURE 4a $2 \times 5 \times 1$ **feedforward net performance Model I: Case A.**

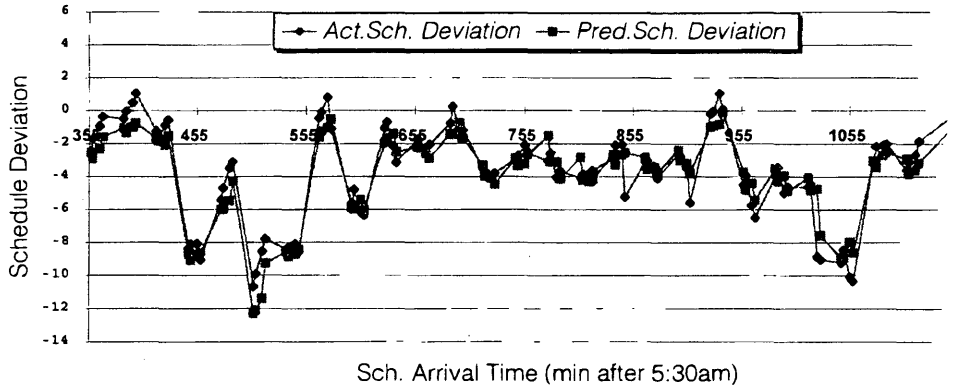


FIGURE 4b $3 \times 6 \times 1$ **feedforward net performance Model II: Case A.**

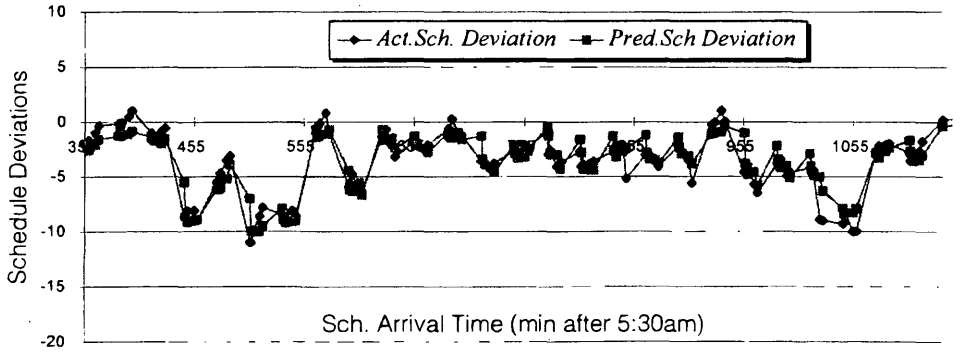


FIGURE 5a $2 \times 5 \times 1$ **Elman net performance Model I: Case A.**

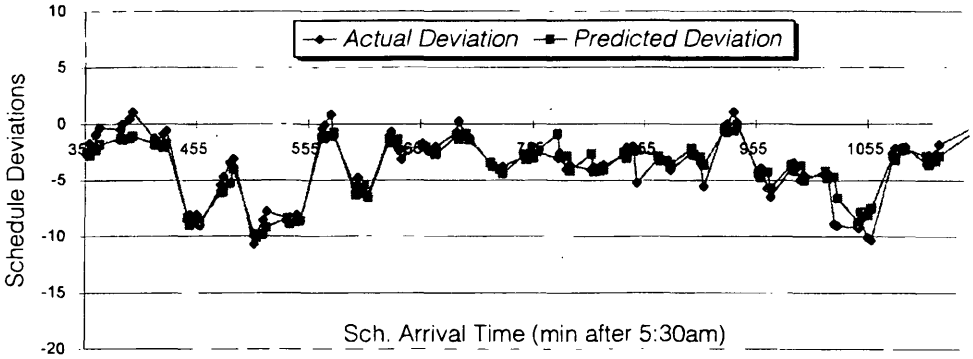


FIGURE 5b $3 \times 6 \times 1$ **Elman net performance Model II: Case A.**

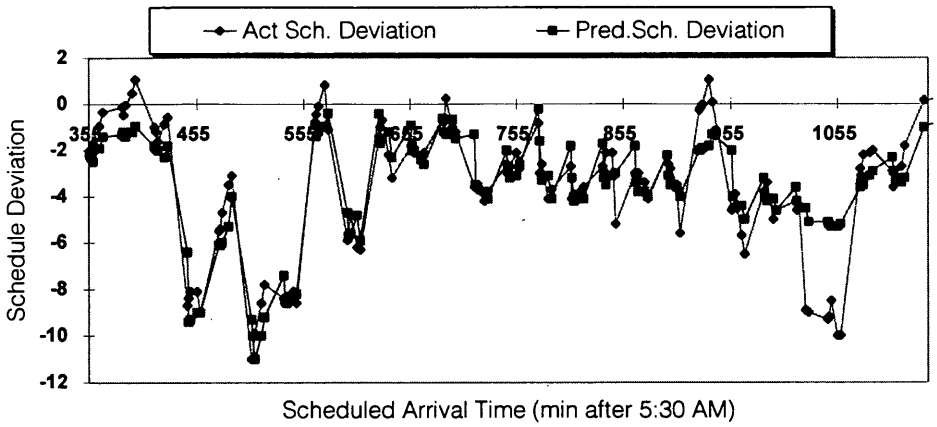


FIGURE 6a $2 \times 5 \times 1$ **Jordan net performance Model I: Case A.**

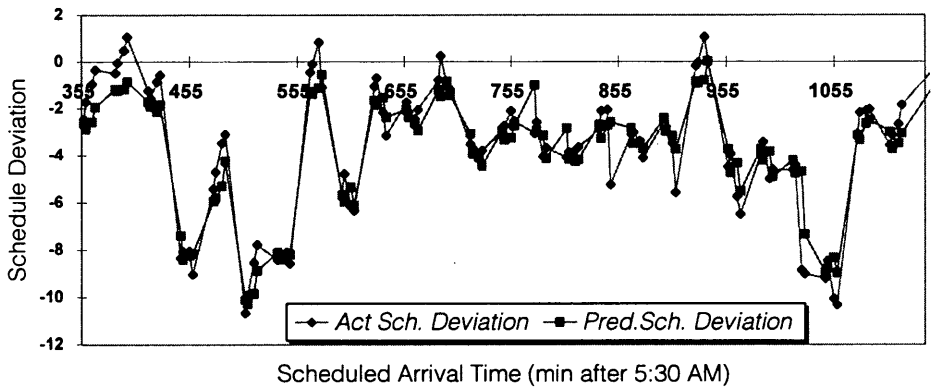


FIGURE 6b $3 \times 6 \times 1$ **Jordan net performance Model II: Case A.**

AVL system. This provides a time window for dispatchers and supervisors to evaluate the system performance online, implement any service-control strategy such as headway adjustments, and schedule adjustments appropriately. In addition, the models are useful for evaluating the effectiveness of any service-control strategy implemented. If the supervisors take appropriate control actions to offset any increase in schedule deviations estimated by the model at a downstream timepoint [$(k + 1)$, $(k + 2)$, . . . etc.], then the control strategies can be evaluated for effectiveness by comparing the actual schedule deviations at timepoints $(k + 1)$, $(k + 2)$, and so forth, with the model-predicted values to see whether there was a decrease in the schedule deviations. Currently, there are no procedures available for evaluating the effectiveness of service-control strategies in real time. Thus, the schedule behavior modeling approach proposed in this study can provide bus transit operators with an automated, on-line performance analysis and evaluation tool.

In summary, the ANN approach provides two distinct advantages over conventional statistical techniques for developing and implementing schedule-behavior models in real-world operations. First, the modeling process can incorporate the concept of spatiotemporal sequencing and short-term memory. Second, the models can first be developed off-line using historical data, and then used with current and new data for on-line updating of the models. This enables transit operators to deal with large amounts of data and a dynamic database in real time and thus can be useful in developing automated decision-support systems to assist dispatchers and supervisors with real-time service-control problems. Initial efforts are focused on investigating the development of schedule behavior models using ANN techniques.

CONCLUSIONS

The results from this case study indicate the suitability of the schedule behavior modeling methodology using ANNs. ANNs have the ability to incorporate short-term memory data about schedule deviations at consecutive timepoints on a route. While the results are encouraging, no definitive conclusions can be made regarding their performance unless a comparison is made between these results and other applicable techniques such as statistical methods. The methodology discussed herein for schedule behavior modeling can be used when applying other modeling techniques including statistical methods, among others. Ongoing research is aimed at investigating the modeling and prediction of schedule behavior of buses using conventional statistical techniques such as the Box-Jenkins model. Also under development are ANN models using longer input time series: $SD(k - 6)$, $SD(k - 5)$, . . . , $SD(k - 1)$. The development of schedule behavior models using the schedule deviation of buses on different routes arriving at a timed-transfer location is also being studied. Modeling the schedule behavior of buses on different routes arriving at a timed-transfer location will be useful for more efficient control of the arrival times of buses on various routes at the transfer location, and thus will minimize the num-

ber of missed transfers. In addition, the models can be useful for designing an optimal time window at timed-transfer locations.

ACKNOWLEDGMENTS

The authors wish to thank the Tidewater Regional Transit, Mid-Atlantic Transportation Centers Program, and the Virginia Department of Transportation for their support. The authors also wish to thank the Institute for Parallel and Distributed High Performance Systems, University of Stuttgart, Germany, for making available the Stuttgart Neural Network Simulator for our modeling experiments.

REFERENCES

1. Lapedes, A., and R. Farber. *Nonlinear Signal Processing Using Neural Networks: Prediction and System Modeling.* LA-UR-87-2662. Los Alamos National Lab Technical Report, Los Alamos, N.M., 1987.
2. Vemuri, V. R. and R. D. Rogers, eds. *Artificial Neural Networks: Forecasting Time Series.* IEEE Computer Society Press, Los Alamitos, Calif., 1994.
3. Su, H.-T., J. T. McAvoy, and P. Werbos. Long-Term Predictions of Chemical Processes Using Recurrent Neural Networks: A Parallel Training Approach. *Industrial Engineer and Chemical Research,* Vol. 31, 1992, pp. 1338–1352.
4. Mehra, P., and B. W. Woh. *Artificial Neural Networks: Concepts and Theory.* IEEE Computer Society Press, Los Alamitos, Calif., 1994.
5. Hertz, J., A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation.* Addison-Wesley Publishing Company, Redding, Mass., 1991.
6. Weigend, A. S., B. A. Huberman, and D. E. Rumelhart. Predicting the Future: A Connectionist Approach. *International Journal of Neural Systems,* Vol. 1, No. 3, 1990, pp. 193–209.
7. Jordan, M. I. Attractor Dynamics and Parallelism in a Connectionist Sequential Machine. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society,* Hillsdale, N.J., 1986, pp. 531–546.
8. Elman, J. L. Finding Structure in Time. *Cognitive Science,* Vol. 14, 1990, pp. 179–211.
9. Fahlman, S. E. Faster-Learning Variations on Backpropagation: An Empirical Study. In *1988 Connectionist Models Summer School* (T. J. Sejnowski, G. E. Hinton and D. S. Touretzky, eds.), Morgan Kaufmann, San Mateo, Calif., 1988.
10. Burke, L. Assessing a Neural Network: Validation Procedures, *PC AI,* March/April 1993, pp. 20–24.
11. Sharda, R., and R. B. Patil. Neural Networks as Forecasting Experts: An Empirical Test. *Proceedings of the IJCNN Meeting,* Washington, D.C., 1990, pp. 491–494.
12. Tang, Z., C. de Almeida, and P. A. Fishwick. Time Series Forecasting Using Neural Networks vs. Box-Jenkins Methodology. *Simulation,* Vol. 57, No. 5, Nov. 1991, pp. 303–310.
13. Weigend, S. A., and A. N. Gershenfeld, eds. Time Series Prediction: Forecasting the Future and Understanding the Past. *Proceedings of the NATO Advanced Research Workshop on Comparative Time Series Analysis,* Santa Fe, N.M., May 14–17, 1992.
14. Ulbricht, C. *Multi-Recurrent Networks for Traffic Forecasting.* Austrian Research Institute for Artificial Intelligence, Vienna, Austria, 1994.

Publication of this paper sponsored by Committee on Artificial Intelligence.

Development of Neural Signal Control System—Toward Intelligent Traffic Signal Control

JIUYI HUA AND ARDESHIR FAGHRI

This study describes the process of developing a traffic signal control for isolated intersections using artificial neural networks. Currently existing signal control systems are briefly discussed and their shortcomings presented. Subsequently, a new multilayered neural network architecture is presented that diminishes many of the shortcomings in existing controllers. The new control system, called neural signal control system (NSCS), is more adaptive to the changes in traffic patterns that take place at isolated intersections. It also provides the traffic engineer more flexibility in terms of optimizing different measures of effectiveness. After describing the architecture and operation of this new system, a comparative analysis is performed by simulating different traffic patterns at a hypothetical intersection using both NSCS and a commonly used software—signal operations analysis package (SOAP). Measures of effectiveness produced by both NSCS and SOAP indicate NSCS's superiority in all aspects. This study concludes with some thoughts for further research in this area.

After disregarding static factors such as geometric conditions, the traffic signal control process for an isolated intersection deals primarily with two issues: the allocation of green time to each traffic movement and the duration of this allocation. From a pattern processing perspective, the signal control process is supposed to find the "best" phasing patterns and their display durations. In a natural sense, the phasing is not necessarily cyclic. Moreover, displaying phases according to traffic fluctuations makes more efficient use of green time from a microscopic point of view.

In the early development of traffic signal control studies, because even traffic flows were assumed, the most important control parameters for isolated intersection control were cycle length and green time split. The first computer traffic signal control concept was established by Webster (*1*), and it still influences current control operations, although many variations and enhancements have been made for different situations (*2–6*). However, there have been no substantial changes in the concept. After observing the variations in traffic flow at isolated intersections, some researchers (*7*) have tried to achieve better control performance by applying stochastic models to build control logics in order to accommodate the fluctuation in traffic flow caused by random arrivals. Conceptually, there is still some awkwardness. In stochastic models the distribution of random arrivals must be preassumed, however, this assumption may be proven to be unrealistic.

Current traffic signal control systems are primarily pretimed, semiactuated, fully actuated or volume-density responsive. The traffic signal operations currently used at isolated intersections fall short in the following ways:

1. In pretimed traffic signal control, the traffic arrival patterns are assumed to be uniform. Though this assumption provides the possibility of vigorous mathematical formulation to optimize the traffic conditions at an isolated intersection, it is obvious that the assumption is far from realistic. Thus, the control effects may not be achieved in a practical sense.
2. In actuated traffic signal controls including volume-density responsive operations, the control logics are made relatively reasonable for a specific range of traffic conditions. The traffic signal control system is then able to accommodate the changes in traffic conditions. The effects of the accommodation made by actuated control are, however, very limited because of the following reasons:

- The phases are displayed in order. The actual traffic conditions may not necessarily demand the preset phase sequence, though enhancements such as phase skipping, multisettings of maximum/minimum green intervals and multiring can be made.
- All decision making processes are based on thresholds such as maximum/minimum green time, passage time, and critical gap. These thresholds may not always lead to a satisfactory decision as traffic conditions change. More critically, in some cases fully actuated and semiactuated operations may fail, for instance, in oversaturated traffic conditions.

Almost all existing traffic signal systems are facing the problem of how to adapt operations to changes in traffic conditions using reasonable assumptions and the capabilities of on-line adjustments. Also, they all need a relatively large presetting effort for real-time implementation.

To achieve a signal control operation which is adaptive to traffic conditions, the neural signal control system (NSCS) has been developed. The NSCS identifies the most suitable phase pattern from among a set of prepared patterns, yet determines the duration of the pattern based on a "human-like thinking" mechanism. The performance of the proposed system is examined with a comparative analysis of computer simulation results obtained from NSCS and the signal operation analysis package (SOAP).

DEVELOPMENT OF NEURAL CONTROLLER

Artificial neural networks (ANNs) are mostly accredited in pattern processing tasks such as recognition and classification. The performance of ANNs in pattern processing strongly suggests their suit-

Department of Civil Engineering, University of Delaware, Newark, Del. 19716.

ability for addressing isolated intersection traffic control problems. In some ANN paradigms, such as Kohonen networks (*8*) and ART networks (*9,10*), the competition process among nodes is analogous to that of traffic demands of different movements. Based on the ANNs concepts, a paradigm suitable for isolated traffic signal control operation has been developed.

Concepts of NSCS

Inasmuch as it has been determined that phase patterns are associated with corresponding traffic patterns, signal operation can be treated as a pattern association process that includes traffic pattern recognition and classification procedures. There are theoretically *N!* combinations of possible phase patterns at any given time, if there are *N!* traffic movements at the intersection. However, there may be only several phase patterns out of the *N!* combinations necessary or applicable due to the constraints of conflict points, traffic patterns, and other mandatory control issues and policies. Therefore, in normal practice fewer phase display patterns will be applied for the different traffic patterns. In other words, traffic patterns can be classified into a number of types and each type of traffic pattern requests one specific phase pattern.

NSCS operates in such a way that approaches to the intersection are grouped to share one phase pattern in order to increase efficiency of green time usage and avoid conflict points. For example, in the five leg intersection shown in Figure 1, approaches can be grouped into two groups, A, B, and E form group 1; C, D, and E form group 2.

Thus, the determination of green time assignment is broken down into two steps:

1. Assign green time to an approach group which has a larger demand; and
2. Display suitable phase patterns according to the traffic patterns of this approach group.

These procedures reflect the principle that the phase pattern currently displayed is always, from among all available patterns, the one which carries the most traffic demand measured on a link group or a phase basis. These procedures will provide reasonable heuristics by increasing the efficiency of green time usage. Figure 2 presents the concept of green time assignment.

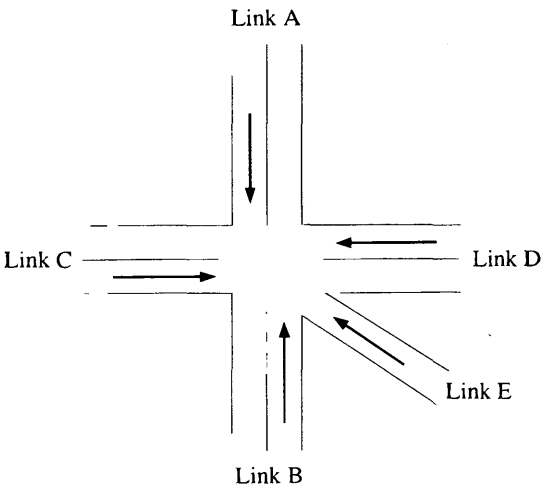


FIGURE 1 Link grouping process for a five-leg intersection.

Duration of Phase Pattern

A selected phase pattern may not continue to be appropriate for traffic patterns which change over time and should therefore be reviewed for adjustment. There is, however, the possibility that the demand for changing a phase pattern occurs too frequently. In actual operations, too frequent changes in phase patterns are not acceptable because of time loss and disruption to the continuity of the traffic flow.

To avoid too frequent phase changes, the mechanism of human visual nerves is employed. The sensitivity of the human visual nervous system adapts to the circumstances it encounters. Suppose an individual has been outdoors on a sunny afternoon. When that person moves quickly into a dark room, he may temporarily lose his sight. At this point, his visual nerves may sense only luminous objects. As time passes, his visual nerves gradually begin to sense other objects in the room which are not luminous. Obviously then, the brightness sensoring range of the human visual nerves varies. Assuming all objects are competing to present themselves to the person's sight, the most luminous object will be the winner at the moment that the person walks into the room. However, the person may find all the other objects in an order of the degree of the brightness of those objects.

A similarity to the aforementioned scenario is also found in human traffic guidance. A human guide would first assign the phase which carries the largest demand. As time passes, the expectation to change to a new phase becomes stronger, even if the old phase is still carrying the largest traffic demand. Such a mechanism is employed in NSCS to determine the duration of a particular phase pattern. This phase pattern determination process is observed as a competition among different phase patterns. When a particular phase has just been chosen, the competition strength of this pattern will be at the strongest level and that of the others at the weakest level. As time

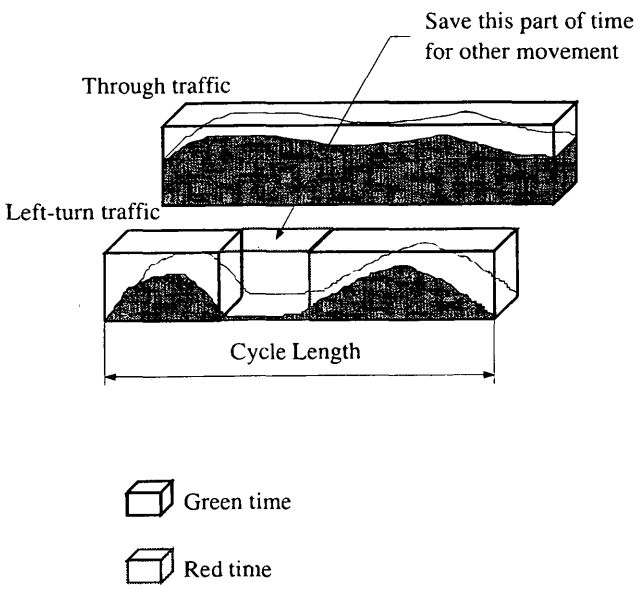


FIGURE 2 The green time assignment.

passes, the competition strength level of the winning phase decreases gradually and that of the losing phases increases gradually. This changing process continues until a new winner is found.

This "decay-enhance" process is conducted through an artificial neural network. The system in this case simulates the traffic control process of the human traffic guide with the ability to adapt to different traffic conditions without incorporating any of the conventional traffic signal operational parameters such as cycle length and split. The green time assignment is based only on the patterns of the traffic approaching the intersection.

Architecture of NSCS

The architecture of NSCS must be, as in most artificial neural networks applications, developed for the subject case to accommodate the geometric conditions of the intersection including number of approaches, number of lanes within each approach, and turning movements. Special attention should be given to avoid conflict points of the multiple traffic movements.

In its architecture, the NSCS contains a total of five layers, each layer containing a number of processing units, also called nodes or processing elements as shown in Figure 3. The first layer is used to hold traffic information about different movements. This layer works as an input buffer, and in normal cases, the information being fed into this layer is the number of cars. This layer must be divided into several slabs, each dedicated to one group of links.

Following this layer, there is a special layer containing several nodes, called "*G* nodes," each connected to all the nodes in a slab of the input layer. Thus each node in the second layer will take the traffic information of one link group. The connections between nodes in the second layer and those of the first layer can change during the operation.

The third layer is a mirror of the first layer except that the nodes in this layer must use transfer functions which are user-specified to compute the measures of effectiveness (MOEs) as their outputs. Nodes in the third layer are connected to the corresponding nodes in the first layer and also to those in the second layer.

The fourth layer contains a set of nodes that compute a value indicating the MOE, if a specific phase pattern was applied (this is discussed in more detail later). The number of nodes in this layer is adjustable by the user so that a variable number of phase patterns can be applied. The connections between this layer and the third layer are also changeable during the operation.

The fifth layer is simply the output of the phase pattern. All nodes in this layer connect to one of the fourth layer nodes.

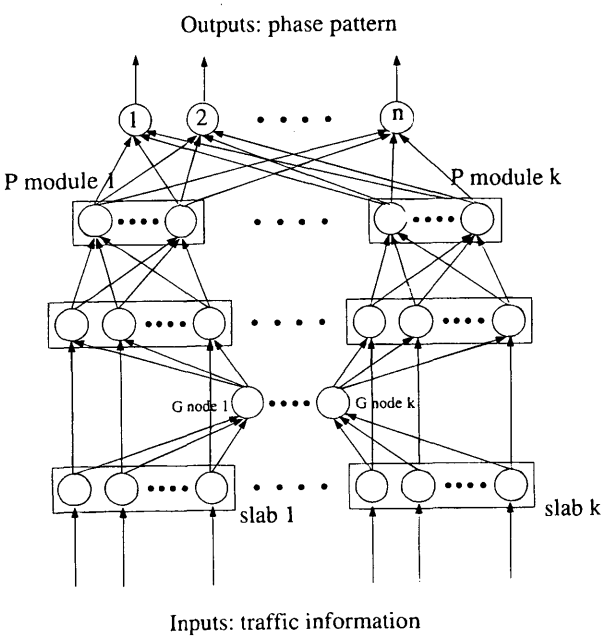


FIGURE 3 General architecture of NSCS.

Operation of NSCS

The NSCS is characterized as being highly dynamic in the sense of phase pattern changes. No specific or predetermined parameters such as cycle length and green time split ratio are required for the operation.

When a vector of traffic information is fed into the first layer, a value of output for every *G* node is first computed. This value implies the traffic demand for the link group. In the second layer, a competition takes place on a "winner takes all" basis. That is only the one whose output value is the greatest among the *G* nodes wins the competition, all other *G* nodes lose. The following simple equation is used to compute the output of *G* nodes.

$$Y_j = \sum w_{ij}^G X_{ij}$$

where

Y_j = output of *G* node *j*
W_{ij}^G = weight of connection from *I* node of *j* slab to *G* node *j*
X_{ij} = input value of *I* node of *j* slab; $X_{ij} = f(n)$, where n is the number of vehicles detected in the pocket.

The winning *G* node fires with a value of 1 and distributes this value to the nodes of its corresponding module in the third layer. The losing *G* node fires with a value of -1 and also distributes this value to the nodes of its corresponding module in the third layer.

In the third layer, the output of the node is computed as follows:

$$Z_{ij} = Y_{ij} H_j$$

where Z_{ij} is the input to the node i of slab j in third layer, and H_j is the output of *G* node j.

Adjustment of the Connection Weights

When competition takes place, the strength of the competition of each *G* node and *P* node changes as time passes. The competition strength of each node is reflected in the value of the weights of the connections from the nodes in the prior layer to this node. The weights of the connections change in an incremental or decremental manner as time is counted on a small incremental basis. The changing range of the weights is from 0 to 1. The step size of the weights adjustment is computed every time a new winner is found.

The initial determination of the step size of weight adjustment for each node is determined as follows:

$$\Delta = 1/(G_{max} - G_{min}) + M$$

where

Δ_k = step size of adjustment for G weights
G_{max} = maximum green time interval specified by the user
G_{min} = minimum green time interval specified by the user
M = an infinitely small real number used for the stability of the system
k = an index denoting the winning node.

Suppose there are n nodes in the phase layer and m nodes in its prior layer, and the current winning node in the phase layer is k. The procedure for adjusting the weights during the operation is determined as follows:

$$w_{ij}^{new} = \begin{cases} w_{ij}^{old} - \Delta_k & \text{for } i = k \\ w_{ij}^{old} + \Delta_k & \text{for } i \neq k \end{cases}$$

where

w_{ij}^{new} = new weight,
w_{ij}^{old} = old weight,
Δ_k = step size of weight adjustment for link group k, and
$i = 1, 2, \ldots, n; j = 1, 2, \ldots, m$.

The same procedures are applied to determine the competition strength of P nodes.

DISCUSSION

NSCS's algorithm is flexible in many aspects. These flexibilities ensure the ability of the controller to perform in various traffic conditions. The operation can be fine tuned by adjusting several factors:

Expandability

The capability of NSCS, in terms of the number of phase patterns prestored, can be virtually expanded to the maximum number of possible combinations of the traffic movements, namely $N!$ phase displays. The expansion is reflected in the increase of the number of P nodes. But, this expansion does not increase the computational cost (computing time) if a parallel computing process is implemented on the hardware being used. This is an advantage of the neural network architecture. With the expansion, the controller should be able to perform a finer function in the sense that the phase pattern more precisely fits the traffic pattern.

Flexibility

NSCS offers the traffic engineer the flexibility of changing the optimization objective. For instance, mathematical formulae can be provided to optimize only the environmental MOEs. Here, we describe the optimization of two objectives.

If traffic demand is defined as the throughput, traffic demand should be calculated as follows:

$$U = \sum WF(X)$$

where

U = output vector
W = weights vector
$F()$ = vector of functions of x's
X = traffic volumes

In the algorithm of our system, the traffic volume is determined by using the Greenshield model.

If traffic demand is defined as the maximum queue length, it should be calculated differently as in the following equation:

$$U = \sum WN$$

where

U = output vector
W = weights vector
N = a vector of number of vehicles detected in the pockets.

Adaptivity

The system adapts to changing traffic patterns. This capability to adapt to different traffic behavior is accomplished through the competition strength function, the minimum/maximum green time thresholds, and the system sensitivity. Each is described below.

Competition Strength Function

The competition strength parameter, which is a function of time, can affect the operation significantly. It can also be given in numerous forms. For example, connection strengths of the losing processing units could be

$$w = t^u$$

where t is time elapsed and u is a real number.

The parameter u can be changed, by using the historical information of traffic arrivals to make signal operations more adaptive to traffic patterns.

Minimum/Maximum Green Time Thresholds

Minimum/maximum green time thresholds can affect the entire signal operation as well. Figure 4 illustrates that by changing the minimum/maximum green time thresholds, the signal flipping point will be different. In fact, these thresholds determine the proportion of stable operation of the signal. The signal display is fixed for the stable portion of a particular phase.

Adjustment of System Sensitivity

The system can be adjusted to be more or less sensitive to the fluctuation of traffic for both assignment of green time to link groups and traffic movements. The operators can fine-tune the system by adjusting the system sensitivity to accommodate their own traffic patterns.

The adjustment of the system sensitivity is accomplished simply by assigning the value of s, where s is the sensitivity index shown in Figure 5. The variable s has value ranging from 0 to 1. When $s = 0$, the system is most insensitive and when $s = 1$, it is most sensitive. The sensitivity can be adjusted in both competition among G nodes and P nodes.

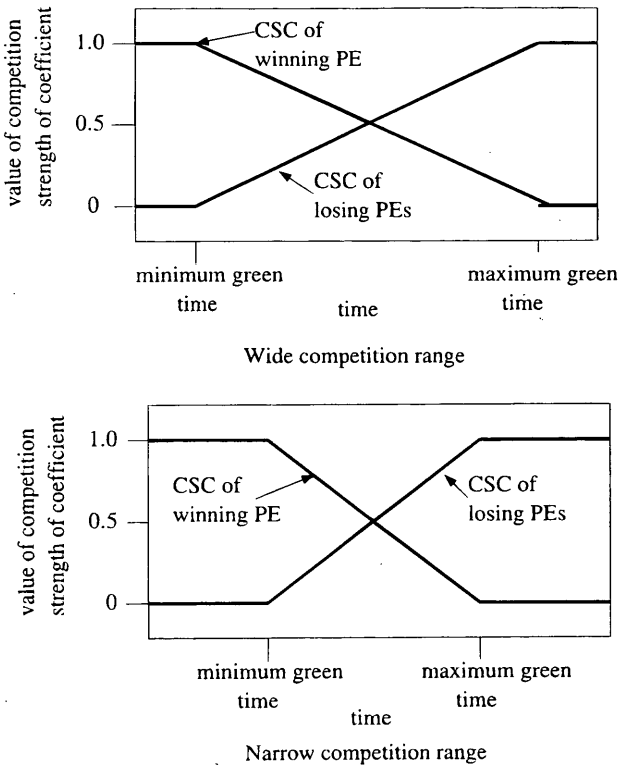


FIGURE 4 Adjustment of maximum/minimum green time interval.

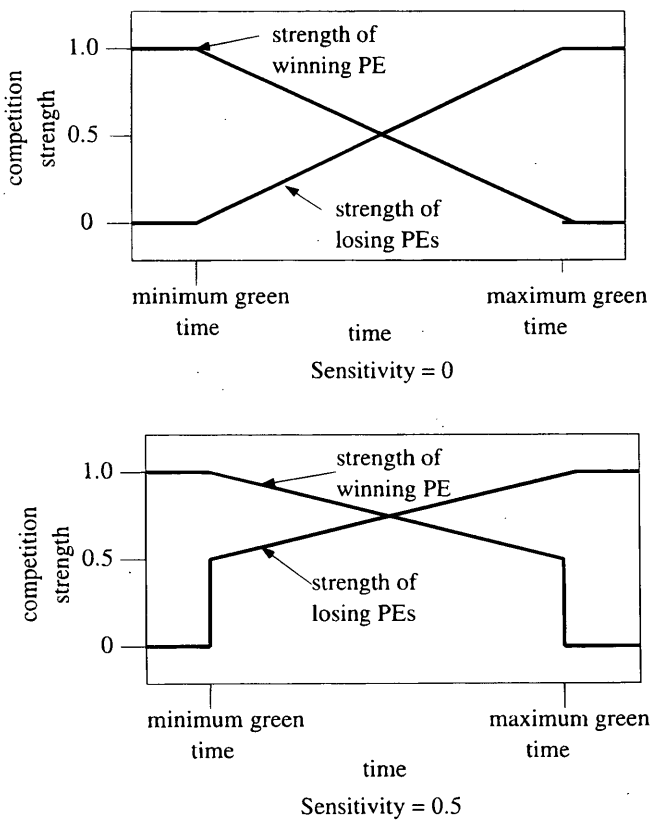


FIGURE 5 Adjustment of system sensitivity.

Unlike conventional traffic control strategies, NSCS approaches the problem by considering the traffic fluctuation microscopically. The optimal control strategy used in NSCS provides a set of sequential phase patterns in a heuristic manner in order to optimize the user specified MOEs, instead of a conceptually defined exact optimal phase sequence. The competition features of both G and P node groups used when assigning green time is a process of simulating the human traffic guide thinking process that is assumed in this study. Thus, when different traffic movements demand green time, the assignment is accomplished on a relative competition basis instead of on simple threshold criteria.

The entire operation of NSCS is conceptually more natural for isolated intersection control. It is expected that an adaptive control process can be realized through NSCS's operation and less effort required for NSCS operators in practical applications.

COMPUTER SIMULATION CASE STUDY

Case Study

A case study is made for examining the performance through comparative analyses of simulation results by NSCS and a popular commercial software, SOAP.

A hypothetical case containing an isolated intersection and different traffic conditions is used for this case study. The intersection has four approaches. All approaches are made with two pockets for exclusive left-turn and through traffic movements. Because it is easier for right-turn traffic to be carried in signal control problems, this study assumes no right-turn traffic on all approaches. A schematic drawing of the hypothetical intersection is shown in Figure 6. A flowchart of the simulation program based on NSCS is presented in Figure 7.

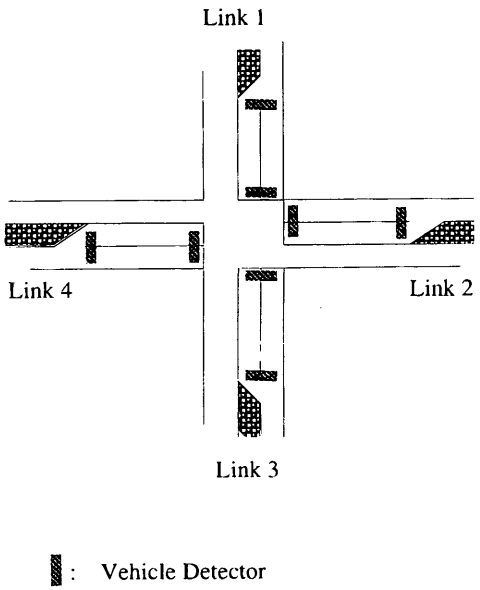


FIGURE 6 Subject intersection of case study.

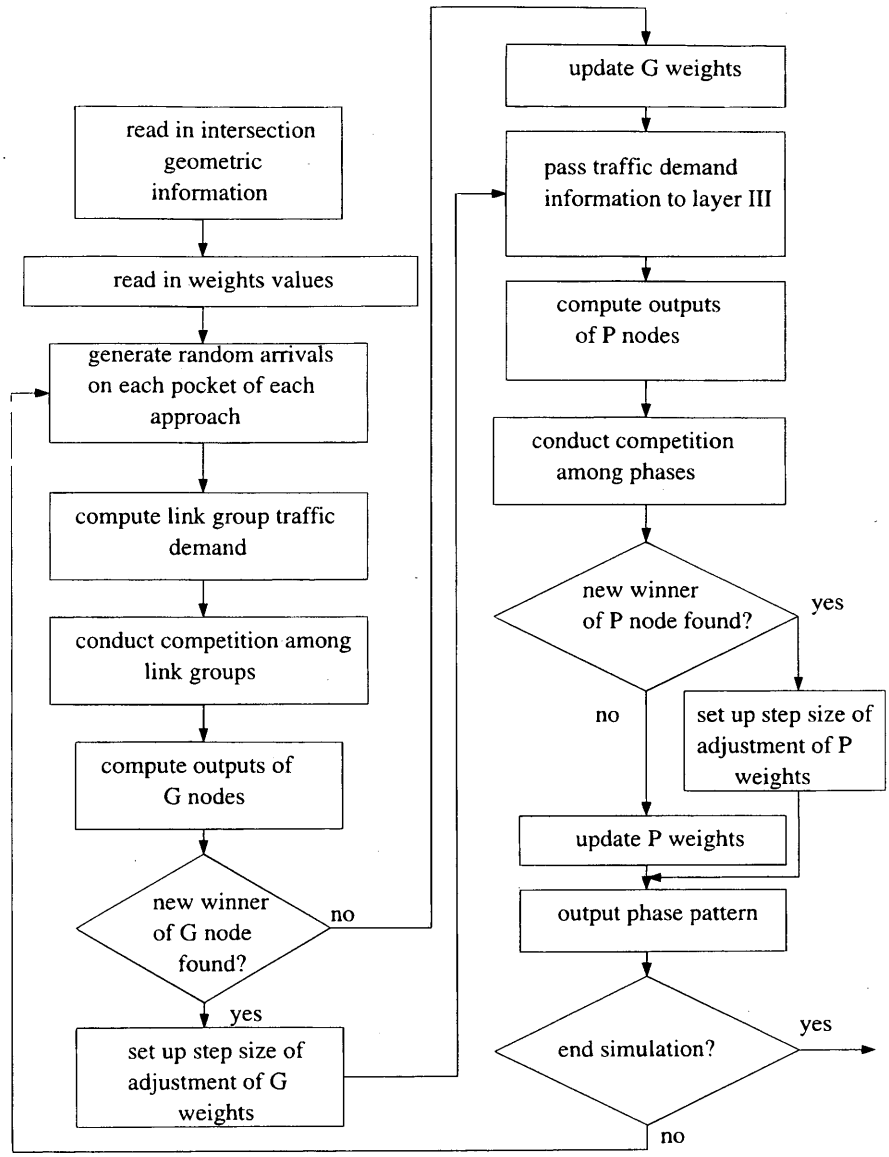


FIGURE 7 Simulation process.

Comparative Analysis of Simulation Results

The foregoing study case is input into both NSCS and SOAP for seven runs under different traffic conditions. When comparing the results generated by the two computer simulation programs, several interesting findings are discovered. The nature of NSCS is also explored.

Simulation Results

Table 1 lists a collection of MOEs recorded from the simulation results by NSCS and SOAP. It shows that in almost all MOEs, NSCS's performance was superior to SOAP's fully actuated mode. Emission pollutant MOEs are not available in SOAP. By investigating NSCS's outputs of emission pollutants, the user can adjust the operation by changing the system sensitivity, the minimum/maximum green time, the competition strength coefficient curves, or the optimization objective to obtain the best environmental MOEs.

Three comparative plots illustrating MOEs of average delay, percentage of stopped vehicles, and fuel consumptions are presented in Figures 8, 9, and 10 respectively.

Findings of the Case Study

By comparing the simulation results produced by NSCS and SOAP respectively, several findings can be summarized as follows:

1. At extremely low traffic volumes, the performance of both SOAP and NSCS are almost the same. This is so, because if waiting traffic is always low enough to be cleared within a single time interval, there should not be any difference in operation between the

TABLE 1 Simulation Results by SOAP and NSCS

Simulation run	Average delay (secs/veh)		Number of stops (%)		Fuel consumed(gals/hr)		Level of service	
	SOAP	NSCS	SOAP	NSCS	SOAP	NSCS	SOAP	NSCS
1	4.68	4.60	54.7	38.3	5.00	2.56	A	A
2	4.93	5.39	66.9	44.1	12.37	7.14	A	B
3	8.68	5.89	79.3	43.5	22.50	12.25	B	B
4	19.66	6.83	91.4	45.6	39.72	22.08	C	B
5	72.40	8.03	97.7	53.3	87.36	37.68	F	B
6	96.21	11.50	97.4	82.8	123.71	51.03	F	B
7	144.86	12.24	100.0	103.0	191.20	53.11	F	B

two programs. The comparative competition of NSCS does not make sense at this point, because such competition often takes place between one nonvoid traffic movement and other void movements. The left portion of Figure 8 indicates this fact.

2. In this particular case, NSCS seems to perform better when the number of stopped vehicles falls within a range of traffic volume from 400 VPH to 1000 VPH. This is in accordance with the basic concept of NSCS analyzed earlier. NSCS always assigns green time to the group of movements in which the traffic is heaviest. Therefore, the number of stopped vehicles can be significantly reduced. If the traffic is too low, the phenomenon described above in Finding 1 takes place. If the traffic is too high, the queue length of vehicles will exceed the detectable length so that a maximum number of vehicles is recognized by NSCS as the maximum detectable number of vehicles. Consequently, the difference among varied movements of traffic is less recognizable by NSCS. This may affect the correct competition strength computation of NSCS. Under very heavy traffic conditions, the NSCS loses its adaptivity and the operation remains constant.

3. NSCS can basically follow the arriving pattern of traffic to assign green time. Figure 11 demonstrates all phase patterns displayed for through traffic on link 1 during the simulation. As observed, the green time pattern is basically a projection of a traffic arrival pattern.

4. The performing range of NSCS is broader than that of SOAP's fully-actuated mode. In Figure 9, when traffic is in excess of 800 VPH, SOAP's actuated mode fails in optimization. Average stopped-delay in SOAP increased dramatically, whereas NSCS maintained a reasonable increase in average stopped-delay.

5. Because of producing lower stopped-delay in heavy traffic conditions, NSCS's fuel consumption is much less than that of SOAP, as displayed in Figure 10.

6. Environmental factors are taken into consideration in NSCS. This will provide practitioners more power in improving the environment through traffic operation.

7. In Table 1, the total number of stopped vehicles produced by NSCS is higher than that of SOAP at the 7th run. This conforms with the adaptivity of NSCS to changing traffic conditions. In cases that generate maximum throughput, NSCS will sacrifice the continuity of traffic flow whereas in SOAP if a preset gap is not detected then the phase will not change from one ongoing traffic movement to another.

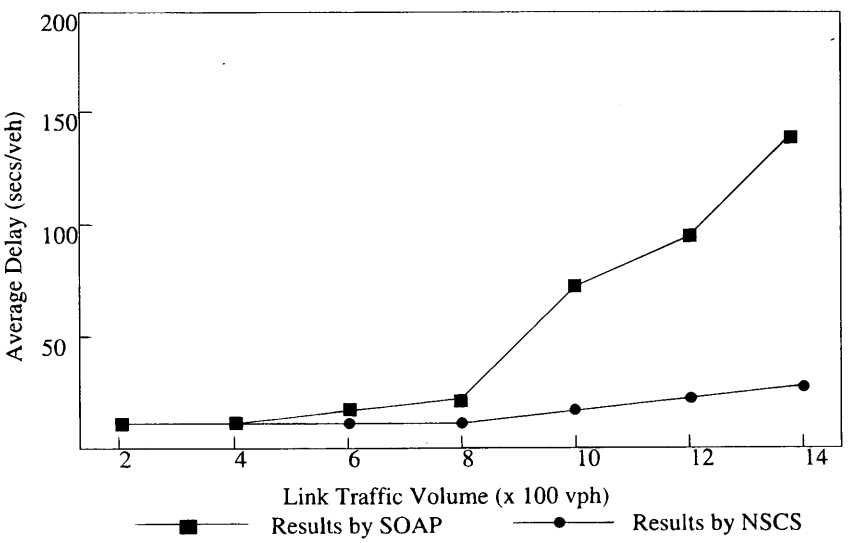


FIGURE 8 Comparison of average stopped delay. Results by SOAP (■) and (●) NSCS.

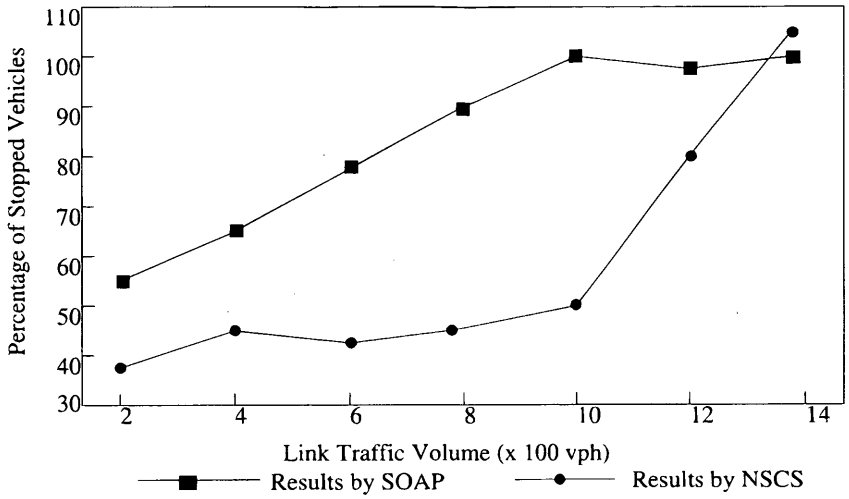


FIGURE 9 Comparison of percentage stopped vehicles. Results by SOAP (■) and (●) NSCS.

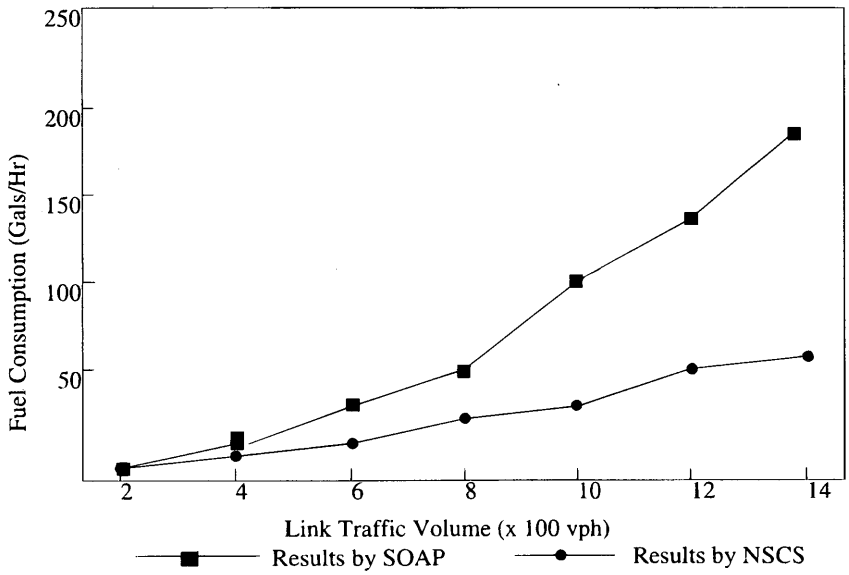


FIGURE 10 Comparison of hourly fuel consumption. Results by SOAP (■) and (●) NSCS.

8. Less presetting effort is needed to operate NSCS. In general, HCM claims that three types of information, geometric conditions, signalization conditions, and traffic conditions, must be understood by human signal designers for setting a conventional signal control system. NSCS requires only the geometric information of the intersection.

CONCLUSIONS

After describing the shortcomings of existing controllers, this paper presented a new concept based on ANNs for optimizing timing at isolated intersections. The chief advantage offered by this multilayered ANN is its adaptiveness to the constantly changing traffic pattern. As a result, this new system does not function through a cyclic operation, instead, it provides the right-of-way with the same approach that a human traffic guide would. Because of the system's adaptivity and flexibility for optimizing different MOEs, less presetting effort is required. There is virtually no need for traffic and signalization information. Therefore, NSCS is expected to overcome malfunctions of traffic signals due to missing or false information provided in the presetting procedure.

In multiple MOEs comparisons, NSCS's performance was superior to a fully actuated control operation simulated by SOAP. The operation offered by NSCS appeared to be adaptive to the fluctuations of traffic approaching the intersection. This feature will be suitable for isolated intersections where larger traffic fluctuations take place.

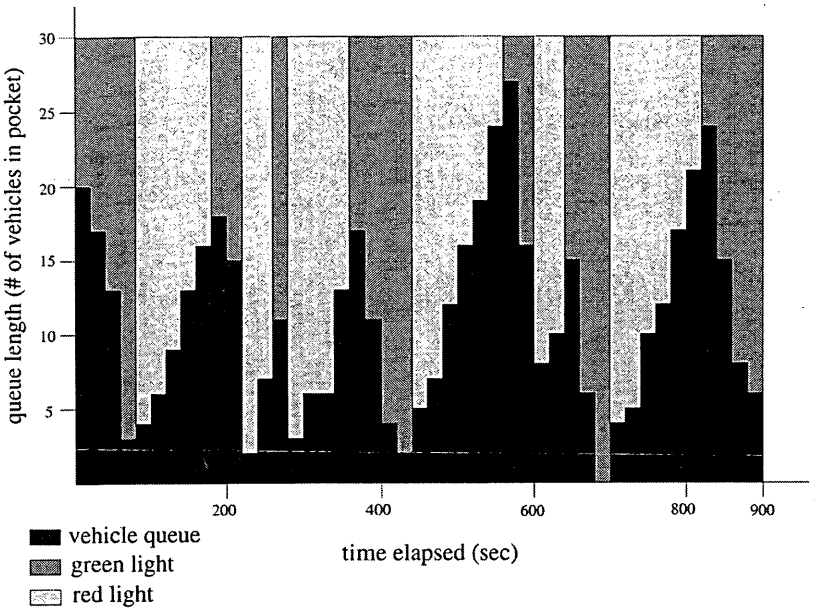


FIGURE 11 Projection of phase patterns to traffic patterns.

The performance investigation of NSCS by computer simulation confirms the methodologies presented in this study. The simulation results also indicate that NSCS stresses several features in its operation:

- Adaptiveness to traffic conditions;
- "Human thinking-like" process;
- A generic ability that allows the process to be applied to virtually any type of isolated intersection; and
- Self-organizing control logic.

It is clear that the NSCS has a much wider operable range and performs better than a fully actuated mode simulated by SOAP for heavier traffic flow.

The control system presented in this paper is a completely new non-cyclic approach to intersection control operations. The preliminary validation tests explained in this study indicate promising results. However, evaluation of the system's capabilities in real life conditions as well as comparison with other sophisticated controllers such as a demand-responsive one are recommended for further research in this area.

REFERENCES

1. Webster, F. V. Traffic Signal Settings. *Paper No. 39*, Road Research Laboratory, London, 1959.
2. Dunne, M. C., and R. B. Potts. Algorithm for Traffic Control. *Operations Research*, Vol. 12, 1964, pp. 870–991.
3. Allsop, R. E., and J. A. Charlesworth. Traffic in a Signal Controlled Network: An Example of Different Signal Timings Inducing Different Routing. *Traffic Engineering and Control*, Vol. 18, 1977, pp. 262–264.
4. Sheffi, Y., and W. B. Powell. Optimal Signal Settings over Transportation Networks. *Journal of Transportation Engineering*, Vol. 109, 1983, pp. 824–839.
5. Peirce, J. R., and P. J. Webb. Mova Control of Isolated Traffic Signals. *IEE Conference Publication on Road Traffic Control*, No 320, 1990, pp. 110–113.
6. Peterson, A., and T. Bergh. LHOVRA—A New Traffic Signal Control Strategy for Isolated Intersections. *Compendium of Technical Papers, 56th Annual Meeting—Institute of Transportation Engineers*, Washington, D.C., 1990, p. 8.
7. Saka, A. A, G. Anandalingam, and N. J. Garber. Traffic Signal Timing at Isolated Intersections Using Simulation Optimization. *Proc., Winter Simulation Conference, IEEE*, New York, 1986, pp. 795–801.
8. Kohonen, T. An Adaptive Associative Memory Principle. *IEEE Transactions on Computers*, Vol. C-23, 1974, pp. 444–445.
9. Grossberg, S. Adaptive Pattern Classification and Universal Recoding: Parallel Development and Coding of Neural Detectors. *Biological Cybernetics*, Vol. 23, 1976, pp. 121–134.
10. Grossberg, S. Adaptive Pattern Classification and Universal Recoding: Feedback, Oscillation, Olfaction, and Illusions. *Biological Cybernetics*, Vol. 23, 1976, pp. 187–207.

Publication of this paper sponsored by Committee on Artificial Intelligence.

A Genetic Algorithm Approach for Solving the Train Formation Problem

DAVID MARTINELLI AND HUALIANG TENG

The train formation plan is one of the most important elements of railroad system operations. Although mathematical programming formulations and algorithms are available for solving the train formation problem (TFP), the computational time required for their convergence is usually excessive. At the same time, shorter decision intervals are becoming necessary given the highly competitive operating climates of the railroad industry. Thus, new techniques are needed for generating efficient solutions for the TFP. In this study, we present the development of a genetic algorithm (GA) as a possible technique for this problem. The calibration and validation of the GA model are carried out for three different complexity levels of objective functions. It is found that the optimal solutions can be found for all the different formulations while consuming only a small amount of computation time.

Railroad system operating plans are developed to perform the sequential decision process of: car block decisions, train formation decisions, train schedule decisions, and empty car distribution decisions. These are made under the consideration of engine power, maintenance, service level requirements, and other competing criteria. Car block decisions determine which blocks the cars will be assigned to, or which demand each block will carry. Train formation decisions determine which train the blocks will be assigned to, or which block each train will carry. Train schedule decisions determine when trains will be released from their origin station and arrive at their destination station. Finally, empty car distribution decisions determine where the empty cars will be sent. In this study, the train formation problem (TFP) is defined as: assign the traffic demand, in terms of cars, to available trains in a network environment so as to minimize the cost incurred in the whole production process.

Despite the substantial quantity and diversity of rail operating decision models, a common element exists in that they all require a substantial investment of computational effort and, subsequently, implementation time. Experience with these models indicate that the computational time required to obtain an optimal (or near optimal) solution varies with formulations.

A common approach for the industry in handling dynamic demands has been to shorten the time period between successive modeling updates. Unfortunately this introduces a tradeoff between longer central processing unit (CPU) time requirements for more realistic solutions and the added resources necessary to provide more frequent model updates. In light of this tradeoff, new approaches such as artificial intelligence are necessary and may prove quite fruitful if shorter implementation times can be achieved without a substantial loss in solution integrity (*1*). One such artificial intelligence technique is genetic algorithms (GA). In employing GA models, the intelligent optimal solution searching process alleviates the impacts of the inputs by directly going to the feasible solution region instead of stumbling on the restrictive constraints. This is the primary reason that genetic algorithms demonstrate promise as a solution technique for the TFP.

In comparison with conventional models, GA demonstrate several distinct advantages. First, they employ an efficient optimal solution searching technique which can be described as multi-hill climbing. The global solutions can be easily found for both linear and nonlinear formulations. Second, the optimal solution searching process is independent of the form of the objective function. Unlike conventional techniques, in which the algorithms usually rely on the structure of the formulation such as the conditions for the decomposition algorithms, GA models can be implemented without such considerations. Third, conventional algorithms are often sensitive to the input patterns such as the conditions set forth by Monte Carlo techniques.

There have been several transportation research efforts in which genetic algorithms are employed to deal with a combinatorial explosion associated with many optimization problems. Xiong and Schneider (*2*) integrated an artificial neural network model into a GA model to solve the traffic network design problem. Foy et al. (*3*) used a GA model to determine the optimal signal timing decisions in a simulation environment for on-line decision making. Chan et al. [unpublished data; cited by Xiong and Schneider (*2*)] applied a GA to road maintenance planning.

BINARY INTEGER PROGRAMMING FORMULATIONS FOR THE TFP

An example railroad network having 6 nodes (representing yards) and 10 links (representing line segments) is represented in Figure 1. Trains are usually divided into long and short distance service. The short distance trains are those whose origin and destination yards are adjacent; whereas long distance trains are those whose origin and destination yards are not adjacent. Normally, short distance trains are always provided for each link, whereas the existence of long distance trains is determined by the train formation plan. Short distance demands are those whose origin and destination are connected directly by one link. Long distance demands are those whose origin and destination are not directly connected. In general, short distance demands are carried by short distance trains and long distance demands are carried by a combination of short and long distance trains.

In railroad networks, there are always a number of different physical routes available for a given demand. On a certain route, there are always a high number of possible itineraries (or assignments). These itineraries are distinguished from each other by the number

D. Martinelli, Department of Civil and Environmental Engineering, West Virginia University, P.O. Box 6103, Morgantown, W.Va. 26506. H. Teng, 1284 Civil Engineering Building, Purdue University, West Lafayette, Ind. 47906-1284.

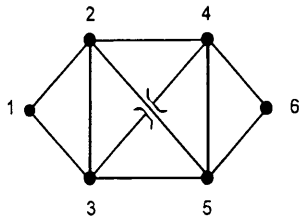


FIGURE 1 Example railroad network.

and types of trains. For example, for demand from Yard 1 to Yard 6, there might be four different physical routes possible: (1, 2, 4, 6), (1, 2, 5, 6), (1, 3, 4, 6), and (1, 3, 5, 6). Further, additional combinations exist for each route. For example, along physical route (1, 2, 4, 6), there might be four itineraries possible as represented in Figure 2. Referring to Itinerary i_2, the demand for Yard 1 to 6 will be relayed from Yard 1 to Yard 2, and then to Yard 6.

The designation of long distance trains and the route they follow are presented in Table 1. The corresponding demand matrix is represented in Table 2. The short distance trains are denoted such as T12 in Figure 2, where 1 and 2 are the train's origin and destination, respectively, whereas the long distance trains are in the form of T*ijk*, where *j* and *k* are the train's origin and destination, respectively, *i* is the sequence of the possible roads the train can follow between *j* and *k*.

It is a common practice for the sake of convenience that, when managing the traffic flow on the railroad network, each demand is usually confined to only one itinerary. If for each demand, a set of 0-1 variables are defined for the choice of itinerary, the TFP could be formulated as a 0-1 integer program. If the objective is minimizing the delay times including the travel times of the cars incurred in the railroad system, subject to demand routing deviation restriction as described above, then the TFP can be formulated as follows.

$$MIN \sum_{l=1}^{2L} t_l Y_l + \sum_{j=1}^{N} v_j x_j \tag{1}$$

Subject to:

$$\sum_{k \epsilon R_i} x_{i,k} = 1. \tag{2}$$

where:

$$x_j = \sum_{i=1}^{M} \sum_{k \epsilon s_j} r_i x_{i,k} \qquad Y_l = \sum_{i \epsilon p_l} x_i$$

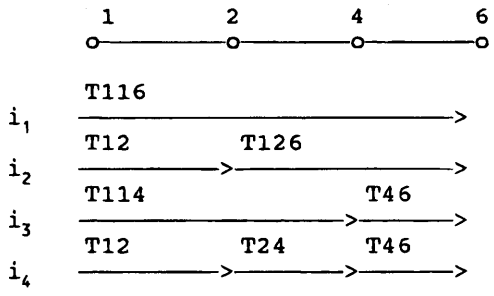


FIGURE 2 Itinerary representation for route (1, 2, 4, 6).

TABLE 1 Designation for the Long Distance Train

Yard	Train	Train Route
1	T116	(1,2,4,6)
	T216	(1,2,5,6)
	T316	(1,3,4,6)
	T416	(1,3,5,6)
	T114	(1,2,4)
	T214	(1,3,4)
	T115	(1,2,5)
	T215	(1,3,5)
2	T126	(2,4,6)
	T226	(2,5,6)
3	T136	(3,4,6)
	T236	(3,5,6)
4	T141	(4,2,1)
	T241	(4,3,1)
5	T151	(5,2,1)
	T251	(5,3,1)
6	T161	(6,4,2,1)
	T261	(6,4,3,1)
	T361	(6,5,2,1)
	T461	(6,5,3,1)
	T162	(6,4,2)
	T261	(6,5,2)
	T163	(6,4,3)
	T263	(6,5,3)

Here X_j denotes the volume of cars in Train j, Y_l the volume on Link l. Here, t_l is the average travel time on Link l, and v_j is Train j's operating time at its destination yard. Also, $x_{i,k}$ is a binary integer variable representing the demand-itinerary choice. $x_{i,k}$ will be 1 if demand i is carried by itinerary k, otherwise zero. R_i is the amount of traffic of Demand i. R_i denotes the set of itineraries by which Demand i was supposed to be carried, S_j is the set of itineraries which include Train j as one part of their line haul, and P_l the set of trains which pass through Link l. L is the total number of links, M the total number of demands, and N the total number of trains possible provided. In this study, $L = 10$, $M = 30$ and $N = 44$. All the t_ls have values of 10 hr/car and the v_js take values around 13–15 hr/car.

Equation 1 is the objective function, in which the first summation is for the travel times incurred on line segments, the second summation is for the times incurred at yards. Equation 2 is the demand-route restrictions. The demand flow conservation and balance constraints usually appear in transportation network models, but are automatically satisfied by this formulation. This is the first case we will investigate in this study.

TABLE 2 Demand Matrix

	1	2	3	4	5	6
1		64	94	121	150	150
2	78		87	27	54	107
3	72	95		4	14	150
4	150	61	19		10	34
5	136	38	89	87		99
6	150	150	140	67	26	

In this constraint formulation, it is assumed that the times in which the traffic is incurred at yards and on line segments are independent of the traffic volume. However, in reality, the times are always dependent on the volume. The relationship between times and the traffic volume is nonlinear. Modifying the objective function accordingly, we have:

$$MIN \sum_{l=1}^{2L} t(Y_l)\, Y_l + \sum_{j=1}^{N} v(X_j)\, X_j \tag{3}$$

The formulation for this case is the objective function of Equation 4 plus the constraints in Case 1. This is denoted as Case 2 in this study.

Furthermore, in practice, it is likely to impose constraints on some variables such as link flow and train load. These constraints can be formulated as:

$Y_l \leq b_1$ for $l = 1, 2, \ldots, 2L$,
$X_l \geq b_7$ for $j = 1, 2, \ldots, N$.

The first indicates that the traffic volumes on links should be less than b_1. The second indicates that the trains can be provided only when the loads on them are larger than b_2. The formulation of Case 3 for this problem is that of Case 2 plus these two additional constraints.

Referring to the railroad networks in Figure 1, there are 10 long distance demands. For demand from 1 to 6 and from 6 to 1, each is assumed to have 16 possible itineraries. For the remaining 8 long distance demands, each is assumed to have 4 possible itineraries. For these conditions, the overall combinations of demands and itineraries is around 10^{17}. All of the formulations in these three cases are binary integer programs. For Case 1, some algorithms such as branch-and-bound, cutting plane and Lagrangian relaxation have been proved to be effective conventionally. The common point of these algorithms might be the use of the linear characteristics of the objective function. In each operation of "branch," for example, relaxed linear programming can be efficiently solved. However, in Cases 2 and 3, the objective functions are not linear. To some extent, this makes the conventional approaches in Case 1 ineffective. Furthermore, these nonlinear functions are convex in nature. This makes the approximation approach almost impossible. With these difficulties, a GA approach is demonstrated in the following sections.

INTRODUCTION TO GENETIC ALGORITHMS

GAs are search algorithms based on the mechanics of natural selection and natural genetics. They combine survival of the fittest among string structures with a structured, yet randomized, information exchange to form a search algorithm with some of the innovative flair of human search. In every generation, a new set of artificial creatures (strings) is created using bits and pieces of the fittest of the old. An occasional new part is tried for good measure. Although effective, GAs can be quite simple in their application, they efficiently exploit historical information to speculate on new search points with expected improved performance (*4*).

In the following section, the GA framework is introduced through a simple optimization problem (SOP):

$$\text{MAX} \quad 1024 - (x - 31)^2 \tag{4}$$

TABLE 3 Solution Strings and Fitness Values

No.	x	String	Fitness	% of Total	Expected Count
1	5	000101	348	9.4	0.566
2	42	101010	903	24.5	1.469
3	53	110101	540	14.6	0.878
4	38	100110	975	26.4	1.586
5	61	111101	124	3.4	0.202
6	16	010000	799	21.7	1.300
Total			3689	100	
Average			614.8		

Subject to $0 \leq x \leq 63$ (5)

where x is an integer.

First, The bit string representation is implemented by the widely used list of 0's and 1's. Table 3 shows the binary strings for six solutions: 5, 42, 53, 38, 61, and 16. The evaluation function is the same as Equation 5, and fitness values for the six solutions are also listed in Table 3. Second, these six solutions are assumed to be the first generation. Third, to generate the offspring generation, three commonly used operators are employed: reproduction, crossover, and mutation. These three operators are applied, in turn, to the solutions in the current generation during the search process.

The first operator, reproduction, is a process in which good solutions survive and are retained and bad solutions die. The number of solutions reproduced by each original solution is proportional to its fitness value. For instance, referring to Table 3, String 1 has a fitness of 348, which represents 9.4 percent of total fitness of the population of solutions. Therefore, its expected count is 0.566 which is obtained through dividing its fitness by the average fitness. Hence, using the population shown in Table 3 as parents, a possible population generated by reproductions is shown in Table 4, in which String 4 in Table 3 produces two solutions whereas String 5 produces none.

The second operator, crossover, is generally performed on the population newly generated by reproduction. The crossover proceeds in two steps. First, members of the reproduced strings are mated at random. Second, the two solutions in each solution pair exchange their "chromosomes" which are represented by an alphabetic string. Suppose in Table 4, String 1 is mated with String 2, String 3 with String 4, and String 5 with String 6. If k is 1, 4, and 2 for these three pairs, respectively, the population generated will be as shown in Table 5, in which the crossover sites are denoted by "|."

After crossover, a mutation operator is used. This operation works on a bit-by-bit basis. It simply changes every bit (or character) in every solution string in the population to its opposite bit

TABLE 4 Population Generated by Reproduction

No.	x	String	Fitness
1	5	000101	348
2	42	101010	903
3	53	110101	540
4	38	100110	975
5	61	111101	124
6	16	010000	799

TABLE 5 Population Generated by Crossover

No.	Old String	Mate	New String	x	Fitness
1	0|00101	1	001010	10	583
2	1|01010	1	100101	37	988
3	1101|01	2	110110	54	495
4	1001|10	2	100101	37	988
5	10|0110	3	100000	32	1023
6	01|0000	3	010110	22	943
Total					5020
Average					836.7

(or other character) with a very small probability. Suppose that, in this SOP, the mutation probability is .001. In this particular case, it is very likely that no mutation will be made to any bit and the resulting population will not differ from that shown in Table 5.

Finally, after the operations of reproduction, crossover, and mutation, the population of a new generation becomes those presented in Table 5. The average fitness value for the SOP has been increased from 614.8 in Table 3 to 836.7 in Table 5, and the maximum fitness has also increased from 975 to 1023, respectively.

GA FORMULATION TO THE TFP

Referring first to the conventional formulation, train formation decisions are represented by 0-1 strings which are illustrated in Figure 3. In this figure, $x_{i,k}$ is the decision variable from Equation 1 where each route for each demand has two itineraries.

Second, the evaluation functions derived for the three cases are the following:

$$BM - \left[\sum_{l=1}^{2L} t(Y_l)\, Y_l + \sum_{j=1}^{N} v(X_j)\, X_j\right] \tag{6}$$

$$BM - \left[\sum_{l=1}^{2L} t_l Y_l + \sum_{j=1}^{N} v_j X_j\right] \tag{7}$$

$$BM - \left[\sum_{l=1}^{2L} t(Y_l)\, Y_l + \sum_{j=1}^{N} v(X_j)\, X_j + \sum_{l=1}^{2L} f(Y_l - b_l) + \sum_{j=1}^{N} g(X_l - b_2)\right] \tag{8}$$

where BM is used to convert the minimizing objective functions to maximizing. The variables f and g are penalty functions.

Third, the GA operations are designed as follows: 1) The initial set of solutions are generated randomly and 2) The operations of reproduction and mutation are the same as those demonstrated in the SOP. However, with regard to the constraints represented in Equation 2, the mate sites in the crossover operation are selected uniformly at random from a specific set of positions, instead of from a set of consecutive numbers like that in the SOP. In this way, the constraint in Equation 2 can be guaranteed automatically in the genetic algorithm operation.

CALIBRATION OF THE GA MODEL

The calibration process for the GA model is to find the appropriate parameters by which the best solutions of the GA model can be obtained. These parameters include the size of the population, the number of generations, the crossover probability, and the mutation probability. In order to quicken the calibration, it is decomposed into two steps. The first step considers only the first two parameters. When generating the schemes, only the first two parameters vary within certain ranges, whereas the last two parameters are fixed at 0.9 and 0.03, respectively. From this step, the optimal number of generations and population size are determined. Given these determined values, the second step generates schemes by varying the last two parameters in a certain range. From this step, the optimal values for crossover and mutation probability are obtained. This process is conducted for all three cases. The details of the validation are described as follows.

In the first step, the generations are set at 100, 200, . . . , 1000, respectively. The population sizes are set at 10, 20, . . . , 100, respectively. Then, for each case, 100 schemes will need to be generated and evaluated. After a rough scanning of all the results, it is determined that the generation of 1000 is the most appropriate to evaluate the performance of the GA model. Then, the remaining task is to investigate the influence of the population size on the search process. The results are plotted in Figure 4 for Case 1. Similar plots were generated and used for Cases 2 and 3. The population sizes are determined by two criteria: the time the GA model uses to decrease the objective function values to the best solution and the stability after the best solutions have been achieved. For some processes, the convergence from the initial objective function value to the optimum is rather quick. On the other hand, other processes will fluctuate around the optimum value. The optimal population size is found to be 10 for Case 1, 100 for Case 2, and 70 for Case 3.

Following the procedures for Step 2, the results listed in Table 6 are obtained. The crossover probabilities are set at 0.6, 0.7, 0.8, 0.9, and 1.0 respectively. The mutation probabilities are set at 0.01, 0.02, 0.03, 0.04, and 0.05, respectively. For ease of analysis, the solution

Demand i				Demand i+1			
Route 1		Route 2		Route 1		Route 2	
itinerary k	itinerary k+1	itinerary k+2	itinerary k+3	itinerary k+4	itinerary k+5	itinerary k+6	itinerary k+7
$x_{i,k}$	$x_{i,k+1}$	$x_{i,k+2}$	$x_{i,k+3}$	$x_{i+1,k+4}$	$x_{i+1,k+5}$	$x_{i+1,k+6}$	$x_{i+1,k+7}$

FIGURE 3 GA string representation of train formation decision.

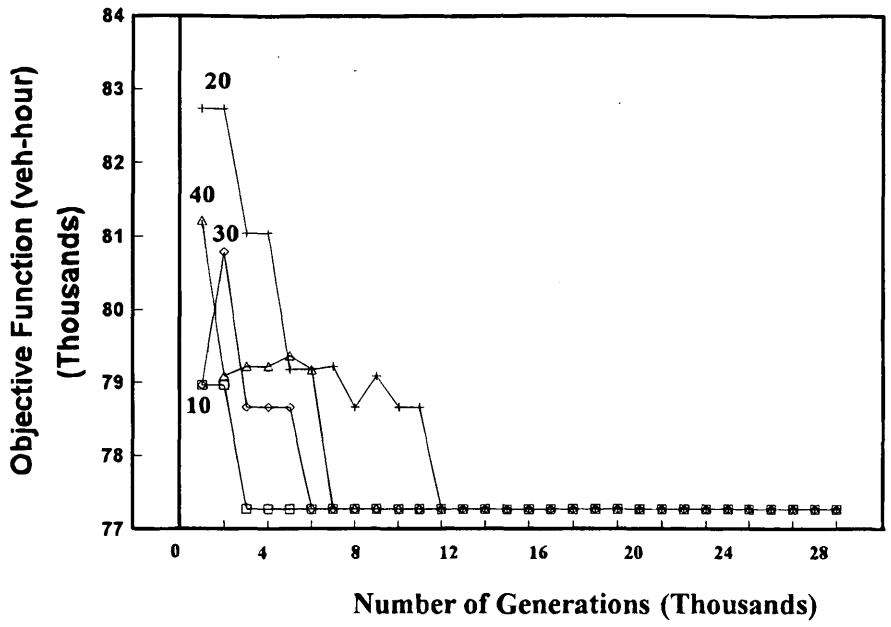


FIGURE 4 Determination of population size for Case 1.

TABLE 6 Calibration for Crossover and Mutation Probability in Cases 1, 2, and 3

Mutation Probability	Crossover Probability Case 1				
	0.6	0.7	0.8	0.9	1.0
0.01	3	2	2	2	2
0.02	3	2	2	2	2
0.03	4	3	3	3	3
0.04	4	2	3	3	3
0.05	8	2	4	4	10
Mutation Probability	**Crossover Probability Case 2**				
	0.6	0.7	0.8	0.9	1.0
0.01	19	10	16	16	(3)
0.02	46	(3)	186	(2)	19
0.03	(2)	160	73	(2)	(3)
0.04	853	117	73	48	81
0.05	(3)	(4)	(2)	(3)	(4)
Mutation Probability	**Crossover Probability Case 3**				
	0.6	0.7	0.8	0.9	1.0
0.01	(3)	(2)	46	(3)	(3)
0.02	(4)	92	32	(3)	(3)
0.03	72	43	277	529	834
0.04	(2)	121	(2)	123	(2)
0.05	(3)	(4)	(2)	(2)	(2)

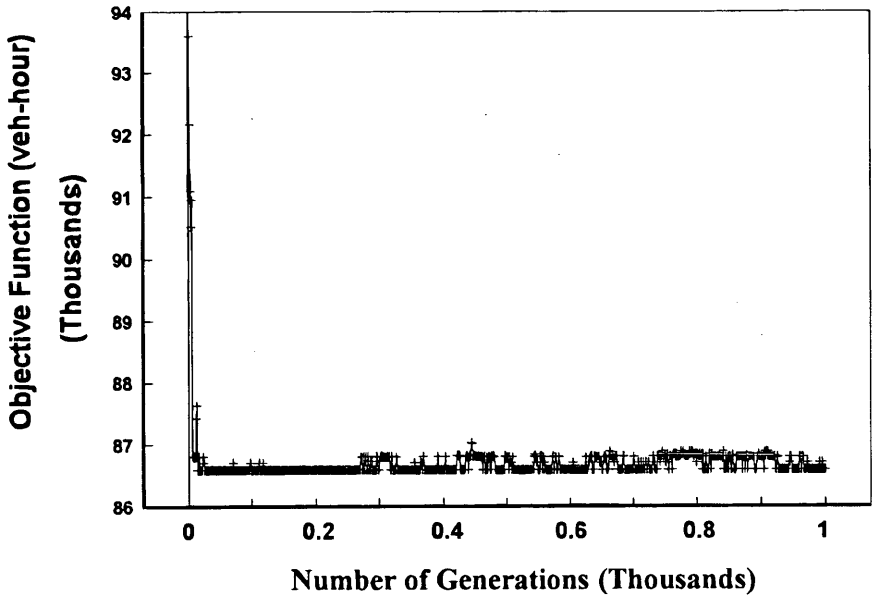


FIGURE 5 Search process: Pattern 2.

searching processes are classified into four patterns. In Pattern 1, the searching processes are stable after the smallest values are found. The generation at which the smallest values are found is called the stable generation. This pattern is viewed to have the best performance. Pattern 2 and Pattern 3, which are represented in Figures 5 and 6, respectively, are similar in the solution search processes. Both patterns indicate that the search processes will fluctuate after the smallest objective function value is achieved. However, in Pattern 2, the search process stays at the convergence status for a longer time than that in Pattern 3. Further comparing with Pattern 4, which is represented in Figure 7, the extent of fluctuation in Pattern 2 and 3 is smaller than that in Pattern 4. Among these four patterns, Pattern 1 shows a strong ability to keep the smallest values they achieved. Pattern 4 is the worst condition.

Referring to Table 6, the number outside parentheses represents the stable generation in Pattern 1, and the numbers in parentheses represent the designation of patterns. The crossover and mutation probabilities are determined by the corresponding row and column values of the cell which have the smallest stable generation. In Case 1, the crossover and mutation probabilities are determined to be .7 and .01, respectively. For Case 2, corresponding the stable generation of 10, they are determined to be .7 and .01, respectively. For the Case 3, corresponding to the stable generation of 32, they are determined to be .8 and .02, respectively.

Referring to Table 6, it can be seen that the linear case (Case 1) involves fewer generations to obtain the optimal solution than the nonlinear cases (Case 2 and Case 3). In Case 1, regardless of the parameters, the GA model always obtains the optimal solutions. In

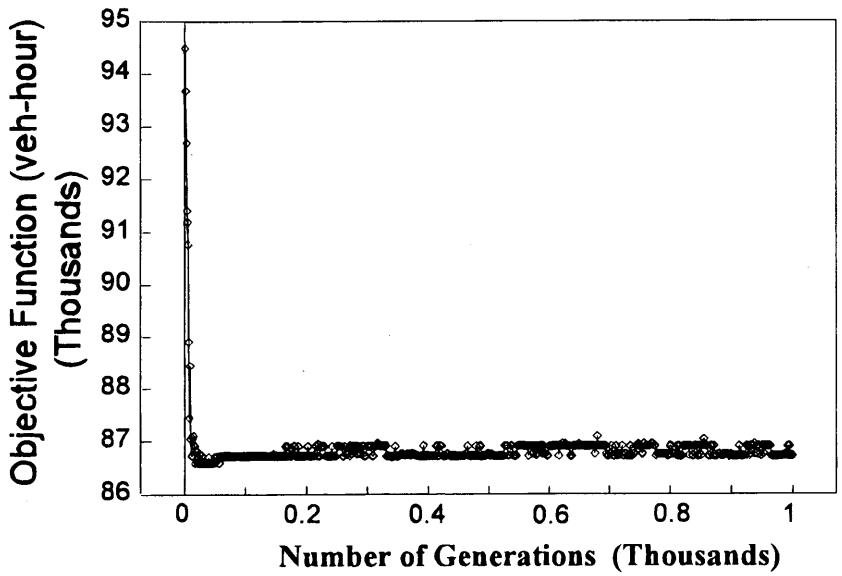


FIGURE 6 Search process: Pattern 3.

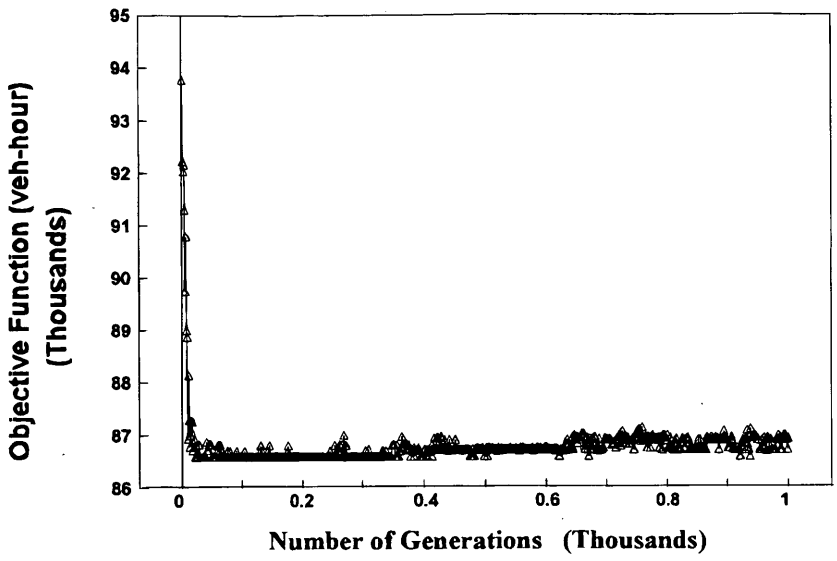


FIGURE 7 Search process: Pattern 4.

Case 2, there are 11 schemes which cannot yield the optimal solution; in Case 3, there are 15 such schemes. Furthermore, from Table 6, it can be seen that additional constraints in Case 3 make the solution searching process longer. In Case 2, there are 14 stable schemes, and the average stable generation is 123, whereas in Case 3, there are 10 such schemes with an average of 217.

VALIDATION OF THE GA MODEL

The task of the validation process of the genetic algorithm model is to test whether the obtained solutions are optimal. In this study, different validation methods are employed for different cases. For Case 1, because the optimal solutions can be obtained from available software developed by conventional algorithms, the solutions from the GA model are compared with them. For Cases 2 and 3, the optimality of the solutions from the GA model are evaluated by observing whether the solutions have obeyed the constraints imposed by the nonlinear delay time functions and imposed by some practical considerations. After the feasibility of the solutions is determined, a variety of combinations around the obtained solutions are tested to see whether smaller objective values can be achieved.

In Case 1, the optimal solution is obtained by using Quant Systems (Version 2.1). This solution is compared with that obtained from the GA model. The solutions are the same, except that GA models can also produce other optimal solutions when multiple solutions exist. In each case, all the solutions are optimal, whereas the itinerary choices are different. The train loads and link volumes from one of the optimal solutions are listed in Tables 7 and 8.

In Case 2, the link volumes should be as small as possible, because convex links and train performance functions are used. From the point of view of networks, however, the link volumes should not fluctuate dramatically. Observe Table 8, comparing with the solution of Case 1, the link volumes in Case 2 are indeed evenly distributed. The largest value in Case 1, 508, disappeared in Case 2. Referring to Figure 3, the load on the trains should be as much as possible. In Table 7, all of the trains are assigned as large a load as possible. From these two observations, it is apparent that the solutions are feasible and the nonlinear functions are effective. Following this analysis, the solutions are analyzed by providing all possible and comparable schemes. It is concluded that the solutions are truly optimal in terms of minimizing the total delay in the railroad system.

In Case 3, beside the nonlinear objective function, the remaining two constraints are added, that is, the volume on each link should

TABLE 7 Train Loads of the Three Cases for Validation

	The Load Assigned for the Long Distance Train											
	T114	T214	T115	T215	T116	T216	T316	T416	T126	T226	T136	T236
Case 1	121	0	0	150	150	0	0	0	0	107	150	0
Case 2	121	0	150	0	0	0	150	0	107	0	0	150
Case 3	0	0	0	0	0	0	0	0	0	0	0	0
	The Load Assigned for the Long Distance Train											
	T141	T241	T151	T251	T161	T261	T361	T461	T162	T262	T163	T263
Case 1	0	150	0	136	0	0	0	150	150	0	140	0
Case 2	0	150	136	0	150	0	0	0	0	150	0	140
Case 3	0	0	0	0	0	0	0	0	0	0	0	0

TABLE 8 Link Volumes of the Three Cases for Validation

	Traffic Volume on Each Link of Each Direction (car)									
	link 12	link 21	link 13	link 31	link 24	link 42	link 35	link 53	link 25	link 52
Case 1	335	78	244	508	298	211	164	375	161	38
Case 2	335	364	244	222	255	211	164	229	204	324
Case 3	335	364	244	222	255	211	164	229	204	324
	Traffic Volume on Each Link of Each Direction (car)									
	link 34	link 43	link 46	link 64	link 56	link 65	link 23	link 32	link 45	link 54
Case 1	154	319	334	357	206	176	87	95	10	82
Case 2	154	169	291	217	249	316	87	95	10	82
Case 3	154	169	291	217	249	316	87	95	10	82

be no more than 300 cars (which is realized by setting b_1 equals 300), and the load on each long distance train should be more than 200 cars (which is realized by setting b_2 equals 200). These two constraints establish that the possible load on each of the long distance trains are not sufficiently large to justify the provision. In Table 7, the loads are really zero, whereas in Table 8, it appears that the link volume constraints are not effective. However, after careful calculation, the overall demand through those links where the volumes exceed 300, it is found that there is no way to distribute these volumes without avoiding the penalty of the violation of the constraints. Thus, the solution is truly the optimal.

GA MODEL COMPUTATIONAL PERFORMANCE

In Case 1, the Quant Systems consumes 1.17 sec of CPU to produce the optimal solution. However, the GA model uses less time.

In Cases 2 and 3, for the number of generations equal to 1000, the GA model requires approximately 10 min of CPU. Because both cases can obtain the optimal solutions in less than 40 generations, the computation time should be about 20 sec. Comparing with the size of the problem (10^{17}), this computation time is quite satisfactory.

Using the calibrated parameters, the GA model is used for varieties of demand patterns. In Table 2, the long distance demands are varied in the range of 100 to 150 cars; there are almost no computation time variations. For all the demand patterns, the GA model produces the optimal solutions within 40 generations.

CONCLUSIONS

Several conclusions can be derived from this study. First, a GA model is able to produce optimal solutions for the formulations which might be difficult conventionally. Also, the computation time is satisfactory. Second, a GA model is not as sensitive to the input patterns compared to Monte Carlo algorithms. Third, the implementation process for a GA model is straight forward. In all three cases, the implementation simply involves the adjustment of the objective function formulations. There is no need to give the structure of the formulation a special consideration. The calibration and validation process are also straight forward. Fourth, the binary representation for the binary integer program (BIP) is especially effective.

Based on the principle introduced in this study, GA models can likely be effective when applied to large railroad networks. The patterns recognition of the solution searching process needs to be analyzed quantitatively instead of qualitatively, however. To this end, some statistical model might need to be developed.

ACKNOWLEDGMENT

The authors wish to thank Lijuan Su, who assisted in the data generation and processing tasks of this research.

REFERENCES

1. Martinelli, D. R., and H. Teng. A Neural Network Approach for Solving the Train Formation Problem. Presented at 73rd Annual Meeting of the Transportation Research Board, Washington, D.C., 1994.
2. Xiong, Y., and J. B. Schneider. Transportation Network Design Using a Cumulative Genetic Algorithm and a Neural Network. Presented at 71st Annual Meeting of the Transportation Research Board, Washington, D.C., 1992.
3. Foy, M. D., R. F. Benekohal, and D. E. Goldberg. Signal Timing Determination Using Genetic Algorithms. In *Transportation Research Record 1365*, TRB, National Research Council, Washington, D.C., 1993, pp. 108–115.
4. Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, Mass., 1987.

Publication of this paper sponsored by Committee on Artificial Intelligence.

Evolutionary Neural Network Model for the Selection of Pavement Maintenance Strategy

MAHMOUD A. TAHA AND AWAD S. HANNA

Neural networks are attracting an enormous amount of attention in many civil engineering disciplines, including transportation, because they represent a class of robust, nonlinear models capable of learning relationships from data. However, in the development of such models for a particular application, various parameter settings are left to the judgment of the network developer. The net result of poor parameter settings will be slow convergence and/or bad performance on unseen cases. Recently, genetic algorithms have emerged as a potential searching technique to design a neural network model that performs best on a specified task according to explicit performance criteria. Genetic algorithms are search algorithms based on the mechanics of natural selection and natural genetics. In this paper we present a genetic algorithm method that evolves a neural network model for the selection of the optimum maintenance strategy for flexible pavements. A hybrid evolutionary-learning system using gradient descent learning as well as a genetic algorithm to determine the network connections weights is described. The developed neural network model has an input vector of seven components and an output vector of seven components. The input vector represents the factors affecting the maintenance strategy selection, whereas the output vector represents the different pavement maintenance strategies available. Brainmaker Professional, a commercially available neural network simulator, was used in the development of the neural network model. The performance of the developed neural network model was validated by testing it using 100 unseen cases. The validation results showed that the system misclassified only six cases with an average error rate of 0.024.

Decision-making in most civil engineering disciplines, including transportation, frequently encounters complicated and unstructured problems for which solutions are devised based on analogy with previous cases with a mixture of intuition and experience. Selection of the appropriate pavement maintenance strategy represents one such problem. The ability of decision makers to find an adequate solution to this problem depends primarily on their accumulated experience. To lessen the dependency on experienced personnel and to improve the consistency of the decision-making process, decision-making aids are required. Neural networks have been recommended by many researchers as a suitable tool for developing such decision aids (*1*,*2*). This is attributable to their ability to learn mappings from a set of inputs to a set of outputs based on training examples and to generalize beyond the examples learned. However, experience with neural networks for learning different tasks has demonstrated the difficulty of selecting an appropriate functional structure for a network as well as appropriate values for learning rule parameters. This bottleneck may seriously impair neural networks progress in the coming years if it is left unaddressed.

Recently, genetic algorithms (GAs) have emerged as a potential searching technique to craft a neural network application that performs best on a specified task according to certain explicit performance criteria. According to Austin (*3*), the GAs can be defined as:

> [A]n iterative procedure maintaining a population of structures that are candidate solutions to specific domain challenges. During each temporal increment (known as generation), the structures in the current population are rated for their effectiveness as domain solutions, and on the basis of these evaluations, a new population of candidate solutions is formed using specific "genetic operators" such as reproduction, crossover, and mutation.

In this paper we present a genetic algorithm approach that evolves a neural network model for the selection of the optimum maintenance strategy for flexible pavements. The backpropagation learning method (*4*) is combined with genetic algorithms to evolve the optimum interconnection weights. BrainMaker Professional, a commercially available neural network simulator, was used in the development of the neural network model.

THE PROBLEM OF NEURAL NETWORKS DESIGN

The process of developing a neural network model for a particular application usually involves four basic stages. First, a network developer selects a problem domain, such as pavement maintenance, based on his or her theoretical, empirical, or applied interests. Next, a network architecture is designed for capturing the underlying criteria from the problem domain. This architecture forms the configuration of the network including the number of units used, their organization into layers, learning parameters, and error tolerance. Third, given this chosen architecture and a chosen task, a learning paradigm is applied to train the network and develop the interconnection weights. Finally, the developer evaluates the trained network according to objective performance measures such as its ability to solve the specified task and its ability to predict the outcome of unseen cases.

Learning takes place in neural networks by adjusting the connection weights between simple processing units. This kind of computation is best understood as a kind of relaxation system (*4*). Most learning procedures perform a search over the weight space to minimize some performance function of the network. The "Boltzmann Machine," a widely used learning technique, uses simulated annealing. Unfortunately, simulated annealing is very slow. A much faster learning procedure is backpropagation. This method is getting the most attention in the neural network research community. However,

M. A. Taha, Construction Administration Program, University of Wisconsin-Madison, 460 Henry Mall Street, Madison, Wis. 53706. A. S. Hanna, Department of Civil and Environmental Engineering, University of Wisconsin-Madison, 1415 Johnson Drive, Madison, Wis. 53706.

this method also has limitations such as slow training and the possibility of getting a solution that is a local minimum, among others. This paper proposes an alternative technique to adjust the network's weights by using a biologically based optimization method.

MERGING GENETIC ALGORITHMS WITH NEURAL NETWORKS

Genetic algorithms are search algorithms based on the mechanics of natural selection and natural genetics (*5*). They imitate nature with their Darwinian survival-of-the-fittest approach. This approach allows genetic algorithms to speculate on new points in the search space with expected improved performance by exploiting historical information. A simple genetic algorithm responds to some function evaluation. It does not associate an output with an input (*6*). Neural networks, on the other hand, are good at associating different patterns once they have developed an internal representation. The problem with neural networks is developing the "proper" internal representation among the connections. Because GAs are best viewed as a function optimizer, it is suited to finding the set of weights that allow the neural network to solve a given problem.

The advantages of using GAs to evolve the neural network weights are twofold. First, GAs represent a global search method. Backpropagation, which is the most widely used search method to develop the network internal representations, is a local optimization method (*7*). Backpropagation involves a gradient descent in sum-squared error that minimizes the squares of the differences between the actual and the desired output. Second, GAs may be capable of much faster optimization. The procedures for evolving an optimum neural network's set of weights for a particular application are outlined in Figure 1 (*8*).

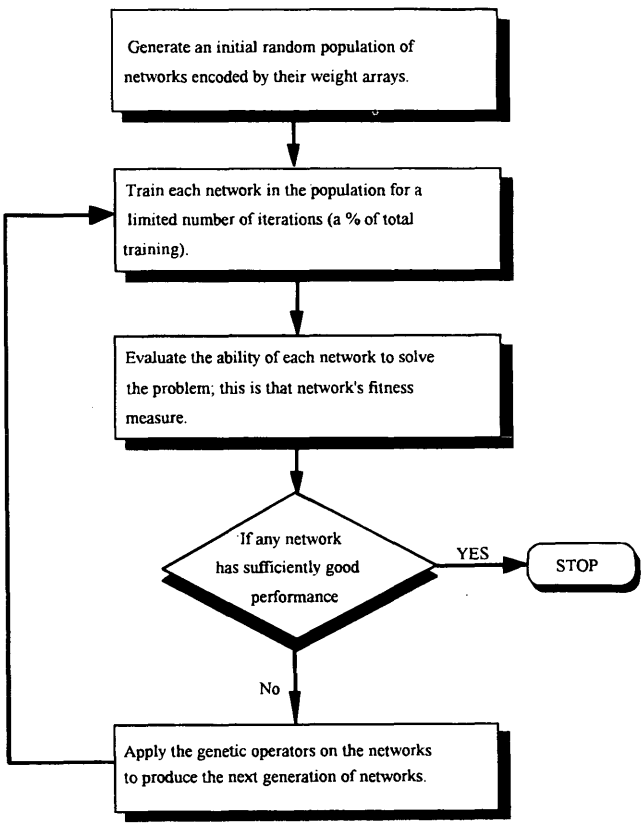


FIGURE 1 Procedures for optimizing weights of neural networks using GAs.

MODEL DEVELOPMENT

The following methodology has been used to illustrate the development of the present model. It includes three main phases: (a) problem definition; (b) model evolution; and (c) running the model for direct problem solving. Each of these phases is described in the following section.

Problem Definition

Pavement maintenance is defined as an action taken to correct deficiencies that are potentially hazardous and to repair defects that seriously affect serviceability to maintain or keep the pavement within a tolerable level of serviceability (*9*). Maintenance of most asphalt pavement involves repairing localized problem areas to prolong the pavement life. The proper maintenance of any highway system depends on the methods that the responsible agency uses to meet the climatic conditions in its jurisdiction.

Determining the best maintenance strategy starts by identifying the type, severity, and density of pavement distress. Pavement distress is defined as the condition of pavement structure that reduces serviceability or leads to a reduction in serviceability (*10*). Severity is measured using a three-point scale (slight, moderate, and severe) and density using a two-point scale (few and extensive). Typically, the decisions regarding the selection strategies are made by experienced senior practitioners. In order to lessen the dependency on these experts and to help less experienced practitioners to participate in this decision-making process, a neural network system is very useful.

Neural networks can fit this problem because of their adaptable structure. This is because of the existence of hidden layers and the nonlinear activation function in their structure permits them to make reasonable generalization. This important property of neural networks enables the performance of complex multiattribute, nonlinear mapping for the selection of the optimum pavement maintenance strategy.

Model Evolution

The process of evolving the neural network model comprises five main aspects: (a) developing network training and testing examples; (b) development tool; (c) genetic operators; (d) fitness evaluation; and (e) training and testing (validating). Each of these aspects is described in the following sections.

Developing Network Training and Testing Examples

Training and testing examples represent the most important ingredient in the development of a neural network model. They consist of all input and output data used by the neural network's learning algorithm. Without those examples, the network would not be able to learn anything about the problem. A given training example con-

sists of two components: (a) a data case consisting of a set of attributes, each with an assigned value, and (b) the corresponding correct class membership or the classification decision made by a domain expert according to the given data case. These examples can be assembled in several ways: (a) they may come from an existing database that forms a history of observations; (b) they may be a carefully culled set of tutorial examples prepared by domain experts; (c) they may be obtained from simulation results; or (d) they may be obtained from literature available from and interaction with domain experts.

In the present work we used the readily available knowledge acquired by Hanna et al. (*9*). This knowledge was acquired to build PMAS, a knowledge-based expert system for assisting highway engineers in planning effective flexible or asphalt concrete pavement maintenance strategies. This knowledge is available in the form of IF-THEN-ELSE rules. The examples used for developing the neural network model were compiled from these rules. The antecedents of the rules make the inputs, and the consequent of each rule makes the outputs of the examples.

In these examples the components of the input vector represent the factors affecting the maintenance strategy selection, which are identified as (a) distress type, (b) density of distress, (c) riding comfort index (RCI), (d) traffic volume, (e) climate, (f) crack type, and (g) severity of distress. Each of the first five input factors is represented by two binary neurons. Each of these neurons takes only one value, either 1 or 0. The sixth and the seventh input factors are represented by three binary neurons. Table 1 shows the possible values for the input neurons corresponding to each input factor. The components of the output vector represent the different pavement maintenance strategies available, which are identified as (a) do nothing, (b) crack seal coating, (c) route and seal, (d) cold mix patching, (e) hot mix patching, (f) hot mix recycled patching, and (g) reconstruction. Seven binary neurons were chosen to represent each output as shown in Table 2.

A total of 335 examples were developed from the available knowledge base. The whole set of examples was divided randomly into a training set of 235 (about 70% of the cases) examples and a test set of 100 examples (about 30% of the cases). A portion of the training and test examples is shown in Table 3.

Development Tool

The commercially available neural network simulator BrainMaker Professional was used for the development of the proposed neural network application. The Brainmaker package forms a complete

TABLE 1 Representation of Input Factors

Input Factor	Input Case	Representation		
		Neuron #1	Neuron #2	Neuron #3
(1)	(2)	(3)	(4)	(5)
Distress Type	Single	0	1	N/A
	Combined	1	0	
Distress Density	Few	0	1	N/A
	Extensive	1	0	
RCI*	< 4	0	1	N/A
	≥ 4	1	0	
Traffic Volume	≤ 2000 VPL**	0	1	N/A
	> 2000 VPL	1	0	
Climate	Coastal (DDI*** < 600 C*days)	0	1	N/A
	Inland (DDI ≥ 600 C*days)	1	0	
Crack Type	Alligator	1	0	0
	Rutting	0	1	0
	Traverse	0	0	1
	Alligator + Rutting	1	1	0
	Rutting + Traverse	0	1	1
	Alligator + Traverse	1	0	1
	Alligator + Rutting + Traverse	1	1	1
Distress Severity	Slight	1	0	0
	Moderate	0	1	0
	Severe	0	0	1

* Riding Comfort Index ** Vehicle Per Lane *** Degree Days Index

TABLE 2 Output Representation

Maintenance Strategy (1)	Representation (2)						
Do Nothing	1	0	0	0	0	0	0
Crack Seal Coating	0	1	0	0	0	0	0
Rout and Seal	0	0	1	0	0	0	0
Cold Mix Patching	0	0	0	1	0	0	0
Hot Mix Patching	0	0	0	0	1	0	0
Hot Mix Recycled Patching	0	0	0	0	0	1	0
Reconstruction	0	0	0	0	0	0	1

TABLE 3 Portion of Training and Testing Examples

Distress Type (1)		Distress Density (3)		RCI (2)		Traffic Volume (4)		Climate (5)		Crack Type (6)		Distress Severity (7)		Maintenance Strategy (8)
(a) Training Examples														
0	1	0	1	1	0	1	0	0	1	0	1	0	1	0	Do Nothing
1	0	0	1	0	1	1	0	1	0	1	0	1	0	0	Do Nothing
0	1	1	0	1	0	0	1	1	0	0	0	1	1	0	Crack seal coating
1	0	0	1	1	0	0	1	0	1	0	1	1	0	1	Crack seal coating
0	1	0	1	0	1	1	0	1	0	0	0	1	0	1	Rout and Seal
0	1	1	0	0	1	0	1	1	0	0	0	1	0	1	Rout and Seal
0	1	0	1	1	0	1	0	0	1	1	0	0	1	0	Cold Mix Patching
0	1	1	0	1	0	0	1	0	1	0	1	0	0	1	Cold Mix Patching
0	1	0	1	1	0	0	1	0	1	0	0	1	0	1	Hot Mix Patching
1	0	1	0	0	1	1	0	0	1	1	1	1	1	0	Hot Mix Patching
0	1	1	0	1	0	0	1	0	1	0	1	0	0	1	Hot Mix Recycled Patching
0	1	1	0	1	0	1	0	0	1	0	1	0	0	1	Hot Mix Recycled Patching
0	1	1	0	0	1	0	1	0	1	1	0	0	0	1	Reconstruction
1	0	1	0	0	1	0	1	1	0	1	1	0	0	1	Reconstruction
(b) Test Examples														
0	1	1	0	1	0	1	0	1	0	0	0	1	1	0	0	Do Nothing
1	0	0	1	0	1	1	0	0	1	0	1	1	0	0	1	Crack Seal Coating
0	1	0	1	1	0	0	1	1	0	0	0	1	0	0	1	Rout and Seal
0	1	1	0	0	1	1	0	0	1	0	1	0	0	1	0	Cold Mix Patching
0	1	0	1	1	0	0	1	0	1	1	0	0	0	0	1	Hot Mix Patching
1	0	1	0	0	1	0	1	1	0	0	1	1	1	0	0	Hot Mix Recycled Patching
1	0	1	0	1	0	1	0	0	1	1	1	0	0	0	1	Reconstruction

system for designing, building, training, testing, and running neural networks on IBM personal computers and compatibles. The Genetic Training Option (GTO) of Brainmaker is used to optimize the weights of the developed neural network model. It applies Darwin's theories of genetic mutation and natural selection. GTO creates a large number of subtly different networks to do the same job. It then tests, trains, and ranks them to find the network(s) that perform(s) the best overall according to the user definition of the "best." It has the capability to keep up to 10 best networks at the end of each run.

The GTO "genetic evolution" works on the neuron connections of a trained network using genetic operators (*11*).

Genetic Operators

The genetic operators used in this study were mutation and crossover. The mutation requires only one parent. The mutation operator allows new genetic sequences to be introduced. This is

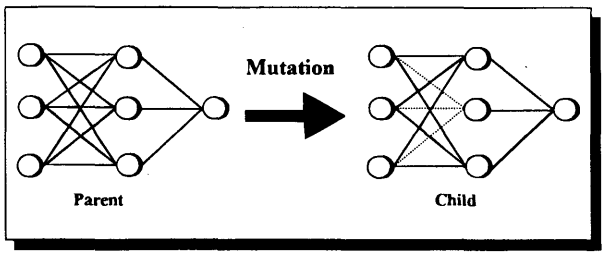
done by changing a random percentage of the neurons by modifying the weights associated with them. In the example shown in Figure 2*a*, the child is produced from the parent network by changing the connection weights from the input layer to the second hidden neuron. On the other hand, the crossover operator allows "sexual" reproduction by combining two parents to produce an offspring. It is implemented by taking some neurons from one "parent" network and some from another to produce an offspring. In the example shown in Figure 2*b*, the child receives the second hidden neuron and output neuron from the second parent and the first hidden neuron from the first parent.

Fitness Evaluation

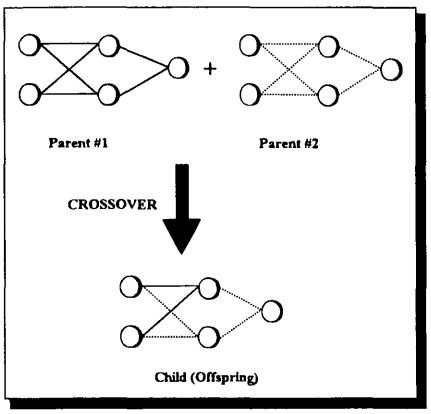
The GAs use an objective function to evaluate the performance of the members of each generation. In our case the members of each generation are neural networks, and the task is to select the optimum pavement maintenance strategy. The objective function is a mapping from the weight space of a particular neural network to a single value. This objective function should possess some degree of smoothness in the region about the solution point in the weight space (*7*). This means that any change in the weights in the direction of the optimum should yield a higher performance value.

In our work the fitness of each network is measured by testing its performance on unseen cases. The network with the lowest average error on all test cases will have a higher chance to survive and to produce the next generation. This average error is calculated using the following equation:

$$\text{Average Error} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{o}|O_{ij} - P_{ij}|}{N} \quad (1)$$



(a) Mutation Operator



(b) Crossover Operator

FIGURE 2 Genetic operators.

where

O_{ij} = Calculated output value
P_{ij} = Desired output value
N = Number of test cases
o = Number of output neurons

Training and Testing

Before any particular run, the network topology is specified. A neural network model with 16 input neurons in the input layer, 16 hidden neurons in the middle layer, and 7 neurons in the output layer is used. The number of hidden neurons is determined based on the rule of thumb that suggests that this number is equal to the number of neurons in the input layer. The learning rate parameter is assumed to be constant over all of the network connections with a value of 1.0.

The training procedure starts by generating an initial randomized weight matrix using BrainMaker Professional. This initial weight matrix is loaded into GTO. The genetic evolution process is used to save the best 10 networks over 50 generations. During this evolution process, the mutation and crossover operators are applied to 10 and 70 percent of the hidden neurons, respectively. The evolution process is started by applying the mutation operator to the initial network to create another parent network. The two parent networks are then crossed over to produce an offspring. These networks are trained using backpropagation for 100 epochs (an epoch is one presentation of the whole training set). Because network generalization is a main criterion for optimization, the networks are tested using 100 unseen cases. The test results are presented in the form of average error rate. If the child network outperforms its parents, it will be saved and used to replace its parents. Otherwise, the best parent will be saved and used to start a new generation.

Because the program can keep up to 10 networks, the best networks in the first 10 generations will be saved automatically. For the following generations, the average error rate on unseen cases of the best network of each generation (ϵ_{new}) will be compared with the highest average error rate on unseen cases of the 10 saved networks (ϵ_{old}). If ϵ_{new} is found to be lower than ϵ_{old}, the new network will replace that one. This procedure will be repeated until no improvement in the average error rate is achieved. In our case no improvement was achieved after 50 generations.

The runs were conducted on a powerful DX2-66 MHz IBM compatible. The results of these runs are summarized in Figure 3, which shows the best-of-generation average error for three independent runs with different initial populations. The plot indicates that the error value decreases with successive generations. Moreover, the best 10 networks generated in this evolutionary process are tested by presenting test examples to them. The test results are summarized in Table 4. These results show that network number 5 provided a better prediction ability (average error rate attributable to test on unseen cases was 0.024; misclassified only 6 out of 100 unseen cases) than the other nine networks. On the basis of these results, this neural network model was chosen as a final solution to our problem.

Also, the initial neural network model used in the evolution process is trained using backpropagation only using the same training set. After presenting the training examples 1706 times, the training error converged to 0. This network is saved, and its generaliza-

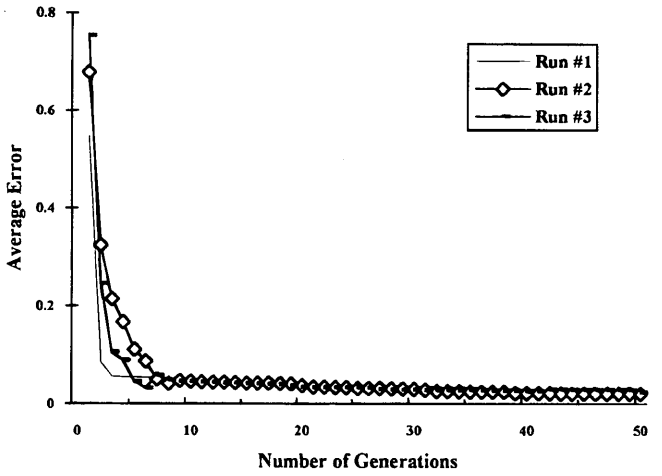


FIGURE 3 Best-of-generation average error rate versus generation number.

tion ability is checked using the same test cases used above. The test results show that the generalization ability of this network was worse than that of the evolutionary network; its average error rate was 0.043 (about 2 times more), and it misclassified 20 out of 100 test cases.

Running the Model for Direct Problem Solving

The developed neural network model is just a set of interconnection weights that are simple real numbers. It can be used to select the optimum maintenance strategy for new cases in two ways:

- If the user owns a copy of the Brainmaker simulator, he/she can prepare a fact file and then run Brainmaker and use this network to predict new cases included in the fact file.
- For a user who does not own a copy of this simulator, a query-and-answer program was developed, using C programming language, to help the end user to run the neural network model. The program will first prompt the user to input the factors affecting the selection process. The program then will translate these factors into the network input values as 1 or 0, propagate these inputs through the developed neural network with the obtained interconnection weights, and prompt the end user with the final recommendation. A copy of the data input screen is given in Figure 4.

SUMMARY AND CONCLUSIONS

Genetic algorithms are search procedures that are able to locate near-optimal solutions after examining only portions of the search space. They are gaining increasing popularity as a valuable tool to optimize many engineering problems. In this paper a simple genetic algorithm made up of mutation and crossover was combined with the backpropagation algorithm to find the optimum weights for a neural network model. The objective of the model is to select the optimum pavement maintenance strategy. The use of backpropagation in conjunction with genetic algorithms is not seen as just a way of fine tuning the rough solution generated by a genetic algorithm but, rather, as a way of sustaining diversity in the population in a nondestructive fashion. This is because the use of backpropagation will change the connection weights and thus introduce new genetic material that can be exploited by the genetic algorithm. The genetic training option of BrainMaker Professional neural network simulator was used in the development process. The details of the development and implementation of the model were presented. The results showed that combining genetic algorithms and backpropagation to develop the neural network weights helps in improving the network generalization ability compared with that obtained using backpropagation alone.

This paper provided a solution for the problem of finding the interconnection weights such that the network can compute a desired input to output mapping. It suggests that a fundamental area of research involves the problem of finding a suitable network topology. This includes number of hidden layers, number of hidden neurons per hidden layer, type of activation function, and learning parameters. Genetic algorithms may also be a good candidate for this problem. Finally, the approach described in this paper could be applied to the development of neural networks in other civil engineering domains.

TABLE 4 Validation Results (Unseen Cases)

Network Number (1)	Missclassified Cases (2)	% of Missclassified Cases (3)	Average Error (4)	Root Mean Square Error (5)
1	8	8	0.0250	0.1065
2	6	6	0.0244	0.1039
3	8	8	0.0252	0.1060
4	8	8	0.0252	0.1063
5*	**6**	**6**	**0.0240**	**0.1035**
6	8	8	0.0258	0.1079
7	6	6	0.0248	0.1044
8	12	12	0.0284	0.1255
9	7	7	0.02449	0.1051
10	8	8	0.0550	0.1062

*** Best Network**

Evolutionary Neural-Based System for
Selecting Optimum Pavement Maintenance Strategy

Please input the following information. You can type only the capital letter(s) included between the brackets for each piece of information. Upon the completion of each input hit <Return> key.

1. The type of distress observed is [(S) for single or (C) for combined] : ▮
2. The density of distress observed is [(F) for few or (E) for extensive] : ▮
3. Riding Comfort Index (RCI) is [(P) if RCI < 4 or (G) if RCI ≥ 4] : ▮
4. The traffic volume (AADT) is [(L) if < 2000 VPL or (H) if ≥ 2000 VPL]: ▮
5. The climate is [(C) for coastal or (I) for inland] : ▮
6. The type of crack observed is [(R) for rutting, (A) for alligator, (T) for traverse, (RA) for Rutting plus Alligator, (RT) for Rutting plus Traverse, (AT) for Alligator plus Traverse, or (RAT) for Rutting plus Alligator plus Traverse] : ▮
7. The severity of distress observed is [(S) for slight, (M) for moderate, or (V) for sever] : ▮

FIGURE 4 Data input screen.

REFERENCES

1. Flood, I., and N. Kartam. Neural networks in Civil Engineering. I. Principles and Understanding. *Journal of Computing in Civil Engineering*, Vol. 8, No. 2, 1994, pp. 131–148.
2. Moselhi, O., T. Hegazy, and P. Fazio. Neural Networks as Tools in Construction. *Journal of Construction Engineering and Management*, Vol. 117, No. 4, 1991, pp. 606–625.
3. Austin, S. An Introduction to Genetic Algorithms. *AI Expert*, March 1990, pp. 49–53.
4. Rumelhart, D. E., G. E. Hinton, and R. J. Williams. Learning Internal Presentations by Error Propagation. In *Parallel Distributed Processing: Explorations in the Micro Structures of Cognition* (D. E. Rumelhart and J. L. McClelland, eds.), Foundations, MIT Press, Cambridge, Mass. 1986, pp. 318–362.
5. Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Publishing Co., New York, 1989.
6. Whitley, D. *Applying Genetic Algorithms to Neural Network Learning*. Technical report CS-88-105, Department of Computer Science, Colorado State University, Fort Collins, 1988.
7. Smith, T. R., G. A. Pitrey, and D. Greenwood. Calibration of Neural Networks Using Genetic Algorithms with Application to Optimal Path Planning. *First Annual Workshop on Space Operations Automation and Robotics* (SOAR 87) (Sandy Griffin, ed.), 1987, pp. 519–526.
8. Caudill, M. Evolutionary Neural Networks. *AI Expert*, March 1991, pp. 28–33.
9. Hanna, P. B., A. S. Hanna, and T. A. Papagiannakis. Knowledge-Based Advisory System for Flexible Pavement Routine Maintenance. *Canadian Journal of Civil Engineering*, Vol. 20, 1993, pp. 154–163.
10. Thomas, W. K., L. R. Freddy, and J. B. Rauhut. Distresses and Related Material Properties for Premium Pavements. In *Transportation Research Record 715*, TRB, National Research Council, Washington, D.C., 1987, pp. 15–25.
11. California Scientific Software. *Getting Started With Brainmaker: User's Guide and Reference Manual*. California Scientific Software, Nevada City, 1993.

Publication of this paper sponsored by Committee on Artificial Intelligence.

Hybrid Artificial Intelligence Approach to Continuous Bridge Monitoring

DAVID MARTINELLI, SAMIR N. SHOUKRY, AND S. T. VARADARAJAN

A framework for an intelligent fusion of outputs from a combination of the state of the art nondestructive testing techniques of highway bridges and parking garages is described. A hybrid artificial intelligence system that automates the decision-making process is employed for the examination of the results obtained from various nondestructive testing methods. Neural networks, fuzzy logic, as well as other computational paradigms are employed for a preliminary pattern recognition and sensor data reduction from individual nondestructive testing techniques. An expert system is used for providing a user interface, selecting alternative paradigms for data analysis guided by the user input, as well as the information obtained from the system's knowledge base. The final decision-making process is performed via the expert system through the interpretation of the numerical results based on heuristic and knowledge-based reasoning. The system is divided into many independent modules that interact with each other. Any of the modules can be individually modified, retrained or replaced altogether without affecting any of the other modules.

The U.S. Department of Transportation classifies more than 40% of the U.S. bridges as structurally inefficient or functionally obsolete. Therefore, minimization if not prevention of bridge deterioration is a critically important task in avoiding unsafe conditions and in reducing repair cost. Recent attempts on bridge maintenance and rehabilitation are shifting toward early diagnosis and repair of flaws in bridges using nondestructive testing (NDT) techniques. The emphasis is to have a continuous bridge monitoring system in order to obtain quantitative information from in situ measurements (*1*) and use reliable data analysis techniques to arrive at rational and objective maintenance decisions.

NDT techniques have become an integral part of continuous bridge monitoring systems. NDT techniques such as ultrasonics, acoustic emission, ground penetrating radar, impact-echo, and infrared thermography are being increasingly used, and often a combination of various NDT methods is necessary for a complete and effective monitoring system (*2*). Each NDT technique has its own advantages and limitations, especially when used to detect defects in concrete (*3*). Ultrasonic testing of concrete is characterized by the inevitable presence of secondary reflections due to the heterogeneous nature of concrete. Acoustic emission (AE) signals are often contaminated with extraneous noise that makes interpretation extremely difficult and may mask the emission from the defect. Impact-echo is a point testing technique in the sense that its capability to detect a vertical crack or a crack which does not fall exactly below the sensor location is limited. Although radar is capable of testing large areas in a short time, radar signals suffer from multiple reflections and scatters which arise because of the presence of rebars close to the defects. In addition, possible variations in material properties and thickness of the asphalt overlay complicates the process of signal interpretation. The results obtained using infrared thermography are affected by the presence of oil on the road surface. In addition, the depth of an internal void or crack cannot be evaluated. Both radar and thermography results are also influenced by the environmental conditions. It can be observed that data collection and analysis techniques that allow the simultaneous use of information from two or more NDT techniques would be very useful in effective bridge monitoring.

Extraction of useful information from data obtained by the use of NDT techniques requires complex signal analysis and interpretation. The data from the sensors are often obscure and noise prone. Artificial intelligence (AI) techniques such as artificial neural networks (ANN) and expert systems are very useful in pattern recognition, classification, and qualitative interpretation of data obtained from NDT methods. In addition, they also help in synthesizing facts and heuristic knowledge to provide useful tools for problem solving and decision making.

In this study, we propose a hybrid AI system that automates analysis and interpretation of data from different NDT techniques. Neural networks and other computational paradigms are used for complex numerical pattern recognition and data classification tasks. An expert system is used for providing a user interface, selecting alternative paradigms for data analysis [guided by a knowledge base (KB) and the user input], and interpretation of the numerical results based on heuristic and knowledge-based reasoning. The system is divided into many independent modules that interact with each other; these modules can be modified or new modules added without affecting the rest of the system.

NEURAL NETS VERSUS KB SYSTEMS

Neural networks offer a high potential alternative to the traditional data processing and interpretation techniques. Neural networks are useful for mapping problems which are tolerant of high error rates and for problems to which mathematical relationships cannot be easily derived. Most often this is the case with sensor data obtained from bridges using NDT techniques which are affected to a great degree by environmental conditions such as weather and traffic. In most of the cases there are no mathematical relationships that relate NDT data to physical parameters, and therefore a data analyst normally uses some type of empirical modeling with neural networks or other paradigms. The ability to learn from ambiguous or contradictory information is also another important advantage of using neural networks in processing data obtained from NDT techniques.

D. Martinelli and S. T. Varadarajan, Department of Civil and Environmental Engineering, West Virginia University, Engineering Sciences Bldg., Room 651, P.O. Box 6103, Morgantown, W.Va. 26505-6103. S. N. Shoukry, Department of Mechanical and Aerospace Engineering, West Virginia University, Engineering Sciences Bldg., Room 505, P.O. Box 6106, Morgantown, W.Va. 26505-6106.

In addition, neural networks have great potential for parallelism since the computations of each component are independent of each other. The inherent parallelism allows rapid parallel search and best-match computations. Therefore, much of the computational overhead can be alleviated when applying the new techniques to data interpretation problems. However, a neural network requires a large number of example data from which it can learn and generalize.

In contrast to neural networks, rule-based expert systems use a symbolic computational approach to encoding intelligence. A rule-based system consists of three key parts: an inference engine, a knowledge base, and a collection of known facts and knowledge about the problem to be solved. Rule-based systems are useful when expert knowledge is available and when there is a need for an interactive system that provides explanation for answers and decisions made.

HYBRID SYSTEMS

A hybrid system, which is a combination of expert systems and neural networks, has the capabilities of both systems while minimizing the problems of each (*4, 5*). Rule-based systems alone cannot always handle large applications requiring complex numerical computations, such as the Continuous Bridge Monitoring system. Their reasoning is not adaptive and their performance does not increase with experience. In addition, they sometimes require too much human input and long experience development. But these hurdles can be overcome by coupling knowledge-based systems with complementary AI approaches like neural networks. Thus, by combining two complementary systems we can obtain effective and complete solutions to difficult problems. Hybrid systems can enhance performance as follows:

1. Neural networks provide pattern-recognition functionality and KBs perform analysis and interpretation of data.
2. Hybrid systems improve user-system interaction by explaining to users how a neural network arrived at a solution to a problem. This is a result of the ability to store within the KB, system data related to environment, traffic load, design properties, maintenance history, and so forth.
3. The rule-based systems can help train neural networks by providing intelligence to create the training and test sets. An important step in training neural nets, is a decision on the type of signal preprocessing applied to the row signal before it can be effectively used for training a neural net. A large number of signal processing approaches are available. Based on the accumulated knowledge regarding the performance of different preprocessing techniques and the achieved consistency of the correct recognition rate of the neural network, the system may suggest a retraining of the network or the use of a particular preprocessing technique for a specific application.
4. The neural network can develop implicit knowledge that supplements the knowledge-based system's explicit rule-based knowledge.

HYBRID SYSTEM ARCHITECTURE

The hybrid system architecture, shown in Figure 1, was developed as a part of the research undertaken at West Virginia University to develop a continuous bridge-monitoring system. One of the objectives of this research is to develop and implement a system for interpretation and synthesis of acquired NDT data toward the development of an effective system for bridge management. The work was divided into two primary research components:

1. Interpretation and synthesis of acquired sensory data.
2. Integration of NDT data with field conditions (e.g., traffic and weather) and maintenance and repair policies.

The acquisition of accurate information about the physical characteristics of materials and the condition state of the structural members is key to maintenance and repair of bridges. The NDT techniques are sensitive to normal variations in environmental and traffic conditions. Therefore, the use of multiple NDT techniques will enhance the reliability of information obtained and result in a more comprehensive assessment of the state of bridge elements. For example, traffic activities on a bridge may obscure the signal obtained from AE sensors, whereas ground penetrating radar used for crack detection in bridge decks under the same condition could provide more reliable information. Thus, the architecture of the proposed Hybrid AI System is guided by the following considerations:

1. The system is to be interactive with the user being able to decide the course and progress of data analysis.
2. The system is to be equipped with the capability of pattern recognition, and computational and numerical tasks that an expert system alone cannot perform.
3. The system should incorporate methods for integrating context dependent data like temperature, humidity, traffic condition, and so forth, in order to provide a robust and comprehensive method of data analysis and interpretation.
4. The determination of the critical physical parameters of the structural elements of the bridge is to be done based on data acquired from different NDT techniques.
5. Based on the physical conditions of different bridge elements, a safety factor for the members and for the whole bridge is to be developed. Using heuristic and expert knowledge guided by maintenance and repair policies, suitable maintenance and repair priorities and decision alternatives are to be made available to the user.

An overall view of the hybrid system architecture with the main functional modules is shown in Figure 1. The sensor data from different NDT methods serve as inputs to the data analysis and preprocessing module. Environmental data such as temperature and humidity are also measured simultaneously and stored so that their effects can be considered during preprocessing of the sensor data. The expert system data analysis and interpretation module performs data interpretation using the numerical results obtained from the AI-based preprocessor module. It also predicts the condition state and feasible action based on the interpreted results and the database that provides information about the facility that is being inspected for defects.

DATA ANALYSIS AND PREPROCESSING MODULE

This module shown in Figure 2, consists of various routines which perform specific tasks in processing NDT data that has been acquired and stored. An ANN that performs classification or pattern

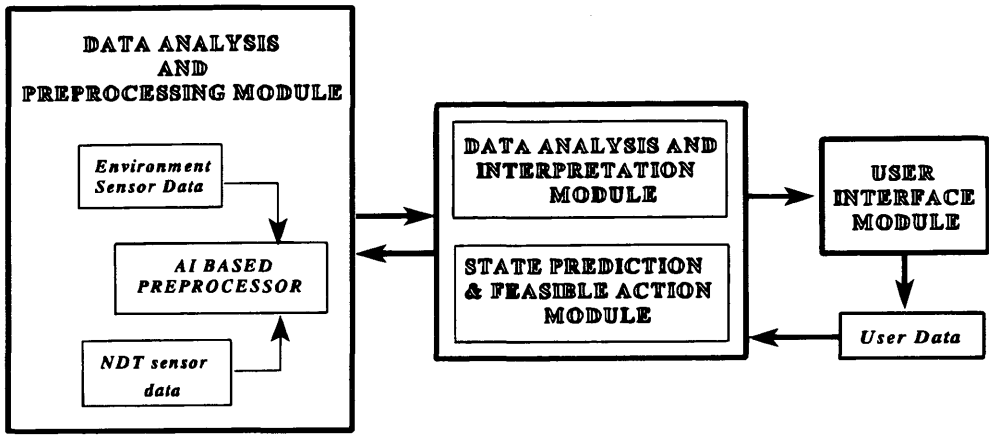


FIGURE 1 Hybrid system overview.

recognition of data obtained using a particular NDT technique, may be one of those routines. Other methods of analysis such as time or frequency domain techniques, fuzzy sets, and so forth, may also be used. The use of neural nets to interpret NDT signals provide a better alternative to the expert system approach. A neural network would perform pattern recognition of signals much more easily and more efficiently than an expert system. The expert system requires the formulation of a set of questions that are applied to a pattern, and a set of rules in order to work out answers to questions (*6*). There is enormous difficulty in both the formation of the questions and the rules. Moreover, provided that such problems are successfully solved, the resulting system might not be flexible to changes as new data and information become available. Neural nets not only can be retrained easily; but also retraining could be embodied in their structures. A brief description of each of the preprocessing modules is described below.

Acoustic Emission (AE)

An ANN network that uses a back propagation algorithm was employed in processing the AE parameters (*7*). It gives estimates of fracture parameters such as the stress intensity factor from which the presence, strength, and intensity of active cracks can be determined. The network architecture is shown in Figure 3. It is trained with AE signal sample wave forms acquired by conducting fatigue and bending tests on steel beams. The AE parameters of interest are the amplitude distribution, energy rate, event rate, deviation, ring-down counts, and rise time. The network is trained to detect signals that result from active propagation of cracks by differentiating it from the noise that is present in the signal.

Ground Penetrating Radar (GPR)

An algorithmic radar signal processing scheme (*8*) was used to interpret data obtained from simulated concrete blocks. The specimens consisted of a set of 15 laboratory-cast concrete bridge deck specimens with and without different types of internal defects (internal cracks) and with and without reinforcement.

The data obtained from the above set of specimens were used to develop a neural network for the interpretation of radar signals (*9*). The data obtained from only 11 specimens were used to train a learning vector quantization neural network. The data from the remaining four specimens were used to test the recognition performance of the trained network. The test set consisted of specimens with and without defects. The trained network was able to correctly classify three out of the four test specimens. This level of recognition was considered satisfactory in view of the small number of examples (*11*) used for training the network.

Ultrasonic and Impact-Echo

A considerable potential for the application of various AI techniques exists in ultrasonic or impact-echo testing. A neural net which automates the signal interpretation from impact-echo tests has been developed (*10*). The network receives the normalized spectrum of the time domain signal as an input. The output is the depth of the crack whenever one exists.

Ultrasonic testing is a knowledge-based activity highly dependent on the expertise of the testing personnel. The inspection process includes the selection of an appropriate testing method, testing procedures, and the interpretation of the measured response for the detection of internal defect and/or material deterioration. The correct choice of testing arrangement, procedures, and the interpretation of the received signal requires a high level of expertise. A rule-based expert system was used (*11*) to answer the following questions:

"Which ultrasonic testing arrangement is most suitable?" and

"What are the most appropriate testing procedures?"

Fuzzy sets have been applied for the evaluation of concrete quality using a set of measured ultrasonic parameters. The pulse velocity, attenuation, and main frequency of the reflected ultrasonic signal were modeled using fuzzy mathematics and multivariate mathematical statistics were employed in determining the concrete strength or flow estimation (*12*).

The mechanism of ultrasonic wave propagation in concrete is very complicated and the ultrasonic reflections are a mixture of coupled responses from longitudinal, shear, and surface waves. The received signal amplitude is also dependent on the amount of coupling between the ultrasonic transducers and the surface of concrete. This makes the signal interpretation, in either time or frequency domain, highly difficult and it would appear plausible to employ

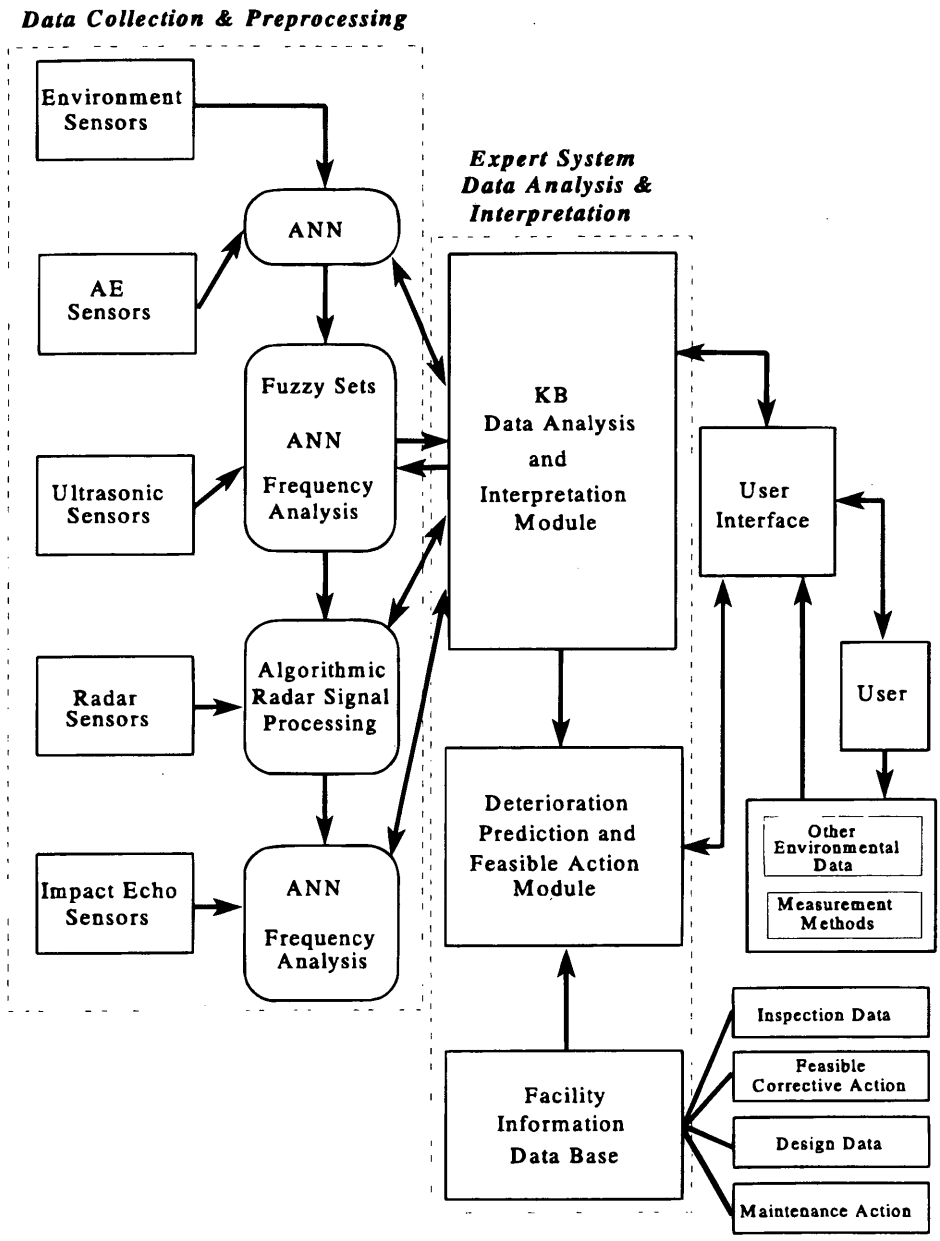


FIGURE 2 Hybrid system components.

neural nets for the interpretation process. A backpropagation neural net for the interpretation of ultrasonic signals obtained from simulated concrete bridge deck specimens has been developed (S. N. Shoukry and D. Martinelli, unpublished work). In this work the same set of 15 concrete specimens used in radar testing was again tested using the ultrasonic pitch and catch method (both the ultrasonic transmitter and receiver are laid on the same side of the specimen). Two 125 kHz transducers, each 2.5 cm in diameter, were used. Nine measurements were taken from each specimen at different distances of separations between the transmitter and receiver. The time domain signals were processed using time dependent Fourier transformation before presentation to the network. After training the network using data from only 11 specimens, it was presented with 9 measurements obtained from a single new specimen that was not used in training. The network was able to correctly classify at least 7 out of the 9 signals obtained from any of the 4 specimens used as test sets. The best performing network architecture consisted of 125 nodes in the input layer, 30 nodes at the hidden layer and 2 nodes for the output layer. The network results were interpreted as 100 percent correct recognition rate, because it was successful in correctly classifying a majority (7 out of 9) of the signals obtained for any concrete specimen within the test set.

DATA ANALYSIS AND INTERPRETATION MODULE

This module uses embedded hybrid system architecture. It is a rule-based system that uses a trained neural network or an algorithmic

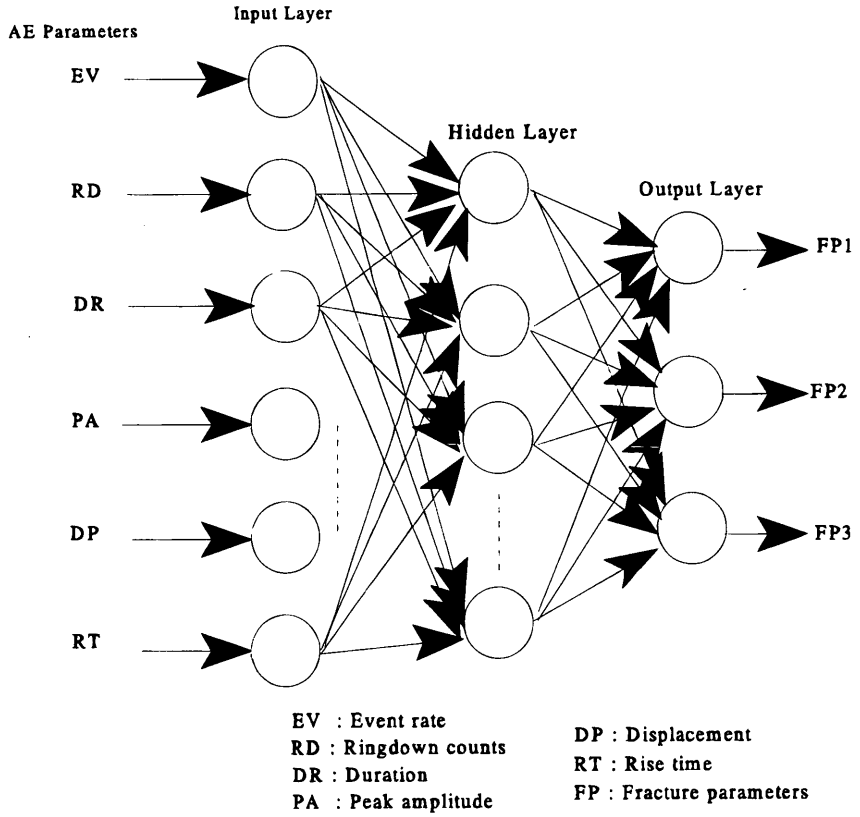


FIGURE 3 Architecture of neural network for AE analysis. EV, event rate; RD, ring-down counts; DP, displacement; DR, duration; RT, rise time; PA, peak amplitude; FP, fracture parameters.

signal processing routine as a subroutine to the "THEN" clause of some rules. The network is called to perform some specific action whenever the rule fires. The action can be a classification, a pattern matching or a regression fit of the data. The network passes the response back to the rule-based system. This module also makes corrections to the quantitative results based on the measurement environment information provided by the user and also attaches degrees of uncertainty to the results depending on the NDT technique that was used to acquire the data and the environmental condition during the acquisition of the data in the field. If conflicting results arise from different NDT techniques, this module would be responsible for their resolution. It provides reasoning for why such a conflict has occurred. The output is quantitative in terms of the location coordinates, size, and type of defect.

STATE PREDICTION AND FEASIBLE ACTION MODULE

This module defines the possible conditions for each unit of the bridge based on the synthesized results and interpretations obtained from the Data Analysis and Interpretation module. Although arriving at this conclusion, the uncertainties attached to the predicted physical parameters are taken into account. Based on the expert knowledge and heuristic reasoning, the degree of safety for each bridge element is determined and the remaining life is predicted. Based on the critical nature of the bridge elements and the condition that they are in, a degree of safety for the bridge as a whole is also determined. All of these results are conveyed to the user, thus the Interface module. Based on maintenance and repair policies, a feasibility action for timely repair and maintenance of the bridge is suggested.

USER INTERFACE MODULE

The User Interface module is a rule-based expert system shell that guides the user through a session of qualitative reasoning and condition state assessment of the bridge components and the bridge for which data was collected. Continuous bridge monitoring involves periodic assessment of bridge condition over fixed intervals. The data for different critical bridge elements, as chosen by the inspector, is stored in a specific file format corresponding to the NDT technique used. These data are later used by the data-processing routines which are invoked by the Data Analysis and Interpretation module. The interface module provides the flexibility to the user to request for the condition state assessment for only certain bridge components and not necessarily for the bridge as a whole. This module also provides the user with a convenient way to input data related to measurement methods and measurement conditions. It also performs consistency checks on the user input data. The other important function of this module is to present the results of the compu-

tation and heuristic reasoning of the other modules to the user in an understandable form, and provide explanations for the results and conclusions when queried by the user.

CONCLUSIONS

A comprehensive and effective system for continuous bridge monitoring involves several tasks ranging from computational, heuristic, qualitative, and quantitative analysis. Moreover, the interpretation and synthesis of acquired data from multiple NDT techniques involve complex signal processing that cannot be done in only one way. The expert system alone lacks the capability of supporting computation and performing pattern recognition and classification tasks. This limitation is overcome in the hybrid system by providing an environment that supports a combined numerical and symbolic processing. Relative weighing of different sources of information from different NDT techniques helps in obtaining an accurate prediction of the state and remaining life of bridge structures. This information is integrated with the context information such as temperature, moisture, and traffic conditions that are provided by the user to make more reliable predictions. Periodic acquisition and analysis of data using the proposed hybrid system will enable the bridge engineer to know the maintenance and repair priorities well in advance, undertake timely repair that will prevent bridges from being in an unsafe condition, and reduce the cost of repairs undertaken.

ACKNOWLEDGMENT

The authors acknowledge the financial support received from FHWA.

REFERENCES

1. Maser, K. R. Automated Interpretation of Sensor Data for Evaluating In-Situ Conditions. *ASCE Journal of Computing in Civil Engineering*, Vol. 2, No. 3, 1988, pp. 215–238.
2. Manning, D. G., and F. B. Halt. Development of Deck Assessment by Radar and Thermography. In *Transportation Research Record 1083*, TRB, National Research Council, Washington, D.C., 1987, pp. 13–20.
3. Malhotra, V. A., and N. J. Carino. *CRC Handbook on Nondestructive Testing of Concrete*. CRC Press, Boca Raton, Fla., 1990.
4. Cardill, M. Expert Networks. *BYTE Magazine*, Vol.16, No. 10, Oct. 1991, pp. 108–116.
5. Martinelli, D. R., and U. B. Halabe. Bridge Monitoring through Sensor Data Synthesis and Interpretation. Presented at ASCE Specialty Conference on Infrastructure Management and Planning, Jun. 1993, pp. 61–65.
6. Roddis, W. M. K., and J. L. Martin. CRACK: Qualitative Reasoning about Fatigue Fracture in Steel Bridges. *IEEE Expert*, Vol. 16, No. 10, Aug. 1992, pp. 41–48.
7. Chen, H. L., and C. L. Chen. Applying Neural Network to Acoustic Emission Signal Processing. Presented at Fourth International Symposium on Acoustic Emission from Composites Materials, Seattle, Wash., Jul. 1992, pp. 273–281.
8. Chen, R. H. L., U. B. Halabe, V. Bhandarkar, and Z. Sami. *Radar Signal Processing and Analysis for In Situ Evaluation of Reinforced Concrete Bridge Decks*. Report No. CFC-93-167 submitted to WVDOH under Contract No. WVDOH RP#90, Phase I. Constructed Facilities Center, West Virginia University, Morgantown, Oct. 1993.
9. Shoukry, S. N., D. Martinelli, S. T. Varadarajan, and U. D. Halabe. *Journal of Material Evaluation*. In press.
10. Sansalone, M., and D. Pratt. Impact-Echo Signal Interpretation Using Artificial Intelligence. *ACI Materials Journal*, Vol. 89, No. 2, Mar.–Apr. 1992, pp. 178–87.
11. McNab, A., and H. S. Young. Knowledge-Based Approach to the Formulation of Ultrasonic Nondestructive Testing Procedures. *IEEE Proceedings*, Vol. 136, Part A, No. 3, May 1989, pp. 134–140.
12. Han Ji-hong, and L. Wei-du. Fuzzy Cluster Analysis and Comprehensive Evaluation of Concrete Quality with Ultrasonic Multi Parameters. Presented at International Conference on Nondestructive Testing of Structural Materials, 1993, pp. 267–276.

Publication of this paper sponsored by Committee on Artificial Intelligence.

Identification of Hazardous Highway Locations Using Knowledge-Based GIS: A Case Study

GARY S. SPRING AND JOSEPH HUMMER

The work described in this paper was conducted at North Carolina Agricultural and Technical State University and North Carolina State University. The study used the increased capabilities offered by geographic information systems (GISs), along with the detailed mapping (which contains highway features and geometrics) available for Guilford County, North Carolina, to demonstrate the use of engineering knowledge regarding accident causation to identify hazardous locations. The general approach taken was that of a pilot study, in which a subset of information is used to demonstrate how a new technology (in this case GIS) may be used to solve a particular problem or problems. The mapping data available from Guilford County, along with various other data files, the MapInfo GIS, and North Carolina's Accident Records System (ARS) were used to conduct the study. The project provided valuable information regarding the limitations and advantages of using engineering knowledge about accident causation to identify hazardous highway locations, and demonstrated the utility and difficulties of applying the GIS to ARS. The overall approach is focused on, and difficulties associated with implementation are discussed.

Accident Records Systems (ARSs) represent the first component of the Highway Safety Improvement Program (HSIP), which was established in 1979 by the FHWA (*1*) as part of its mandate (as set forth in the 1966 Highway Safety Act) to help states to develop safety programs.

ARSs are data bases that contain accident information, as well as traffic and physical information such as road inventories, traffic counts, pavement condition, railroad grade crossings, bridge locations, traffic signal and sign inventories, and traffic permit files. ARSs can help provide fast, safe, and high-quality service to the motoring public on both state and local facilities if they are used efficiently and effectively. This requires accurate and complete data input, as well as consistent and highly accessible data files.

This paper describes work completed at North Carolina Agricultural & Technical (A&T) State University and North Carolina State University. This project used the increased capabilities offered by geographic information systems (GISs), along with the detailed mapping (which contains highway features and geometrics) available for Guilford County, North Carolina, to demonstrate the use of engineering knowledge about accident causation to identify hazardous locations. The paper's focus is on issues associated with implementing knowledge-based GISs for identifying hazardous highway locations. Details on knowledge base development and evaluation of the system can be found in the Project Final Report (*2*).

G. S. Spring, Department of Civil Engineering, North Carolina A&T State University, Greensboro, N.C. 27411. J. Hummer, Department of Civil Engineering, North Carolina State University, Box 7908, Raleigh, N.C. 27695.

MOTIVATION

Accidents are random and rare events. High accident experience has been since the inception of the HSIP the most common way to determine hazardous locations, probably because the statistics are easily generated and are readily available. The use of accident data alone, as is the practice in the majority of states (*3,4*), to focus resources on locations that may be hazardous has several problems and limitations. The problems and limitations are discussed in detail by Zegeer (*5*). Most are related to the poor quality, incompleteness, or inaccessibility of the data, such as data errors and inconsistencies in accident records; inaccurate location of accidents, particularly in rural areas; outdated accident data; and inconsistent referencing systems.

North Carolina's MERGE system (*6*) has two primary deficiencies that were of interest for this project and that hinder the effectiveness of North Carolina's HSIP. These are inaccurate location of accidents and incomplete information on highway features and geometrics. The latter deficiency, along with budgetary constraints, is an important reason why state agencies such as North Carolina's opt for the use of statistics to identify hazardous locations, which essentially equates high accident experience to hazardousness. The sole use of statistics for this purpose presents several problems, one of which is the "regression to the mean" phenomenon. High accident levels may be due to this statistical anomaly, as discussed by Pendleton and Morris (*7*), and not to a roadway problem. Various techniques have been discussed that attempt to overcome this and other problems associated with using statistics (*8–10*).

High-accident locations often represent problems on the roadway, but other locations may have equally high potential for a catastrophic event, even though accident experience is not yet abnormally high. Moreover, on a systemwide basis a particular element may have a high accident experience; thus, it may be more cost-effective to make a systemwide correction of a common element than to correct a high accident location. The need for a comprehensive program to address hazardous roadway elements is discussed by Zegeer (*5*). A commonly identified example of a potentially hazardous element is a roadway section with a low friction number.

> This condition can be identified in two ways: by searching for sites with high wet-weather accident experience and then checking skid resistance properties of those sites, or by friction testing sites throughout the highway system and listing sections with low friction numbers. If only the "high-accident" sections with low friction numbers are selected and improved, the problem has only been partly corrected. The likely result is that other sections with low friction numbers will develop high-accident experience in the future. Thus, the ideal solution is to systematically identify all of the problem sections and improve those with the greatest need (*5*).

To summarize, North Carolina makes exclusive use of statistical analyses applied to its ARS to focus its limited resources on areas of the highway system identified as "hazardous." Primary reasons for this are the problems (typical of problems faced by other state departments of transportation) associated with its ARS data, primarily the lack of highway feature and geometric data. Although the incompleteness of accident location data is a problem in North Carolina, as it is in other states, it was beyond the scope of this project and therefore was not addressed further.

A GIS SOLUTION

Because of the spatial character of ARS data, GIS technology greatly simplifies their extraction and presentation, provides a higher degree of user friendliness, and provides better access to the data. GISs also provide a means to integrate data from many sources (e.g., U.S. Census data, U.S. Geological Survey [USGS] data, accident records, pavement conditions, etc.).

The GIS is a computerized data base management system that provides graphic access (capture, storage retrieval, analysis, and display) to spatial data. The most visually distinctive feature of GIS software is a map display that allows thematic mapping and graphic output data overlaid on a map image. The key element that distinguishes GISs from other data systems is the manner in which geographic data are stored and accessed. GISs store geographic data using topological data structures: objects' locations relative to other objects are explicitly stored and therefore are accessible. These data structures allow analyses to be performed that are impossible using traditional data structures. Standard GIS functions that are useful for this application include thematic mapping, statistics, charting, matrix manipulation, decision support systems, modeling algorithms, and simultaneous access to several data bases.

In order for a GIS to be useful, a set of detailed base maps, of an acceptable scale and precision for ARS applications, must be available. Guilford County, North Carolina, has mapped (planimetric information only) approximately 90 percent of the county at a 1 in 2,400 scale to USGS mapping standards, which is adequate for locating accidents.

The general approach taken for the project was that of a pilot study, in which a subset of information is used to demonstrate how a new technology (in this case GIS) may be used to solve a particular problem or problems. The mapping data available from Guilford County, the MapInfo GIS, and North Carolina's MERGE system (along with other data described later) were used to accomplish the project's objective. A workstation-based prototype knowledge-based GIS (KBGIS) was developed for the project's demonstration purposes. The system was anticipated to be accessible through microcomputers tied to the system. This was not in fact possible, due to the limitations of the software used. However, that software does allow direct transfer of data between various hardware platforms. Thus, a data base developed on a UNIX-based workstation platform is directly transferable to Windows or Macintosh platforms. Limitations are discussed later in the paper.

Hardware Setup

Given the diversity of ARS users' needs, goals, and current activities, the best hardware configuration to implement the small pilot study's ARS in GIS was determined to be a local area network (LAN) at North Carolina A&T, consisting of two workstations, two IBM PCs, and three Macintosh Quadras. The LAN, an ethernet network using X-windows and the TCP/IP communication protocol, allows evaluation of a decentralized system; that is, users may access the workstation network, which contains ARS data, from their stand-alone PCs or from a workstation. The LAN allows software to be run under the DOS, MacOS, and Windows environments, thus providing maximum flexibility in accessing the data base. In addition to providing flexibility, the LAN satisfies the special needs associated with geographic information systems: large mass storage capacity, portability, rapid and powerful computing abilities, precision digitizing and quality plotting capabilities, and high-quality graphics. The LAN also provides access to the Internet, which facilitated the sharing of large data files among the project team.

Software

Because of unforeseen problems with Ultimap, the GIS software originally chosen for use on the project, the Guilford County map data were converted to MapInfo format. Ultimap Corporation abandoned the version of software originally intended for use in this project and entered bankruptcy. MapInfo was chosen as the replacement for several reasons, the most important of which were as follows:

- It allows access to true object geometry, thus allowing queries about, for example, highway curves' radii.
- It provides a built-in full-featured structured programming language, MapBasic, which was used to interface the GIS and knowledge-based components (which were written in MapBasic).
- MapInfo (Version 2.1) allows access to a wide variety of data base formats.
- The same version of the software runs across multiple platforms, which makes it a strong choice for multi-user environments.
- MapInfo is widely used across North Carolina. Additionally, several large GIS packages provide "hot links" to MapInfo files.

Data

The project team decided to take a feature-based approach to hazardous site identification. Features chosen were curves, bridges, and intersections. Therefore, data for each of these in addition to average daily traffic (ADT) and accident data were collected as part of this effort as summarized in Table 1. Deficiencies in and difficulties associated with the data are described later in this paper.

Accident data, obtained through the University of North Carolina's Highway Safety Research Center from the state of North Carolina's MERGE system, were coded to a 1:24,000 centerline map of Guilford County. The coding was performed by personnel at the University of North Carolina's Institute for Transportation Research and Education GIS laboratory. Both data sets (the accident data and the road centerline data) were transferred to the A&T network via the Internet. The data were then imported to the MapInfo software.

The North Carolina Department of Transportation (NC DOT) bridge maintenance unit provided project personnel with its federal bridge file for 751 bridges in Guilford County, dated 1992. The file was provided in ASCII format and was imported first into the FoxBase data base manager to provide data structure (the data were

TABLE 1 Project Data Summary

Description	Source	No. of Variables	No. of Records
Accident location data	MERGE	13	11,554
Accident attribute data	MERGE	77	63,899
Roadway centerline network (1:24K)	ITRE GIS laboratory	15	23,524
Bridges	NCDOT Bridge Maintenance Unit, Federal Bridge file for 751 bridges	21	448
Average daily traffic	NCGIA - derived from NCDOT HPMS	52	3,726
Planimetric data	Guilford County GIS Unit	1	500,000
Centerline data (1:150K TIGER files)	MapInfo StreetInfo+	6	36,201

space-delimited, which is not supported by the MapInfo software) and then into the MapInfo software. The bridge file contained 21 variables, including latitude and longitude. Bridges were geocoded using these values instead of milepost information because the latitude-longitude data were more complete.

Intersection data consisted solely of location and street names. They were created from enhanced TIGER files (called StreetInfo Plus) that were purchased from MapInfo Corporation, for Guilford County. No other data, such as turning movement counts, were available.

Curve data were in the map data obtained from Guilford's planimetric data base. The data base was digitized from digital orthophotos created by flying the county in 1991. The resulting maps are at a 1 in 2,400 scale and therefore provide a ±60-unit precision as mentioned earlier. Because of the change from the Ultimap software to MapInfo, it was necessary to transfer the data from the Apollo Domain workstation on which they were created in Ultimap to a UNIX-based workstation on MapInfo runs. The project team transferred more than 300 files via the DXF format. The resulting file contains more than 500,000 records and requires 4.5 megabytes of storage.

THE KBGIS MODEL

Figure 1 depicts the model's structure. The system's engine is the GIS. The user initiates sessions by choosing the type of site for which an analysis is to be performed. The GIS engine, through a series of graphic SQL queries (such as buffering) and traditional SQL queries of its data bases, provides the information listed in Table 2, in the form of "Danger Tables," to the knowledge-based component of the system. This information, along with user-provided information, is used for each site type (curve, intersection, or bridge) to identify and rank sites based on their hazardousness. Conceptually, the process consists of calculating an accident rate based on data and a rate based on models that have been developed by others: intersection models by Hauer et al. (*11*), bridge models by Turner (*12*), and curve models by Zegeer and Council (*18*). Adjustments to the rates and levels of confidence in the rates, extracted from the knowledge base, are used to calculate a combined accident rate and a level of confidence in that rate. Hazardousness level is then determined using a function depicted by Figure 2, derived from expert interviews.

In the Guilford County case study, only accident rates based on data for bridges were possible, because the bridge file had ADT data as one of its fields. The ADT file obtained from NC DOT for the other site types had several difficulties associated with it, which are described in a later section.

Accident Frequency Calculations

The procedures used to determine accident frequency are generic and are used for all three site types. This is essentially a buffering problem, the objective of which is to find and count all objects of Type 1 (e.g., accidents) that lie within some specified radius of

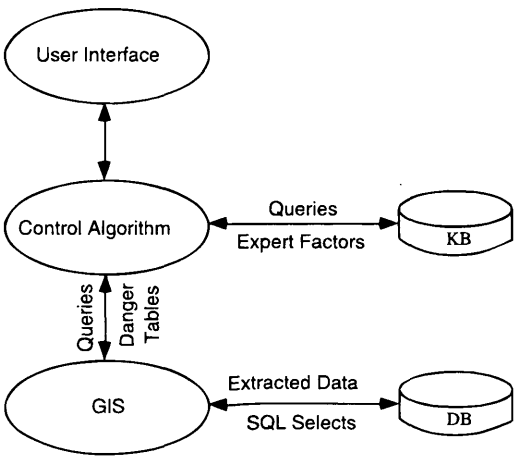


FIGURE 1 Model structure.

TABLE 2 Information Passed To Kb

Category	Derived Data Item
Curves	Length of curve Degree of curve Number of accidents occurring within specified distance
Bridges	Accident rate based upon data Accident rate based upon model
Intersections	Number of accidents occurring within specified distance

objects of Type 2 (e.g., intersections). The buffering function provided by MapInfo results in one count for this query, namely, the total number of accidents that lie within some radius of *all* intersections. The count of interest, however, is the number of objects of Type 1 that lie within a specified radius of *each* object of Type 2. Thus, it was necessary to write special code that would do the latter. Using the intersection example, the code has two parts:

1. Convert intersection point objects to circles with radius as specified.
2. Count the number of accident objects that fall, geographically, within the circle objects just created.

Figure 3 provides an example of high-accident locations selected based on a criterion of greater than four accidents within a specified radius from intersections. Figure 4 depicts a thematic map generated from this information.

Accident Model Rate Calculations

As mentioned earlier, bridges were the only site type for which it was possible to calculate model rates, given the limitations of the data available for this project. To generate the bridge model rate tables shown in Figure 5, the following steps were used.

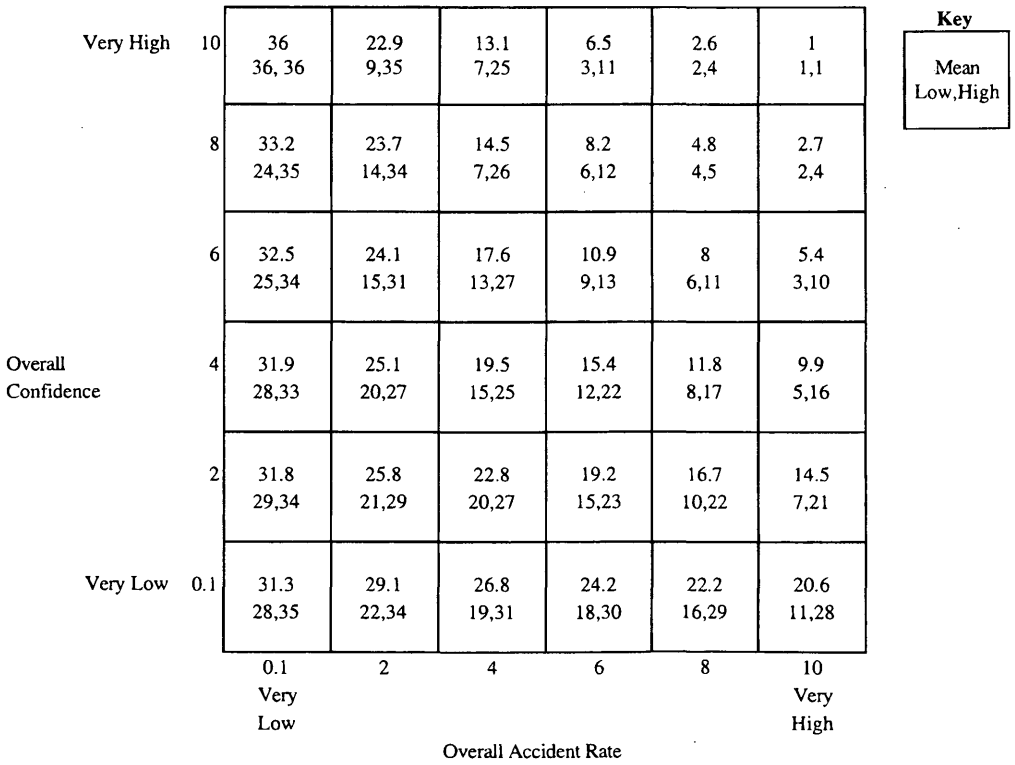


FIGURE 2 Hazardousness function.

High Accident Locations (> 4 Accidents)

InterName	Accidents
W MARKET ST & S EUGENE ST	5
W MARKET ST & N EUGENE ST	5
W MARKET ST & S EDGEWORTH ST	6
W MARKET ST & N EDGEWORTH ST	6

FIGURE 3 Sample table used to identify high-accident intersections.

1. Select a unique field and structure number, and create a derived field using the bridge rate model.
2. Normalize the model rates in Figure 5 to a scale of 10.

The curve rate model uses curve parameters that may be calculated from curve radius and delta. It also uses road width, which is not available in the data and therefore must be obtained directly from the user. Radius and delta values were calculated using the MapInfo-supplied access to true curve geometry along with a series of complex geographic selections.

Data used for these calculations are summarized in Table 2. The information depicted in Figures 3 through 5 are passed back to the knowledge-based components of the system. The system then prompts the user for more detailed information on a site-specific basis, for example, ADTs on curves, turning movements at intersections, and so on.

Data Integration

Although the use of the MapInfo software was not initially anticipated, the choice proved to be serendipitous. MapInfo provides a fairly full-featured, structured programming language called MapBasic, which provides transparent access to MapInfo's functionality. Therefore, all programming for this effort was performed in MapBasic. Additionally, problems associated with integrating GIS, knowledge base rules, and inferencing mechanism were avoided in this way. This has been the case in a great many other KB GIS efforts described in the literature (*14–18*). As was described earlier, all data were imported to the MapInfo format for use.

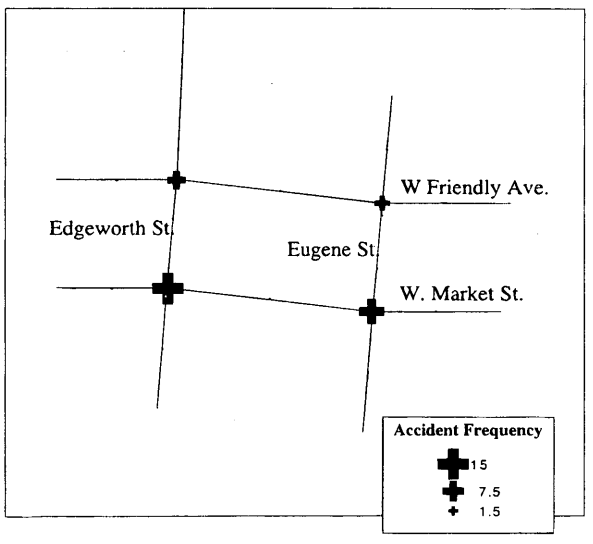


FIGURE 4 Sample thematic map of high-accident locations.

THE KNOWLEDGE-BASED SYSTEM ELEMENT

The key feature of knowledge-based systems (KBS) with regard to data base applications is their ability to use relations among objects stored in an existing data base to infer, with varying degrees of certainty, other objects. Since they infer new data using existing data, queries for which there are no data explicitly stored may be posed and answered. Consider, for example, a family tree. To store the tree in a conventional data base would require that *all* family relationships be explicitly defined as fields. That is, for each person in the family, the data base would have to contain explicit information regarding his or her relationship to all other family members. One way to do this would be to have each row represent a family line and each field represent a relationship. This method requires full specification of all relationships for which queries are to be made. For example, if one were to ask the data base "Who is Fred's cousin?" the answer could only be answered if the cousin relationship were stored in the data. Using a KBS in combination with the data base would require specification of only one relationship (such as parent), along with a set of rules representing all other relationships. This simplifies the data base and allows new relationships to be added easily (or existing ones to be changed or deleted) without disturbing the data. Relationships that have some degree of uncertainty associated with them may also be used in this process. For example, inheritance rules governing eye color or congenital defects could also be included if these types of information were of interest.

Information regarding accident causation derived from previous studies [for example Zegeer and Council (*18*), Harwood and Warren (*19*), and Spring (*20*)] were used in concert with GIS analysis strategies to identify hazardous highway locations using engineering knowledge rather than pure statistics. The project team conducted a series of interviews with state and local traffic engineers regarding the forensics of hazardous highway locations. Many of the questions used for the interviews were derived from a thorough review of literature such as that mentioned previously. The principles used for construction of the knowledge-based prototype were described by Mouradian (*21*). He advocated early prototype development using generally numerical (ratio) data. He also advocated using a simple scoring method to weight different pieces of information. This was the approach taken for this project. A set of models was taken from the literature, along with traffic engineers' confidence in those models, and were used to quantify hazardousness.

EVALUATION

The objective of the evaluation method chosen was to assess the validity of the underlying system model of the real world, which requires that evaluation criteria and acceptable levels of performance be established. A common evaluation criterion, which was used for this study, consists of a simple comparison of system conclusions with human conclusions. How closely they agree may be used as a measure of performance. There are problems intrinsic to this approach, described by Spring (*22*). However, it provides a gross but, for the purposes of this preliminary system, sufficiently

STRUCTURE_NO	MR (Accidents/yr)	STRUCTURE_NO	MRAdjusted
40185	2.00295	40185	7.65E-05
40187	1.4051	40187	5.37E-05
40188	0.96556	40188	3.69E-05
40189	79.1197	40189	0.0030237
40193	0.420459	40193	1.61E-05
40194	9.463	40194	0.000361645
40195	364.573	40195	0.0139328
40197	12404.01	40197	0.474041
40198	821.483	40198	0.0313944
40202	204022.53	40202	7.79708
40203	39.1525	40203	0.00149628
40204	0.478173	40204	1.83E-05
40205	48.2972	40205	0.00184576
40206	92.1084	40206	0.00352009
40207	72.6338	40207	0.00277583

FIGURE 5 Sample bridge model rate tables.

accurate assessment of system performance. The approach is common for preliminary evaluations such as this one (23). A list of locations that were programmed for improvement in the period after 1992 by NC DOT was obtained and compared to the list of sites, ranked by level of hazardousness, output by the KBGIS. Table 3 presents a summary of these results, which were subjected to a simple χ^2 test, which tested whether or not the two sets of conclusions are related. It was concluded that, at the 95 percent level of confidence, they are related. Given several limitations associated with this study (and perhaps in light of the limitations), overall agreement was deemed adequate. A complete description of the evaluation process and its results may be obtained from Spring and Hummer (2).

DIFFICULTIES

Difficulties encountered during project implementation fall essentially into two categories, software and data. As is explained in the following paragraphs, most of the major difficulties faced were due to data problems or limitations. Although these may seem daunting, they are the same difficulties that any GIS development effort faces, and so are not specific to this particular application. They would be addressed as part of an agency's GIS implementation program.

Software

In the process of using the Ultimap and MapInfo (including MapBasic) software packages several difficulties were encountered, some of which were bugs in the software and some of which were simply missing or poorly designed features. In the case of Ultimap, difficulties may be attributed to its lack of robustness; perhaps this is one reason why the company was bankrupted. MapInfo, although adequately robust for most applications, was not originally designed as a full-featured GIS package and therefore lacked robustness for the purposes of this project. Another function that would be important if agencies implemented the system developed for this project is networkability, which MapInfo does not support. Three bugs in MapInfo were discovered that caused delays and had a negative impact on the Project's final product:

- For certain graphic objects, the MapBasic object geography function, which provides access to object attributes, does not consistently return correct values for those attributes.
- Result tables from multiple selections sometimes do not retain derived data columns, within MapBasic. (Using MapInfo for this yields acceptable results.)
- The MapInfo version that runs on the workstation (Version 2.0.2) does not allow the user to choose a subset of layers from AutoCAD DXF translation files.

Additionally, MapBasic does not provide access to many of the excellent built-in functions provided by MapInfo. For example, MapInfo has a powerful address matching capability to which only minimal access is given through MapBasic. Even when some built-in function is available in MapBasic, it may be difficult to find. Many tricks and processes that would smooth the way are not given anywhere in the written documentation. Additionally, MapInfo does not support a networked environment, eliminating one of the advantages of the LAN described earlier.

TABLE 3 Evaluation Format

	NCDOT	
KBGIS	Not Hazardous	Hazardous
Not Hazardous	7	16
Hazardous	14	4

Data

Table 4 summarizes the difficulties encountered with the data used for the project. Of the 23,524 road segments contained in the 1:24,000 mileposted centerline file used to geocode accidents, only 3,126 had names. Essentially all of the named segments were mileposted. Additionally, only about 31 percent of the accidents obtained from the MERGE data base had milepost information (that is, were locatable). This resulted in only 11,554 of the 38,157 acci-

TABLE 4 Data Limitations

Data File	No. of Elements	No. Usable	Reason
Roadway centerline network (1:24K)	23,524	3,126	Unnamed segments
Accidents	38,157	11,554	Not mileposted
Bridges	753	448	Missing lat/long information
ADT	52	0	Inconsistent road name conventions

dents that occurred in Guilford County in the last 3 years being geocoded. This is a major limitation to the validity of project results and is certainly one of the biggest obstacles facing implementation of a GIS-based ARS. However, given that most of the mileposted accidents are in rural areas and that the majority of named roads happened to be in rural areas, the validity of project results was enhanced. Locating accidents and other features (ADT and so on) on a map requires a match between referenced road names. The apparent absence of a consistent convention for naming rural roads, among the various data sources used for this study, therefore also contributed greatly to the difficulties in tying accidents and other feature data to the map. This was especially critical for the ADT data file. Road names used in the file were inconsistent with any other naming system that could be found. This prevented the project team from locating the ADTs and using them for rate calculations.

There were additional problems encountered in translating planimetrics from the original format to the MapInfo format via DXF. These all were related to bugs in the Ultimap GIS software in which the map data were originally stored. The platform on which it was anticipated programming would be done, Apollo/Domain, was not used due to the abandonment of Ultimap, nor was the A&T network completely functional. It was necessary for a second workstation to be purchased, as well as a PC-based computer, and for all components to be connected for a successful transfer to take place. Both software packages have bugs that contributed to the lower quality of these graphic data. Line and curve segments, after conversion to DXF and transfer to UNIX-based workstation, are discontinuous and some curves came through as full circles. This is due to a bug in the Ultimap software which creates these anomalies in the data during the conversion to DXF.

Another difficulty faced was due to the third MapInfo bug described earlier. The original data files had approximately 25 to 30 layers of data. Ultimap handled these as features that could be turned on and off and so all data were contained in the original map files. The MapInfo system is a layering system which requires separation of those features during the importation process. If all layers were included in the same data file, there would be no way to differentiate among the various data items such as bridges, roads, census tracts, and so on. The size of the resulting data file would be unwieldy as well. Unfortunately, the MapInfo Corporation, at the time of this project, had issued an upgrade to Version 2.1 only for its IBM PC Windows product, due to the bug described earlier. This required that all 300 files be transferred to an IBM PC that had MapInfo Version 2.1, which allows this selectivity. After being imported to MapInfo, the files were assembled into one large file and transferred back to the workstation. The resulting map file, even with only road edge information, contained over 595,000 elements for Guilford County alone.

CONCLUSIONS

Given the spatial nature of accident data, the use of GIS for ARS makes good sense. Within GIS software, different types of data are easily related, either graphically or in report form, thus making the data more easily accessible and providing a friendlier and more flexible user interface. For example, pinpointing problem spots on the highway network by displaying sites whose signs have reflectivity values below a certain level may be done with ease. These qualities also help to provide better quality data. The use of accident causation information will enhance hazard elimination programs by avoiding, or in some cases eliminating, problems associated with the use of accident records alone in identifying problem locations. The GIS also provides a link between the various ARS data files. Presently, inconsistent data files—that is, data files that use inconsistent referencing systems—make it extremely difficult to fully utilize the ARS as an accident analysis tool (i.e., all available information cannot be used together in one analysis). Often, data files, when created, were intended for purposes other than accident analysis and therefore often have different referencing systems (e.g., mile marker versus link node). The fact that in GIS all locations are referenced to the same map eliminates this problem.

The knowledge-based approach described in this paper creates a synergy by providing consistent access to a common pool of engineering knowledge. This also provides an excellent means of computer-based training for novice traffic engineers.

With these features, a KBGIS can provide more cost-effective, safer, and more efficient highway systems for the user community. This project demonstrates the process of integrating GIS with ARS and, it is hoped, will demonstrate that, with the dramatic advances in small computer hardware technology over just the last few years, the limits to what can be done are imposed by what users are willing to do, rather than by technology.

ACKNOWLEDGMENT

The authors thank the traffic engineers who participated in the interview process for this project, and members of the Project Advisory Group whose valuable input made the happy outcome of the project possible. This work was sponsored by the U.S. Department of Transportation's Southeastern Transportation Center.

REFERENCES

1. Zegeer, C. V. *Highway Safety Improvement Program.* Report FHWA-TS-81-218. FHWA, U.S. Department of Transportation, 1981.
2. Spring, G. S., and J. Hummer. *Identifying Hazardous Locations Using Geographic Information Systems.* Final Report to Southeastern Transportation Center, U.S. Department of Transportation, Raleigh, N.C., 1994.
3. Spring, G. S., J. Collura, and P. W. Shuldiner. Analysis of High Hazard Locations: Is An Expert Systems Approach Feasible? In *Transportation Research Record 1111,* TRB, National Research Council, Washington, D.C., 1987, pp. 7–17.
4. Zegeer, C. V. *Highway Accident Analysis Systems.* National Cooperative Highway Research Program Synthesis 91, NCHRP, 1982.
5. Zegeer, C. V. *Methods for Identifying Hazardous Highway Elements.* National Cooperative Highway Research Program Synthesis 128, NCHRP, 1986.
6. Fischell, T., and E. Hamilton. Local MERGE—Bringing the Data Back Home. *Proceedings of the 14th International Forum on Traffic Records Systems,* National Safety Council, San Diego, Calif., 1988.
7. Pendleton, O. J., and C. N. Morris. Introducing the BEAST—A New Method for Accident Analysis. Presented at the 69th Annual Meeting of the Transportation Research Board, Washington, D.C., 1990.
8. Lau, M. Y., and A. D. May. Applications of Accident Prediction Models. Presented at the 68th Annual Meeting of the Transportation Research Board, Washington, D.C., 1989.
9. Higle, J. L., and M. B. Hecht. Comparison of Techniques for the Identification of Hazardous Locations. Presented at the 68th Annual Meeting of the Transportation Research Board, Washington, D.C., 1989.
10. Hauer, E., and B. N. Persaud. Problem of Identifying Hazardous Locations Using Accident Data. In *Transportation Research Record 975,* TRB, National Research Council, Washington, D.C., 1984, pp. 36–43.
11. Hauer, E., J. C. N. Ng, and J. Lovell. Estimation of Safety at Signalized Intersections. In *Transportation Research Record 1185,* TRB, National Research Council, Washington, D.C., 1988, pp. 48–61.
12. Turner, D. S. Prediction of Bridge Accident Rates. *Journal of Transportation Engineering,* Volume 110, Number 1, ASCE, New York, N.Y., 1984, pp. 45–54.
13. Leug, Y. Fuzzy Logic And Knowledge-Based GIS—A Prospectus. *IGARSS '89—Twelfth Canadian Symposium on Remote Sensing,* Vancouver, B. C., Canada, Part 1 (of 5), July 10–14, 1989.
14. McKinney, D. C., D. R. Maidment, and M. Tanriverdi. Water Planning Using an Expert GIS. Water Resources Planning and Management: Saving a Threatened Resource—In Search of Solutions. *Proceedings of the Water Resources Sessions at Water Forum.* ASCE, New York, N.Y., 1992, pp. 219–224.
15. McKinney, D. C., D. R. Maidment, and M. Tanriverdi. Expert Geographic Information System for Texas Water Planning. *Journal of Water Resources Planning and Management,* Vol. 119, No. 2, March–April 1993, pp. 170–183.
16. Stonebraker, M. The Integration of Rule Systems and Database Systems. *IEEE Transaction on Knowledge and Data Engineering,* Vol. 4, No. 5, Oct. 1992 p 415–423.
17. Gronlund, A. G. *A Knowledge-Based GIS Methodology for Forest Fire Management.* Master's thesis, University of North Carolina at Charlotte, 1993.
18. Zegeer, C., and F. Council. *Cost-Effective Geometric Improvements for Safety Upgrading of Horizontal Curves.* FHWA-RD-90-021. FHWA, McLean, Va., 1991.
19. Harwood, D. W., and D. L. Warren. Operational and Safety Effectiveness of Passing Lanes on Two Lane Highways. Presented at the 84th Annual Meeting of the Transportation Research Board, Washington, D.C., 1985.
20. Spring, G. S. An Expert System to Analyze Hazardous Intersections. *Microcomputers in Civil Engineering,* Vol. 3, No. 3, 1988, pp. 299–309.
21. Mouradian, W. H. Knowledge Acquisition in a Medical Domain. *AI Expert,* Vol. 5, No. 7, July 1990, pp. 34–38.
22. Spring, G. S. Validating Expert System Prototypes Using the Turing Test. *Transportation Research, Part C,* Vol. 1, No. 4, Pergamon Press, Ltd., Great Britain, 1993, pp. 293–301.
23. Spring, G. S., J. Collura, J. Shuldiner, and P. W. Watson. Testing, Verification and Validation of Expert Systems: A Case Study. *ASCE Journal of Transportation Engineering,* Vol. 117, No. 3, 1991, pp. 350–360.

Publication of this paper sponsored by Committee on Transportation Data and Information Systems.

Knowledge-Based Geographic Information System for Safety Analysis at Rail-Highway Grade Crossings

SRIRAM PANCHANATHAN AND ARDESHIR FAGHRI

The development of a knowledge-based geographic information system for managing and analyzing safety-related information for rail-highway grade crossings is discussed in this paper. The allocation of federal funding for safety improvements at public, at-grade rail-highway crossings is made based on the performance of the states with respect to accident reduction. The motivation behind the work was to establish guidelines and to develop an integrated system that would ultimately result in accident reduction through better access and management of safety information. This was accomplished by using geographic information system (GIS) technology and decision support tools through integration of the GIS application with a statistical model and a knowledge-based expert system (KBES). The work continued an ongoing project that resulted in the integration of rail-highway grade crossing safety data from various sources, such as the FRA and Delaware Department of Transportation (DelDOT), into a data base management system. The selection and integration of the U.S. Department of Transportation (USDOT) accident prediction model into the system was also required. This paper describes the conversion of rail-highway grade crossing safety attribute data into a GIS-acceptable format, the development of the GIS application including the spatial analysis, visual display, and query capabilities, and the development of a KBES to account for site-specific and qualitative factors. The KBES is also capable of suggesting safety upgrade action(s) at the crossing. Interfacing of the KBES and the program for the USDOT model with the GIS and the framework of the complete package developed for safety analysis at rail-highway grade crossings are also discussed in the paper.

Transportation agencies are faced with an ever-increasing need for not only readily available information but also the capability to analyze, manipulate, access, and operate on that information on a broad basis to enable the basic functions of planning, safety, and operations management.

Developments in the field of information technology such as GIS, Expert Systems, and Database Management Systems are especially relevant to the transportation field because of the spatially distributed nature of transportation data, the use of engineering judgment, the often approximate and uncertain nature of transportation information, and the large amounts of data involved.

The application of GIS, in particular, has relevance to transportation due to the essentially spatially distributed nature of transportation-related data, and the need for various types of network-level analysis, statistical analysis, and spatial analysis and manipulation.

Most of the analysis of transportation data is related to quantifiable and statistically significant factors, and most of the models, routines, and algorithms process mathematically quantifiable factors only. The use of KBES in conjunction with mathematical models, algorithms, and procedures will allow for more complete analysis with the incorporation of engineering judgment and heuristic knowledge into the decision-making process and by accounting for qualitative and statistically insignificant, yet important, data.

This paper discusses the development of an integrated package for the management of rail-highway grade crossings safety information. The GIS application was developed using TransCAD (*1*), and the KBES was developed using CLIPS (*2*) shell, which provides a forward-chaining inference engine. The paper focuses on the development of the above-mentioned components, interfacing of the components, and the working mechanism of the integrated system.

RAIL-HIGHWAY GRADE CROSSING SAFETY PROGRAM

The State Transportation Assistance Acts of 1978 and 1982 mandate the provision of federal funding to states for safety improvement projects at rail-highway grade crossings. States are required to establish procedures to rank crossings and to use the rankings in an allocation process to achieve effective use of funds and the greatest possible accident reduction.

In the state of Delaware there are 548 rail-highway grade crossings, of which 265 are public. During the period from 1981 to 1991 there were 71 train-automobile accidents, 10 of which resulted in fatalities (*3*). The responsibility for identification and inventory of rail-highway grade crossings is currently assigned to the Delaware Department of Transportation (DelDOT). An earlier part of the ongoing study resulted in the identification of the most feasible of several existing empirical models and the development of a data base management system that includes all safety-related inventory information for public at-grade rail-highway crossings in Delaware (*3*).

Limitations in the Existing Rail-Highway Grade Crossing Program for Delaware

The existing rail-highway grade crossing safety program for Delaware consisted of a data base management system that maintained safety-related attribute information for all public, rail-highway grade crossings. The USDOT accident prediction model was chosen out of several nationally recognized empirical models for the prediction of accident hazard at a crossing to be implemented for the existing program (*4*). The USDOT model is documented in the work of Hitz and Cross (*5*). The data base management system

Department of Civil Engineering University of Delaware, Newark, Del. 19716

also included a computer program to execute the USDOT model on the data base and append an accident index value to it. The data base was, however, not in a location-referenced format that would enable it to be compatible for a GIS application. Additionally, the site-specific and qualitative information that may be used by the KBES to comprehensively evaluate safety and provide decision support was not available and was maintained in a written form (*3*).

The existing program included the ability to update, modify, and prioritize the data base for rail-highway grade crossings and to calculate the accident hazard index. However, this accident index does not present a true picture of accident hazard at any crossing as it does not account for several site-specific and qualitative factors such as sight distances, truck and hazardous materials-carrier traffic, land use in the crossing vicinity, and other factors. The program also does not have data analysis and manipulation capability and no indication of what action has to be taken at the crossing.

Lack of decision support and inability to consider heuristic and judgmental information were the main deficiencies of the program. The expert, when presented with a text file output of the data base, cannot have a good perception of the existing conditions at individual crossings as he/she has to look up a map every time he/she desires to relate the information to the geographic location of the crossing.

APPLICABILITY OF GIS TECHNOLOGY

Since the crossings safety attribute data is spatially distributed, this problem is suitable for GIS. The main advantage of using GIS is its ability to access and analyze spatially distributed data with respect to its actual spatial location overlaid on a base map of the area of coverage that allows analysis not possible with other data base management systems. The main benefit of using the GIS is not merely the user-friendly visual access and display, but also the spatial analysis capability and the ability to apply standard GIS functionalities such as thematic mapping, charting, network-level analysis, simultaneous access to several layers of data and overlayment of same, as well as the ability to interface with external programs and software for decision support, data management, and user-specific functions.

The existing data base does not allow the user to manipulate, access, and query the data base other than in a very limited way. The user is limited to textual queries only, and selection and viewing of crossings attribute data with respect to spatial and topological relationships is not possible. Other related data, such as land use, population, and the road network characteristics of the area in the crossings vicinity, cannot be accessed in the present data base. This ability of the GIS, along with the final presentation of results on a digital base map, will allow the user a better perception of the problem, enable better decisions, and allow a better understanding of what is to be achieved in a broader sense. The ability to define conditional queries, perform statistical analysis, create thematic maps, and provide charting enhances the crossings safety program by allowing for better understandability of data.

Furthermore, the ability of most GIS software to provide many basic transportation models and algorithms may also be useful in specific situations. The ability to link up to external procedures and software also provides flexibility, as these procedures can access data within the GIS and present the results of analysis to the GIS for viewing and analysis.

DEVELOPMENT OF GIS APPLICATION

Source of Graphical Data

In general, there exist several sources for spatial data from which the digital base map for the application can be created. The base map can be created by scanning or digitizing from hard copy maps. The other approach is to use map data from some primary source such as the Bureau of Census Geographic Base Files (GBF), the United States Geological Survey (USGS), Digital Line Graphs (DLG), and the United States Census TIGER files.

TIGER was found to be a suitable source for building the base map for this project primarily because of the level of detail available about the street centerlines and the rail lines and, also, because of the valuable attribute information tied into it. Since the main requirement of the application was for data retrieval, spatial analysis, and visual display of analysis results from external procedures, and not network-level analysis, the inaccuracies of TIGER would not affect its use in the development of the GIS application.

Nature of Attribute Data

The existing program consisted of data from various sources integrated into one data base. This included inventory data maintained by the FRA, accident data from the DelDOT Bureau of Traffic, and data from other sources, such as railroad companies, in tabulated inventory sheets. Each crossing has a unique DOT-AAR (Association of American Railroads) identification number that serves as a key to access each crossing record. Additionally, site-specific and qualitative information was not in the data base but was maintained in written form (*3*).

The attribute information is the data required as input for the USDOT model (*3*). The location reference is given in the form of city and county codes, railroad ID (identification) number, road and railroad names at the crossing, and the milepoint on the approach road.

Location Referencing of Attribute Data

If the location reference is in the form of latitude and longitude coordinates, the input data for a point data base needs to be contained in a single data file. TransCAD (*1*) has the ability to convert data referenced in any other coordinate system when provided with the local and world coordinates of any three points. However, in the case of the current data base, the only location reference information consisted of city, county, street and railroad names, and street milepoint information. Milepoint information on the approach roads could not be used for location referencing because of the unavailability of accurate milepoint information. Interactive referencing was done to tag records to their associated spatial locations.

Some of the crossings records could be location referenced by a simple program that searched for a match in the strings consisting of the railroad and street names for a crossing record with the respective fields in the node layer data base. If a match was found the coordinate information from the node layer record was attached to the crossings record in question. Many of the records could not be geocoded in this manner because of the inconsistencies in the rail and road name strings between the TIGER files and the crossings records for the Delaware rail-highway grade crossings data base.

These records were location referenced by locating the approximate location on the base map, zooming into the location of the crossing, keying in the attribute data into the identified point in the node layer, and finally, downloading the information into an ASCII file and building the crossings data base.

The advantage of location referencing to latitude-longitude coordinates is the referencing of data to a global and universal coordinate system that can be accessed and accepted by most other GIS software.

Analysis and Interpretation of Data

This system provides several options for analysis and interpretation of data and results. Procedures and conditional query abilities are available, including the ability to classify information by theme through color coding and representative icons. Another important spatial analysis capability is buffering. The user can specify a circular area around a crossing point and count the number of objects of a specific type that fall within that area. It is possible for the user to do selective spatial querying to estimate the number of point objects within a specified area, for instance, the determination of the location of another crossing within a quarter mile of a particular crossing, which is essential to the decision to close a crossing. Other objects that can be identified are accident locations, the presence of an intersection within a specified range of the crossing, and the presence of a certain type of land use in the vicinity of the crossing. Overlaying can be performed to gather land use data in the vicinity of crossings; this is one of the inputs to the expert system developed for the program.

Some of the conditions created for query are shown in Figure 1. Conditions were created to reflect the resource allocation strategies to be considered and the strategies for identification of deficient, hazardous, or significant attributes, as the case may be. The user could create thematic maps to classify crossings based on hazard index, traffic, or accidents.

Linkage to external procedures allows for flexibility and the selection of records from data base based on the specific nature of requirements of the type not possible within the functional capabilities of the GIS. An external class of procedures, including a program for executing the USDOT model and a KBES, for decision support were created using the external procedure interface in TransCAD (*1*).

DEVELOPMENT OF THE KNOWLEDGE-BASED EXPERT SYSTEM

Though the USDOT model gives a reasonably accurate estimate of the hazard potential, it does not consider several site-specific and qualitative factors that are important in determining the overall safety status of a crossing. Furthermore, the process of identification of the cause for deficiency or hazard at a crossing and the associated remedial action, is, by and large, one involving engineering judgment on the part of the field expert as well as compliance to codes, mandates, and procedures established for this purpose. The objective of this study was to develop a KBES that accounts for site-specific and judgmental factors that are not considered by the USDOT model and that also provides a guideline for selecting the most feasible combination(s) of actions that can be considered to

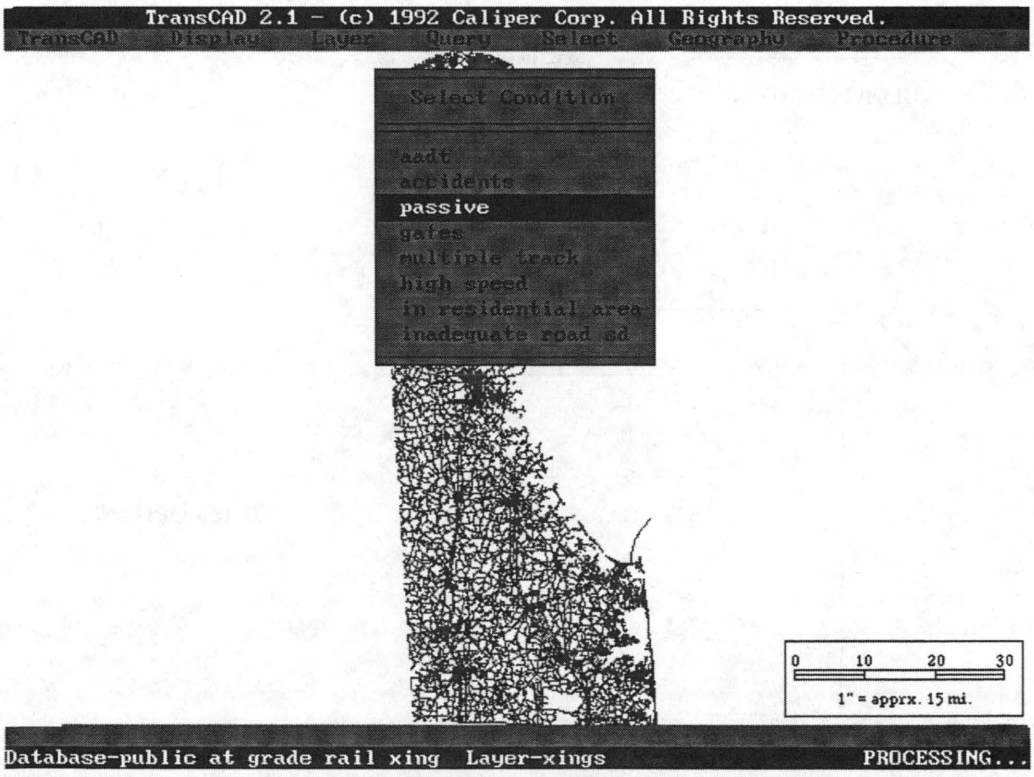


FIGURE 1 Some conditions created for query on attribute data fields.

TABLE 1 Site-Specific and Qualitative Factors Considered by KBES in Safety Decision Making

Track sight distance ratio (actual/available)
Road sight distance ratio (actual/available)
Night visibility
Illumination
Type of land use at the crossing
School bus traffic
Hazardous materials carrier traffic
Pedestrian traffic
Signing near the crossing and sign visibility
Speed limit compliance for the approach road
Proximity to intersection on approach road
Proximity to another crossing (for closure warrant)
Occurence of fatalities
Accident history specifics (train or nontrain related, severity, etc.)
Percentage of trucks at crossing

bring about accident reduction at the crossing. Table 1 consists of the significant site-specific and qualitative factors that are considered by the KBES.

The KBES uses these site-specific factors in conjunction with the USDOT accident index and the basic safety and inventory data (*3*) to assign an indicator of the danger level at a crossing. It further suggests a remedial action that could be considered for safety improvement at the crossing. Fifteen possible basic safety improvement alternatives were identified and established with the help of interviews with field engineers and from the codes and relevant literature establishing guidelines and procedures for safety improvement at crossings (*6,7,8,*). Some of these are standard safety device installation mandates that cover the three basic types of devices, namely, passive protection, flashing lights, and gates. Several other options, which are nonstandard and do not classify as safety devices but are nevertheless important for safety improvement and accident reduction, are installation of Stop signs, reduction of speed limits, improvement of pavement condition, illumination at the crossing, warrant for a pedestrian overpass, and finally grade separation and closure leading to elimination of crossings. Each of these options has a cost and effectiveness factor associated with it that the KBES considers for the particular case being evaluated.

Figure 2 shows the inference network for the KBES. The system is a forward-reasoning system developed in CLIPS (*2*) that goes through a phase-by-phase evaluation. The system modifies the accident hazard index for the crossing and finally suggests a set of possible action(s) for safety improvement. Some of these actions are due to mandates such as the closure of a crossing that has time-table speeds greater than 125 mph, categorized as a high-speed line, and the mandatory installation of gates at every multiple track crossing.

Extensive documentation and on-line help is also available to the user. Documentation includes providing the user with the inference

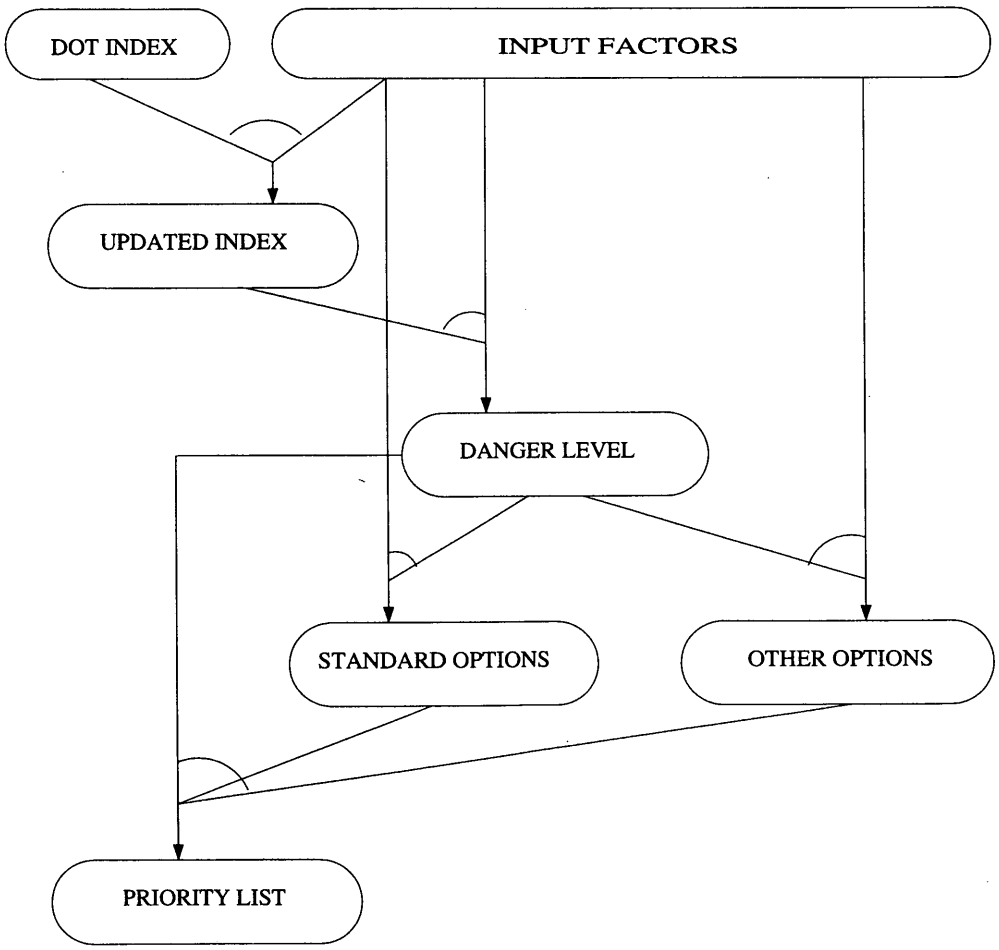


FIGURE 2 Inference mechanism of the rail-highway grade crossing KBES.

chain that the KBES uses in selecting a particular decision alternative. On-line help includes screens containing information on how to input data, run the system, view results, and the options available to the user.

The KBES can run in two modes. It can be user-interactive, with the user responding to queries by the system and supplying the input data, or the system can directly extract the input data from the GIS data base. Results of the system are also provided to the user in detailed descriptive form or can be converted into a text file that can be accessed and viewed within the GIS environment.

The user-interactive mode can be used to evaluate the effect of change in conditions for any single crossing while the data-acquisitive mode can be used to analyze several crossings at a time. This would be especially useful when processing a large crossings data base. Figure 3 shows a sample result of the KBES for a crossing giving the danger, current protection, and improvement, recommendations at the crossing.

Verification and Validation

"Verification is the process of building the system right, validation is the process of building the right system" (*9*).

Verification of the system was performed on similar lines to that done for conventional software programs. Verification in this case is restricted to that of the knowledge base, as the inference engine used is that of the CLIPS shell (*2*), verification of which is thus assumed. The two major steps involved in the verification process were checking the system for compliance with requirements, and checking for syntactical, logical, and other errors that affect the consistency and completeness of the knowledge base.

The interface between the system components was checked for proper functioning through passing of data with the GIS data base. The specification was for a user interface that permitted user interaction, automated extraction of data from the GIS data base, and finally, the presentation of results within the GIS environment. Compliance of the developed user interface to this specification was found to be satisfactory. Other aspects that were considered in system verification included the adequacy of explanation and help facilities and the modularity and maintainability of the system. The system was also checked for syntactical errors in the rule base including conflicting, redundant, subsumed, circular, dead-end, missing, and unreachable rules.

The validation methodology established was the presentation of a set of test cases and analysis of the results provided by the system in response to these test cases. Some of these test cases were from the crossings data base, and the results for these were compared with those of a human expert's evaluation. Other test cases were random and hypothetically formulated to test for theoretically possible and extreme combinations of input data. For the problem under consideration there is no clear-cut solution. The KBES suggests a set of options, as seen in Figure 3. Agreement of the experts with these options provided by the KBES for a series of test cases was the criterion for validation. In the majority of the cases tested on a random basis, the expert was satisfied with the results provided by the KBES. Some minor modifications were made as a result of the validation process.

The system was also tested for adequacy and completeness of the knowledge base by presenting random test cases with different combinations of input data to catch any errors of omission. Omission is a situation in which the system fails to reach a conclusion for a given input case, indicating incompleteness of the domain knowledge.

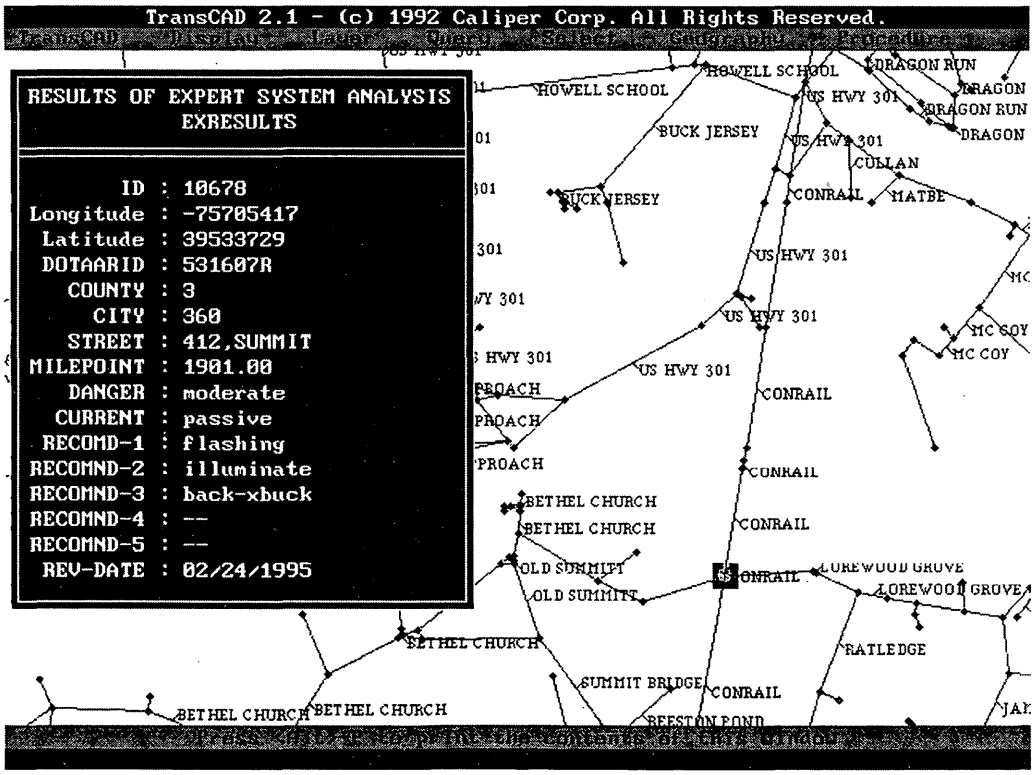


FIGURE 3 Results of KBES analysis for the crossing in response to the changes, viewed in the GIS.

INTEGRATING THE COMPONENTS OF THE CROSSINGS SAFETY PROGRAM

The previous sections described the development of the different components of the rail-highway crossings safety program for Delaware. The different components developed for the program that have been described are:

- A GIS application for rail-highway grade crossing safety in TransCAD (*1*),
- A location-referenced, expanded crossings safety data base that can be accessed by the various components of the program,
- the USDOT model for calculating accident index, and
- A KBES for updating the crossings hazard index, identifying hazards, and suggesting improvement action(s) at each crossing.

The interfacing of these components is the basis of the integrated functioning of the rail-highway safety management program. The format of input and output for each of the components of the system must be compatible with the others in order to interface them. Table 2 shows the input and output format and content specifications for each of the components. The framework of the integrated system is shown in Figure 4. The interface can either be from within the GIS or from outside the GIS depending on the requirements and the level of access desired. For the integration of these components, the required data is accessed by each of the procedures from the GIS data base and the results are passed back to the GIS. The integrated working of all these components is required for complete analysis of the crossings problem. The interface was achieved in two ways:

1. The user can access the KBES and the USDOT model program from within the GIS application environment by selecting the appropriate option from the procedures menu as shown in Figure 5. The GIS environment is used to display crossings, view the results of analysis, and provide input data to the interfaced components.
2. The control can be passed to the user from outside the GIS, as shown in the custom menu in Figure 6, that is, at the operating system level so that the user can access the different components including the data base builder. It also presents the user with documentation and help before using any of the components, assuming the user has minimal knowledge of the software/shells in question.

The crucial part of the interface is to maintain the location reference on any record being processed at every stage of the process. Figure 3 shows a sample result of a typical analysis for a crossing by the KBES. The external program takes in data from the record for that crossing and appends its location reference data to the data file containing the results of the analysis so that it can be passed back into the GIS for viewing and analysis.

When any menu item is selected from the menu shown in Figure 6, the appropriate chain of command substitutions and path changes are performed. Data are converted from the GIS data base to the format required. On-screen messages and documentation are provided to the user about the system component to enter, the specific function to perform, and options to execute within that environment.

TABLE 2 Input/Output Format and Requirements for Interfacing GIS, KBES, and External Procedures

Comp.	Input Specification	Input Contents	Output Specification	Output Contents
KBES	ASCII Format, User Screen Inputs, Worksheet Form	Hazard Index, Safety Factors, Site-Specific, Qualitative Information	ASCII Format, Screen Output, Worksheet Form	Danger Index, List of Decision Alternatives
External Procedure (USDOT Model)	ASCII Format	Crossings Safety Factors, Accident Data	ASCII Format	Updated Hazard Index, Priority List of Crossings
GIS	Spatial Data: TIGER, DLG, Scanned, Digitized Hard Copy Maps, Orthophotos etc.	Point, Line and Area Layers, Roads, Land Use, Census Blocks etc.	PCX, Other Graphic Format	Screen Image at any Display
	Attribute Data: ASCII, Text Files, Worksheet files etc.	Safety, Accident Data etc.	Worksheet, Text, ASCII, Other File Formats	Specific Attribute Data Dumped on Requirement

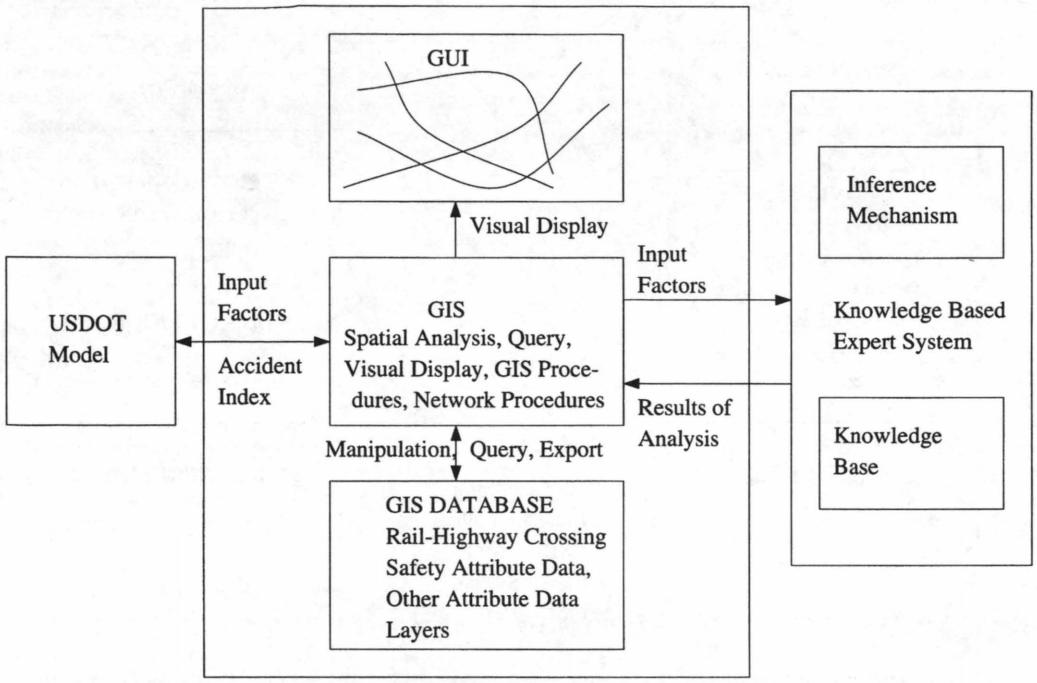


FIGURE 4 Framework of the knowledge-based GIS for safety analysis at rail-highway grade crossings.

FIGURE 5 Interface menu for KBES in CLIPS and USDOT model from within TransCAD rail-highway application.

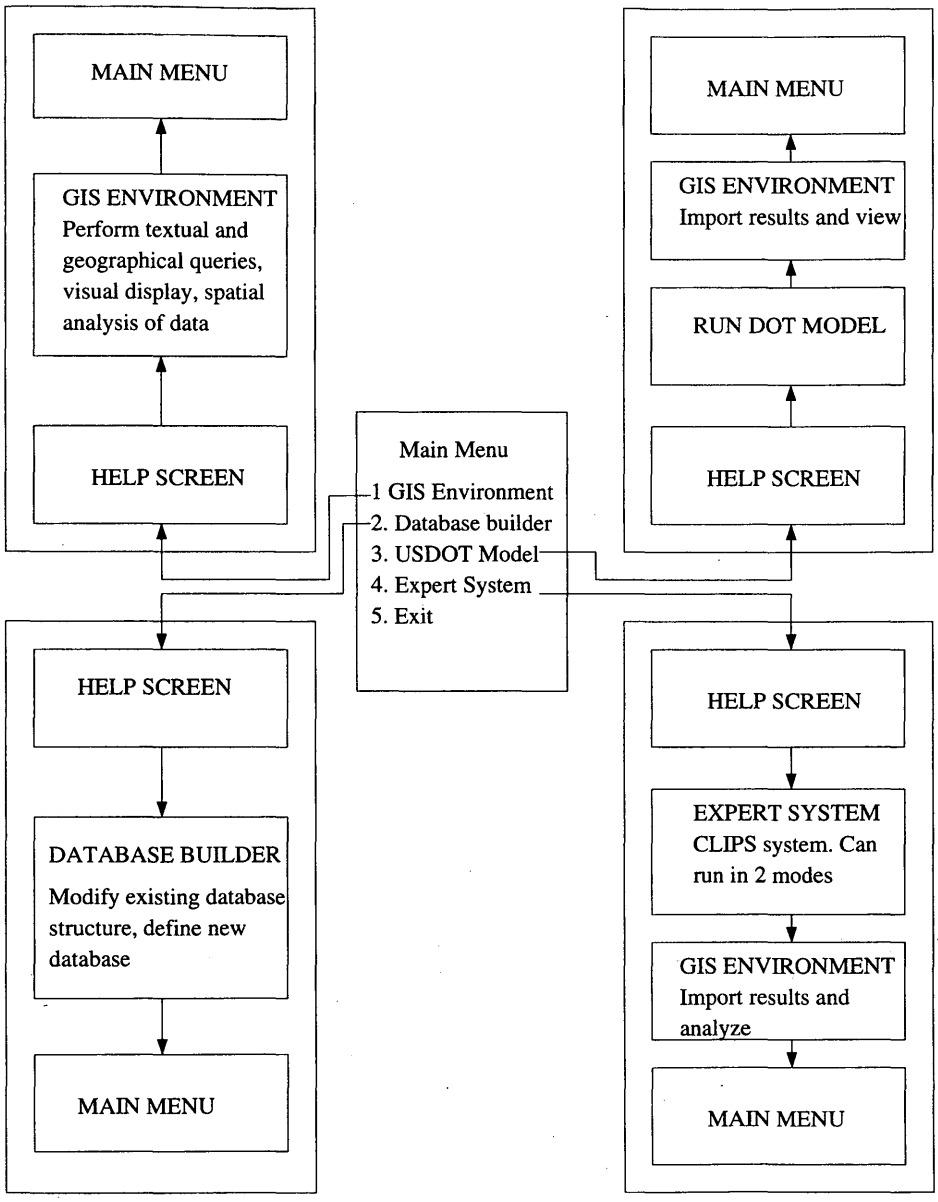


FIGURE 6 Custom menu of the interface, chain of processes in the execution of each menu item.

Figure 6 shows a view of the screens and environments presented to the user after execution of each of the menu items.

CASE STUDY

The case study consists of a rail-highway grade crossing for which some of the safety attributes have changed. The case study is presented to show how the system components could be used to analyze a particular crossing. The case shown here is an example of how the crossings data base can be updated, the USDOT model and the KBES run, and the results passed back to the GIS.

Some of the current safety attribute values on the crossing are seen by querying on the crossing location as presented in Figure 7. Consider a hypothetical change due to a new residential development and an increase in local traffic, as well as in hazardous material-carrier traffic caused by some rerouting in the vicinity of the crossing. The land use changed from predominantly insignificant use to residential use because of a new development in the vicinity of the crossing. This information is directly obtained from the land use layer, if it is regularly updated, or must be updated by keying in the information. The changes due to this new development are an increase in AADT, a change in land use type, and an increase in hazardous materials-carrier traffic.

The changes were updated in the attribute data base. Figure 8 shows some of the changed attribute values upon query at the crossing. The KBES was then run, and the results were passed back into a separate data layer created for that purpose, as can be seen in Figure 3. The danger level at the crossing after the KBES evaluation changed from low to moderate. The current protection is passive.

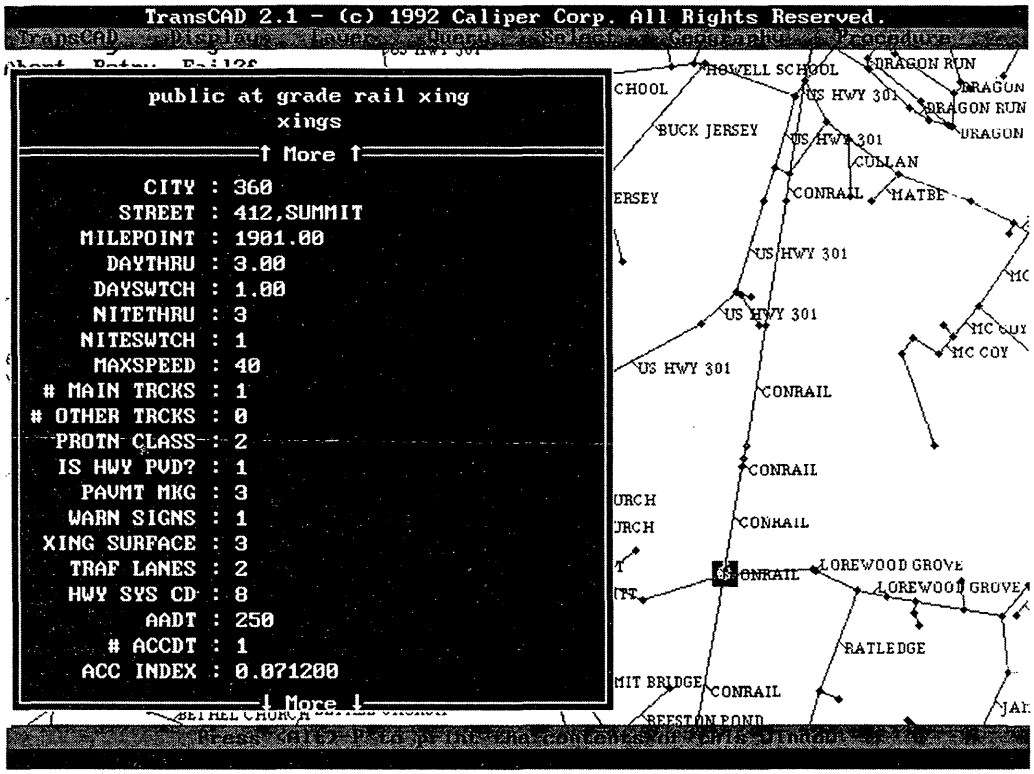


FIGURE 7 Query on crossing showing safety attribute data before changes.

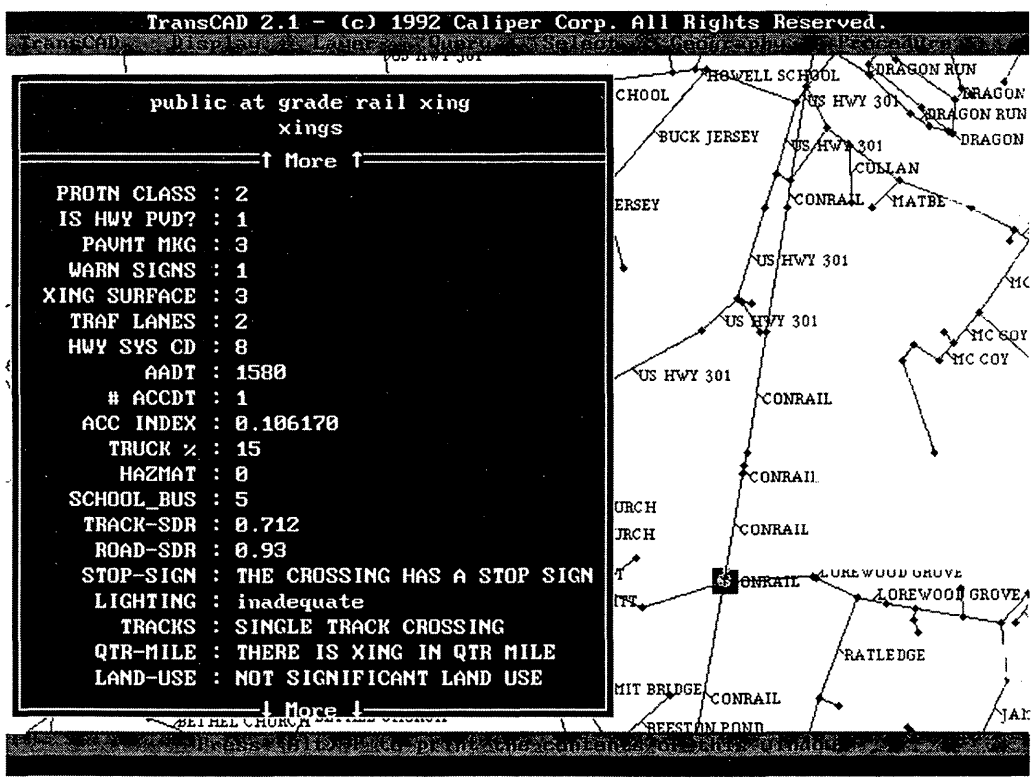


FIGURE 8 Query on crossing after changes in attribute data and USDOT accident index.

The system suggests the installation of a flashing light at the crossing and the improvement of illumination at the crossing. The installation of a backup crossbuck from sight distance consideration is also recommended. Note that all these recommendations only serve as a guideline, and that the field engineer needs to use his or her judgment to make the final decision. For example, in this case, the installation of flashing lights may be examined if the new traffic pattern, because of rerouting, is likely to become a permanent feature. Other options, such as installing a back-up crossbuck, may be considered more strongly because of the lower cost associated with it.

SUMMARY AND CONCLUSION

The work effort described consists of development of an integrated software package for management of safety-related data for rail-highway grade crossings. This involved developing a GIS application for visual display and spatial analysis of safety data and remedial actions; incorporating a program for calculation of an indicator of the accident potential using the USDOT model; and finally, the development of a KBES for modifying and prioritizing the indicator, and suggesting action(s) for safety improvement at the crossing. These components are integrated in such a way that the results and data required by each are compatible with the others, in order to enable visual access and presentation of results and data within the GIS environment.

The benefits of this work are the automation and efficient handling of large amounts of data, typical of a crossings management program for a large state, which can have a few thousand crossings. The use of GIS allows not only better display, spatial analysis, and an overall better visual perception of the problem, but also better data access and management, including access of related data from other layers and sources, easier editing and updating of data, and enabling all analysis to be performed by a user with minimal knowledge of the system.

The complete system results in an aid to resource allocation for safety improvement at rail-highway grade crossings. A more complete and meaningful analysis of the factors involved is achieved through the incorporation of the considerable amount of heuristic reasoning and engineering judgment that goes into the resource allocation process through the KBES. The overall benefits are not only the result of the better analysis and presentation capability of the GIS, but also of the incorporation of a decision support mechanism into the system.

ACKNOWLEDGMENT

Funding for this project was provided by the Delaware Department of Transportation through the Delaware Transportation Institute, Department of Civil Engineering, University of Delaware.

REFERENCES

1. *TransCAD—Transportation GIS Software, Version 2.0, Reference Manual.* Caliper Corporation, Newton, Mass., 1990.
2. *CLIPS Reference Manual, Version 5.1.* Lyndon B. Johnson Space Center, Houston, Tex., Sept. 1991.
3. Faghri, A., and N. Vukadinovic. *Evaluation of Rail-Highway Grade Crossings Program in Delaware.* Report 91-DTC-1. Delaware Transportation Center, University of Delaware, Sept. 1991.
4. Faghri, A., and M. J. Demetsky. A Comparison of Formulae for Predicting Rail-Highway Crossing Hazard. In *Transportation Research Record 1114,* TRB, National Research Council, Washington, D.C., 1980, pp. 152–155.
5. Hitz, J., and M. Cross. *Rail Highway Crossings Resource Allocation Procedure User's Guide.* Report FHWA-IP-82. Transportation Systems Center, U.S. Department of Transportation, Cambridge, Mass., 1982.
6. U.S. Government. *Code of Federal Regulations, 23.* Washington, D.C., 1991.
7. *Railroad-Highway Grade Crossings Handbook.* FHWA, U.S. Department of Transportation, 1986.
8. *Manual on Uniform Traffic Control Devices.* FHWA, U.S. Department of Transportation, 1978.
9. Gonzalez, A., and D. Dankel. *The Engineering of Knowledge Based Systems—Theory and Practice.* Prentice Hall, Engelwood Cliffs, N. J., 1993.

Publication of this paper sponsored by Committee on Artificial Intelligence.

Knowledge Acquisition, Representation, and Knowledge Base Development of Intelligent Traffic Evaluator for Prompt Incident Diagnosis

SOMPRASONG SUTTAYAMULLY, FABIAN C. HADIPRIONO, AND ZOLTAN A. NEMETH

Incident-related congestion on freeways costs the United States billions of dollars a year in loss of productivity, property damage, and personal injuries. Congestion on rural freeways is even worse than that on urban freeways because the resources needed for appropriate incident responses are not always nearby and high-tech equipment, such as closed-circuit televisions, is not available to detect and verify the incidents. Furthermore, incident responses are based only on the judgment of a patrol officer at the scene. Unfortunately, highly experienced officers may not always be available for managing such a situation. A relatively inexperienced officer may overreact or, with even more detrimental results, fail to call for sufficient response. Thus, to provide quick and suitable responses, an expert system for incident management (IM) is needed. The INtelligent TRaffic Evaluator for Prompt Incident Diagnosis (INTREPID) is being developed as a knowledge-based IM system to help a dispatcher manage an incident with the proper responses. INTREPID is part of the Advanced Rural Traffic Management Systems, which is a component of the Intelligent Vehicle Highway System. Unlike other systems, users can directly enter key information gathered from eyewitnesses to obtain prompt responses from the proper agencies and request the proper equipment without delay. The development of INTREPID is discussed and includes the following steps: (*a*) knowledge acquisition, including interviewing and literature searching, (*b*) knowledge representation, which involves the development of a decision tree, and (*c*) knowledge base development in a multimedia environment.

Traffic delays on urban, suburban, and rural highways throughout the United States have become a significant problem. The delays, which impede mobility and increase travel costs for road users, may be caused by recurrent or non-recurrent incidents. Recurrent incidents are events that always happen around the same time and place, such as traffic congestion during a peak period. Non-recurrent incidents, such as traffic accidents, are random events that can happen at any time and any place. Traffic accidents alone cost $70 billion annually (*1*). The toll is especially high in rural areas where the collision speeds are higher, increasing the likelihood of fatalities. Approximately 57 percent of fatal accidents occur in rural areas (*2*).

Congestion directly causes both inefficient movement of traffic and poor quality of environment. Its consequences cost the nation approximately $100 billion annually in loss of productivity (*1*). More specifically, each year the congestion amounts to more than two billion vehicle-hours of delay, more than 7 billion L of wasted fuel, and almost $16 billion in user costs. The FHWA predicts that by the year 2005 incidents will cause 70 percent of all urban freeway congestion, with a road users' cost of $35 billion (*3*).

Substantial research has been conducted to enhance incident management (IM) techniques. The conventional techniques range from reliance on eyewitness reports and IM agencies to the use of automatic incident detection systems and central control operators. Many of the techniques are inefficient, even the automated systems have high false alarm rates due to deficient incident detection algorithms. However, the detection problems are not significant if the technology of two-way communications is widely spread. The IM problems that still need to be worked out include the accuracy of incident verification and the application of appropriate responses.

The INtelligent TRaffic Evaluator for Prompt Incident Diagnosis (INTREPID) was proposed to speed up the IM decision making process and provide suitable responses. The functions of INTREPID are: (*a*) promptly verifying the nature of incidents and (*b*) applying appropriate IM strategies quickly to alleviate traffic delays. INTREPID, which employs expert system techniques to fulfill its goals, was developed as a knowledge-based system using an expert system shell and a multimedia technique. By combining these applications, the development of an intelligent traffic management system as part of the Intelligent Vehicle Highway System (IVHS) can be expected eventually to lessen congestion problems and reduce unnecessary costs to the nation.

The stages of developing INTREPID, namely, knowledge acquisition, knowledge representation, and knowledge base development, are discussed, and an illustration of a consultation process is presented.

KNOWLEDGE ACQUISITION

In the knowledge acquisition process, information is obtained through interviews and literature searches. Interviews were conducted with experts from the Ohio State Highway Patrol (OSHP) and the Ohio Department of Transportation (ODOT). The criteria for the selection of experts are discussed in the next section.

Criteria for Selecting an Expert

Because the selection of an expert is a difficult task in the development of a traffic management system, the following guidelines were established:

S. Suttayamully, Department of Civil Engineering, Suranaree University of Technology, Korat, Thailand 30000. F. C. Hadipriono and Z. A. Nemeth, Department of Civil Engineering, The Ohio State University, Columbus, Ohio 43202.

- Experts may be provided by related organizations, such as OSHP and ODOT. This implies that they have a significant expertise within the area of interest. For example, they must have extensive experience in managing incidents at the incident site or building an IM system.
- The expert should have an excellent record and be recognized as superior to others performing the same task.
- Experts need to be available and willing to participate throughout the system development processes.

After several meetings with OSHP and ODOT individuals who deal with freeway incident management in the state of Ohio, two experts in the following fields of expertise were selected:

- From the field of operations, OSHP's Lieutenant Harold E. Nease was selected by his organization and colleagues as a prominent incident manager whose proficiency is outstanding.
- From the field of traffic engineering, ODOT's George E. Saylor, whose expertise is in congestion management.

The following sections discuss the process of knowledge acquisition, which was separated into four phases: preliminary, intermediate, advanced, and organizational. The interviews during these phases were conducted by knowledge elicitors, who in the preliminary phase reviewed the relevant literature, selected domain experts, and gathered general information concerning the impact of incidents, IM strategies, current IM plans, and equipment needed in IM. The present Ohio freeway IM was found to have many shortcomings, including the unnecessary repetition of incident verification processes (causing errors during the information collecting processes) and assigning complete responsibility to only one officer.

In the intermediate phase, more specific information was requested from the experts concerning the type and nature of incidents, and the protocol in handling incidents. The questions used in the interview were divided into two groups: antecedent and consequence. They were formulated to facilitate the establishment of If-Then statements. For example, the case of an overturned truck that blocks all travel lanes and causes personal injury can be formulated as follows:

Antecedent: If (Type of Incident is Major) AND (Personal Injury is Yes) AND (Lane Blockage is All) AND (Fire is No).
Consequence: Then Action 1 and Action 2.

Action 1 may involve dispatching a patrol vehicle to the scene, dispatching an emergency medical services (EMS) team to the scene, or dispatching a Type C tow truck to the scene. Action 2 might involve notifying ODOT for possible rerouting and calling a radio station to broadcast the incident information. The questions were prepared by reconstructing all of the non-recurrent incidents that occurred on Ohio rural freeways I-70 East during the past several years. Each scenario that was developed helped elicit the experts' knowledge in handling real-life incidents.

In the advanced phase, the consistency of the information acquired in the intermediate phase was checked and If-Then statements were formulated. If conflicting information was found in the acquired knowledge due to misunderstandings or different thought processes, clarification was sought. Most of the conflicts were related to the procedure for managing an incident, which was not a difficult task to correct.

The fourth phase, organizational, focused on the creation of the knowledge structure. Having completed the formulation of all If-Then statements and having arranged the flow of thought processes, all the If-Then statements were rewritten using Level5 Object language, which is similar to standard English. The statements were then stored in the INTREPID knowledge base.

Experts' Criteria for Incident Classification

In this research, freeway incidents on I-70 East between Columbus and Zanesville, Ohio, were classified as minor or major, based on the judgment of the experts. This classification is useful because minor and major incidents call for different responses. In addition, this division allows for easy maintenance of the knowledge base.

Minor Incidents

In general, a minor incident involves a vehicle that has had a flat tire, run out of gas, stalled, overheated, or been involved in a fender-bender even if the vehicle is located on the shoulder and poses no hazard. According to the experts, any incident that does not involve a blocked travel lane, personal injury, fire, spilled hazardous material, or an area considered dangerous, is minor and requires no urgent response. An incident that involves one or more of those situations is regarded as a major incident. However, any minor incident that occurs in a hazardous area should be considered an urgent minor incident.

Management of Non-urgent Minor Incidents

As stated, a non-urgent minor incident does not involve personal injury, fire, or travel lane blockage. Such an incident is investigated by a patrol officer to determine its causes. The owner of the involved vehicle is notified and ordered to remove it as soon as possible. The involved vehicle is allowed to remain in a safe area on the highway for 48 hr before further action is taken. The common practice for OSHP is that any vehicle left at the incident scene for more than 48 hr is to be towed at the owner's expense.

Management of Urgent Minor Incidents

An urgent minor incident is one that occurs on the shoulder of a hill, a sharp curve, a bridge, or an on-off ramp. Such an incident must be cleared promptly to avoid severe consequences. For example, a vehicle stalled on a sharp curve, even on the shoulder lane, may cause a sudden slow-down in approaching traffic and could lead to a head-on collision. After a minor incident has been detected and deemed urgent, a patrol officer must secure the area with emergency markers.

Minor Incidents

Major incidents on rural freeways are similar to major incidents on urban freeways in that both have the potential for severe consequences. For example, an overturned truck on a freeway that has already caused a delay in existing traffic, may cause a secondary accident due to an unexpected stop or a stop-and-go situation. Without the proper IM system, such a situation may lead to the unnecessary loss of life and property.

Although it is easy to recognize a major incident, it is quite difficult to make a decision about managing it. For investigating officers, a major incident differs from a minor one in that it is multijurisdictional. Faulty communication about the incident can endanger a responding crew or motorists. This problem may be caused by misjudgment or the inconsistent management techniques of an inexperienced patrol officer. The following section proposes guidelines for recognizing a major incident and applying the proper management techniques to resolve it.

According to experts, a major accident is classified as a major incident if it consists of the following elements.

1. It may involve personal injury, usually the result of a serious collision or an overturned vehicle.
2. It tends to block one or more travel lanes.
3. It sometimes involves a fire, requiring the response of a fire unit.
4. Such an incident always draws the attention of motorists, which may result in a secondary accident and further block travel lanes.

Acts of nature and hazardous material spills may also be involved in major incidents. Ice, flooding, or fog on a roadway, strong winds, or a landslide, can create a negative impact on traffic. The results may include roadway closure, multiple-car accidents, and traffic rerouting, which can lead to delays, loss of productivity, waste of fuel, and more. Similarly, a hazardous material spill can complicate travel for motorists. However, this study is limited to major accidents. The system currently used to manage major accidents on I-70 East is described in the following sections.

Management of Major Incidents

Unlike minor incidents, major incidents are harder to manage and require several responding agencies and management techniques.

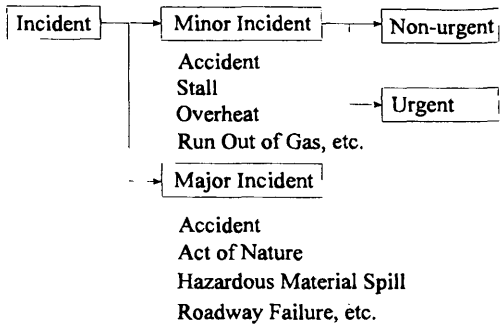


FIGURE 1 Two principle types of non-recurrent incidents.

Incident clearance takes longer and more types of equipment are needed since vehicles cannot be removed from the scene as easily as they can in a minor incident. Thus, the IM for major incidents is a complex task that demands great effort from every responding agency. If the agencies and equipment needed for response and clearance are known, over- and under-responses can be avoided.

Major incidents on I-70 East often receive less-than-proper responses because of difficulties with verification and inadequate response systems. The present system, therefore, requires revision.

KNOWLEDGE REPRESENTATION

The process of knowledge representation involves the use of a decision tree to represent the knowledge acquired from experts. The decision tree presents the structure of knowledge in a way that is relatively easy to view and understand. It is simple, so that one can directly check the consistency of acquired knowledge. Finally, it is helpful for developing the knowledge base and maintaining operations.

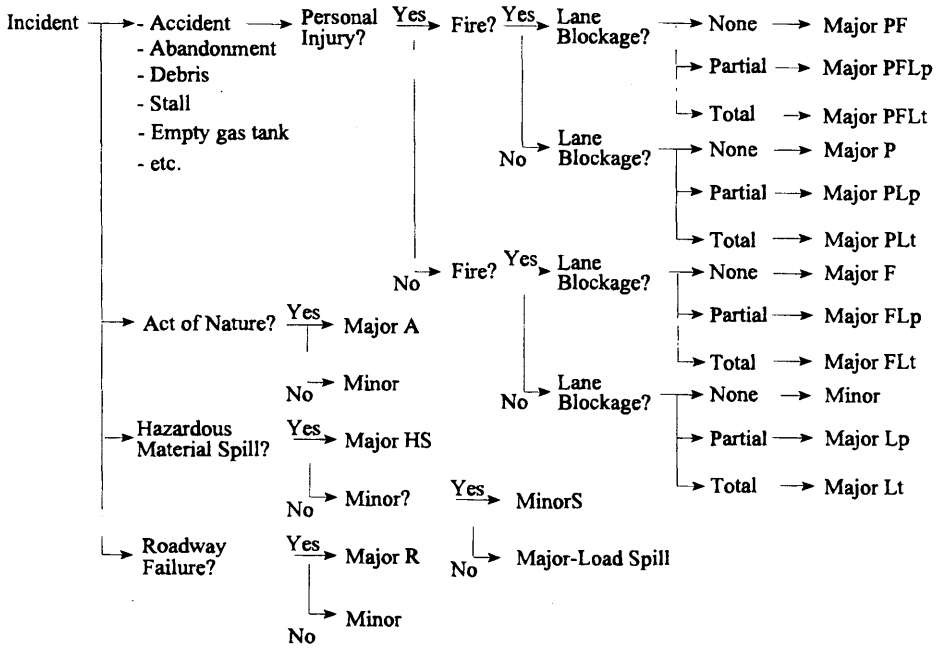


FIGURE 2 A decision tree used in the development of knowledge representation.

For the purposes of system development, incidents have been classified into two principal groups, minor (both urgent and non-urgent) and major. These two groups are shown in Figure 1. Figure 2 represents all possible causes of incidents on I-70 East. The tree consists of several nodes, such as accident, act of nature, hazardous material spill, and so on. Each node represents a question that requires input from users. For example, "Personal Injury?" represents the question "Is there any personal injury?" If the answer is "Yes," the tree leads the decision to "MajorP," which indicates that an accident involves a personal injury. The other branches will be explained in the following sections. In Figure 2 of the decision tree, nodes arise at either major or minor incidents. The development of the decision tree is continued until conclusions are reached.

Representation of Minor Incidents

The minor incident branch of the tree is shown in Figure 3. The tree proceeds with the question, "Hazardous Area?" If the answer is "Yes," an incident is considered an urgent minor incident. Otherwise, it is regarded as a non-urgent minor incident. In the following sections, the representation of the acquired knowledge on both non-urgent and urgent minor incidents is discussed.

Representation of Non-urgent Minor Incidents

When a hazardous area variable is "No," the flow of a decision tree follows the non-urgent minor incident branch, shown in Figure 3. After the type of minor incident is identified, the responses from the vehicle owner from the variable "Immediately Respond" must be obtained. In both urgent and non-urgent minor incidents, if the variable "Immediate Respond" is "Yes," the tree will reach its end node, which is a set of response actions or recommendations from the system. Otherwise, the tree, which represents additional branches of non-urgent minor incidents, continues further.

The tree provides specific response actions, such as MNUAB000 and MNUAB001. MNUABxxx is a file name containing a series of suitable responses that constitute recommendations for each incident. As an example, MNUAB001, which represents a response file for a non-urgent minor incident, recommends several actions, including dispatching a patrol vehicle to the scene, notifying the vehicle owner to remove the vehicle within 48 hr, and so on.

Further actions and a continued branch of this tree are shown in the lower part of Figure 3, which charts the characteristics of the rest of the non-urgent minor incidents. The type of vehicle and the severity of damage ("Towable?" and "Position?") are two other variables given in the tree.

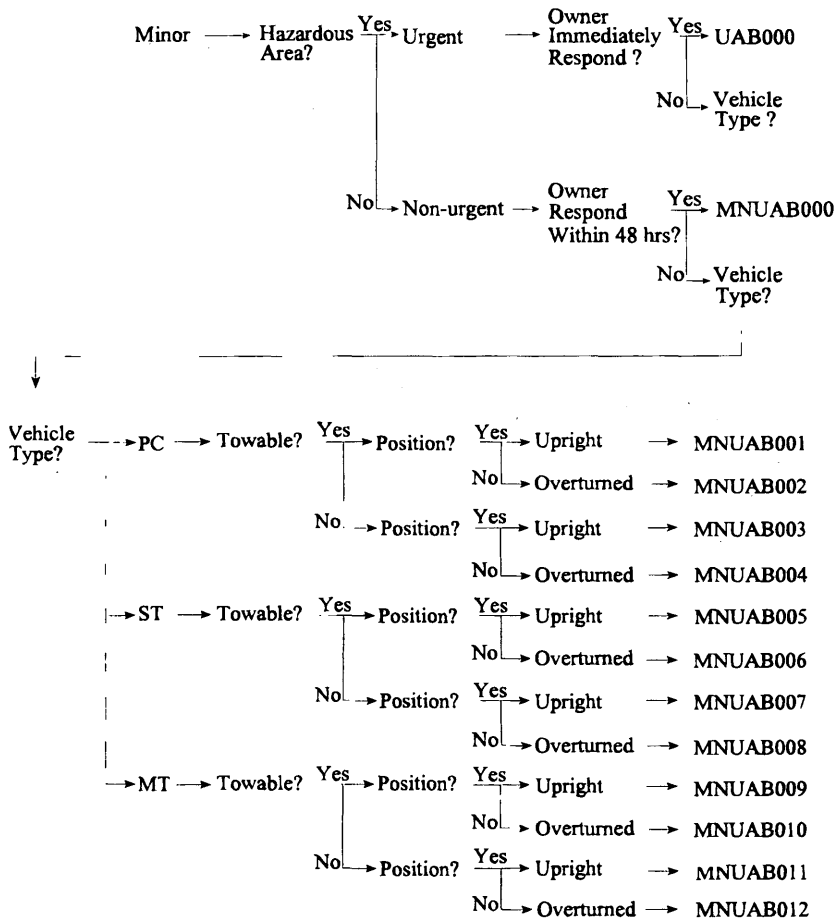


FIGURE 3 A minor incident branch of the decision tree with urgent and non-urgent minor incident and the continuation of a non-urgent minor incident branch of the decision tree.

Representation of Urgent Minor Incidents

The urgent minor incident branch is similar to a non-urgent minor incident branch of the decision tree, except for some differences in the details of responses, as shown in Figure 4. According to the experts, the urgent minor incident must be cleared as soon as possible for urgent responses. UAB001, an example of an end node, consists of the actions that should be taken in response to an urgent minor incident; those actions may include dispatching a patrol vehicle to the scene or dispatching a tow truck to remove the involved vehicle to a safe area, etc.

Representation of Major Incidents

In Figure 2, if the variable "Personal Injury" is "Yes," the next two variables, "Fire" and "Lane Blockage," must be identified. If the identification of these variables is "No," the tree will be classified in Figure 3 as a minor incident. Otherwise, the incident is classified as major. If the identification of those three variables is "Yes," this expanded portion is displayed in the upper branch of the tree. Otherwise it will be shown in the following branches of the decision tree.

In this study, MajorP is an incident that involves personal injury, and, as a branch of the tree indicates, it may be accompanied by other variables, such as "Fire," "Lane Blockage," and "Through Traffic." The variable "Lane Blockage" can be none, partial, or total, which indicates whether the lane is blocked by involved vehicles, spilled loads, accident debris, or injured persons. In Figure 2, "No Lane Blockage" would mean that the involved vehicles did not block a travel lane when an eyewitness reported the incident. "Partial" would mean that some travel lanes were obstructed but others were still open to traffic. Finally, "Total Blockage" would mean that all travel lanes were blocked by the incident. However, total blockage does not always mean that through traffic is impossible.

After the variables "Fire" and "Lane Blockage" are defined, the next variable to be evaluated is "Through Traffic" (Figure 5). If the variable "Through Traffic" is "Yes," this implies that through traffic is possible. Traffic may be diverted to any open lane or around the incident scene in the area adjacent to the shoulder lane without any potential danger to motorists. If the answer is "No," travel lanes may be totally closed due to hazardous material spills, load spills, removal of injured persons, lane blockage, geographical constraints, and so on. This expanded branch is shown in Figure 5, in which the variable "Time of Day" will be evaluated. The continuation of each branch of the tree is denoted by a circled letter.

In Figure 5, the time of day is divided into several intervals, midnight to 05:00 a.m., 05:00 a.m. to 08:00 a.m. and so on. After the time of day is known, the type of vehicles involved in the incident must be determined. In the development of INTREPID, vehicles are assigned to three major groups: passenger car, single-unit truck, or multiple-unit truck. Each type of vehicle requires different equipment for the clearance processes. After the type of vehicle has been identified, its position is then requested. The position is important for incident clearance, especially if the vehicle is a truck, which can be removed from the scene only when it is upright. If overturned, it must be made upright before removal from the scene.

The lower part of Figure 5 records the degree of vehicle damage, which may fall into one of three categories: functional, nonfunctional, or disabled. After inquiring about the vehicle's position, the tree then addresses the possibility of a load spill. If the condition of "Load Spill" is "Yes," the magnitude of delay is increased. If a load spill is involved, its type (e.g., sludge, live stock, construction material, etc) must be determined. By identifying the type of cargo spill, the proper agencies and equipment can be dispatched in a shorter period of time. This part of the extended decision tree is shown in Figure 6. If there is no load spill, the tree will reach its end node, which is MJP-000, or conclusion of the problem. In the lower part of Figure 6, if the variable "Owner Immediately Respond" is "Yes," the tree will reach its end node: MJP-MI00. If the variable is recorded as "No," the type of cargo spill must be provided, as shown in Figure 6. If the load spill represents a minor incident, the tree is constructed as shown in Figure 7, which is the continuation of Figure 2.

MJP-xxxx is a file of response activities, recommendations, or conclusions about the proper type of equipment needed for a par-

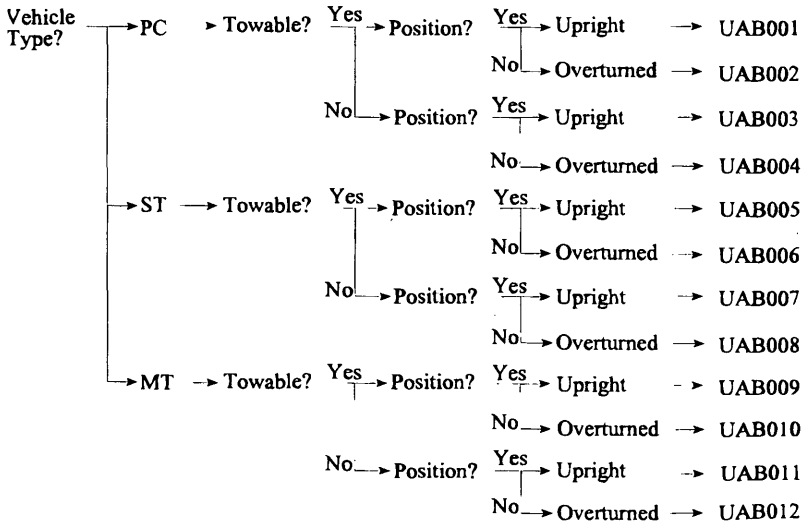


FIGURE 4 The continuation of an urgent minor incident branch of the decision tree.

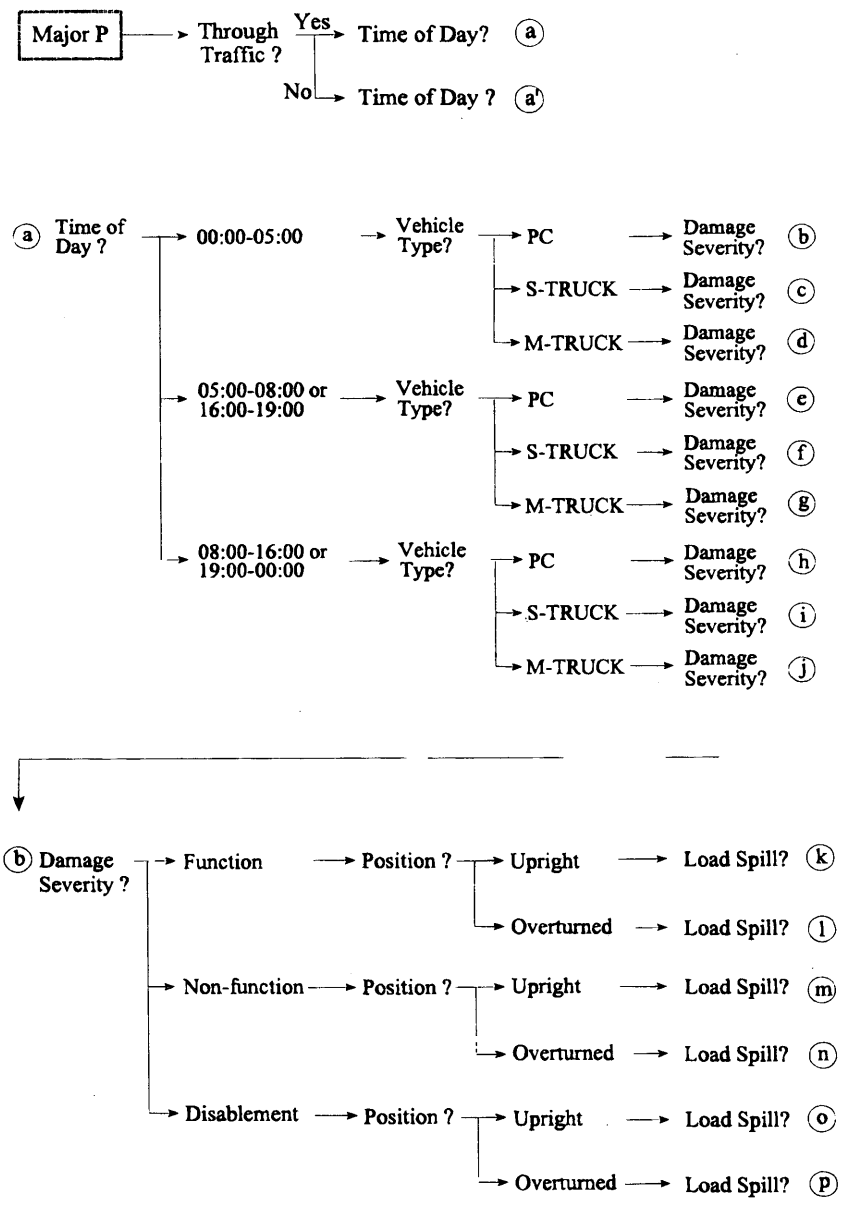


FIGURE 5 The decision tree of a major accident ("majorP") and continuation.

ticular major incident. Details of the recommendations are provided under each file name. Recommendations for addressing a particular problem can be accessed after the user has provided all requested information. Other branches of a major accident will not be mentioned in this discussion. In general, decision trees for other major incidents are similar, with some minor modifications.

THE KNOWLEDGE BASE DEVELOPMENT

The knowledge representation introduced in the previous section was used in constructing the knowledge base of INTREPID. In the following sections, the processes of knowledge base development and system implementation are explained.

Knowledge Base of INTREPID

Figure 8 shows the system architecture of INTREPID. We have selected Level5 Object (*4,5*) as the expert system shell. INTREPID was divided into two main portions: main program and supporting facility. The main program, which is developed using the expert system shell, consists of the first four components mentioned in the previous section. The first component, knowledge base, contains the knowledge acquired from the experts and is represented in the form of production rules. The second component, the inference engine, serves as the inference and control mechanism of INTREPID. The next component, the user interface, enables user-friendly fact entering or input, and controls and formats all output or end results for the user. This component also provides the user

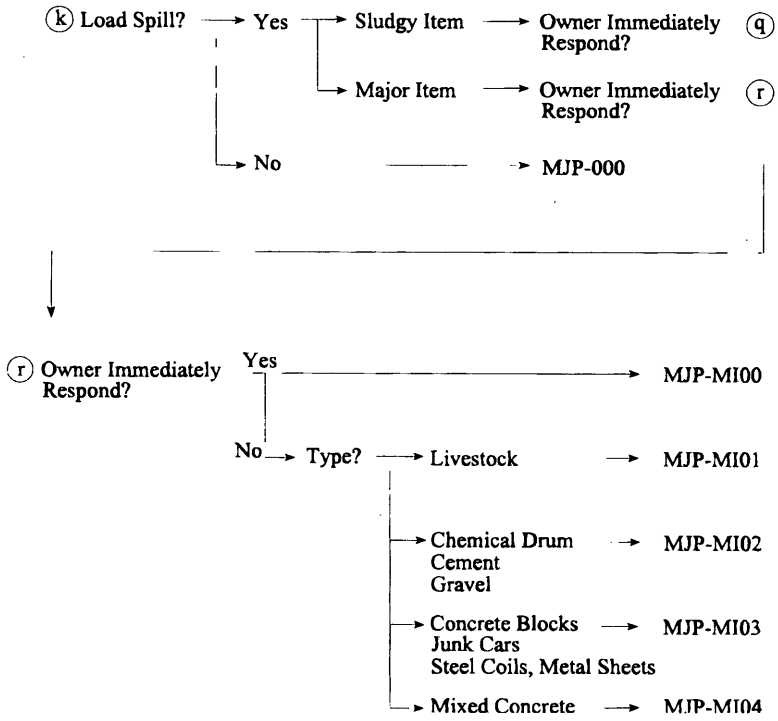


FIGURE 6 The continuation of the decision tree of load spill and the major item spill.

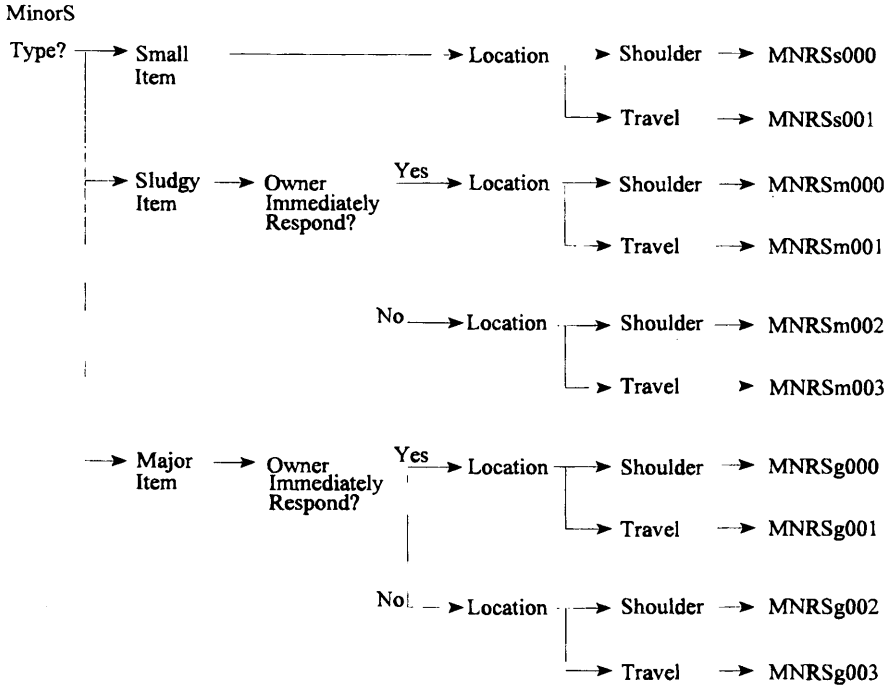


FIGURE 7 The continuation of the decision tree of major load spill.

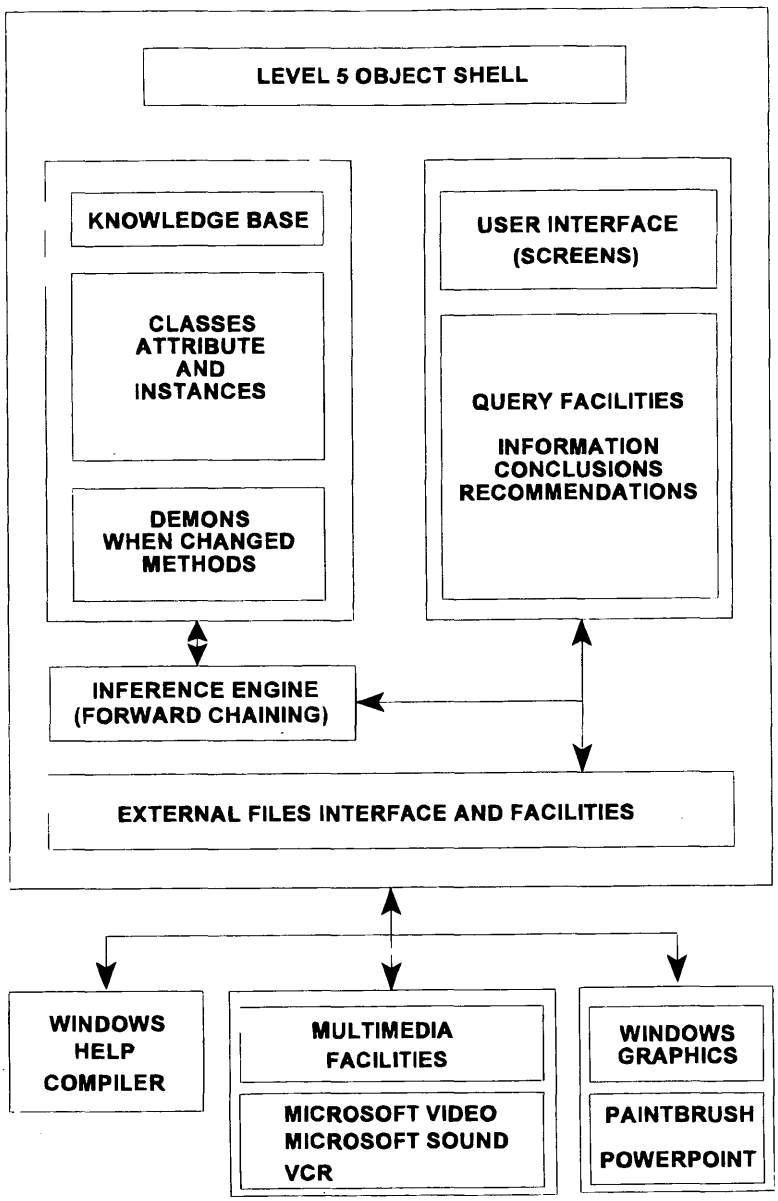


FIGURE 8 The architecture of INTREPID.

with an explanation facility. The fourth component, the external file interface, helps enhance the ability of INTREPID to interface with external computer programming, such as graphic, multi-media, and text files.

An example of INTREPID's screens and a user's consultation process are shown in Figures 9 to 13. The first screen (Figure 9) welcomes users to INTREPID and ask them to define the project route number and location. Figure 10 show users the incident management screen. Selecting "Abandonment" triggers the screen shown in Figure 11. The user then must provide the type of vehicle and the numbers and types of damaged vehicles. After the user inputs this information, the screen shown in Figure 12, which is a conclusion screen, will appear. Figure 12 gives all the necessary information in IM, such as the type and number of responding agencies and the type of equipment needed. Figures 13 shows a typical multi-media component in INTREPID, used for a demonstration purpose.

SUMMARY AND CONCLUSION

In an effort to reduce incident-related congestion, INTREPID has been developed as a comprehensive knowledge-based IM system. It is designed to assist a dispatcher in diagnosing incidents and initiating quick and appropriate responses on the rural freeway I-70 from Columbus to Zanesville, Ohio. Unlike other systems, INTREPID offers recommendations based on key information provided by users.

Knowledge acquisition and knowledge representation were the first two tasks undertaken in INTREPID's development. During the knowledge acquisition process a knowledge elicitor obtained key information to develop INTREPID's control screen and knowledge base. In addition, experts from OSHP and ODOT with experience managing minor and major incidents at the scene and developing the IM system in Ohio, participated in the study. Many incident

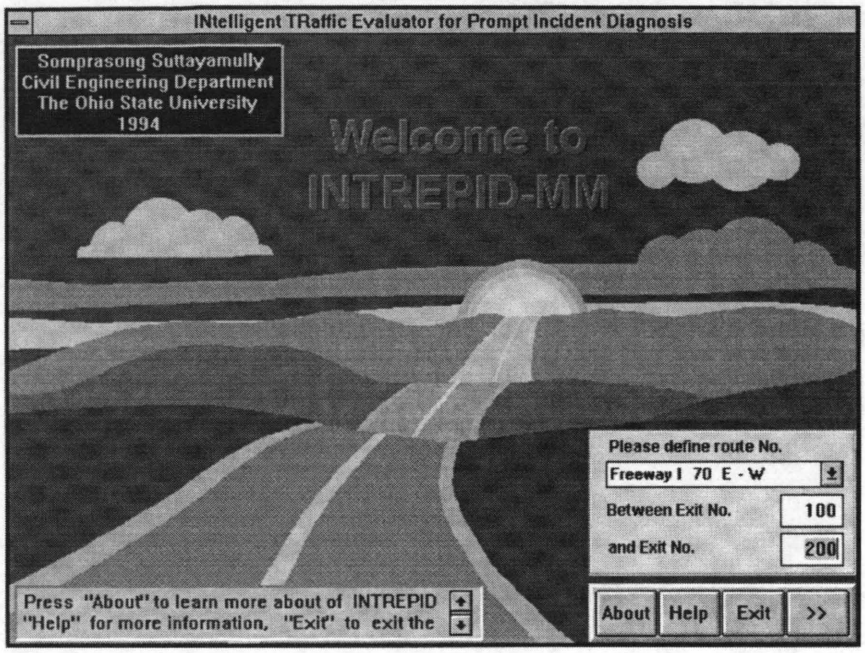


FIGURE 9 INTREPID's welcome screen.

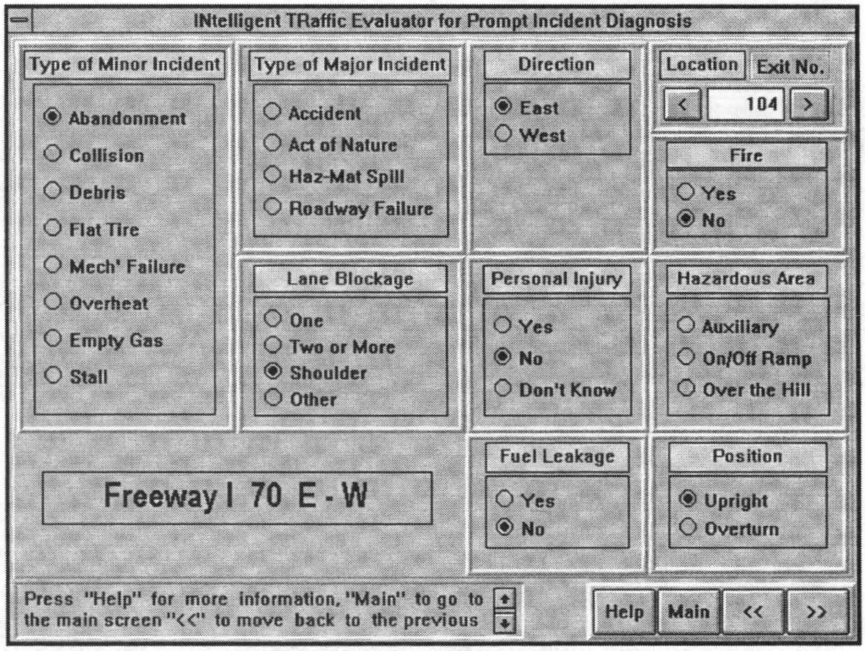


FIGURE 10 The incident management main screen.

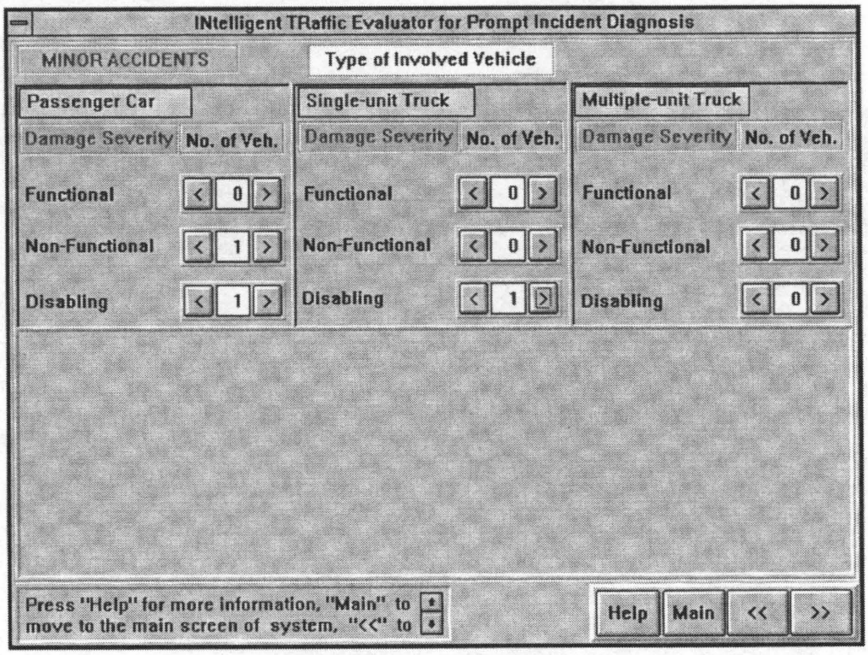


FIGURE 11 The severity assessment screen of minor incident.

cases were reconstructed to cross-check the consistency of the acquired knowledge. Furthermore, dispatchers, the potential users of INTREPID, were also interviewed to help the knowledge engineer build a user-friendly intelligent system.

The acquired knowledge of INTREPID was represented in a decision tree that is easily understood and transformed into production rules. The tree helps the knowledge engineer maintain the knowledge base, which makes INTREPID a robust system. With the decision tree, the knowledge engineer can easily construct the knowledge base and complete the development of a limited, simple-to-use INTREPID. However, like other intelligent systems, much still needs to be done in the refinement of INTREPID's knowledge base.

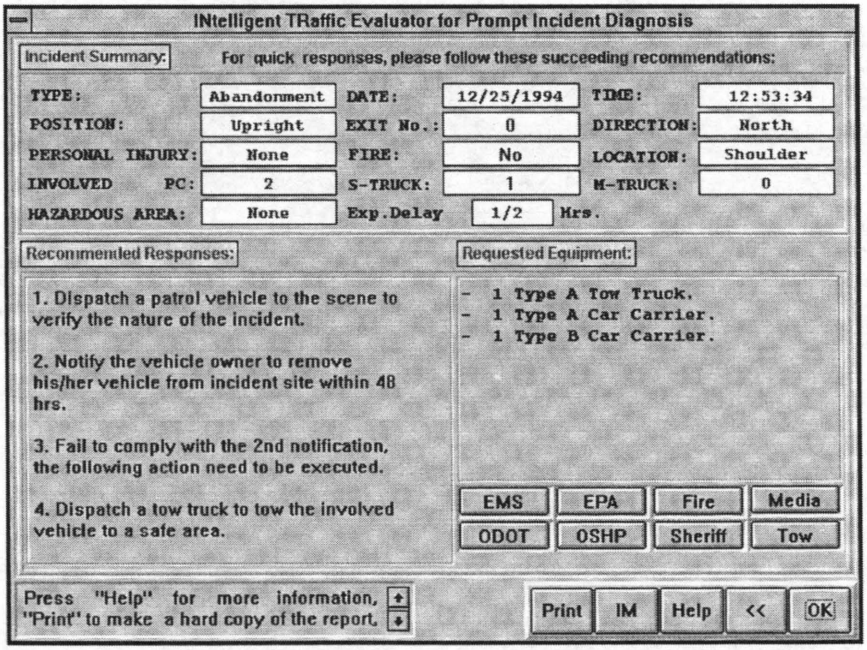


FIGURE 12 A recommendation screen of INTREPID.

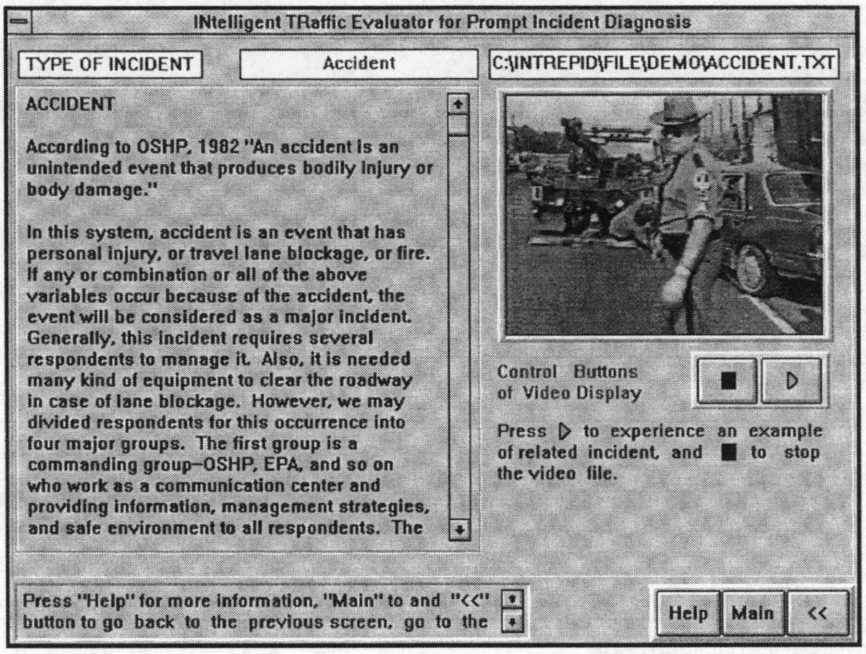


FIGURE 13 A multi-media component of INTREPID.

ACKNOWLEDGMENTS

The authors would like to thank Douglas W. Kullman and George E. Saylor of ODOT for their time in providing valuable information. The authors would like to give special thanks to Lieutenant Harold E. Nease and Lieutenant Joseph R. Montgomery of OSHP for their participation as experts throughout the process of knowledge base development.

REFERENCES

1. IVHS AMERICA. *Strategic Plan for IVHS in the U.S.*, Ohio Department of Transportation, Columbus, Ohio, 1992.
2. The Demonstration Project No. 86. *Incident Management Workshop*. Department of Transportation, Columbus, Ohio, 1993.
3. Mannering, F. L., M. Hallenbeck, and J. Koehne. *A Framework for Developing Incident Management Systems*. Washington State Transportation Center, University of Washington, Seattle, Wash., 1992.
4. *Level5 Object Reference Manual*. Release 3.0. Information Builders, Inc., New York, 1993.
5. *Level5 Object User's Manual*. Release 3.0. Information Builders, Inc., New York, 1993.

Publication of this paper sponsored by Committee on Artificial Intelligence.

Geographic Information System Inventory Data Preparation: Assigning Spatial Properties to Highway Feature Files Using Independent Data Sources

SCOTT A. KUTZ

Many transportation organizations are considering the use of Geographic Information System (GIS) technology for part of their overall approach to managing infrastructure data. They will often find that existing data sets developed and maintained over long periods of time may lack spatial properties needed for inclusion into the GIS environment. Even though these data sets represent the core inventory of features (highways, bridges, signals, and so on), the pre-GIS uses for the data typically never required any type of spatial information. However, most of the data sets did incorporate some type of location reference such as street and address range in urban areas or route and milepost in nonurban areas. This paper discusses the procedures used to assign spatial properties to features in one data set of highway features using an independent data set as its source of spatial information. The highway features were in an existing, nongraphic, address-delimited data set for the city of Chicago. The data set included street centerlines, bridges, viaducts, intersections, traffic signals, and vertical clearance/underpasses. The data set that contained the spatial properties was an independent base map data base of right-of-way, "midlines." The overall approach to reconciling these data sources and assigning the spatial properties of Illinois State Plane X,Y coordinates, network connectivity, and graphic representation to the highway Street Centerline features is discussed. The successful conversion achieved to date has produced a large number of valuable and useful highway features that have been loaded into the geographic data base. The difficulties encountered in matching between these two independent data sources developed by two different organizations are also presented.

The applicability of Geographic Information Systems for Transportation (GIS-T) is well documented in the literature, with three of the more prominent publications cited here (*1–3*). Establishing a geographic base map that serves as a common geographic reference is an essential step for implementing a GIS-T. In the context of this paper, the geographic base map is intended to be some representation of the base highway network. The location referencing system(s) used by an organization relate any stated location (route and milepost, control section and milepoint, street address, and so on) back to the adopted geographic base map. It is through consistent references to this common geographic base that a GIS-T is able to integrate information that may be physically stored in several different data bases (*4*).

The meaning of the "geographic base map" will be different for each implementing organization. The source and content of the data selected to form the "base" can be expected to vary with the organization. These variations are typically in response to differing needs for such factors as accuracy, required highway classes, the need to represent divided highways or multi-tiered highways as a single or multiple centerlines, and length of time needed to acquire, create, or correct inaccuracies in the data. Some possible sources from which base map data is derived include Topographically Integrated Geographic Encoding and Referencing (TIGER) files, aerial photography, or locally digitized data from original hardcopy source documents. Content variation often takes the form of determining which classes of highway are included versus those that are excluded from the data base. One approach may be to include any class of highway for which the implementing organization has some type of responsibility. This approach works well from the Federal or State perspective. However, it can result in some conspicuous gaps in network coverage when this approach is used at the local level.

It is often the case that any combination of problems can arise when actually creating a geographic base map. Some of the problems encountered include: difficulties in associating geographic coordinates to highway segments delimited with control section and milepoint references (*5*); inconsistencies between the locations of network links and road segments in a region (*6*); and difficulties when matching two data sets that are both encoded at the street block level in an urban area (*7*). Another problem can be that the organization does not have too few data but may actually have too many data. This is particularly a problem when redundant data sets exist. For example, the city of Los Angeles was identified as having two different versions of its 300,000-link street network during discussions at a recent GIS-T workshop (*8*).

This paper discusses a project that, although successfully converting a large portion of its data, has encountered its own set of problems when trying to construct the geographic base map for the street network in the city of Chicago. The problems are conceptually similar to the types of problems described above (*5–8*).

Constructing the geographic base map for the Chicago Department of Transportation (CDOT) entailed use of the existing public rights-of-way base map (the official city base map controlled by the Department of Planning and Development, DPD) as the basis for assigning spatial properties to the CDOT street inventory. The CDOT street inventory has been developed over many years as a tabular data set containing street and address range information but without the spatial properties of Illinois State Plane X,Y coordinate data, network connectivity data, or graphic representation data. Without these spatial properties, the CDOT data could not be integrated into a GIS data base.

Camp Dresser and McKee Inc., The Sears Tower, Suite 450, 233 South Wacker Drive, Chicago, Ill. 60606-6306.

The DPD Base Map features represented public right-of-way midlines ("centerlines") that did have spatial characteristics and also contained street name and address range data. However, data sets created and maintained by different organizations with different missions typically are not the same because the organizations each have their own approach to data structures and information management objectives. For example, the perspectives of the two organizations relative to the "streets" are somewhat different. These different perspectives are contrasted in Table 1. The resulting differences between the two data sets being discussed here complicated the process of matching the highway features to the Base Map right-of-way features.

This work was done as part of creating a Physical Inventory data base for the Chicago Citywide Infrastructure Management System (CWIMS), a GIS-based decision support system being developed for planning and coordinating capital improvement projects. Constructing the Physical Inventory portion of the CWIMS data base requires merging data from independent sources, mostly from various city departments that have developed these data sources independently over the years. Specifically for the CDOT data, the process involved the use of address and address range matching. Various techniques were attempted before adopting an address range-matching approach that yielded the highest match rate.

PROJECT OVERVIEW

The CWIMS project was started in May 1992, and its first release is scheduled for June 1996 (followed by a 1-year warranty period).

Objective

The objective of the project is to implement a computerized data base, mapping, and decision support system for planning and coordinating capital improvement projects. It is a GIS that will improve infrastructure planning capabilities within the city's public rights-of-way. CWIMS is intended to improve the city's ability to perform the following tasks:

- Identify and coordinate capital improvements to the water, sewer, street, and bridge infrastructure;
- Allocate capital resources for more effective management and maintenance of the $30 billion infrastructure network; and
- Identify potential infrastructure problems/conflicts before they become critical.

The CWIMS project is being performed by Camp Dresser & McKee, Inc., (CDM) for the Mayor's Office of Budget and Management. Participating departments include: Department of Water (DOW), Department of Sewers (DOS), Department of Transportation (CDOT), Department of Streets and Sanitation (DSS), the Department of Planning and Development (DPD), and the Department of Management Information Services (MIS). Several subconsultants are also participating in the project.

Geographic Coverage

The geographic area covered by the CWIMS project (the limits of the city of Chicago) is completely contained within the Illinois-East zone (number 1201) in the U.S. State Plane Coordinate System. The data base coordinates are based on the North American Datum of 1927 (horizontal datum). The data base units are "0.0305 meters" ("tenths of feet"); that is, a coordinate change of 100 data base units in either the X or Y coordinate represents a distance of 3.05 m (10 ft).

System Architecture

The CWIMS project is divided into the following contract deliverables: Project Design a series of data bases (Base Map, Physical

TABLE 1 Contrasting Perspectives of "Streets" by DPD and CDOT

Department of Planning and Development (DPD) Perspective	Chicago Department of Transportation (CDOT) Perspective
Base Map represents public rights-of-way (ROW) in the city	Inventory represents streets within the public rights-of-way
Includes ROW information for Expressways	Does not include data on Expressways
Does not include information for Ramps	Includes data on Ramps
View is "flat" (planimetric), i.e., ROW which cross also intersect	View is "non-planar", i.e., overpassing streets can cross over other streets without intersecting
Base Map is non-directional	Streets are directional (i.e., recognize a direction of travel)
A ROW "corridor" is viewed as a single entity	A ROW "corridor" can contain multiple instances of street features, such as the east-bound lanes and west-bound lanes of a median-separated boulevard
Instances where there is a short "jog" in the ROW where two adjacent sections of ROW do not exactly intersect are often explicitly represented in the DPD data (but not always)	Instances where there is a "jog" in the street centerline on two sides of a general area of intersection tend to not be explicitly shown in the CDOT data.

Inventory, Condition Assessment, Capital Projects, and Current Replacement Cost); a series of implemented systems (Capital Planning Decision Support, Systems Data Transference, Application Programs, Security, and Mainframe Diagnostics); and training.

The focus of this paper is in the area of the Physical Inventory Data Base deliverable. More specifically, this paper discusses the experience gained while using the official city base map provided by the DPD as the basis for associating spatial properties to the Physical Inventory data provided by the CDOT. The CDOT data was nongraphic with locations specified by street name and address range. The spatial properties that were assigned to the CDOT features are listed below:

- Illinois State Plane X,Y coordinates for the location of CDOT features,
- Network connectivity (which also defined the logical relationship between features), and
- Graphic representation.

The hardware and software environment for the CWIMS project reflects the city's interest in maintaining a central data base as the repository for all infrastructure data while taking advantage of the benefits offered by workstations and personal computers for system users. Data access is provided by use of a city-wide Transmission Control Protocol/Internet Protocol (TCP/IP) network with departmental servers. The central data base resides on the city's IBM mainframe computer in DataBase 2 (DB2) relational data bases. The GIS portion of the infrastructure data base is managed by the IBM product geoManager, part of the IBM Geographic Facilities Information System (GFIS) product set. Many other DB2 relational tables of infrastructure data exist external to geoManager. These tables can be accessed either directly or through geoManager data retrievals using relational "JOIN" operations.

The project plan calls for the delivery of analysis, planning, and decision support capabilities to CWIMS users via the ArcView2 desktop GIS product from Environmental Systems Research Institute, Inc. The data will be extracted from the city-wide corporate data base. A data translator will convert the GFIS data model stored in the geoManager data base into ArcView2 format during data extraction.

CDOT DATA MODEL

The CWIMS project is building its data model consistent with the GFIS architecture. The GFIS data model is a two-dimensional hierarchical (network) model that includes individual feature instances grouped into user-defined layers. Within each layer, each feature's network connectivity is explicitly modelled by permitting one or more nodes to exist at any X,Y location. This approach implements the nonplanar aspect of the data model, permitting multiple features to exist at the same X,Y location without the need for them to be network connected. Graphic representations are also explicitly defined for each feature, with the flexibility of defining any number of scale-dependent pictures that can be displayed at any location required by the application, even if that location is different from the feature's data base coordinate location. The GFIS data model also provides for up to two levels of "child" (dependent) features that inherit their location and their existence from their "parent" features (*9*).

The CWIMS CDOT Physical Inventory data model includes the features outlined in Table 2. A view of the CDOT features converted to date in the Chicago central business district are shown in Figure 1.

METHOD FOR ASSIGNING SPATIAL PROPERTIES TO CDOT DATA

In addition to the tabular address-delimited data provided by CDOT, the spatial properties of Illinois State Plane X,Y coordinates, network connectivity, and graphic representation must be added before the CDOT Physical Inventory data can successfully be incorporated into the CWIMS data base. This section outlines the process used to assign this spatial data.

DPD Midlines (Right-Of-Way Limits)

For the purpose of this conversion, the portion of the DPD Base Map data that provided the basis for assigning spatial properties to the CDOT data is the "midline" of the public rights-of-way (ROW) limits. These ROW midlines are modelled in the CWIMS data base as span features. The midline attributes of interest for matching the incoming CDOT data are listed below.

- Common street name,
- Unique midline identifier, and
- Address ranges: low/high even addresses, low/high odd addresses.

By virtue of its being in the geoManager data base, each instance of a midline feature also had Illinois State Plane X,Y coordinates for its endpoints, network connectivity at each end (expressed as nodes), and its graphic representation consisting of a polyline between the two endpoints. The spatial data associated with the Midline features are depicted in Figure 2*a*.

Corresponding CDOT Street Centerline Data

Although the CDOT Physical Inventory data consist of more feature types than street centerlines, all other CDOT data are located relative to the converted street centerlines. Therefore, the primary focus for converting the CDOT data into CWIMS was to assign spatial properties to the CDOT street centerlines by matching to the DPD midline features representing the official City Base Map already in CWIMS. Similar to the above discussion about the relevant midline attributes, the street centerline attributes that were used in matching to the ROW midlines follow:

- Street name,
- Street ID number (unique for CDOT, different from midline ID), and
- Address range (low/high addresses, *not* separated by even and odd).

Conversion Process for the CDOT Features

At first glance, it appeared that the CDOT street centerline data contained sufficient information to successfully identify the corre-

TABLE 2 Features in the CWIMS CDOT Data Model

Identifier	Type	Description
CSTREET	Span feature	CDOT Street Centerline: typically represents a portion of a street between two intersections (with an accompanying address range). In some cases, a Street Centerline feature may span a distance less than a complete block or cover a distance of multiple blocks depending on the locations of intersecting Streets. Network connectivity: All Streets at an intersection are connected to the same network node. This approach will be modified somewhat to assign different nodes for different "levels" when the conversion effort expands to include multi-level (multi-tiered) streets.
BRIDGE	Point feature	Bridge: elevated structure which "carries" a CDOT Street segment across a body of water, other street segments, railroads, etc. Network connectivity: Attached to the closest network node of the Street Centerline feature which it "carries".
VIADUCT	Point feature	Viaduct: elevated structure which "carries" a railroad, pedestrian walkway, or anything other than a CDOT Street over a CDOT Street segment. Network connectivity: Attached to the closest network node of the Street Centerline feature on which it is located.
CINTER	Point feature	Intersection: point at which two or more streets attach or converge. Network connectivity: Attached to the common network node shared by all Street Centerline features at the Intersection location.
SIGNAL	Dependent feature	Traffic Signal: Child feature of an Intersection which stores data about the number, status, and operation of the signals at the Intersection. Network connectivity: None, not applicable for dependent features.
UNDRPASS	Dependent feature	Vertical Clearance/Underpass: Child feature of a Street Centerline which stores data about the type and clearance between the pavement surface and the restricting structure above the street (such as a bridge or viaduct). Network connectivity: None, not applicable for dependent features.

sponding DPD midline features. Both data sources contained street name and address range on specific streets, so matching that information would provide access to the spatial data needed for the CDOT street centerlines.

It quickly became apparent that the mapping of Street Centerlines to ROW midlines was not one-to-one. Almost 1,600 more DPD ROW midlines existed than CDOT street centerlines. Not surprisingly, there were also several cases in which the spelling of the street names differed between the DPD data and the CDOT data. The spelling inconsistencies were fixed before continuing with the remainder of the conversion process.

The discussion in this paper focuses on converting the street centerline features. Once the centerlines were converted, assigning spatial properties to the remainder of the CDOT features was primarily an exercise in geocoding the point feature's street address.

Converting Street Centerline Features

Three different approaches were evaluated for the street centerline conversion.

Approach 1: Identify Matching Midlines Based on From/To Intersecting Streets

This approach attempted to use the from/to intersecting street information in the CDOT data to identify the midline feature that represented "one end" of the CDOT street centerline and likewise for the "other end." This approach was not selected because it only matched about 65 percent of the streets, took several hours of mainframe CPU time to complete, and could not handle the many-to-one case in which multiple midline features exist along the extent of a CDOT street centerline. This latter limitation could result in assignment of incorrect X,Y endpoint coordinates and a "gap" in the graphic representation.

Approach 2: Identify Matching Midlines Based on Low/High Addresses

The DPD midline data included low/high address ranges for both the even and odd addresses. In contrast, CDOT data included only low/high address ranges without distinguishing between even and

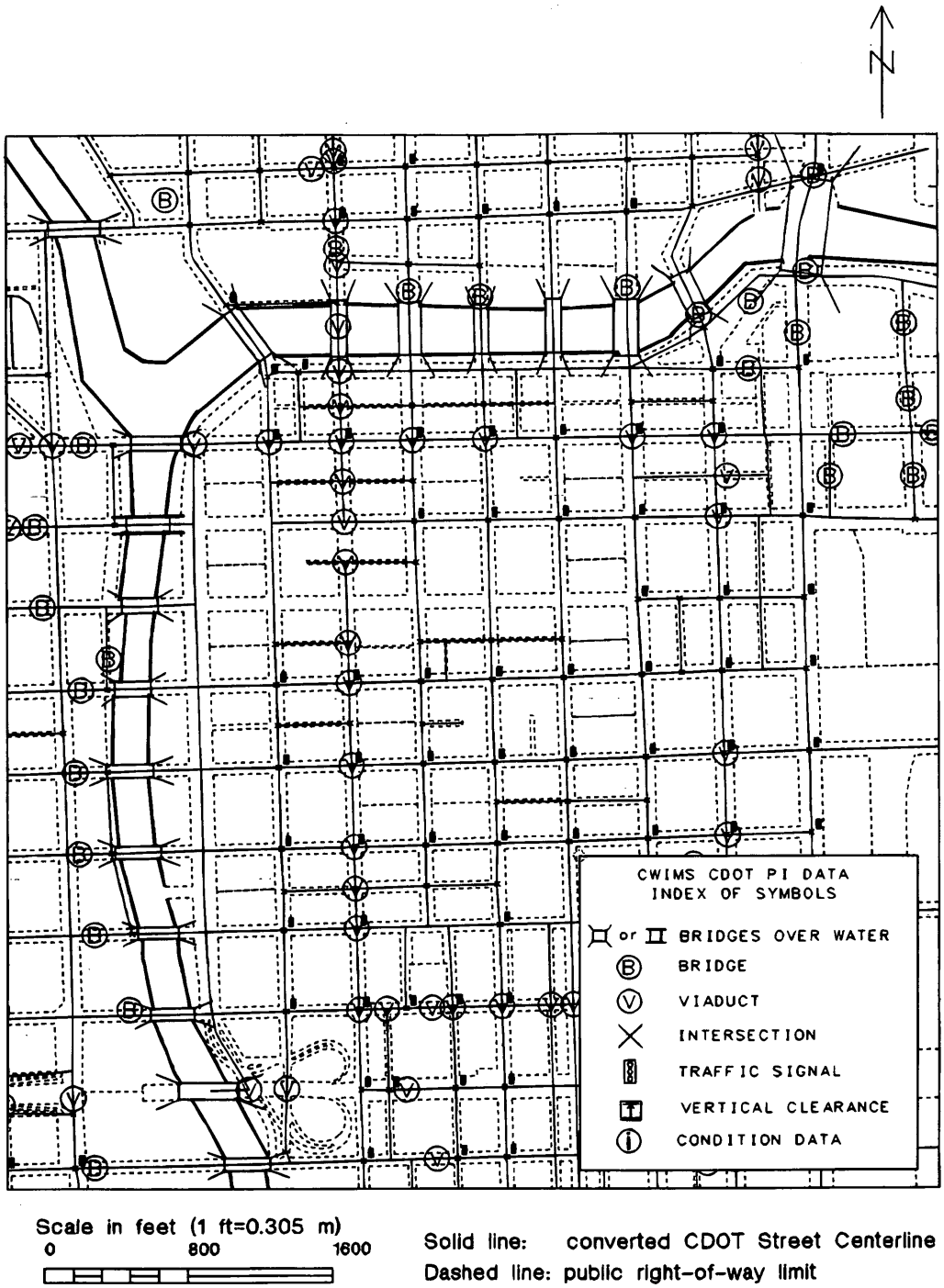


FIGURE 1 Converted CDOT highway features in the Chicago central business district.

odd addresses. Approach 2 relied on pairing streets with midlines based on similar low and high address ranges. The concept of "similar" was implemented in multiple passes through the data, with the second pass relaxing the criteria somewhat in an attempt to improve the match percentage. The match criteria for the two passes follows:

- Pass 1: CDOT street matches a DPD midline with the same name and the street has a low address that matches either the low odd or even midline address and a high address that matches either the high odd or even midline address.
- Pass 2 (for any street centerlines not matched on Pass 1): CDOT street matches a DPD midline with the same name, and all four midline addresses (low/high even and low/high odd) fall into an address range calculated as 5 below the CDOT street low address and 5 above the CDOT street high address.

This approach was more successful, but was not adopted. Pass 1 resulted in a 35 percent match rate, and Pass 2 increased the match percentage to 74 percent. This approach proved to be susceptible to anomalies in the DPD midline address range data. These anomalies

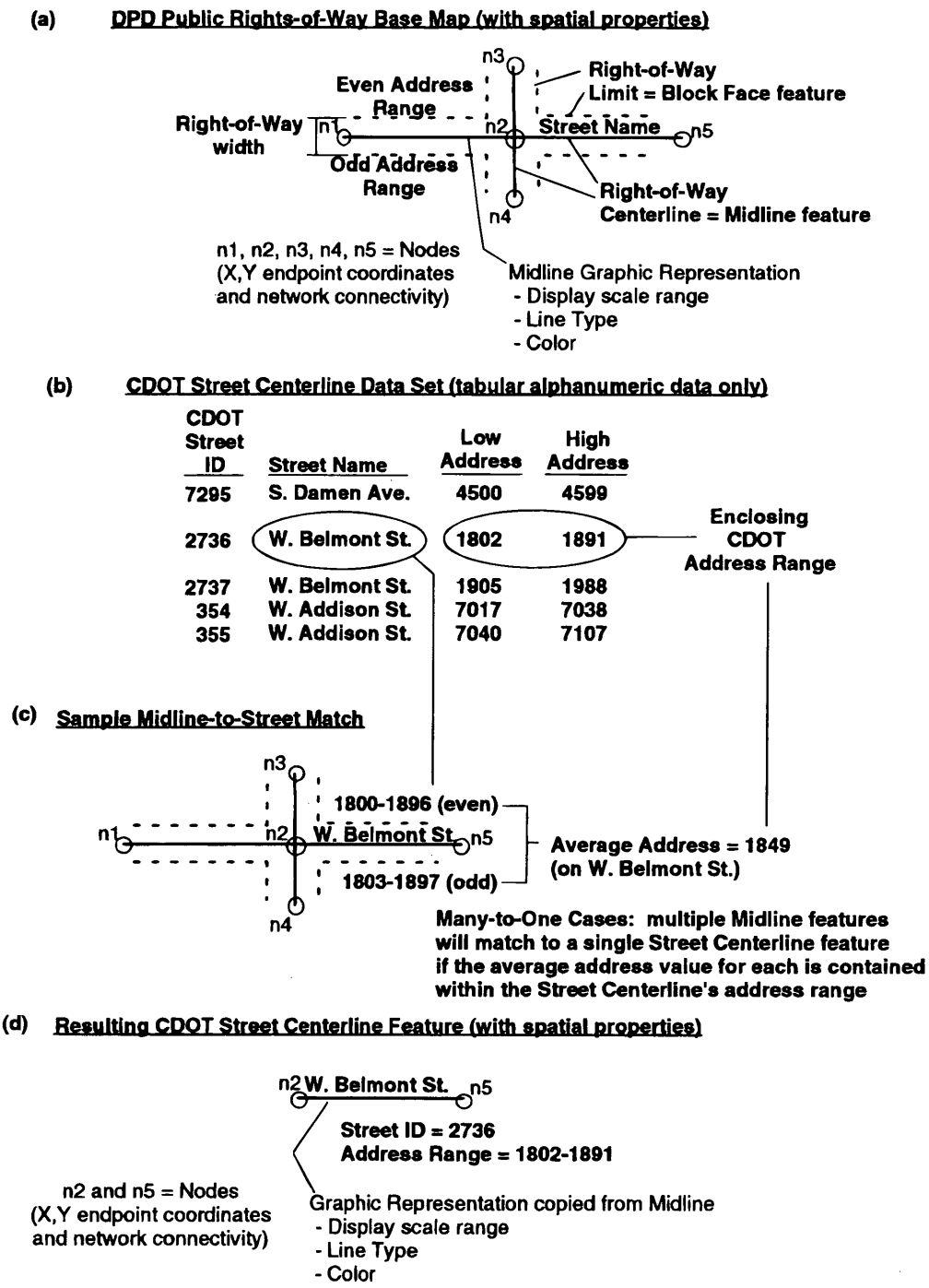


FIGURE 2 Use of average midline address range to match street centerline.

included cases in which the low even and odd addresses were more than 10 addresses apart. For example, the low even address may be 400, but the low odd address might be 425. Similarly, cases existed in which the high even address was different from the high odd address by more than 10 addresses. A few cases were also encountered in which the low even/odd address was larger than the high even/odd address.

Approach 3: Identify Matching Midlines Based on Average Address

This approach first calculated the "average address" for each DPD midline feature across the four values of low/high even and low/high odd addresses. This midline feature was then matched to the first CDOT street centerline that had the same street name and

whose address range included this average midline address. Approach 3 proved to be the most flexible and resulted in a match percentage of 92 percent. One major advantage is that it permitted more than one midline feature to be matched to a single CDOT street centerline, which was the desired action in those cases in which the extent of a street centerline included multiple midline features. Approach 3 was selected for the actual CDOT data loading into the CWIMS Physical Inventory data base. An overview of Approach 3 is shown in Figure 2. Figure 2*a* depicts the existing spatial characteristics of the DPD Base Map data, including the midline features. Figure 2*b* shows that the incoming CDOT street centerline data were tabular alphanumeric data. Figure 2*c* shows the use of the average midline address (along with a common street name) as the method for matching to the CDOT street centerline data. Note that in many-to-one cases in which multiple midline features correspond to a single street centerline, multiple midlines would be matched to a single street if their average address value fell into the range of a single street centerline feature. Figure 2*d* depicts the resulting street centerline feature with its newly assigned spatial characteristics of X,Y coordinates for its endpoints, nodes for its network connectivity, and its graphic representation (copied from its matched midline feature or features).

Large areas of the Chicago street network are characterized by a well defined regular grid. In those areas, almost 100 percent of the CDOT street centerline features were successfully matched to the DPD ROW midline features. The adopted technique (Approach 3) did prove to give some "false-positive matches" in densely developed areas of the city where traffic control medians or triangular-shaped dividers exist.

Street Centerline Spatial Data

The spatial data assigned to a matched street centerline is summarized in Figure 2*d* and were determined as follows:

- The street centerline feature was added to the "C" layer (CDOT).
- The X,Y coordinates for the endpoints were assigned to be the Illinois State Plane X,Y coordinates for the corresponding endpoints of the matched midline feature or features.
- All street centerline features were assigned the same "node value" at their respective X,Y endpoints (on the C layer). This guaranteed a connected street network because all street features with an endpoint at any common X,Y coordinate will occupy the same network node at that location. The node assignment process will be modified to some extent in the future when it becomes possible to process information for multi-tiered (multiple level) streets. In this case, multiple nodes will be required at the same X,Y coordinate to ensure proper network connectivity across the different tiers.
- The graphic representation of the street centerline was obtained by copying the existing graphic representation of the midline feature or features that had been combined to represent the street.

Converting the remainder of the CDOT features was dependent on the successful conversion of the street centerline feature or features with which the bridge, viaduct, intersection, or vertical clearance underpass was associated. Once the corresponding street centerline(s) was converted, assigning spatial data to the other types of CDOT features was accomplished by converting the point feature's street address into an Illinois State Plane X,Y coordinate, assigning it to a node at the near end of the street centerline feature, and assigning a designated type of symbol for its graphic representation. The focus of this paper is on the process and issues associated with converting the street centerline data, so no additional discussion will be provided for the other CDOT features.

ISSUES AND OBSERVATIONS IN STREET CENTERLINE CONVERSION

It would be naive to expect that the algorithmic approach outlined in the previous section would result in a perfect match between the two independent data sets. In fact, several issues were identified as the result of problems encountered during the matching process.

Encountering problems during the conversion process typically meant that one or more CDOT street centerline features could not be converted into the CWIMS Physical Inventory data base. To focus attention on these cases and recommend corrective action, listings of all the "no match" features were compiled into a *Source Data Errors Report*, and that report was provided to CDOT. Recommendations have also been made to the DPD for enhancements to the DPD Base Map representation of midline features to better handle the cases of divided streets at the same level and streets that exist at multiple levels (i.e., tiered streets). The work to resolve and correct these identified issues is an ongoing process within the CWIMS project.

Note that there is also a domino effect whenever matches are not successful for any features that have dependents or relationships to other features. For example, an underpass/vertical clearance is modelled as a child feature to a street centerline. Any time a street centerline cannot be matched, then it is not possible to convert any underpass/vertical clearance features associated with that centerline. Similarly, intersection point features are located by identifying the point where two or more street centerline features come together. Any time street centerline features cannot be converted, then it also will not be possible to convert one or more intersection features. Traffic signals are modelled as child features of an intersection. As a result, any intersection features that are not converted will typically prevent the conversion of one or more traffic signal features.

The sources of inconsistencies when assigning spatial properties to the street centerline features in the CDOT data files follow. To generalize the information, the two different data sets are referred to as the Reference Source (for the DPD data set containing spatial properties) and the Input Source (for the CDOT data set to which spatial properties must be assigned):

1. Street centerline features existed in the Input Source whose street name could not be located in the Reference Source.
2. Street centerline features existed in the Input Source whose street name was located in the Reference Source but for which there were no associated instances of midline features. This is the situation in which the Reference Source has provision for a named ROW (midline feature) in its common street name table. However, there were no instances of features for that ROW midline. As a result, it was not possible to obtain the spatial information needed to convert the corresponding street centerline.
3. A single ROW midline feature existed in the Reference Source and contained within its address range the entire address range of two or more street centerline features in the Input Source, as depicted in Figure 3*a*. Per the project design, it was intended to handle any given midline as a complete feature. There was no intent

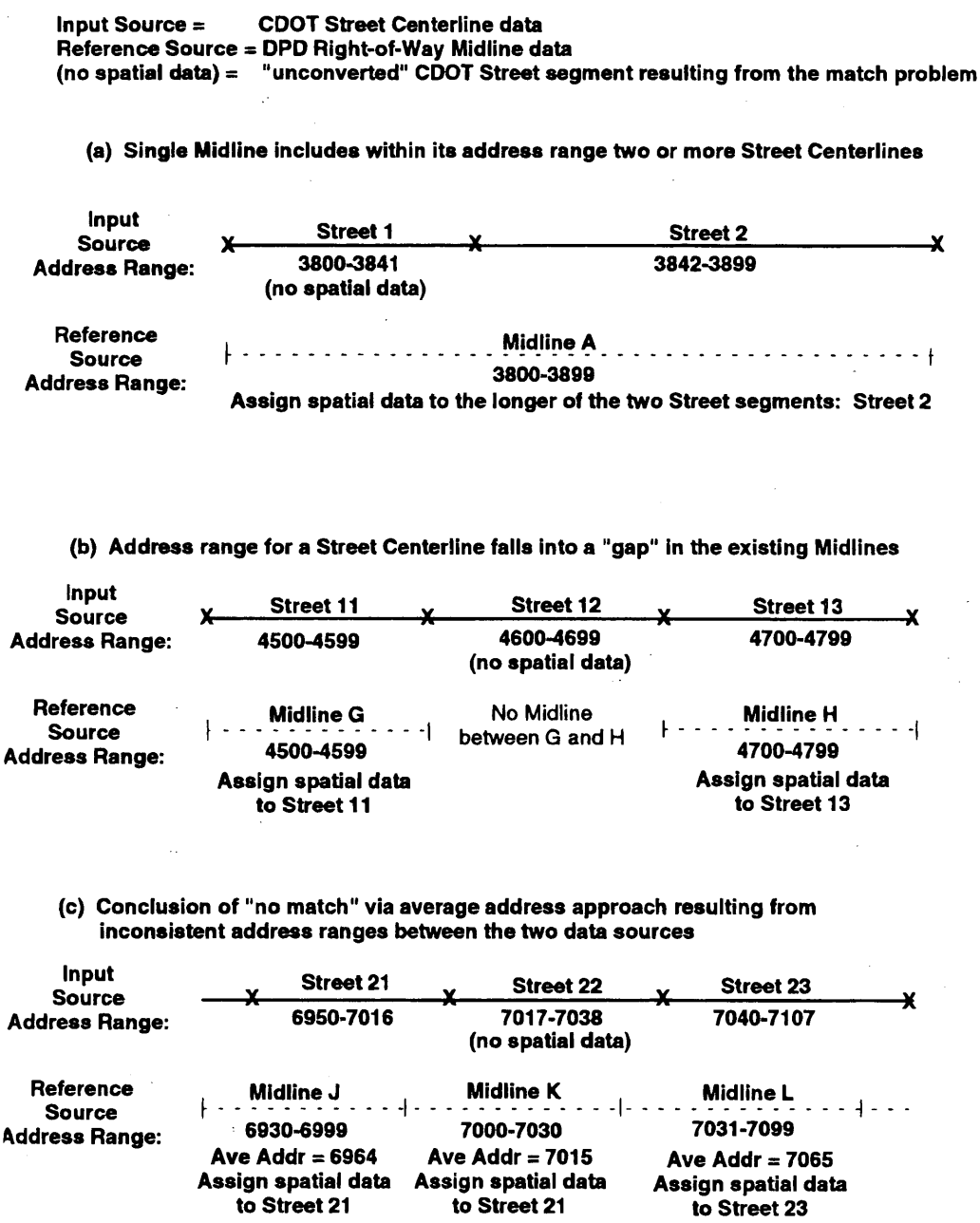


FIGURE 3 Overview of problems encountered with the "average midline address" approach.

to subdivide any midline features when matching to the street centerlines because it is the city's intention to make revisions in the raw data to resolve many of these differences. Accordingly, for this phase of the data conversion, the spatial properties for the entire span of a midline were assigned generally to the longer of the multiple street centerline features. This approach left as unconverted the other (shorter) street centerlines that were also completely contained within the midline address range.

4. Even though the street centerline from the Input Source found a match on a common street name with a ROW midline in the Reference Source, the address range of the street centerline was inconsistent with the address range of the midline in one of the following ways:

a. Address range too high: street centerline had an address range that was greater than the address range of any midlines with the same common street name.

b. Address range too low: street centerline had an address range that was less than the address range of any midlines with the name common street name.

c. Address range fell into a gap: although the street centerline features may have spanned a continuous set of address ranges, the corresponding ROW midline features had gaps such that there

was no midline feature (and consequently no spatial data) corresponding to one or more sets of address ranges along the street centerline. An example of this case is provided in Figure 3*b*.

5. Additional anomalies in address ranges between the street centerlines in the Input Source and ROW midlines in the Reference Source follow:

a. Street centerline segments with very short address ranges seldom resulted in a successful match to a midline. The short address ranges typically resulted when a diagonal street intersected another street and created a very short block. Usually, the conversion process was unable to find a midline whose *average address* was within the address range of the short street centerline segment. This is similar to the case shown in Figure 3*a*, in which *Street 1* represents the short street that was not converted.

b. Address ranges on "even" and "odd" sides of ROW midline features were sometimes very different. When considering the entire address range (across both the even and odd sides) of a midline, it was sometimes the case that the address range for a street centerline segment may be completely contained within the address range for two different adjoining midlines.

c. The approach of using the "average midline address" sometimes resulted in situations in which no match was concluded, even though the high or low address of the midline was somewhat contained in a street centerline feature. This often resulted from inconsistent address ranges in the two different data sets, as shown in Figure 3*c*.

CONCLUSIONS

The approach developed to this point in the project for assigning spatial properties to the CDOT street centerline features from the independent DPD midline features has resulted in the successful conversion of a large percentage of the CDOT data into the CWIMS Physical Inventory data base. Although there is still room for improvement, recall that the main objective for CWIMS is to support the capital planning process. The CDOT street centerline spatial data already converted represent major progress because they support a level of comparative analysis for capital projects planned by different departments in the city in a manner never before possible.

Work continues on the CWIMS project. With the benefit of the initial analysis for using the DPD midline and CDOT street centerline data sets together, the city is evaluating the types of changes that are appropriate in both sets of data to improve data consistency. The following conclusions can be cited as the result of experiences to date:

1. The expected "consistency" between the existing data set that requires assignment of spatial properties and the candidate data sets that can be used as source data should be assessed at the start of a conversion project. Choose the candidate source data set that is expected to have the smallest amount of inconsistency. The more years for which the two data sets were maintained independently, the more likely it is that inconsistencies will exist.

2. It is likely that this process of using one data set to assign spatial properties to another data set will be the first time that any type of "independent analytical work" has been performed on either of the data sets. The process may uncover problems in the raw underlying data. A focused effort is required to ensure that this process is used as a learning experience with the goal of making improvements in the overall quality of the data. Expect an iterative process, with revisions to the conversion process between each iteration.

3. Once inconsistencies are identified, it is important that the reason for the inconsistency be determined and that processing procedures are adopted to avoid the same problems in the future.

4. Jurisdictional considerations may affect the content of one or both of the data sources. This is more noticeable with municipal or local governments in which the agency often has no control over, or responsibility for, federal or state transportation infrastructure features. The federal and state-owned features tend to be the larger features on the transportation network, so their omission can create some conspicuous gaps in the geographic data base at the municipal or local level. It is appropriate to evaluate whether transportation features for which an organization has no jurisdiction or responsibility should at least be accounted for in the municipal or local geographic data, perhaps in the context of read-only data obtained on a periodic basis from a federal or state transportation agency.

5. It can be expected that some amount of data typically will not be successfully converted when using automated techniques. However, the features that are converted still represent significant amounts of usable, valuable data. Further, this successfully converted data provide a framework for future work to resolve discrepancies and to improve the conversion success rate.

6. When using address ranges as a basis for matching between sets of span features, identify if there are any local conventions that can simplify the process of determining which end of a street (or ROW) segment is the "low end." For example, the city of Chicago uses a well defined quadrant system for assigning street prefixes (North, South, East, and West). These quadrants are based on an origin at a designated intersection of two major streets in the downtown area. By convention, all address ranges are assigned such that the "low" address is always the end of a street (or ROW) segment closest to the origin street intersection.

ACKNOWLEDGMENTS

The project described in this paper is being performed by Camp Dresser & McKee, Inc., (CDM) for the Office of Budget and Management, city of Chicago. Carl Johnson and Daniel Hudson of CDM provided many valuable review comments during the writing of this paper, and their insights are appreciated.

REFERENCES

1. Nyerges, T. L., and K. J. Dueker. *Geographic Information Systems in Transportation.* The Center for Urban Studies. Portland State University, Portland, Ore., 1988.
2. *NCHRP Research Results Digest 180: Implementation of Geographic Information Systems in Transportation Agencies.* TRB, National Research Council, Washington, D.C., 1991.
3. *NCHRP Report 359: Adaptation of Geographic Information Systems for Transportation.* TRB, National Research Council, Washington, D.C, 1993.
4. Nyerges, T. L. Locational Referencing and Highway Segmentation in a Geographic Information System. *ITE Journal,* Vol. 60, No. 3, March 1990, pp. 27–31.
5. Guthrie, M. F. The Digital Control Section Atlas at the Michigan Department of Transportation: Implementing Segmentation and Refining Geometry. In *Proc., Geographic Information Systems for Transportation Symposium,* Orlando, Fla., 1991, pp. 163–176.

6. O' Neill, W. A., and B. Akundi. Automated Conversion of Milepoint Data to Intersection/Link Network Structure: An Application of GIS in Transportation. In *Transportation Research Record 1261,* TRB, Washington, D.C., 1990, pp. 27–34.
7. Azad, B., J. Brown, and J. N. Brown. Matching Census and Local Geographic Boundaries: A Practical, Repeatable Approach. *Geographic Information Systems,* Oct. 1991, pp. 40–45.
8. Petzold, R., S. Lewis, and D. Fletcher. *GIS as a Tool for ISTEA Management.* Fourth Annual TRB Workshop on Application of Geographic Information Systems to Transportation, Washington, D.C., Jan. 1994, p. 74.
9. IBM Corporation. *Graphics Program Generator Application Developer's Guide.* Publication Number SH20-6891-02. Mechanicsburg, Pa. Jan. 1990, pp. 25–31.

The views expressed in this paper are solely those of the author and not of the city of Chicago.

Publication of this paper sponsored by Committee on Transportation Data and Information Systems.

Development of a Regional Geographical Information System for ITS/IVHS Network

MUHAMMAD SHAHID IQBAL, CAROLYN S. KONHEIM, AND BRIAN T. KETCHAM

This paper describes the development of a geographical information system for transportation (GIS-T) intelligent transportation systems (ITS) network for the 23-county metropolitan region of New York, New Jersey, and Connecticut. The network was developed as part of a regionwide ITS implementation strategy study. The goal of the study was to design a regional ITS "architecture," a framework for an integrated and multimodal transportation network. The GIS network was developed as a tool to identify critical transit and roadway corridors within the region that can benefit from ITS technology and strategies, and to define and prioritize projects for near-, mid-, and long-term implementation and/or expansion of ITS in these corridors. When built on TransCAD, the GIS-T can perform numerous analyses of the data to identify corridors of recurring and nonrecurring congestion that warrant deployment of ITS. The accuracy of these projections is limited by the static nature and highly variable quality of existing data. As new data are generated, they can be entered onto spreadsheets keyed to the GIS for easy updating of the GIS. Additional layers can be added to the regional GIS, such as local streets, census data, and scheduled construction projects, to assist in evaluating and coordinating proposed ITS measures and transportation plans. Ultimately, the addition of ITS-generated real-time travel data would enable the GIS-T to serve traffic operations centers as the basis for transportation management and traveler information services.

This paper describes the development of a geographical information system for transportation (GIS-T) intelligent transportation system (ITS) network for the 23-county metropolitan region of New York, New Jersey, and Connecticut.

A GIS is described as "a decision support system involving the integration of spatially referenced data in a problem-solving environment" (*1*). The data maintained in a GIS are typically in three forms: point data, such as an intersection or accident location; link data, such as street or highway segments; and area data or polygons, such as political or traffic analysis zone boundaries.

TransCAD was selected for the network development, because it is one of only a few microcomputer-based GIS systems that have network analysis capabilities, such as network allocation algorithms (*2*). TransCAD can be used for planning, managing, and analyzing the characteristics and performance of transportation systems and facilities. Information on highways and streets, railroads, airports, terminals, and ports can be stored, displayed, and analyzed with appropriate procedures in a problem-solving environment by transportation professionals.

The GIS-T network was developed as part of a regionwide ITS Implementation Strategy Study for Transportation Operations Coordinating Committee (TRANSCOM) (*3*). TRANSCOM is a consortium of 15 transportation and public safety agencies in New York, New Jersey, and Connecticut. TRANSCOM's mission is to improve both interagency response to incidents and major events and the coordination of construction activities.

The goal of the study was to develop an ITS "architecture," a framework for an integrated and multimodal transportation network, for the 23-county New York, New Jersey, and Connecticut metropolitan region. The GIS network was developed as a tool to identify critical transit and roadway corridors within the region that can benefit from ITS technology and strategies, and to define and prioritize projects for near-, mid-, and long-term implementation and/or expansion of ITS in these corridors.

The study team was headed by James H. Kell Associates (JHK), and the GIS work was performed by JHK's subconsultant, Konheim & Ketcham (K&K). Other members of the study team were the RBA Group, Clough Harbour & Associates (CHA), Hughes Aircraft, and Robert S. Foote. The boundaries of the study area identified by TRANSCOM extended from Hartford, Connecticut, to central New Jersey.

The network contains the following counties; in Connecticut: Fairfield, New Haven, and Hartford; in New York: Bronx, Kings, Richmond, Manhattan, Queens, Nassau, Suffolk, Westchester, Orange, Rockland, Dutchess, and Putnum; and in New Jersey: Bergen, Essex, Hudson, Passaic, Union, Somerset, Morris, and Middlesex.

Developing a regional GIS has long-term benefits. To make all of the complex ITS/Intelligent Vehicle Highway Systems (IVHS) systems in the 23-county TRANSCOM region work together, there must exist a flow of useful information between the data-gathering hardware and the operators of the individual transportation systems. This necessitates communication strategies that tie together the various systems and traffic management and control strategies in response to changing traffic and transit conditions. In this exchange, a regional GIS is valuable as a data base into which large amounts of information can be integrated into a single network with a common reference for access by multiple agencies.

CRITERIA FOR SELECTION OF CRITICAL CORRIDORS

At the onset of the study, criteria for the selection and identification of critical corridors were developed that reflect the priorities of the TRANSCOM member agencies who were interviewed by the study team (*4*). These criteria are presented in Table 1. Criterion 1 is roadways with recurring congestion caused by high traffic volume for available capacity. It is based predominately on minimum volume-

Konheim & Ketcham, Inc., 175 Pacific Street, Brooklyn, N.Y. 11201.

to-capacity (V/C) ratio and AADT (average annual daily traffic) per lane thresholds. Links with a regional traffic generator (e.g., airport, stadium) or capacity bottleneck (e.g., bridge, tunnel) are also included in the list of recurring congestion corridors. Similarly, Criterion 2, roadways with nonrecurring congestion, is primarily based on frequency of accidents (accidents per lane per mile; where, 1 km = 0.62 mi) and includes AADT per lane threshold as well. Criteria 3, 4, 5, and 6 could not be practically applied because of the sporadic nature of the available data. The inputs and methods by which these conditions (Criteria 1 and 2) were computed and evaluated are explained below.

REGIONAL TRANSPORTATION NETWORK

The regional ITS GIS on TransCAD comprises two subnetworks: a highway network and a transit network. The highway network includes all limited-access highways, major arterials, and major crossings (bridges, tunnels), and contains a total of 452 two-way links with an average link length of 8.06 km (5 mi). The transit network includes commuter rail, subway, ferry, and publicly operated express bus links and contains 1,200 links. Figure 1 shows the regional highway network and Figure 2 shows the regional transit network. Because the study has a regional focus, local streets and local bus lines were not included in the ITS GIS, but, if desired, these could be retrieved from the underlying GIS files and integrated into a regional data base.

Data Sources and Development of the Roadway Network

The steps in the development of the roadway network were (a) identifying the links to be included in the network; (b) identifying data requirements (link attributes); (c) developing data worksheets; (d) researching the data and reporting them on the worksheets; (e) cleaning, refining, geocoding, and customizing the network; and (f) importing worksheet data into the TransCAD network.

Creating the Base Map

For the development of the roadway base map of the New York, New Jersey, and Connecticut metropolitan area, existing GIS networks were used. All GIS networks had the same scale as that of the GIS-T network that has been developed. The MAGIC network, a GIS of all state highways and arterials in New Jersey developed in TransCAD, was obtained from the New Jersey Department of Transportation. For New York, the MAGIS network, a GIS of all state highways and arterials, developed in ArcInfo format, was obtained from New York City Department of Transportation and converted to TransCAD format using TCBuild (a file conversion utility in TransCAD). For Connecticut and counties not covered in the existing networks, the principal arterial roads and highways in the U.S. Census 1990 Tiger files on CD-rom were extracted and converted into a TransCAD line data base.

Maps of each county and state were printed and distributed to those team members most familiar with each state for identification of candidate links. Links are roadway segments between two major intersecting roads, river crossings, tolls, or locations of significant changes in main-line capacity. The RBA Group was responsible for identifying candidate links in New Jersey, Clough Harbor & Associates for New York state outside New York City, and James H. Kell Associates for New York City and Connecticut. A total of 452 two-way links were identified by the team and refined through several iterations.

Travel Characteristics

A comprehensive list of roadway characteristics (link attributes) was developed. The list of 76 attributes for each link and the descriptions of each are presented in Table 2. Of these, 46 are physical and operational characteristics and 30 are features of existing or planned ITS/IVHS installations. The consultant team used Lotus worksheets to report all specified attributes for each link. The sources of information included data available from state departments of transportation and local agencies, large-scale drawings and maps, and any available regional planning reports. The final completed worksheets were merged into a single worksheet that was later exported into the TransCAD network.

Method of Computing V/C Ratios

In addition to the physical characteristics of the roadways cited in Table 2 and the reported AADT, it was necessary to obtain (or make assumptions for) the directional distribution [proportion of the design hourly volume (DHV) in the peak direction of travel], peak hour factor (proportion of AADT in the DHV), and growth factor to compute base year (or design year) design directional hourly volumes (DDHV).

The AADTs from past years (e.g., 1982, 1986, 1991) for each link were projected to the base year (1994) AADT using a linear growth factor of 1.5 percent per year. Computation of capacity for each link was based on the methodology provided in the 1985 Highway Capacity Manual (*5*). Where available, the actual number of lanes, lane width, shoulder width, vertical grade, and percentage of heavy vehicles were used to compute the capacity of each link. The DDHV was computed and divided by the computed capacity to obtain V/C for each link.

Developing the GIS

Geocoding and customizing of the network was performed according to the marked-up maps of candidate links and nodes prepared by each consultant. Because MAGIC had the largest number of links, it was used as the base map. Although all layers that are active can be displayed simultaneously, only one layer can be selected as the "current" or working layer. First (keeping the MAGIC segment layer current), New Jersey links and nodes were modified by splitting the links and joining nodes according to the marked-up maps. TransCAD's geocoding commands were employed for this purpose. Next (keeping MAGIS link and node layers active in the background and the MAGIC segment layer as current), New York links and nodes were geocoded and digitized. Every effort was made to geocode the actual shape of the link by zooming in to close range. Similarly, Connecticut links were geocoded by keeping TIGER link lines active in the background layer. The main purpose of this exer-

TABLE 1 Criteria for Selection of Critical Corridors

1.	Roadways with Recurring Congestion		
	▪ Volume/Capacity ≥0.90 (Level of Service E/F)		
	▪ Minimum traffic volume thresholds, as AADT/lane		
	Interstate/freeways	15,000	
	Divided arterials	10,000	
	Undivided arterials	7,500	
	▪ Regional traffic generator (airport, stadium)		
	▪ Capacity bottleneck (bridge, tunnel)		
2.	Roadways with Non-Recurring Congestion		
		AADT/lane	acc/lane/mi
	▪ With shoulder ≥8.0 ft.		
	Interstates/freeways	12,000	7
	Divided arterials	8,000	11
	Undivided arterials	5,000	14
	▪ With shoulder ≤6.0 ft.		
	Interstates/freeways	9,600	6
	Divided arterials	6,400	9
	Undivided arterials	4,000	11
3.	Potential for Bus Operational Improvement: Significant percentage (minimum of 3%) of buses on roads with recurring or non-recurring congestion		
4.	Major Truck Route: Significant percentage (minimum of 7%) of trucks on roads with recurring or non-recurring congestion		
5.	Presence of Accessible Alternate Facilities (e.g., road with available capacity, HOV lane, transit or intermodal option)		
6.	Major Scheduled Reconstruction Projects		

cise was to maintain a single and uniform link layer for the entire regional ITS/IVHS highway network.

Once the geocoding was completed, the network data customizing was performed using the file conversion utility, TCBuild, while defining new fields and attributes associated with each link. The type of each field was specified and the appropriate width and name assigned. Coded character type fields were given respective codes and descriptions. For example, in Table 2, Attribute 5, Type of Facility, has several codes such as P = Parkway, B = Bridge, or T = Tunnel. Customizing coded characters enables the TransCAD data editor to show the description of each code, even though in the data worksheet (to be imported in TransCAD) only codes are provided.

The Lotus data worksheet was imported into the TransCAD data editor window by assigning a common identification (ID) for each link in both the TransCAD data editor and the worksheet. This was done by selecting each link one by one in the TransCAD map display window, each time switching to the data editor window and entering the same ID for the selected link, which was already assigned in the worksheet. The accuracy of each ID factor was further verified by using the query option in the TransCAD map display window. Next, in the data editor window, the record/import/

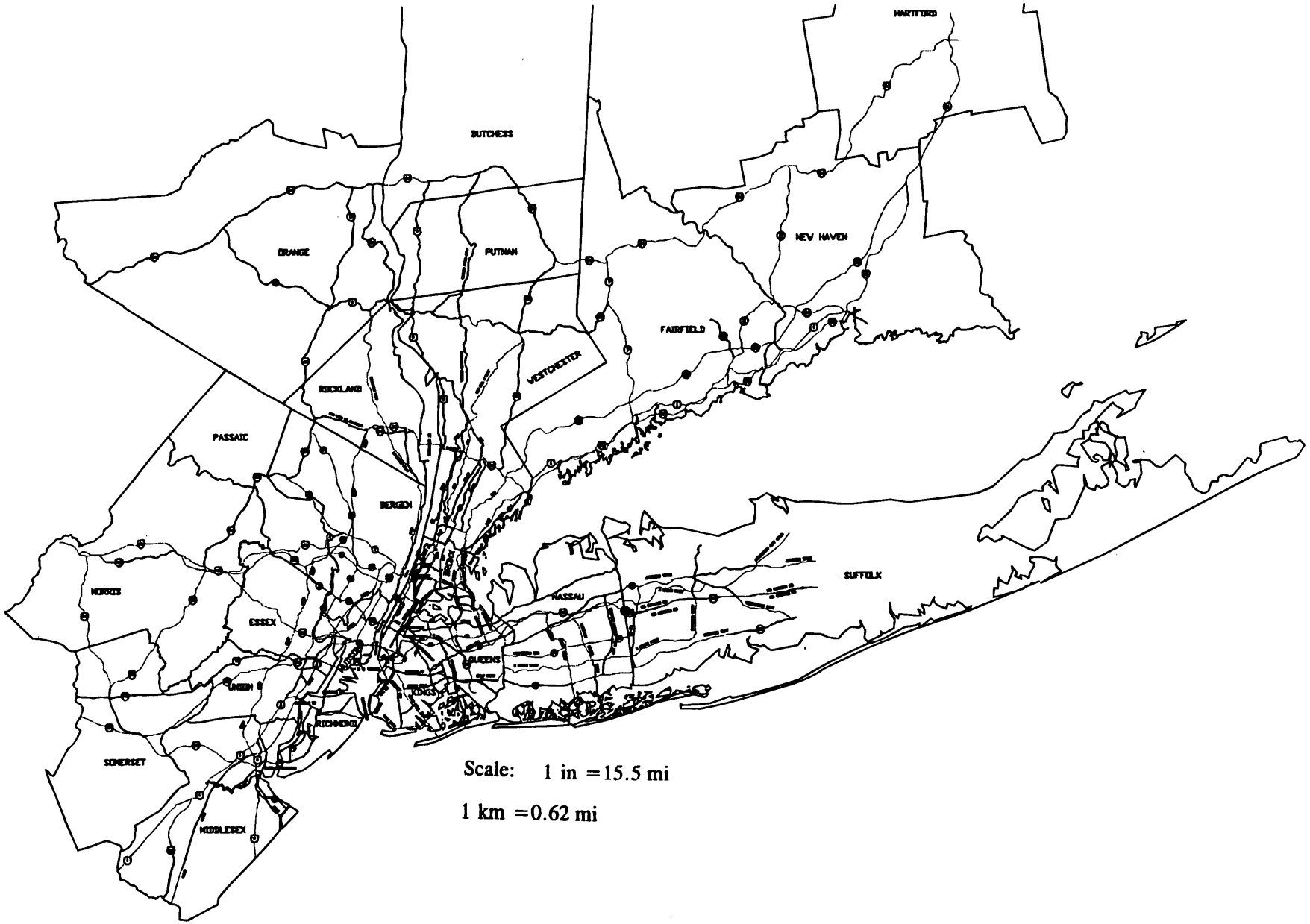


FIGURE 1 Regional highway network.

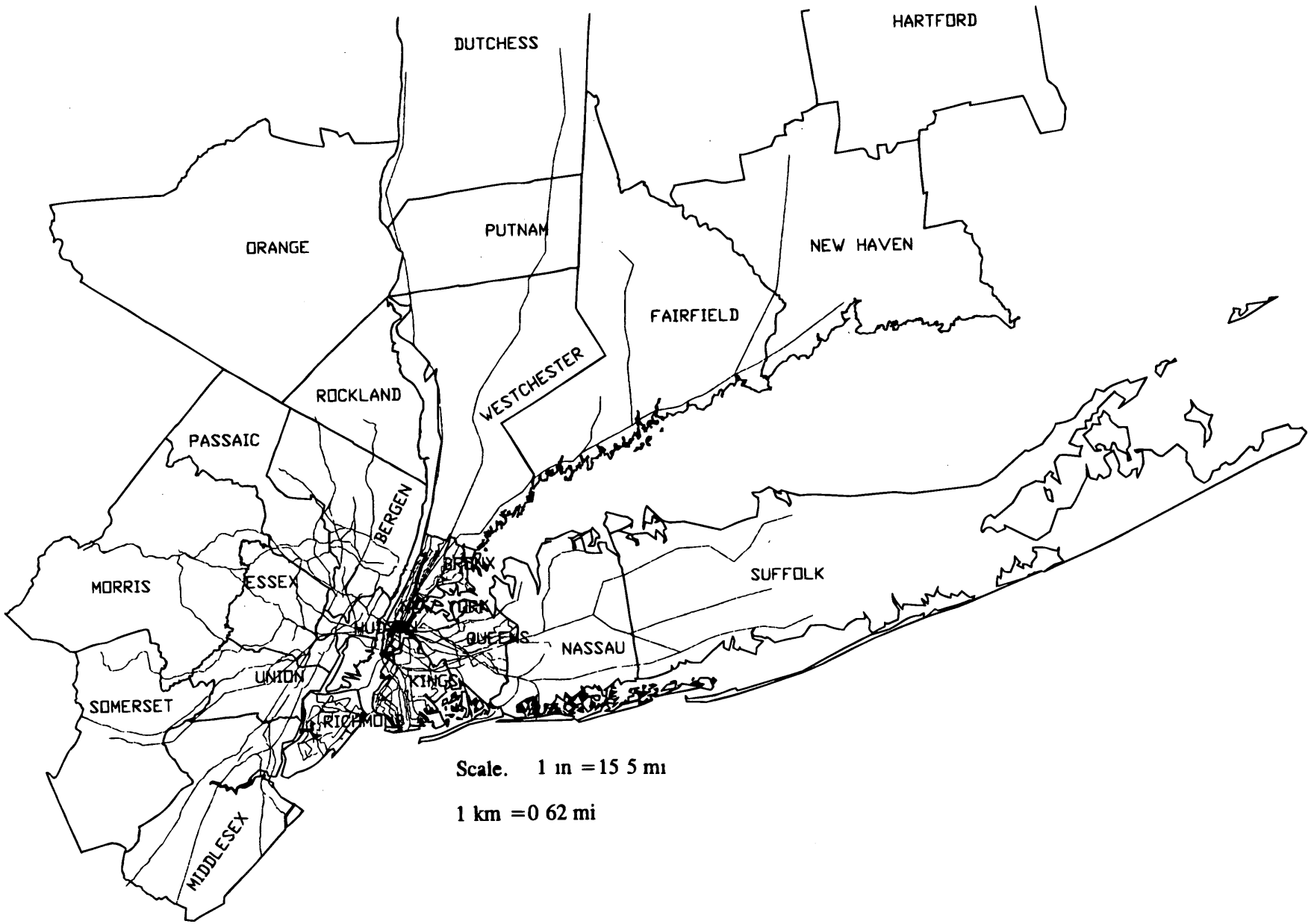


FIGURE 2 Regional transit network.

TABLE 2 List of Highway Network Attributes

COLUMN	DESCRIPTION
1	ID Number for the Segment (Assign an ID # for the Link marked on the Map)
2	Name of the Segment
3	Length of the Segment, in miles
4	Route Designation
5	Type of Facility (P=Parkway/Non−Commercial, C=Commercial, B=BRIDGE, T=TUNNEL, TF=TOLL FACILITY
6	Bridge is Movable (Y=Yes, N=No), if no bridge, leave blank
7	Segment Elevation (E= Elevated, A=At−Grade, D=Depressed EA=Partial Elevated/At−Grade, ED=Partial Elevated/Depressed, AD=Partial At−Grade/Depressed)
8	Starting Location of the Segment (Node A, Give Name)
9	Ending Location of the Segment (Node B, Give Name)
10	Average Daily Traffic/Average Annual Daily Traffic in Direction AB (vpd)
11	Average Daily Traffic/Average Annual Daily Traffic in Direction BA (vpd)
12	Total ADT/AADT, in vpd, (If Directional ADT/AADT not available then leave columns 10 and 11 blank)
13	Directional Distribution/Split in the AM Peak Period (Directions AB/BA)
14	Directional Distribution/Split in the PM Peak Period (Directions AB/BA)
15	Start of AM Peak Period (HH:MM)
16	End of AM Peak Period (HH:MM)
17	Start of PM Peak Period (HH:MM)
18	End of PM Peak Period (HH:MM)
19	Year of ADT/AADT Count
20	Number of Lanes in Direction AB
21	Number of Lanes in Direction BA
22	Width of Outside Shoulder in Direction AB (ft.)
23	Width of Inside Shoulder in Direction AB (ft.)
24	Width of Outside Shoulder in Direction BA (ft.)
25	Width of Inside Shoulder in Direction BA (ft.)
26	Average Width of Each Lane in Direction AB (ft.)
27	Average Width of Each Lane in Direction BA (ft.)
28	Percent of Single Unit Trucks in the ADT/AADT Vehicle Mix
29	Percent of Tractor Trailer Combination in the ADT/AADT Vehicle Mix
30	Percent of Buses in the ADT/AADT Vehicle Mix
31	Speed Limit on the Link, in mph
32	Peak Hour Operating Speed, in mph (in Direction AB, if available)
33	Peak Hour Operating Speed, in mph (in Direction BA, if available)
34	Beginning of Scheduled Construction on the Link (Year)
35	Ending of Scheduled Construction on the Link (Year)
36	AB Segment is Tolled (Y=Yes, N=No)
37	BA Segment is Tolled (Y=Yes, N=No)
38	Toll Rates Based on Vehicle Type, P=Rate, 2−Axle=Rate, 3−Axle=Rate, 4−Axle=Rate)
39	HOV Technique Implemented on the Link (N=No, Y=Yes, P=Proposed)
40	Type of Access Control on the Link (P=Partial, F=Full, N=None)
41	Traffic Signals (Blank=None, I=Isolated Signals, C=Centralized System, L=Closed Loop System)
41a	If column 41 indicates a Signal System, give Status: (E=Existing, D=In Design, P=In Planning)
42	Link Provides Access to a Major Trip Generator (A=Airport, S=Stadium, SC=Shopping Center, N=None)
43	Average Annual Number of Fatal Accidents that Occur on the Segment
44	Average Annual Number of Injury Accidents that Occur on the Segment
45	Average Annual Number of PDO Accidents that Occur on the Segment

(*continued on next page*)

worksheet command was used to transfer all data from worksheet to the TransCAD network.

Data Sources and Development of the Transit Network

The boundaries for the transit network are the same as for the highway network. The transit network was mainly extracted from the Metropolitan Transportation Authority's (MTA) GIS network. The network provided a fairly comprehensive data base and had several fields associated with each link including link name and length, start station, end station, peak hour volume, capacity, load factor, fare, type of mode, and so forth. The MTA transit network contains link information for several modes of transportation including local and express buses (for New York City), subways, commuter rail [Long Island Railroad (LIRR), Metro-North, and Port Authority Trans-Hudson (PATH)], and ferry (New York and New Jersey). The TransCAD condition window was used to select and segregate the

TABLE 2 (*continued*)

COLUMN	DESCRIPTION
	IVHS ATTRIBUTES
46	Variable Message Signs (VMS) on the Link (Blank=None, E=Existing, D=In Design, P=In Planning)
47	Number of Lines on the VMS
48	If column 46 indicates VMS, give the Name of Vendor
49	Highway Advisory Radio (Blank=None, E=Existing, D=In Design, P=In Planning)
50	If column 49 indicates HAR, give Frequency, in MHz
51	Roadway Flow Detection (Blank=None, E=Existing, D=In Design, P=In Planning)
52	If column 51 indicates Roadway Flow Detection, give Type (S=Single Loop, D=Double Loop, R=Radar Detector, V=VIDS)
53	All Lanes are Covered (Y=Yes, N=No)
54	Average Spacing for Single Loop Detectors, in miles (If column 51 is None, leave blank)
55	Average Spacing for Double Loop Detectors, in miles (If column 51 is None, leave blank)
56	Average Spacing for Radar Detectors, in miles (If column 51 is None, leave blank)
57	Average Spacing for VIDS, in miles (If column 51 is None, leave blank)
58	Special Detection (W=Weather Sensors, O=Overhead Detection, A=AVI)
58a	If column 58 indicates special detection, give status: (E=Existing, D=In Design, P=In Planning)
59	HOV Lane (E=Existing, D=In design, P=In Planning)
60	Metering (E=Existing, D=In design, P=In Planning)
61	Close Circuit Television (Blank=None, E=Existing, D=In Design, P=In Planning)
61a	If column 61 indicates CCTV, give Coverage (F=Full, P=Partial)
62	Weigh Stations (WM=Weigh-in-Motion, PO=Pull-off-Lane)
63	Service Patrol (Blank=None, E=Existing, P=Planned)
63a	If column 63 indicates Service Patrols, give Type: (I=In House, P=Private, O=On Call)
64	Type of Vehicle (P=Pick-up, V=Van, W=Wrecker)
65	CI=Call-in, CB=Call Boxes, CP=Cellular Phones, CB=CB Monitoring
66	Enforcement Agency for the Link (Name)
67	Operating/Maintaining Agency for the Link (Name)
68	Communication Medium: Trunk (F=Fibre Optics, T=Twisted-Pair, S=Spread-Spectrum, M=Microwave, C=Coax)
69	Communication Medium: Distribution (F=Fibre Optics, T=Twisted-Pair, S=Spread-Spectrum, M=Microwave, C=Coax)
70	Spare Trunk Capacity for Fibre Optics (Y=Yes, N=No)
71	Spare Trunk Capacity for Microwave (Y=Yes, N=No)
72	Spare Conduit Capacity (Y=Yes, N=No)

required transportation modes and links. Local bus links were excluded from the network. This resulted in a total of about 3,400 transit links. A decision was made to reduce the number of links by including only express stops on subways and limiting commuter rail to transfer stations, major stations, and the ends of lines.

Geocoding was performed in TransCAD by merging old nodes and showing new link and end points including start and end stations, link length, travel time, volume, and capacity. When two links are merged in TransCAD, their associated individual characteristics are lost. Updating of data relevant to the new link was done manually and required some judgment. For example, for subways and commuter rail (LIRR and Metro North) in the original MTA network, local stops were merged to represent express links only, and the travel time attribute was updated by adding travel times on individual links to be merged. Similarly, segment name, start station, and end station were updated and reentered. In the case of several one-way subway or rail links along the same alignment, only the link with the highest V/C ratio was kept, and that V/C ratio was assigned to the whole new link. The direction characteristics were changed to two-way, and the "route" attribute of the link to be maintained was updated by adding all subway and rail numbers using the link. For example, among the Nos. 1, 2, 3, and 9 trains between 34th Street and 14th Street, the subway line with the highest V/C ratio was selected to represent the link. This resulted in reporting numerous links on both subway and rail lines that show V/C ratios greater than one. In the case of New Jersey Transit (NJT) express buses, different routes of variable distances between stops were aggregated, producing V/C ratios that may be overestimated for much of their length. With rail lines, the V/C ratio of the highest links were assigned to an entire line. These localized aberrations were considered acceptable when viewed in a regional context.

NJT rail and Amtrak links within study boundaries were extracted from a GIS network that was obtained from New York Metropolitan Transportation Council. Rail lines were supplemented by express bus route data provided by NJT on peak hour frequency of operation, capacity, ridership, and route numbers segregated by the links in the New Jersey roadway network. These express bus links were entered into the transit network by keeping the ITS/IVHS transit link layer as current, keeping the ITS/IVHS highway network active in the background, and digitizing the links. The pertinent data were then entered manually in the appropriate attribute field in TransCAD.

Many attempts were made to obtain data on privately operated express buses in all three states, but information on routes and passenger volumes were not available from any agency, and the private operators were not forthcoming in response to mailed solicitations for their participation.

APPLICATION OF SELECTION CRITERIA TO THE DATA BASE

The selection criteria for recurring and nonrecurring congestion (Table 1) were applied to identify critical corridors. For example, each type of roadway with the specified minimum AADTs and a capacity bottleneck that equals or exceeds a V/C ratio of 0.90 meets the threshold for a roadway with recurring congestion. Examples include a bridge or tunnel or access roads to a major traffic generator, such as an airport or stadium.

The maps of recurring congestion and nonrecurring congestion links were exported to AutoCAD for route labeling and printing. Maps were not printed directly from TransCAD because its current version lacks sophisticated labeling capability. Figure 3 shows the regionwide links of recurring congestion. Lists of links with recurring congestion were prepared. Based on the reported travel attributes of the 452 roadway links in the TRANSCOM region, there are 278 links, totaling 2145 km (1,330 linear mi), that meet the criterion of recurring congestion. In New York State, these total 576 mi; in New Jersey, 427 mi; and in Connecticut, 327 mi. Other than New York City, nearly all are major limited-access highways. Many roads in New Jersey are critical for almost their entire length, but on other Interstates and parkways, there are segments that are not critical. This is also true in New York State in which several Interstates and parkways meet the critical criterion for their entire lengths within the study area; whereas for other Interstates and major state roads, only certain segments are critical. Major local arterials in New York City, which appear not critical for their entire lengths, may in fact become critical if some of the reported data such as number of effective lanes were changed based on field checks. In Connecticut, all of the Interstates are critical.

In addition, 191 roadway links in the TRANSCOM region, totaling 1406 km (879 mi), meet the criterion of nonrecurring congestion, determined principally by the frequency of accidents. These are shown in Figure 4. In New York State, these occur predominantly in New York City and Long Island totaling 595.2 km (372 mi) with only 33.6 km (21 mi) in upstate New York. In New Jersey, there are 526.4 km (329 mi) of nonrecurring congestion, largely in eastern portion of northern New Jersey. Connecticut has 251.2 km (157 mi) of nonrecurring congestion on its two main coastal routes. Table 3 summarizes the above results.

The availability of alternate facilities for aggregates of links of roads of recurring congestion, either transit or state highways that do not also experience recurring congestion, can be estimated by observation of the maps. It can be seen that, for most roads, there are no uncongested state highway alternatives, and that, in many cases, rail lines are the only alternative. Additional travel capacity could also be created by giving express buses priority treatment on congested roads. An examination of the capacity of local arterial roads to absorb diverted traffic was beyond the scope of this study. The locations of intermodal connections are in the data base but were not used to prioritize corridors for ITS treatment.

An inventory of existing and planned ITS/IVHS elements by roadway link was reported. These include the status of:

- Variable message signs (VMS),
- Highway advisory radio by coverage and frequency,
- Roadway detection by spacing and type of detection,
- Electronic toll collection capability,
- Closed-circuit television coverage,
- High-occupancy vehicle (HOV) lane use,
- Service patrols, and
- Fiber optic cable.

Figure 5 is a sample map that shows status of VMS for the links in Connecticut.

Determination of Critical Corridors for ITS Application

Applying the criteria to the travel attributes, the resulting critical corridors for ITS/IVHS implementation based on recurring congestion comprise 278 links for a total of 2145.16 km (1,330 mi) in the three states. The additional roadways that have nonrecurring congestion are fewer and generally shorter, 191 links for a total of 1408 km (880 mi) in the three states, determined principally by the frequency of accidents. In very few cases are there alternative state highways with available capacity to accommodate diverted vehicles. Transit services are also at capacity on many links.

Nearly all of the critical corridors are designated for ITS treatment; the gaps are identified, and an approximate schedule of implementation is reported. In most cases, studies and demonstration projects are under way or just beginning. In addition, there are ITS plans for roads that are not designated critical.

The two transit agencies in the region, NJT and MTA, have long-term ITS plans for their entire systems. All the transit services provide comprehensive travel information, but none are real-time and they are accessible mostly by telephone. Vehicle tracking systems are being installed, but it will take a number of years before this information is used for centralized control or passenger information. The inherently integrated and centralized nature of transit facilities, and the extent of planning for specific ITS measures that have already occurred, suggests that transit can move forward rapidly with ITS if it is adequately funded. The important determinants of effectiveness of ITS measures, whether roadways or transit, are the schedule of implementation and the level of sophistication. For example, advanced transit systems would offer multiple means of user access to real-time travel information on all routes based on automatic vehicle location. This would be equivalent to enabling drivers to obtain pretrip or en route information on roadway conditions for optional routes.

Prioritizing implementation of ITS will involve policy considerations that are the domain of TRANSCOM and its member agencies. The TRANSCOM agency interviews indicated that improved traveler information is the greatest need perceived by the operating agencies. The value of real-time travel information was substantiated by opinion research, conducted for this study by the Peter Harris Research Group, which found that almost 9 in 10 travelers in the New York, New Jersey, Connecticut area support enhanced travel information systems, and 78 percent from all income groups are willing to pay something for such services (*4*). In view of this user need, facilities that will benefit most from travel information services should be the highest candidates for near-term implementation of ITS in the TRANSCOM region.

FUTURE APPLICATIONS OF THE REGIONAL ITS DATA BASE

The network model provides an extremely valuable tool for continued planning of the deployment of ITS. When more multimodal data are collected, it would be possible, for example, to produce a

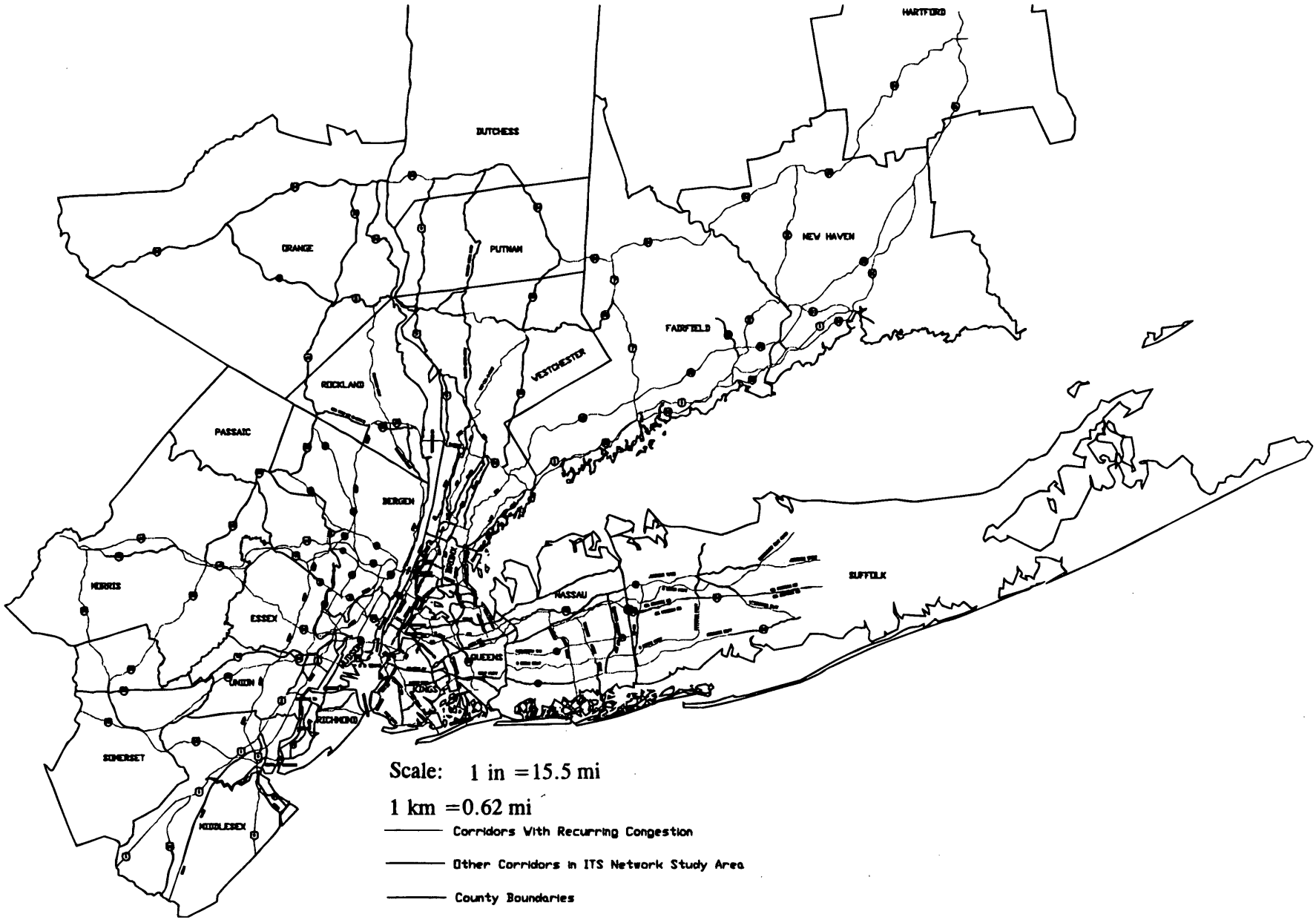


FIGURE 3 Corridors with recurring congestion in TRANSCOM region.

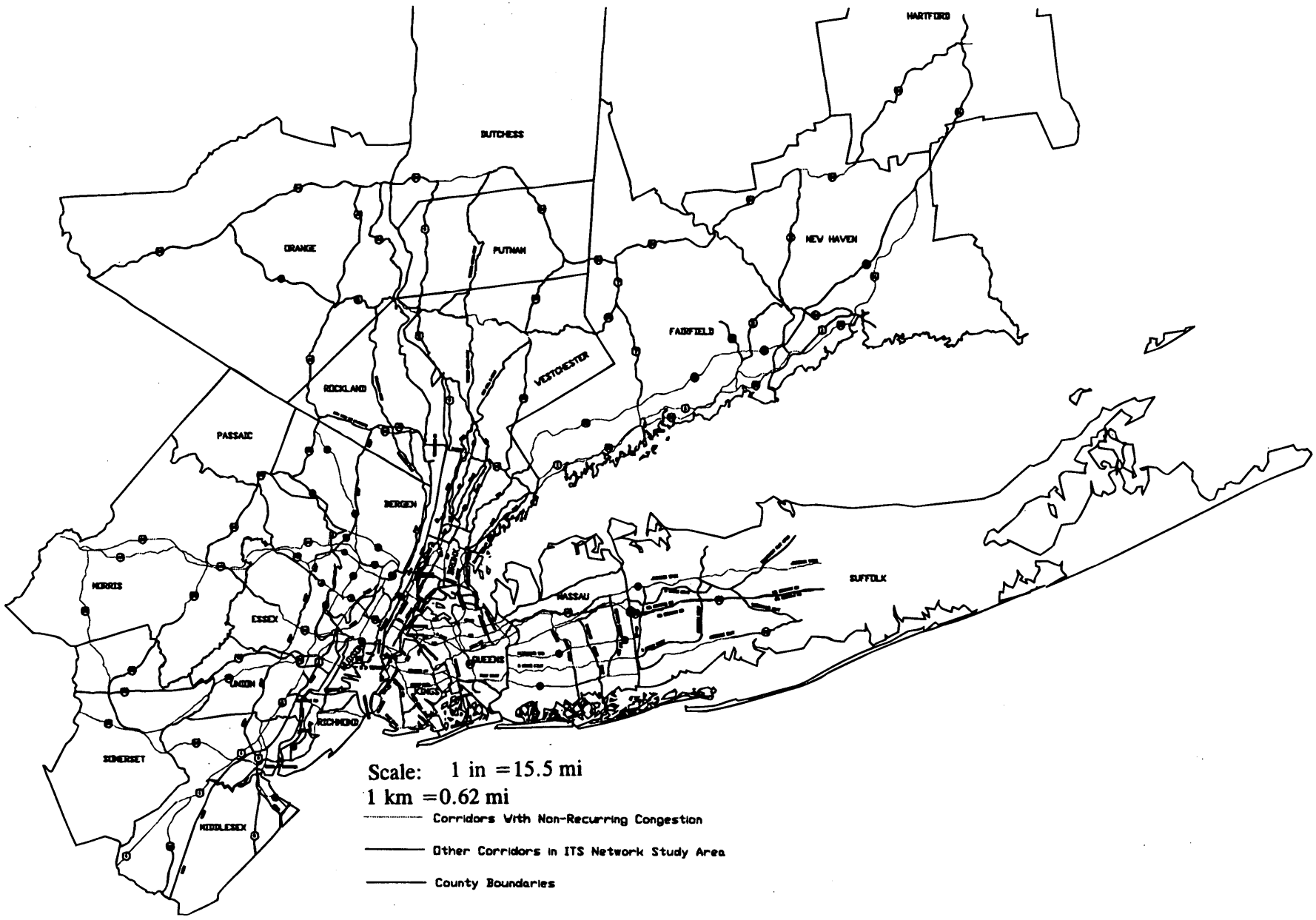


FIGURE 4 Corridors with nonrecurring congestion in TRANSCOM region.

TABLE 3 Overview of Findings

State (County)	All Roadways		Recurring Congestion		Non-Recurring Congestion	
	# of Links	Length (miles)	# of Links	Length (miles)	# of Links	Length (miles)
Connecticut (total) ➢	**28**	**334.03**	**27**	**327.29**	**15**	**156.84**
New York (total) ➢	**274**	**1,231.51**	**147**	**576.43**	**100**	**393.68**
Nassau	*39*	*182.43*	*20*	*87.87*	*19*	*81.29*
Suffolk	*23*	*190.84*	*17*	*145.05*	*5*	*38.59*
Long Island (sub-total) ➢	*62*	*373.27*	*37*	*232.92*	*24*	*119.88*
Kings	*20*	*76.44*	*10*	*41.56*	*10*	*52.41*
Bronx	*23*	*56.53*	*12*	*32.07*	*9*	*26.4*
Queens	*38*	*131.89*	*25*	*69.35*	*28*	*110.58*
Manhattan	*30*	*70.78*	*25*	*60.6*	*17*	*51.68*
Staten Island	*10*	*33.01*	*4*	*7.19*	*3*	*11.3*
New York City (sub-total) ➢	*121*	*368.65*	*76*	*210.77*	*67*	*252.37*
Upstate New York	*91*	*489.59*	*34*	*132.74*	*9*	*21.43*
New Jersey (total) ➢	**150**	**587.16**	**104**	**427.01**	**76**	**328.51**
TOTAL (Region) ➢	**452**	**2,152.7**	**278**	**1,330.73**	**191**	**879.03**

1 km = 0.62 mi

variety of descriptive, thematic maps that show the roadway links that have V/C ratios greater than 0.90, or those bus stops and subway stations that have the greatest number of boardings and alightings during peak periods. Each corridor can be examined in detail for intermodal opportunities. Most important, this process can be used to prioritize ITS investment for maximum benefits during a period of fiscal constraints.

A regional layout of construction schedules could be achieved by entering each state's transportation improvement program into the GIS. A visual display of projects would highlight areas in which schedules need to be optimized to maintain mobility on a regional basis.

The regional ITS GIS provides the primary stage for an ongoing integrated regional transportation information system. At present, it is a static network with a historical data base. However, the spatially referenced indexing of the network provides a common foundation for sharing aggregate information within and across multiple transportation agencies. The GIS provides a ready means to maintain and report frequently updated data.

More links and their related data could be added to the network to incorporate minor arterials, collectors, and local streets in modeling. Any subarea network could be selected for modeling trip generation, trip distribution, modal split, and traffic assignment by developing trip tables and using TransCAD's network analysis procedures, such as the gravity model, all-or-nothing assignment, shortest path analysis, and so forth. Ultimately, the highway and transit networks with added links could be used in TRANSCOM's planned traffic operations centers where real-time traffic data collection sources could provide direct input to the network. These data could then be displayed graphically on screens indicating congested corridors, location of incidents, and travel delay information. Users of the system could access such information via computers in vehicles, homes, or offices by entering their origin and destination and obtaining display and printout of the shortest path in terms of time.

RESULTS AND CONCLUSIONS

A GIS-T, a geographic information system of transportation networks and their related travel characteristics, is a valuable tool for transportation planning, especially for prioritizing investments in ITS intended to relieve congestion. The GIS can serve as a common data base for multiple agencies to share data in comparable formats. When built on TransCAD, the GIS-T can perform numerous analyses of the data to identify the corridors with critical conditions. Applying criteria for recurring and nonrecurring congestion to a data base of the New York, New Jersey, and Connecticut metropolitan area succeeded in identifying roadways and transit routes that warrant ITS deployment. The accuracy of these projections is limited by the static nature and highly variable quality of existing data. As new data are generated, they can be entered onto spreadsheets keyed to the GIS for easy updating of the GIS. Additional layers can be added to the regional GIS, such as local streets, census data, and scheduled construction projects, to assist in evaluating and coordinating proposed ITS measures and transportation plans. Ultimately, the addition of ITS-generated real-time travel data would enable the GIS-T to serve traffic operations centers as the basis for transportation management and traveler information services.

ACKNOWLEDGMENTS

This paper is based on a study sponsored by the New York State Thruway Authority and TRANSCOM, with funds from the FHWA. The authors are grateful to the project managers, Donald Geoffroy of the NYSTA and Louis Neudorff of JHK, and to Philip Riggio and Krishnaveni Venkataswami, also of JHK, for their guidance and input. The authors appreciate the technical assistance of Rizwan Ahmed and Raj Maan of K&K. Special gratitude is owed to all team members, particularly Armando Lepore of the RBA Group and

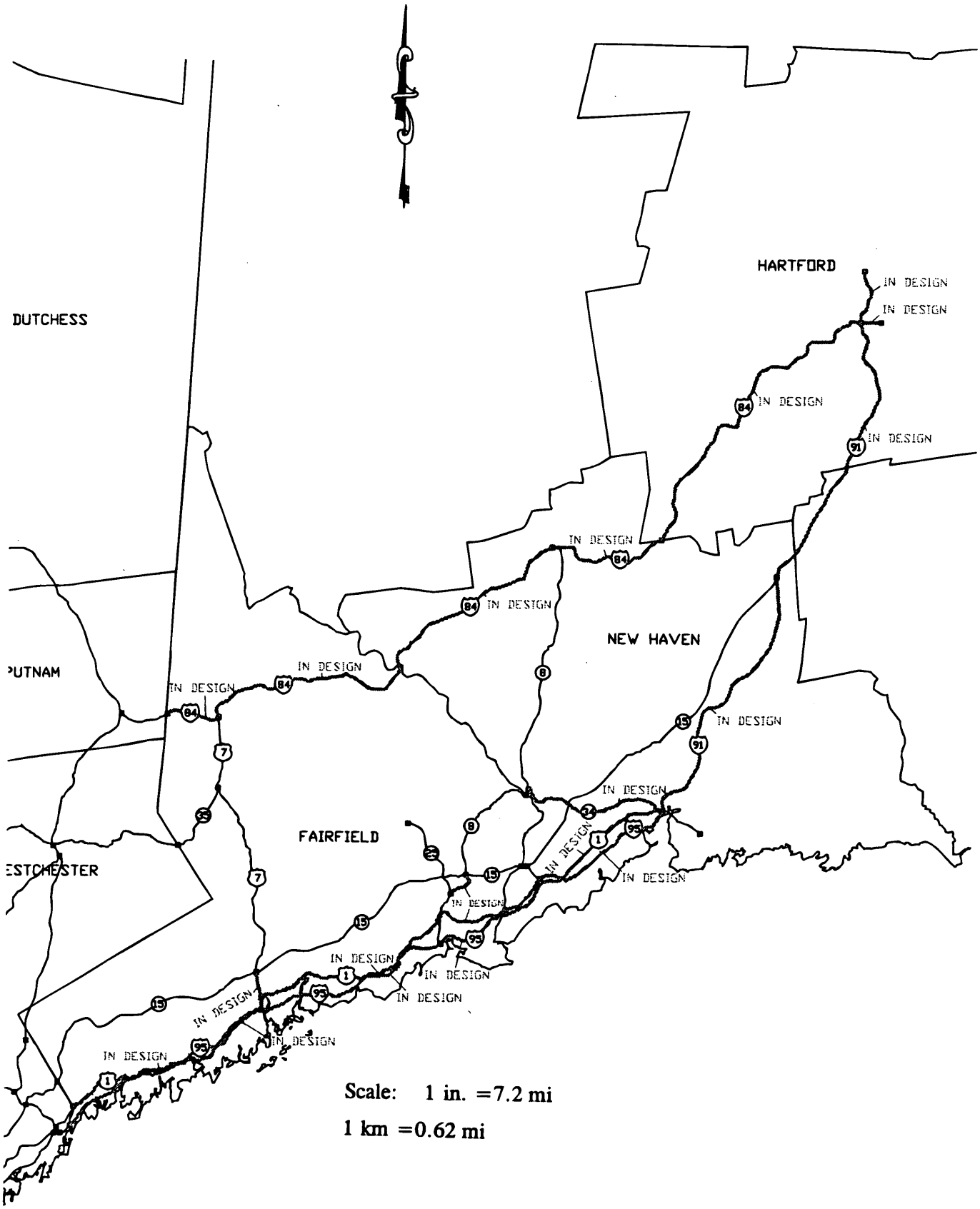


FIGURE 5 ITS/IVHS Elements in Connecticut, status of VMS.

Gary Robinson of CHA, who identified the candidate links and obtained the data for the worksheets.

REFERENCES

1. *Improving Transportation Decision Support, Making Informed Choices Using GIS Technology.* Environmental Systems Research Institute, Calif., 1994.
2. TransCAD. Transportation Workstation Software, Reference Manual, Version 2.1, Caliper Corporation, Mass. 1992.
3. *Alternative Configurations for the TRANSCOM Regional ITS/IVHS Architecture.* A Technical Memorandum. James H. Kell & Associates, New York, 1994.
4. Harris, P. *The Public's Interest In and Willingness to Pay for Enhanced Traveler Information as Provided by IVHS.* A Report for TRANSCOM, 1994.
5. *Special Report 209. Highway Capacity Manual:* TRB, National Research Council, Washington, D.C., 1985.

Publication of this paper sponsored by Committee on Transportation Data and Information Systems.

Geographic Information Systems/Global Positioning Systems Design for Network Travel Time Study

BO GUO AND ALLEN D. POLING

In October 1993, Maricopa Association of Governments (MAG) conducted travel speed and delay studies in the metropolitan Phoenix area. This paper concentrates on a Geographic Information Systems (GIS)/data base systems approach to process the field data collected with Global Positioning Systems (GPS) units and portable computers after a brief discussion of the travel route selection and preparation and a brief comparison of the several different travel time survey techniques using the test vehicle method. The three system modules are described: the Field Data Conversion and Validation module uses the data base approach to combine the GPS data (position and speed) and event data into a data base format and uses ARC/INFO to check GPS run validity; the Link Topology Definition module extracts topological information of the selected travel routes by data base programming and ARC/INFO's Dynamic Segmentation; the Data Analysis and Reporting module structures the continuous GPS data on a link-by-link basis and calculates travel time, delay, and speed information for each link. This module also feeds the results to a SAS program for statistical analysis and to ARC/INFO for generating graphics reports. The system design is based on the interface between ARC/INFO 6.11 on the SUN Sparc Station and FoxPro 2.5 for DOS on PC.

Geographic Information Systems (GIS) are useful analysis and presentation tools in the field of transportation planning, engineering, and management. When combined with the capabilities of Global Positioning Systems (GPS) as a data collection tool and the power of a relational data base system, GIS yields new solutions to areawide travel speed and delay studies.

Travel speed is a direct measure of road network performance. Low speeds are an indication of congestion, delay, increased fuel use, and higher pollution emissions. Decreasing auto travel speeds may signal a need for increased road or transit capacity or for greater attention to travel demand management. Travel speed surveys are an important component of Congestion Management Systems. Quality speed data result in better travel demand and air quality modelling.

TRAVEL SPEED FIELD DATA COLLECTION METHOD

Areawide travel speed studies consist of collecting, processing, and presenting automobile travel speed and delay information for a street and highway network. There are many methods available for conducting travel time surveys. The test vehicle method in its various forms is used most often when a comprehensive evaluation of travel time in a transportation network is needed.

Lee Engineering, Inc., 3240 E. Camelback Rd., Suite 180, Phoenix, Ariz. 85016.

Manual and Distance Measuring Instruments Test Vehicle Method

In its original form, the test vehicle method requires two data collectors, one for driving the vehicle, the other for recording time and events using a stop watch and tally sheets (*1*). This results in labor intensive data collection and data reduction processes. The invention of Distance Measuring Instruments (DMI) offered an automated way to collect and process travel time information using the test vehicle method. In a typical configuration, the DMI mounted on the test vehicle automatically logs travel distance and time, and the driver codes events that occur during the survey process. However, there are two major disadvantages of using DMI. First, it gives no spatial information about the route the vehicle travels on. This makes it difficult to have travel data mapped into a GIS coverage or to check whether the driver followed the designated travel route for quality management purposes. Second, if travel time is needed on a link-by-link basis, the driver may need to keep track of the time the test vehicle passes each major intersection along the route, in addition to recording other events. The more human interface required, the greater the likelihood of errors.

GPS Technology

GPS is a positioning and navigational system. This space-age technology has found a variety of applications in natural resource management, urban development and analysis, agriculture, and social sciences. A GPS consists of three segments: space, control, and the user. The space segment consists of 24 Navstar satellites that broadcast a signal providing information on their orbital path. The control segment is the U.S. Department of Defense, which monitors the movement of the satellites and transmits data to each satellite to ensure that they are broadcasting accurate information on their orbital path to earth. The user segment is a receiver, which receives the signal and uses information contained in the signal to calculate latitude, longitude, and altitude (*2*).

Data collected using a test vehicle equipped with a GPS unit give both temporal and spatial information about the travelling vehicle, thus naturally lending itself to GIS-based postprocessing. The data collection method makes validity checking of the travel runs possible by overlaying GPS runs on the GIS base map, in addition to allowing the driver to concentrate on driving and recording of other incidents. GPS was used in test vehicles to collect travel speed data for the Maricopa Association of Government (MAG) travel speed study.

DATA COLLECTION

Travel Route Selection and Route Data Preparation

The selection of routes for the travel speed survey was based on a procedure agreed on by the Maricopa Association of Governments Transportation Planning Office (MAGTPO) and Lee Engineering. Route selection began with 14 significant routes identified by MAGTPO for inclusion in the 1993 travel speed study. The remaining routes were randomly selected in such a manner to balance the number of roadway miles for each roadway functional classification of each area type. This was accomplished by developing a data entry program that allowed a route to be entered into a data base table in a link-by-link fashion with the MAG functional class. Links on a route were identified using the reference street scheme, in the form of on-street, from-street, and to-street. A data base program was developed that output a matrix of roadway miles for each functional classification of each area type. Using this matrix, routes were selected in an iterative manner to obtain a balance of roadway miles between each cross-classification in the matrix. The routes selected include over 1288 km (800 mi) of arterial roadways and over 161 km (100 mi) of freeway and high occupancy vehicle (HOV) lanes. A total of 61 routes were selected. Each link on the selected routes was then assigned a unique link identification (ID). Other attributes such as the MAG area type, MAG functional class, speed limit, number of lanes, and average daily traffic (ADT), were collected and keyed into the table. There are a total of 1,790 links for the 61 routes selected, each representing a roadway segment for a given travel direction.

GPS Data Collection Design

Travel speed data on the selected routes were collected for representative weekday conditions (Monday through Friday) for the three time periods: morning peak (7:00 am to 9:00 am), off peak (10:00 am to 2:00 pm), and evening peak (4:00 pm to 6:00 pm). Travel speed data were collected for all 61 routes during the evening peak. In addition, 18 routes had morning peak data collected, and 19 routes had off-peak data collected.

The data collection vehicles or test vehicles were equipped with GPS units that were set to record the date, time, vehicle position, and velocity of the vehicle on a 2-sec sampling interval. Each vehicle was also equipped with a portable computer that was used for collecting other data such as identifying the time and type of delays encountered. The data collectors (drivers) were instructed to start each route at a particular location and time. The data collectors were also given a starting direction. Travel speed data for a route were collected for a minimum of three runs for each period. To avoid bias in the data, for any given route during a period, the starting points and the direction of the travel were varied. The data collectors were instructed to record any event that caused a stop delay other than general intersection delay on a portable computer. A stop delay is defined as the time during which the vehicle is travelling at a speed less then or equal to 8 km/hr (5 mph). Such events included traffic accidents, school zones, begin/end construction zones, pedestrians, emergency vehicles, and trains. Careful consideration was given to categorizing the types of delay so that remembering the classifications and corresponding keys would not be difficult.

Accuracy and Validity of Field GPS Data

The positional accuracy that can be achieved with GPS ranges from 100 m to millimeters, depending on the type of receiver used and if the data collected are differentially corrected. To obtain the highest level of accuracy for any receiver requires differential correction, a process of placing a receiver on a known location, called a base station, and using the collected satellite data to adjust GPS positions computed by other receivers at unknown locations during the same time period (2). The differential correction can be done in real time, if the receiver is equipped for real-time correction and a correction signal is being received, or in a postprocess fashion. Postdifferential correction was used for the travel speed study to improve the accuracy from 100 m to 2 to 5 m. Base station data for postdifferential processing was provided by Maricopa County.

The majority of the travel time data collection effort was completed in October and November of 1993. Additional runs were made for those routes requiring more data runs to meet the designated travel time tolerance. By April of 1994, a total of 416 runs had been collected, of which 392 had at least partial useful GPS data. The runs without useful GPS data were mainly attributed to drivers' failure to set up the GPS unit correctly on wire due to unavailability of the base station data for correction. Total loss of data could be reduced or eliminated through better training of the data-collecting personnel and the use of real-time differential GPS. Partial GPS data loss occurred because portions of the data were missing or not correctable against the base station because of unfavorable satellite configurations. Partial GPS data loss also occurred because of obstruction of GPS signals from tall buildings, bridges, tunnels, and so on. A link without complete GPS data in a run was combined with its adjacent links to form a longer link for travel speed data analysis whenever possible.

POSTPROCESSING SYSTEM DESIGN

The field data collection produced more than 40 megabytes of the raw GPS data and event data. Successful reduction and analysis of data in such volume can only be achieved through a carefully designed data analysis procedure that takes full advantage of an integrated GIS, data base, and statistical analysis packages of different platforms to maximize efficiency and minimize errors. Most of the data conversion and manipulation were performed through programming in FoxPro on the PC platform with a PC SAS interface for statistical analysis and an ARC/INFO interface for spatial analysis and graphics reporting. The integrated data analysis system is shown in Figure 1. There are three major components in the system, the functions of which are discussed in the following sections.

Field Data Conversion and Validation

This module prepares the field-collected GPS data and portable computer data for analysis. First, the field-collected GPS and portable computer data are downloaded to the PC. The GPS data are then corrected using base station data to increase accuracy. After correction, the GPS data are exported to an ASCII file using NAD 27 and State Plane Coordinate Systems, a system that the GIS base map uses. A set of data base programs (ASC2DBF) merge the GPS ASCII data with field event data to create a table in the dBase format. The events are put in a temporal order, and the coordinates of the events are determined by interpolating in the newly created dBase file.

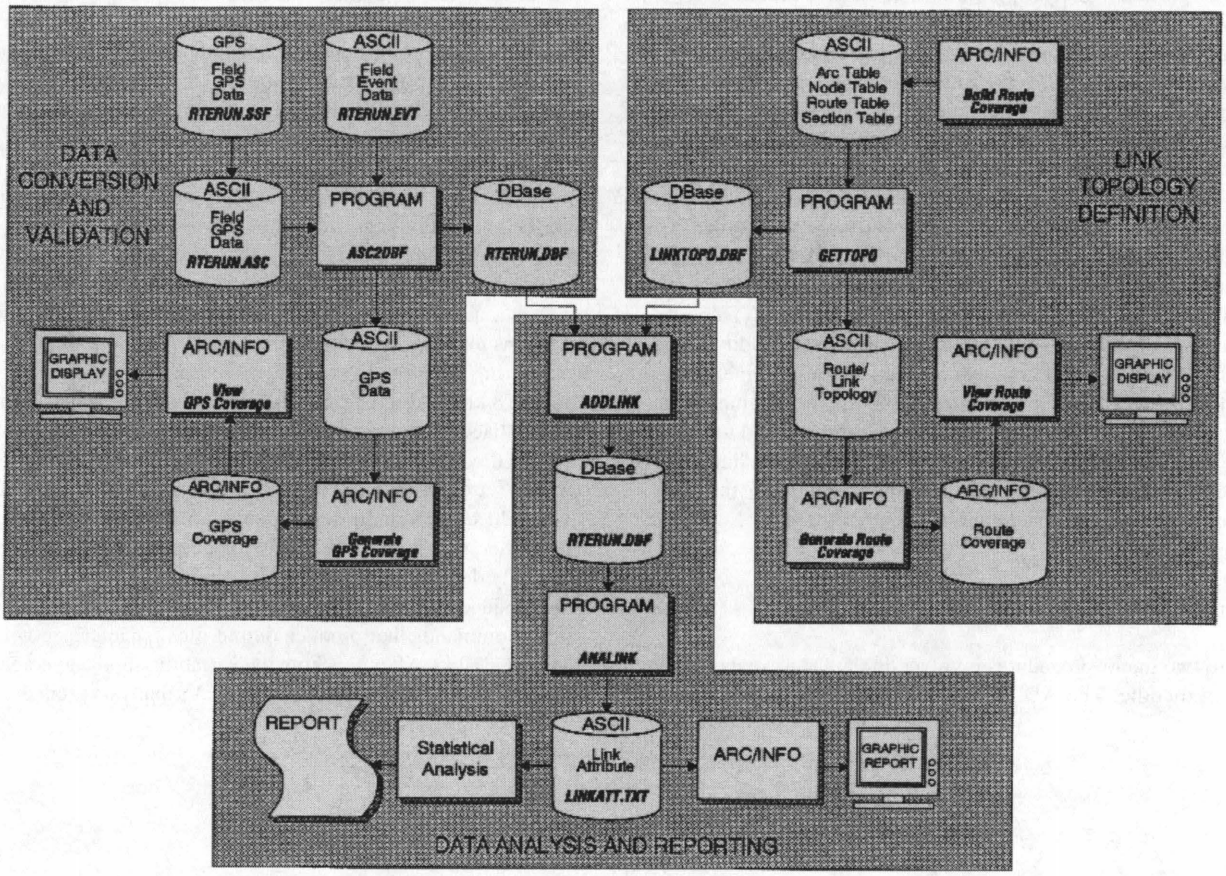


FIGURE 1 Data processing flow chart.

Validation of the GPS data is an important part of the module. The ASC2DBF programs also produces ASCII GPS and event data files with coordinate information, which serve as input files to an ARC/INFO Macro Language (AML) program. This AML program generates two-point coverage, one for the GPS points and the other for the event points for a given run. Once the coverages are generated, a menu-driven AML program overlays the GPS points and event points on the GIS street network, a process used for checking the validity of the runs. Figure 2 shows an ARC/INFO screen in which a GPS run was overlaid on the street network. If the trace of a GPS run showed a different or incomplete path of the designated route for the run, causes for such abnormality, such as driver negligence, bad GPS signal reception, or improper route definition, were identified, and appropriate measures were taken. Except for the part for GPS validity checking, which requires human interaction and judgment, the process is designed for batch processing with log files for error reporting.

Link Topology Definition

There are several notations to define a link on a route. The reference street notation (in the form of on-street, from-street, to-street) is easy for human comprehension. However, computers prefer the segment-node notation (on-segment, from-node to-node, or on-segment from-XY to-XY) or the route-measure notation (on-route, from-measure, to-measure). The task of this program module is to convert links already defined in the reference street form into the segment-node notation and to the route-measure notation. Segment-node notation is needed to structure the GPS data, which are continuous throughout a run, on a link-by-link basis using the coordinates of the control nodes of the links. The route-measure notation is needed to map the link attributes onto the GIS coverage, in addition to obtaining the length of each individual link.

Link topology was defined by extracting information from the base map; therefore, an accurate base map is the key to the success of the process. The base map used for the study was digitized from stereo aerial photography in 1991 at a scale of 1:24,000 by the Maricopa County Department of Transportation in cooperation with the

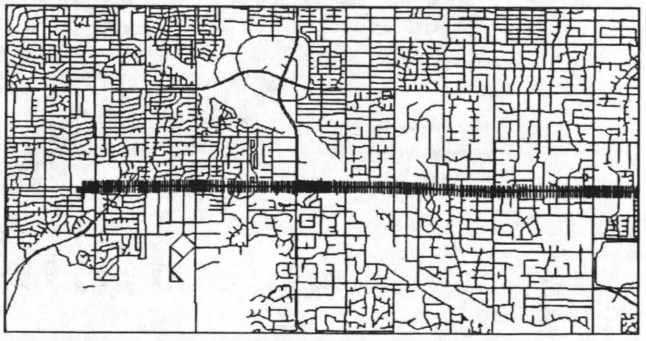


FIGURE 2 A GPS run was overlaid on the street base map for validity checking.

Arizona Department of Transportation. The relative accuracy of the map is 7.5 m (25 ft), which was deemed adequate for the study.

The first step in the module is to build a route feature that contains the travel routes and links for the study. Special attention was given to ensure that each route was continuous and without any unwanted offshoots. Programs were written to detect discontinuity between adjacent arcs on each route in the base map so such gaps could be corrected. Once the route feature was built, various attribute tables relating to the study routes and sections/arcs were exported to the PC, where the data base programs convert the already defined reference street link definition into segment-node definitions by making reference to these tables. The resulting link control nodes, in X,Y coordinate form, were exported to ARC/INFO to check their validity. Once the control nodes passed the validation, their measures or linear positions on the routes were generated using Dynamic Segmentation commands in ARC/INFO. These measures were then used to calculate the from-measure and to-measure and the length of the links. The LINKTOPO table resulting from the procedure contains the three different notations of route-link definition.

Data Analysis and Report Module

There are two major procedures involved in the Data Analysis and Reporting module. The ADDLINK procedure structures the continuous GPS data on a link-by-link basis. It inserts the control node in the GPS data and labels each link with a link ID. The time the vehicle passes through the control node of a link is interpolated based on the times for the two adjacent GPS points and the distance of the control node to the two adjacent GPS points. This procedure generates the second version of RTERUN.DBF.

The ANALINK procedure then calculates travel time, running time, and delays (including different types of stop delay) on each link for each individual run and generates a report of mean travel time, mean running time, mean delay, and types of delay for each travel link. Stop delay is defined as the time that the vehicle is travelling at a speed less then or equal to 8 km/hr (5 mph). The program first scans for delay groups on a given link. A delay group is the time span during which the vehicle is travelling at a speed less than or equal to 8 km/hr (5 mph). It was arbitrarily decided that if the time gap of adjacent delay groups is within 10 sec, these delay groups are combined as one group. The ANALINK procedure then tries to associate a delay group with any event recorded within the time span of 20 sec before the delay group or 5 sec after the delay group. The 20-sec and 5-sec time spans for event association were also arbitrarily decided.

Embedded in these procedures are Structured Query Language (SQL) commands that produce output files for interface with SAS and ARC/INFO. After checking the variability of the average travel speed and average running speed, the SAS analysis procedures cal-

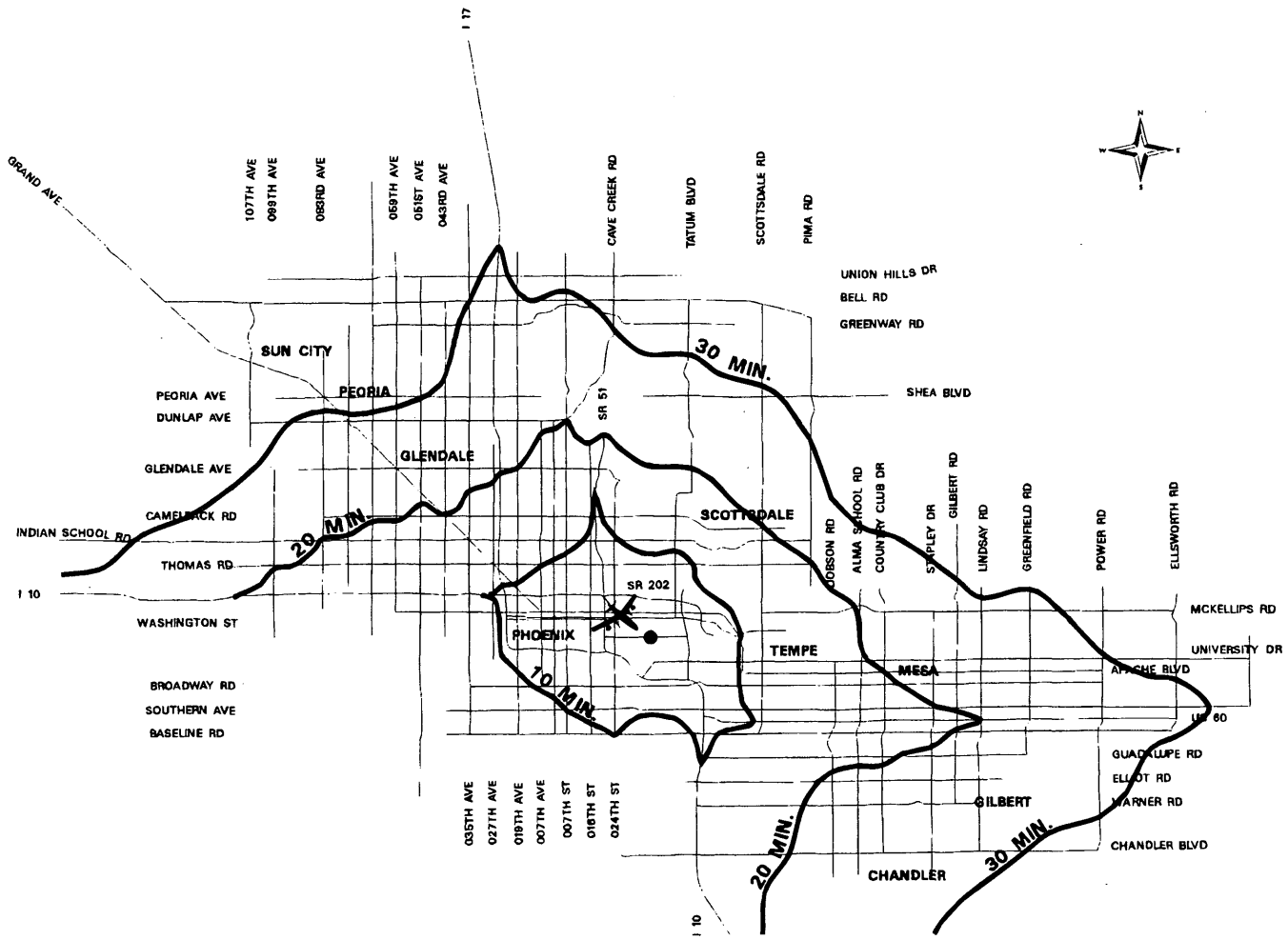


FIGURE 3 PM travel time from Sky Harbor Airport.

culate the mean travel speed, mean recurring and nonrecurring stop delay, and mean free-flow speed for different combinations of area type, number of lanes, and functional classification during a given period. The SAS procedures were also used to calculate miles of roads surveyed by functional class. The resulting data were then used to estimate travel speeds for the links that were not surveyed but shared the same combination of the above mentioned attributes. The tables containing such link-based information as mean travel speed, mean running speed, mean recurring, and nonrecurring stop delay are then passed on to ARC/INFO to conduct network analysis and to generate graphics reports. The preliminary network analysis produced the travel time contour maps and the minimum paths to and from major activity centers during all three periods. A left-turn delay of 51 sec and a right-turn delay of 37 sec resulting from the intersection delay study conducted as part of the project were used as turn impedance for the network analysis. Figure 3 shows travel time contour from Sky Harbor International Airport during typical weekday pm peak hours.

CONCLUSION

GPS technology provides a new way of collecting network travel speed data for processing in the context of GIS. Although it is not the intention of this paper to evaluate the effectiveness of this method, that GPS has its advantages in areas of GIS integration, data collection effort, and data validation. However, GPS units are expensive to purchase or to rent. The postprocessing can be very complex and requires a good electronic base map. If real-time differential GPS is not available, postdifferential correction adds another burden to the postprocessing. Furthermore, a number of factors can affect good GPS reception, many of which are difficult to prevent or even to predict. In summary, the success of an integrated GPS/GIS system in network travel time study relies on good GPS reception, accurate base maps, a well designed post-process system, plus a good planning and quality control for field data collection.

REFERENCES

1. McShane, W. R., and R. P. Roess. *Traffic Engineering*. P.140, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1990.
2. *ProXL System Operation Manual*. GPS Pathfinder Series, Trimble Navigation Limited, Sunnyvale, Calif., 1994.

Publication of this paper sponsored by Committee on Transportation Data and Information Systems.

Design of Routing Networks Using Geographic Information Systems: Applications to Solid and Hazardous Waste Transportation Planning

M. HADI BAAJ, SULEIMAN A. ASHUR, MIGUEL CHAPARROFARINA, AND K. DAVID PIJAWKA

Geographic information systems (GIS) represent a technology with considerable potential for important applications in transportation engineering. This paper focuses on the applications of GIS technology in the transportation routing area. Two case studies are presented: the first focuses on the design and analysis of different Arizona statewide waste tire collection transport networks corresponding to different percentages of total annual waste tires collected. The second case presents ongoing Environmental Protection Agency-sponsored research on the transportation routing and risk management of hazardous waste shipments across the U.S.-Mexico border region. Both routing applications take advantage of the efficiency and productivity of GIS technology and have been implemented on the GIS software TransCAD.

This paper focuses on the applications of Geographic Information Systems (GIS) technology in the transportation routing area. This is demonstrated via two case studies implemented on the GIS platform TransCAD. The first section presents a general overview of geographic information systems and its applications in transportation engineering, an extensive literature review of GIS applications in the transportation routing area, and a brief description of TransCAD. The second section presents the first case study: the design of Arizona statewide waste tire collection networks using data provided by the Arizona Department of Environmental Quality (ADEQ). The problem definition and application significance are revealed, the solution approach is reviewed, and application results are presented. The third section presents the second case study: ongoing Environmental Protection Agency (EPA)-sponsored research focusing on the transportation routing and risk management of hazardous waste shipments across the U.S.-Mexico border region. The interest of EPA in this research study is a direct result of the recent North America Free Trade Agreement (NAFTA). This paper then concludes with an overview of results and directions for further research.

GEOGRAPHIC INFORMATION SYSTEMS

Overview of GIS Systems

GIS is a computerized data base management system for the capture, storage, retrieval, analysis, and display of spatial data. A GIS contains two broad classifications of information: geocoded spatial data and attribute data. Geocoded spatial data define objects that have an orientation and relationship in two- or three-dimensional space. Each object is classified as either a point (such as an accident or a signal location), a line (for example a highway), or a polygon (number of people living within a block) and is tied to a geographic coordinate system. These objects have precise definitions and are clearly related to each other according to the rules of mathematical topology. Moreover, a GIS contains the same attribute data that are found in traditional data bases. Attributes associated with a street segment might include its width, number of lanes, construction history, pavement conditions, and traffic volumes. GIS is preferred over a traditional data base because data attributes are associated with a topological object (point, line, or polygon) that has a position somewhere on the surface of the earth (*1,2*). Spatial considerations are fundamental to most transportation activities. A transportation system network representation consists of nodes, links, and entities distributed in two- or three-dimensional space. Events happen within this system at a point (an accident or a signal location), along a segment (vehicle volumes or pavement deficiencies), or within a geographical area (the number of people living within two blocks of a bus stop or working in an industrial park).

GIS Applications in Transportation Planning

GIS applications can be expected in traditional areas of the highway agency responsibility, namely: pavement management; traffic engineering; planning and research; bridge maintenance; and field office support (*3*). A wide range of prototype and even fully operational GIS applications were identified by a research team from the University of Wisconsin in planning, management, and engineering (*4*). GIS software was used by the Saskatchewan Department of Highways and Transportation to build a regional highway network using their corporate highway data base including traffic count data (*4,5*). A second planning application used the overlay and routing GIS functionalities for hazardous material routing. The overlay GIS function was used to generate estimates of the population within a specified distance of each link in the highway or rail network. This enabled the inclusion of the population exposure in an objective function for route selection (*4*). Other planning applications include evacuation planning, planning for hazardous material release inci-

M. H. Baaj, Department of Civil and Environmental Engineering, American University of Beirut, Beirut, Lebanon. S. A. Ashur and M. Chaparrofarina, Department of Civil Engineering, Arizona State University, Tempe, Ariz. 85287-5306. K. D. Pijawka, School of Planning and Landscape Architecture, Arizona State University, Tempe, Ariz. 85287-2005.

dents, development of new traffic analysis zones from census tracts, and development of new urban highway networks (*4*).

GIS Applications in Routing

GIS is a powerful tool in the analysis and design of transport routing networks. Its graphical display capabilities allow not only visualization of the different routes but also the sequence in which they are built, which allows the understanding of the logic behind the routing network design. A GIS adds a degree of intelligence and sophistication to a transportation data base that has been previously unknown. For a segment on a road, a GIS system knows what routes cross it and whether there is an actual physical intersection. It knows the position of roadside features along the segment and can tell which census blocks are to the right and left of the segment or within any distance of it. Rather than being limited to textual queries, it is possible to perform geographic queries in a straightforward, intuitive fashion. For example, a GIS with the appropriate routing algorithm and data can easily compute and display the route that will result in the minimum population exposure to a shipment of hazardous materials. With the route drawn on the computer screen, the analyst can see immediately how the logic of the model has bypassed certain population centers. The analyst can create a detour by pointing to a road segment and deleting it from the network and then watch the routing algorithm redraw the path. Similarly, the analyst can ask a series of geobased questions and obtain the answers quickly in an easy-to-understand, color-coded display on the screen, hard copy, or disk file.

The interaction between the transportation system and its surrounding environment makes GIS technology ideally suited for hazardous materials-routing design, risk analysis, and decision making. GIS combines information on the transport network, social and demographic factors, weather conditions, topography and geology to assess the likelihood of a spill and its probable consequences. GIS can also be integrated with sophisticated mathematical models and search procedures to analyze different management options and policies.

GIS has been used in the risk analysis of hazardous materials transportation in Arizona (*6*). The main objective of that research was to assess the risk and vulnerability of transporting hazardous materials and waste on the Arizona highway system. A general GIS system called Geographic Information Management System (GIMS) was used. The risk assessment model used was based on four factors: accident rate by segment, shipment frequency by highway segment, population density along the routes, and response time. Each component was evaluated and presented on the GIS map. Although this study was one of earliest and most basic research studies in this area, it lacks a strong modeling of risk assessments and routing evaluations.

A second study aimed to assess the best routes for transporting hazardous materials in an area of about 609 km^2 (235 mi^2) in France (*7*). To assess alternative routes, the transport risk was calculated based on the multiplication of the accident probability by the number of exposed people. The minimum risk route from one origin to a destination was determined using Dijkstra's shortest path algorithm. Different risk values were assigned to different segments of the transportation network based on the type of material transported, its packaging aspects, accident probability (which varies with vehicle and route characteristics, such as road classification, bridges and tunnels, and so on etc.), transportation operational aspects (such as speed limit), law enforcement and weather conditions, and the environmental surroundings along the routes (such as population density, sensitive buildings such as schools and hospitals, general land use, and, in particular, sensitive environmental zones such as water resources areas). An output of the study was a GIS software called INGRES. Although the level of detail in this study was comprehensive, it was specific for the area of study. In addition, the economical component was not considered in this study (e.g., risk profile and the tradeoff between cost and risk).

A third study developed a first-generation GIS-based system for hazardous materials transport routing and risk management (*8*). The system, called HAZ-TRANS, currently under development, can be run using a stand-alone microcomputer. It is designed to accommodate future data collection and model enhancements. This system was used to highlight the conflicts presented by routing decisions, namely, between economic and safety concerns, and among potentially affected populations. Two applications were analyzed: first to minimize trip distance and second to reduce population exposure. It was found that a relatively small increase of about 10 to 15 percent in trip distance will achieve a reduction of 40 percent in population exposure.

TransCAD

TransCAD is a GIS software package for the planning, management, operation, and analysis of transportation systems and facilities (*9*). It can be used for any application that requires digital mapping, spatial analysis, and data retrieval and maintenance. It can analyze all types of geographic and spatial data with extended capabilities for transportation data. In addition it is ideal for many different transportation applications such as highway or transit planning and operations, facility management and inventory systems, accident reporting and analysis, pavement management, maintenance planning, demand modeling and forecasting, market analysis, environmental impact assessment, regulatory and policy analysis, distribution planning, routing and scheduling, and emergency management (*9*).

The reason for selecting TransCAD as the GIS software platform for both routing applications discussed in this paper is that it includes a battery of procedures tailored for transportation applications. Such procedures geocode data, build networks, find shortest paths, create service districts, and perform polygon overlay processing. The first application demonstrating the design of statewide waste tire collection networks uses the "SHORTEST-PATH" and "VEHICLE-ROUTING-PROBLEM" procedures of TransCAD applied repetitively to the GIS-coded map of the state of Arizona. The second application assessing the risk of transporting hazardous waste across the U.S.-Mexico border region uses the "K-SHORTEST-PATHS" procedure of TransCAD applied repetitively to the GIS-coded map of the U.S.-Mexico border region (*10,11*).

ARIZONA WASTE TIRE COLLECTION NETWORK DESIGN

Application Definition and Significance

A waste tire processing facility has been operating in the state of Arizona since 1990. As part of the incentives given to the operating company to start business in the state, Arizona guaranteed to the company that for the first 2 years of operation, all waste tires disposed in the state's county landfills would be processed at its facility. Thus, there was a need for the state to contract with a hauler to transport the waste tires from each waste tire collection site (WTCS) to the waste tire

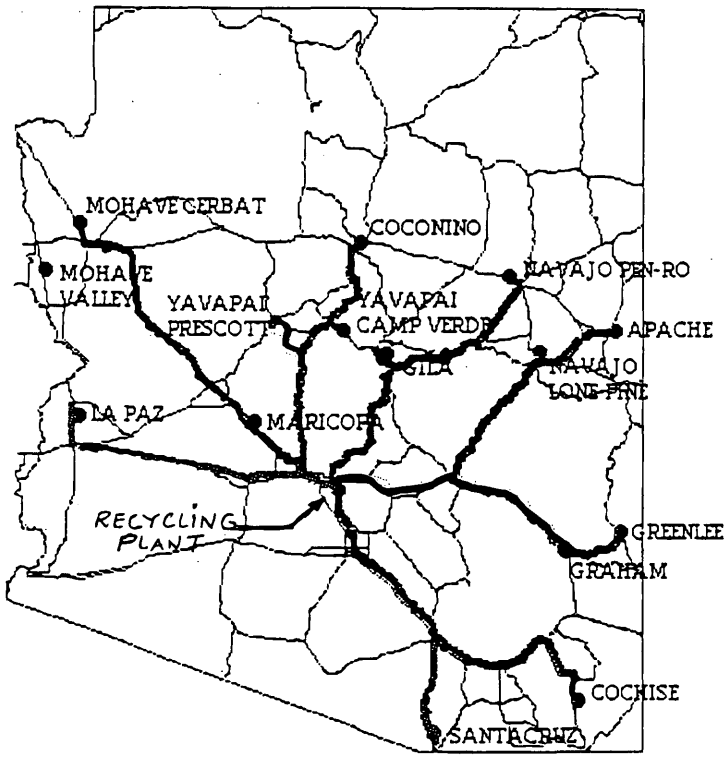
processing facility (WTPF). There are 15 approved WTCS and 1 WTPF in operation in Arizona as shown in Figure 1 (*12*).

The Arizona Department of Environmental Quality (ADEQ) was interested in identifying the optimal network for the collection of waste tires from all the 15 participating WTCS (100 percent demand satisfaction) to the WTPF and its corresponding total annual hauling cost. This information is useful and necessary to ADEQ so that it can use it as a base to compare the hauling bids with, and to better negotiate its contract with, the selected hauler. In addition, ADEQ was interested in obtaining a plot of the total annual hauling cost versus the number of waste tires collected because it increases from 13,678 megagrams/year (15,080 tons/year) (the number generated by the largest WTCS of Maricopa County, representing 75.2 percent of the total annual waste tires disposed) to 18,187 megagrams/year (20,052 tons/year), the total number generated by all 15 participating WTCS. This involves the design of different routing networks corresponding to different percentages of the demand satisfied. This is useful to ADEQ in decision making under budgetary constraints that may affect the amount of funding allocated for the waste tire hauling program. Thus, given a certain maximum funding allocation for tire hauling, ADEQ can then determine the corresponding total number of waste tires that could be collected for processing and the collection network configuration.

The study objectives were to (*a*) design with the aid of the GIS TransCAD the statewide tire collection network and determine its total annual hauling cost; (*b*) provide ADEQ with a plot of the total annual hauling cost versus the percentage of total tires hauled for processing; and (*c*) demonstrate and test the use of TransCAD in the design of collection networks. The data for this study were obtained from ADEQ's tire manifest data base. By law, at each WTCS, any disposer of waste tires is required to fill a manifest indicating the date of disposal and number and size (large or small tires) of the disposed tires. Such manifests are collected by each county authority and are delivered to ADEQ. Thus, ADEQ is capable of maintaining a waste tire data base indicating the number and sizes of tires disposed daily at every WTCS.

Solution Methodology

The waste tire collection network configuration can be either a direct network, an indirect one, or a composite network. In a direct



Annual Number of Disposed Tires in Arizona's 15 WTCS (in Tons/year)

Mohave Cerbat = 793; Mohave Valley = 793; Coconino = 801; Yavapai Prescott = 577; Yavapai Camp Verde = 424; Navajo Pen Ro = 224; Gila = 250; Apache = 91; Navajo Lone Pine = 133; La Paz = 101; Maricopa = 15,080; Greenlee = 13; Graham = 200; Cochise = 424; Santa Cruz = 148. TOTAL = 20,052 Tons/year.

FIGURE 1 Arizona's 15 waste tire collection sites (WTCS) and waste tire processing facility (WTPF). (Also shown is the direct network generated by TransCAD satisfying 100 percent of total demand).

network, all collection trips are performed directly between each WTCS and the WTPF. Thus, there are as many routes as there are WTCS, and each hauling route operates independently with its own fleet of trucks. This network can be easily designed with repetitive application of the SHORTEST-PATH procedure of TransCAD.

In an indirect network, a hauling truck can travel not only between one WTCS and the WTPF but among several locations and the recycling plant. This is suitable for the case of disposal sites with a minimal annual number of tire deposits, which does not justify the expensive cost of direct line hauling. This mode of operation follows that of a multiple-origin-single-destination traveling salesman problem (TSP). Such a network is configured with repetitive application of TransCAD's "VEHICLE-ROUTING-PROBLEM" procedure.

A composite network is one involving a direct network between a subset of the WTCS and the processing facility and an indirect network between the complimentary subset of WTCS and the processing facility. For the case of Arizona, one WTCS (namely, that of Maricopa County) produces 75.2 percent of the total number of annually disposed tires, and the combination of the remaining 11 WTCS produces 25 percent. Thus, the composite network was perceived to consist of one direct haul line between Maricopa County's WTCS and the WTPF and an indirect network between the remaining 11 WTCS and the WTPF (*13*). For this application, a simplifying assumption was made that 1.61 km (1 mi) costs $1 to operate (such costs include fuel, maintenance, driver wages, and depreciation but exclude the fixed costs of acquiring the trucks). Also all fleet sizes were assumed to be the same, consisting of 18.14-megagram (20-ton) trucks.

Results of Application

Computational results using TransCAD indicate that it costs approximately $145,000/year to operate the direct line serving Maricopa County's WTCS (75 percent of total demand satisfied). To serve 80 percent of the demand, there are many possible combinations of WTCS other than Maricopa County's WTCS, which account for approximately 5 percent more of the demand satisfaction. Using TransCAD, the best network configuration (whether it is a direct, indirect, or composite one) is determined for each possible combination. The combination of WTCS whose optimal network configuration has the lowest annual hauling cost is selected as the one corresponding to the given percentage of demand satisfaction. Figure 2 shows the plot of the total annual hauling cost versus the percentage of demand satisfied. As shown in Figure, as the percentage of demand satisfied increases from 75 to 100 percent, the total annual hauling costs increases from $145,000 to $280,000 (a 93 percent increase in the annual cost for a 33 percent increase in demand satisfaction) (*13*).

HAZARDOUS WASTE SHIPMENTS IN THE U.S.-MEXICO BORDER REGION

Application Definition and Significance

Under a 1988 Mexican environmental law, any hazardous waste generated by U.S. companies (Maquiladoras operating in Mexico's border region) must be returned to the U.S. for treatment and disposal. According to the EPA, only 40 percent of the hazardous waste generated in Mexico returns to the U.S. Moreover, in 1992, only 10 percent of the companies in the Mexican states of Baja and Sonora requested official shipments of hazardous waste to the U.S. Most of this waste was either stored in Mexico or was treated in small recycling firms. Currently, there is no complete data base providing information on the pattern of shipments of waste from these industries as well as the shipment routes to disposal facilities. Furthermore, it is widely expected that the amount of hazardous waste will increase substantially as a result of the North American Free Trade Agreement (NAFTA) and the resulting accelerated relocation of U.S. industries to Mexico. In addition, NAFTA's proposed regulations will render illegal dumping very difficult in Mexico (*14*); the permit process for regulating hazardous waste has progressed since 1988.

The goal of the ongoing EPA-sponsored research is to develop an analytical framework for assessing the transportation risk of shipping hazardous waste in the border region using the GIS technology. The research addresses two overriding and major issues; the first issue concerns the lack of a current comprehensive data base that tracks the amounts of hazardous shipments, determines their risks to population and the environment, and identifies the patterns of shipments (their origins, destinations, and transport routes). The second issue is the growing need for a risk assessment model that can assist in determining the transport risks involved and can serve as a valuable tool for formulating different management scenarios aimed at transportation risk reduction and equity (*14*). The research objectives are to: 1) develop a risk assessment model/framework implemented

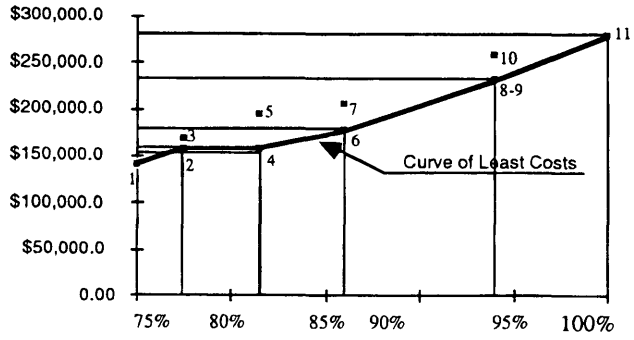


FIGURE 2 Annual hauling costs versus the percentage of demand satisfied.

on the GIS platform TransCAD; 2) test the model; and 3) develop methodologies that can be applied to the U.S.-Mexico border region.

The aims of the framework are to assess the risks of transporting hazardous waste in the U.S.-Mexico border and to examine different transportation and other related scenarios that will facilitate decision making and planning to reduce the impacts of transportation accidents. The framework will be tested and demonstrated by applying it to a specific geographical area along the U.S.-Mexico border region, namely, the Arizona-Sonora area. This area was selected because of the availability of data and because the risks of shipping hazardous waste were considered a dominant environmental and public safety concern (*14*).

Solution Methodology

The following tasks were proposed to achieve the goals of the research:

1. *Literature Review of Risk Assessment Models*: This task has been completed, and results were summarized in the preliminary report sent to EPA (*14*).
2. *Data Collection and Analysis*: This task has been completed. It involves the collection of several data for implementation on the GIS software TransCAD:
   a. Hazardous waste shipment data: These data include the quantities, types, and routes that are used in the shipment of hazardous waste from U.S.-owned industrial plants in the Mexican State of Sonora to the state of Arizona in the U.S. Moreover, these data need to be confirmed through the manifest copies available from the EPA.
   b. Land Use Data: These are data on sensitive properties and population densities in the U.S. and Mexico.
   c. Truck Accident Data: These data are necessary to calculate the probability of accidents of trucks carrying hazardous waste shipments.
3. *Development of Solution Framework*: This ongoing task considers computer modeling by establishing a TransCAD-implemented data base of Arizona and Sonora highways with different attributes (such as length of road segments, speed limits, type, and population along the route), applying the risk assessment model to the GIS map, and developing the necessary computer algorithmic analysis programs (Database IV, FORTRAN, and so on). The use of TransCAD's K-SHORTEST-PATHS procedure will be instrumental in the identification of many alternative routes for every pair of hazardous waste origin-destination.
4. *Development and Analysis of Different Management Scenarios*: A major contribution of this study is to identify the impacts of different risk management scenarios. There are two sets of scenarios; the first set of scenarios focuses on changes in the demand pattern (and amounts) such as: (*a*) dumping some/all of the waste in Mexico; (*b*) the partial closure of the border between California (or Texas) and Mexico; and (*c*) the construction of new recycling facilities in Sonora and/or Arizona. The second set of scenarios focuses on impacts of routing of hazardous waste under different preferences (such as minimum cost, minimum risk, risk equity, or a combination thereof).

CONCLUSIONS

GIS is a powerful technology in the analysis and design of transport routing networks. The key contribution of GIS technology is that it adds a major degree of intelligence and sophistication to a transportation data base that is inherently geographical in nature. The interaction between the transportation system and its surrounding environment makes GIS technology ideally suited for solid and hazardous waste routing design, risk analysis, and decision making. GIS technology combines information on the transport network configuration, social and demographic factors, weather conditions, topography, and geology to assess the likelihood of a hazardous spill and its probable consequences. GIS can also be integrated with sophisticated mathematical models and search procedures to analyze different management options and policies. Two case studies were presented to demonstrate the usefulness of GIS technology in the design of routing networks. The first case dealt with the design of many Arizona statewide waste tire collection networks corresponding to different percentages of the demand satisfied. The second case reviewed ongoing research examining the routing and risk management of transporting hazardous waste across the U.S.-Mexico border region.

REFERENCES

1. Simkowitz, H. J. GIS: Technology for Transportation. *Civil Engineering,* Vol. 59, No. 6, June 1989, pp. 72–75.
2. Simkowitz, H. J. Geographic Information Systems: An Important Technology for Transportation Planning and Operations. In *Transportation Research Record 1236,* TRB, National Research Council, Washington, D.C., 1989.
3. Abkowitz, M., S. Walsh, E. Hauser, and L. Minor. Adaptation of Geographic Information Systems to Highway Management. *Journal of Transportation Engineering,* Vol. 116, No. 3, May/June 1990, pp. 310–327.
4. Vonderohe, A. P., L. Travis, R. L. Smith, and V. Tsai. *NCHRP Report 359: Adaptation of Geographic Information Systems for Transportation.* TRB, National Research Council, Washington, D.C., 1993.
5. Kriger, D., M. Hossack, and M. Schlosser. Integration of GIS with the Transportation Model TModel2 at Saskatchewan Department of Highways and Transportation. *Proc., Geographic Information Systems (GIS) for Transportation Symposium,* Orlando, Fla., 1991, pp. 127–138.
6. Anders, C., and J. Olsten. GIS Risk Analysis of Hazardous Materials Transport. *State and Local Issues in Transportation of Hazardous Wastes Materials: Towards a National Strategy. Proc., National Conference on Hazardous Materials Transportation,* May 1990.
7. Lassare, S., K. Fedra, and E. Weigkricht. Computer-Assisted Risk Assessment of Dangerous Goods Transportation for Haute-Normandie. State and Local Issues in Transportation of Hazardous Wastes Materials: Towards a National Strategy. *Proc., National Conference on Hazardous Materials Transportation,* May 1990.
8. Abkowitz, M., P. D. M. Chang, and M. Lepofsky. The Use of Geographic Information Systems in Managing Hazardous Materials Shipments. In *Transportation Research Record 1245,* TRB, National Research Council, Washington, D.C., 1989.
9. TransCAD. Transportation GIS Software, Reference Manual, Caliper Corporation, Newton, Mass., 1993.
10. Shier, D. R. Iterative Methods for Determining the k Shortest Paths in a Network. *Networks 6,* 1976, pp. 205–232.
11. Shier, D. R. On Algorithms for Finding the k Shortest Paths in a Network. *Networks 9,* 1979, pp. 195–214.
12. *Waste Tire Program Annual Report.* Arizona Department of Environmental Quality (ADEQ), Phoenix, 1992.
13. Chaparrofarina, M. A. *Arizona Statewide Optimal Network for the Collection of Waste Tires Using GIS.* M.S. Thesis. Arizona State University, Tempe, 1994.
14. Ashur, S., M. Hadi Baaj, and K. D. Pijawka. *Risk Assessments of Transporting Hazardous Wastes in the U.S.A.-Mexico Border Region.* Center for Environmental Studies, Arizona State University, Tempe, July 1994.

Publication of this paper sponsored by Committee on Transportation Data and Information Systems.

Modeling Washington State Truck Freight Flows Using GIS-T: Data Collection and Design

KENNETH L. CASAVANT, AMY ARNIS, WILLIAM R. GILLIS, WAYNETTE NELL, AND ERIC L. JESSUP

As part of their planning process, state departments of transportation and metropolitan planning organizations are required to include detailed information on freight and goods movements. However, obtaining comprehensive information on freight truck movements is often difficult. The Washington State Department of Transportation initiated a statewide freight truck origin and destination study in April 1993, completed in March 1994, to meet this challenge. An overview of the research procedures used to conduct the study is presented. Specific emphasis is given to arranging the data to complement the role of GIS-T, which is used as a tool for organizing, analyzing, and presenting information for use by transportation planners, program administrators, and policy makers. A case study of southbound trucks on SR-395 from Canada to Spokane, Washington, illustrates how GIS-T can be used to document and analyze the characteristics of freight truck movements.

The efficient intermodal movement of freight and goods, a primary responsibility of state departments of transportation, metropolitan planning organizations (MPOs), and many local governments, has received increased emphasis because of the federal Intermodal Surface Transportation Efficiency Act of 1991 (ISTEA). ISTEA requires states and MPOs to include a specific focus on freight and goods mobility as one element of their updated plans.

Planning for the efficient movement of freight and goods by trucks is hindered by a lack of information concerning the source and characteristics of freight movements on state and regional highways. Freight movement by rail and water can be tracked adequately through Interstate Commerce Commission waybill samples, the U.S. Army Corps of Engineers Waterborne Commerce data, and other published sources. However, comprehensive information on truck freight movements is much more difficult to obtain because of the large number of carriers and the numerous potential origins and destinations.

To address this information gap, the Washington State Department of Transportation (WSDOT) initiated a statewide freight truck origin and destination study in April 1993. A regionwide freight truck origin and destination study was first proposed in Washington as an element of the Eastern Washington Intermodal Transportation Study (EWITS). EWITS is a 6 year ISTEA planning study to define the multimodal network necessary for the efficient movement of freight and people throughout the section of Washington located on the east side of the Cascade Mountains. Supplemental funding provided by WSDOT enabled the EWITS freight truck origin and destination study to be expanded to include the entire state. Washington State University and the Gillis Group, a private consulting firm, were asked to conduct the study.

K. L. Casavant and E. L. Jessup, Washington State University, Department of Agricultural Economics, Pullman, Wash. 99164-6210. W. R. Gillis and W. Nell, The Gillis Group, 108 North Adams, Ritzville, Wash. 99169. A. Arnis, WSDOT Planning Department, P.O. Box 47370, Olympia, Wash. 98504-7370.

The freight origin and destination study will help the state of Washington comply with ISTEA planning requirements, and it will contribute to the Washington State Transportation Policy Plan and Statewide Transportation System Plan. MPOs and regional transportation planning organizations will use information from the study to evaluate freight and goods mobility needs for their updated plans. Examples of specific contributions that will result from the study include:

- Documented information on freight movements that will target limited resources and pave the way for infrastructure improvements important to Washington's economy.
- A better understanding of the industries most reliant on Washington's highways. This information will be important for predicting future infrastructure demands associated with growing and declining industries.
- Improved pavement management systems arising from more accurate information on the specific routes used by freight carriers and the weights of commodities hauled over those routes.
- Essential routes serving deep-water ports, international airports, and Canadian shippers that will enhance Washington's international competitiveness.
- Improved efficiency of Washington's intermodal infrastructure systems resulting from the delineation of essential highways linked to Washington's rail, air, and barge intermodal centers.
- Information to better define needs on Washington's national highway system.

A discussion of the research procedures used to conduct the Washington State Freight Truck Origin and Destination Study is presented, with specific emphasis given to designing the data to complement the role of GIS-T. GIS-T is a data base that organizes, analyzes, and presents information for use by transportation planners, program administrators, and policy makers.

FIELD DATA COLLECTION PROCESS

An extensive process of collecting information directly from the drivers of freight trucks is a unique feature of the Washington State Freight Truck Origin and Destination Study. Truck drivers are interviewed at 30 locations throughout Washington, including 21 Wash-

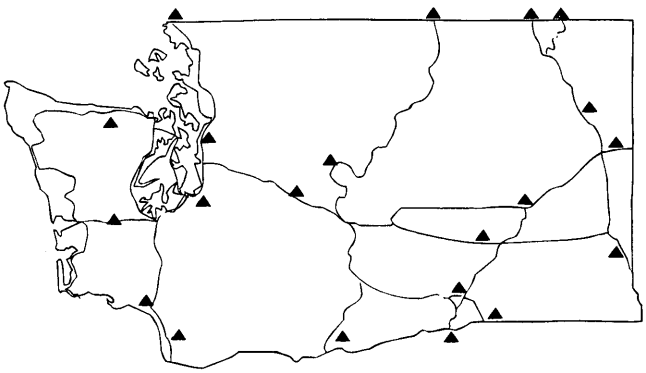


FIGURE 1 Truck interview locations.

ington State Patrol weigh stations, three Canadian border locations, and the Oregon Port of Entry at Umatilla (see Figure 1). The interviews are carried out four times over a 1-year period to reflect seasonal differences in freight movements. The summer, fall, and winter interview rounds are already complete. Interviews at most locations are conducted over a continuous 24-hr period to provide a complete 1-day freight truck movement profile for each season. To obtain median traffic patterns, truck driver interviews are conducted on Wednesdays instead of on Mondays or Fridays, when traffic flows are exceptionally heavy.

Cooperation from the Washington State Patrol Commercial Vehicle Enforcement Office and customs offices from both the United States and Canada is essential to the success of this research effort. Officers at designated locations along the highway flag down trucks and conduct routine enforcement activities. After they are stopped, truck drivers are asked to participate in a brief 2-min interview. Each driver is asked to provide information about the trucking company, vehicle weight, type of commodity carried, and the origin and destination of the vehicle.

To obtain an accurate seasonal profile of truck movements throughout the Washington, it is necessary to conduct interviews simultaneously at more than six sites across the state. Because at least 15 people are needed to cover a 24-hr interview session at each of the sites, a very large, short-term labor force was required to successfully complete the freight truck origin and destination study. To obtain the necessary manpower, members of community service clubs from across the state were hired and trained to conduct the interviews. The constant refinement of interview personnel management systems has helped improve the accuracy and effectiveness of data collection efforts. Overall, the service clubs have provided quality data in a highly professional manner.

Cooperation and participation by truck drivers also has been excellent. Statewide, more than 96 percent of the truck drivers who were asked to participate agreed. A statistical sampling procedure is used to ensure accurate representation of statewide truck freight flows. Approximately 7,000 truck drivers are interviewed during each survey round, providing a data base of approximately 30,000 interviews for the year-long study.

DATA MANAGEMENT, ANALYSIS, AND MODELING PROCEDURES

The framework for data management, analysis, and modeling of Washington state freight truck movements is depicted in Figure 2. Key data management, analysis, and modeling procedures for the study are discussed in this section.

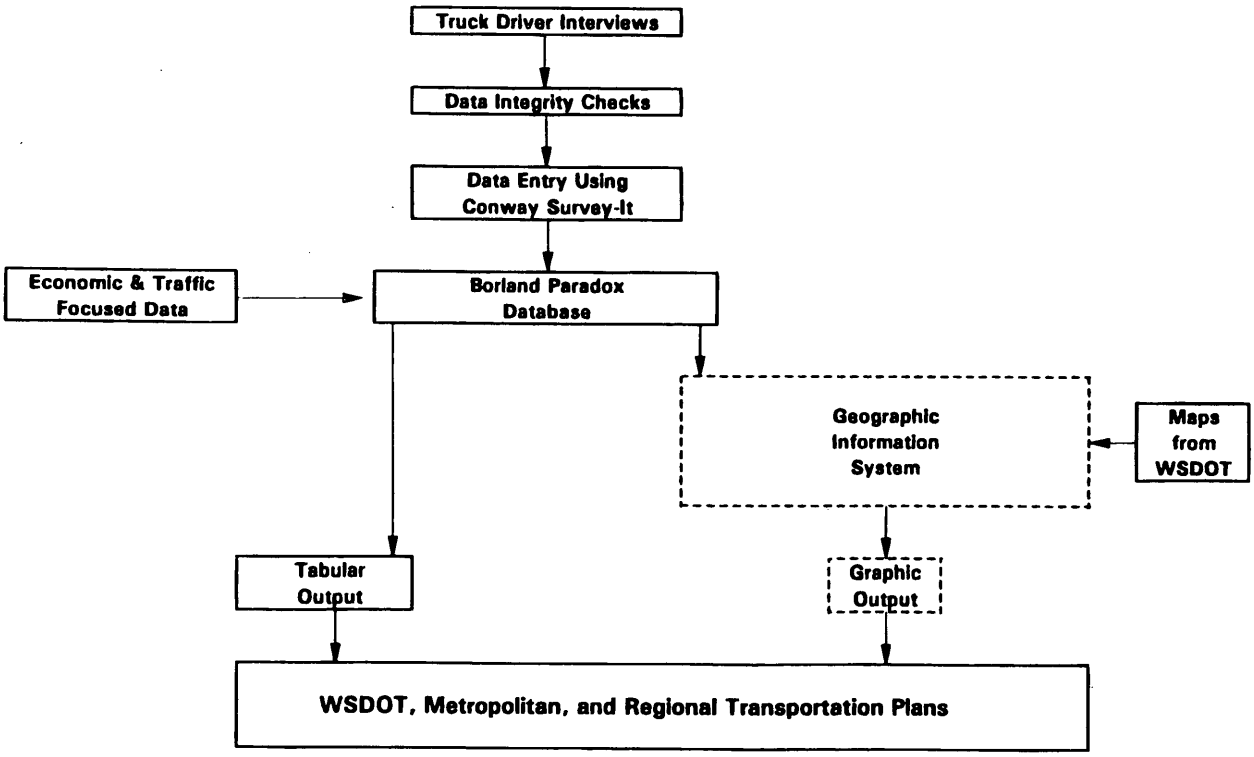


FIGURE 2 Framework for data management, analysis, and modeling.

Data Management

To accurately represent statewide and regional freight truck movements, data base management procedures must be carefully designed and implemented. Effective data management systems help reduce errors made during field data collection and data entry.

At least three potential sources of error are associated with field interviews of truck drivers. They include

1. Systematic problems caused by inappropriately worded questions, interview procedures, and site selection,
2. Inaccurate responses to questions, and
3. Interviewers who may incorrectly record vehicle data or drivers' responses.

Potential systematic errors caused by flaws in the survey methodology were minimized through constant evaluation of and adjustments to the interview questionnaire and site survey procedures. Improving the clarity of interview questions helped reduce errors related to inaccurate responses from drivers. A program of training and supervision for the community service club teams helped reduce the number of incorrectly recorded responses. Despite these safeguards, field data collection errors cannot be eliminated completely.

A data integrity review for each completed questionnaire was implemented before entering the information into the data base. Each questionnaire was reviewed to ensure that the answers were logically presented and consistent. Among the most frequent errors were questionnaires in which the total combined payload and empty vehicle weight given was well above the legal limit for a particular axle configuration. In these cases, the driver was usually providing the interviewer with the gross weight instead of the requested cargo weight. Another common error was the reporting of a truck carrying cargo when it was actually empty. The data integrity review process included the development of specific decision rules to revise incorrect data using other information on the questionnaire. For example, truck drivers who reported a combined cargo and empty vehicle weight in excess 110 percent of the legal limit were assumed to have provided gross weights. Revised payload weights were estimated as gross weight minus the reported empty vehicle weight. Empty vehicle weights were generally reported accurately.

Using these techniques, data recorded incorrectly on the field interview questionnaires were identified and corrected before data entry. The research team used the Conway Survey-It software package for data entry. Survey-It provides a user-friendly, menu-driven data entry screen, but only limited data base capabilities. Data entered into Survey-It were then exported into Borland Paradox. Paradox is used as the primary data base software for the project. Additional data integrity checks were implemented using the cross-tab, edit, and search functions of Paradox.

Data Analysis and Modeling

Data collected through statewide interviews with freight truck drivers provides valuable information for a variety of transportation planning applications. A number of specific examples of potential applications are listed in Figure 3.

Corridor Planning

- Identify highway corridors most critical to key industries
- Pinpoint major freight truck generators for specific corridors
- Document routes most widely utilized for national and international trade
- Provide base data to project freight truck traffic growth and decline for specific corridors
- Provide base data to estimate the economic value of specific commodities shipped on specific corridors

Intermodal Systems Planning

- Delineate essential highways linked to rail, air, deep water and river ports
- Evaluate intermodal systems most critical to key industries and international competitiveness
- Geographic proximity of intermodal facilities relative to origins and destinations of trucks utilizing those facilities
- Provide base data to project changes in highway usage that would result from rail-line abandonment or closing of key river ports

Pavement Management

- Document highway segments with the highest average freight cargo volumes and weights
- Provide base data to project future changes in freight cargo volumes and weights on specific highway segments

Congestion Management and Safety

- Document origins, destinations and routes used by freight trucks traveling through congested urban areas
- Provide base data to evaluate opportunities to reduce freight truck traffic through urban areas during peak commute periods

FIGURE 3 Potential applications for freight truck origin and destination data.

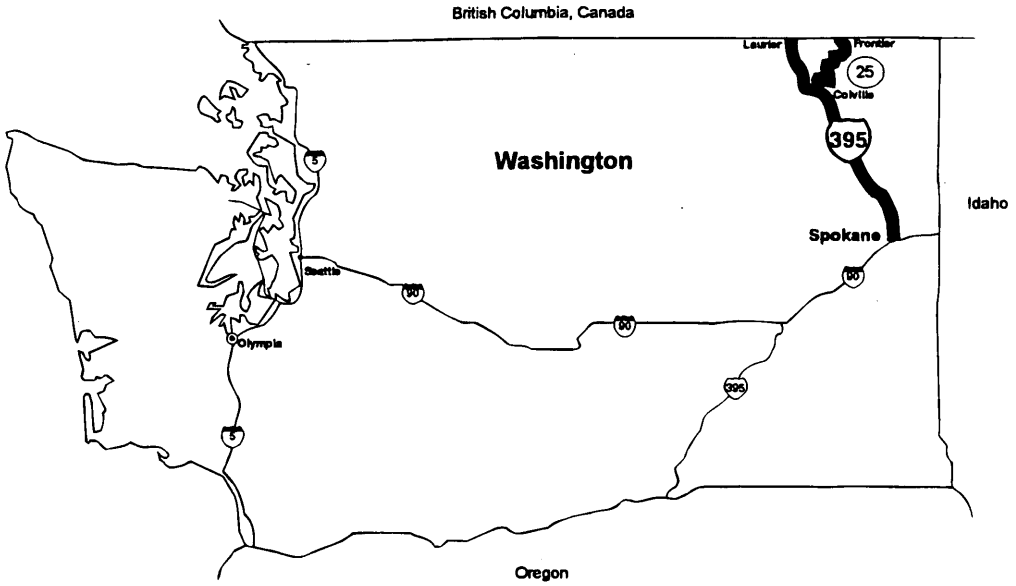


FIGURE 4 Geographic location of case study highway corridor.

The Paradox data base is organized to be as flexible as possible for specific transportation planning applications. Sample data collected from interviews with 30,000 Washington truck drivers will be weighted by the total number of freight trucks passing through each interview location over a 24-hr period. Additional weights based on vehicle route information are applied to eliminate double counting of trucks traveling on a specific highway segment. Using the weighted sample data, statewide truck movements can be accurately profiled and compared for different geographic locations within the state of Washington. After completing all four rounds of data collection, the weighted data will be used to compare differences in 24-hr freight truck movements for each of the four seasons.

For many planning applications it will be necessary to link results from the freight driver interviews with other data bases. GIS's structure provides that dynamic interlink of data. Industry data bases (such as the U.S. Census of Manufacturing and the U.S. Department of Commerce County Business Patterns) and state-level industrial growth projections are particularly useful. Each

TABLE 1 Profile of Cargo Carried by Southbound Trucks on SR-395 from Canada to Spokane and Destinations Beyond

Type of Cargo (by origin)	Number of Trucks	Percent of Total Trucks With Cargo	Percent of All Trucks
	(16 hour traffic sample, July 1993)		
Canadian Origin			
Wood Chips	31	18%	14%
Lumber Products	20	12%	9%
Fertilizer	19	11%	9%
Other Freight	5	3%	2%
Empty	8	NA	4%
Subtotal	**83**	**44%**	**38%**
Northeast Washington Origin			
Wood Chips	14	8%	6%
Lumber Products	24	14%	11%
Logs	8	5%	4%
Agricultural Products	7	4%	3%
Other Freight	42	25%	19%
Empty	43	NA	20%
Subtotal	**138**	**56%**	**62%**

Note: NA indicates not applicable

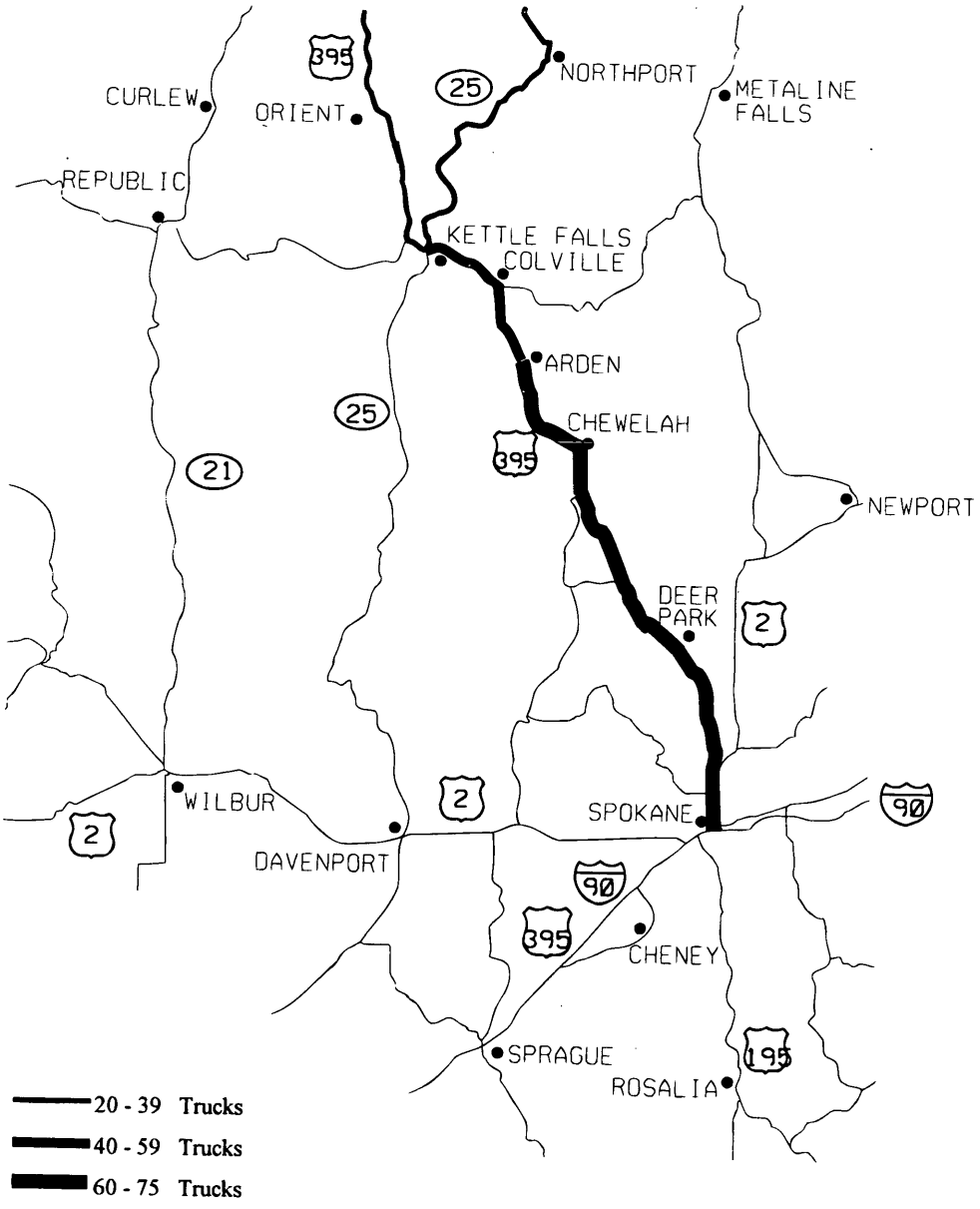
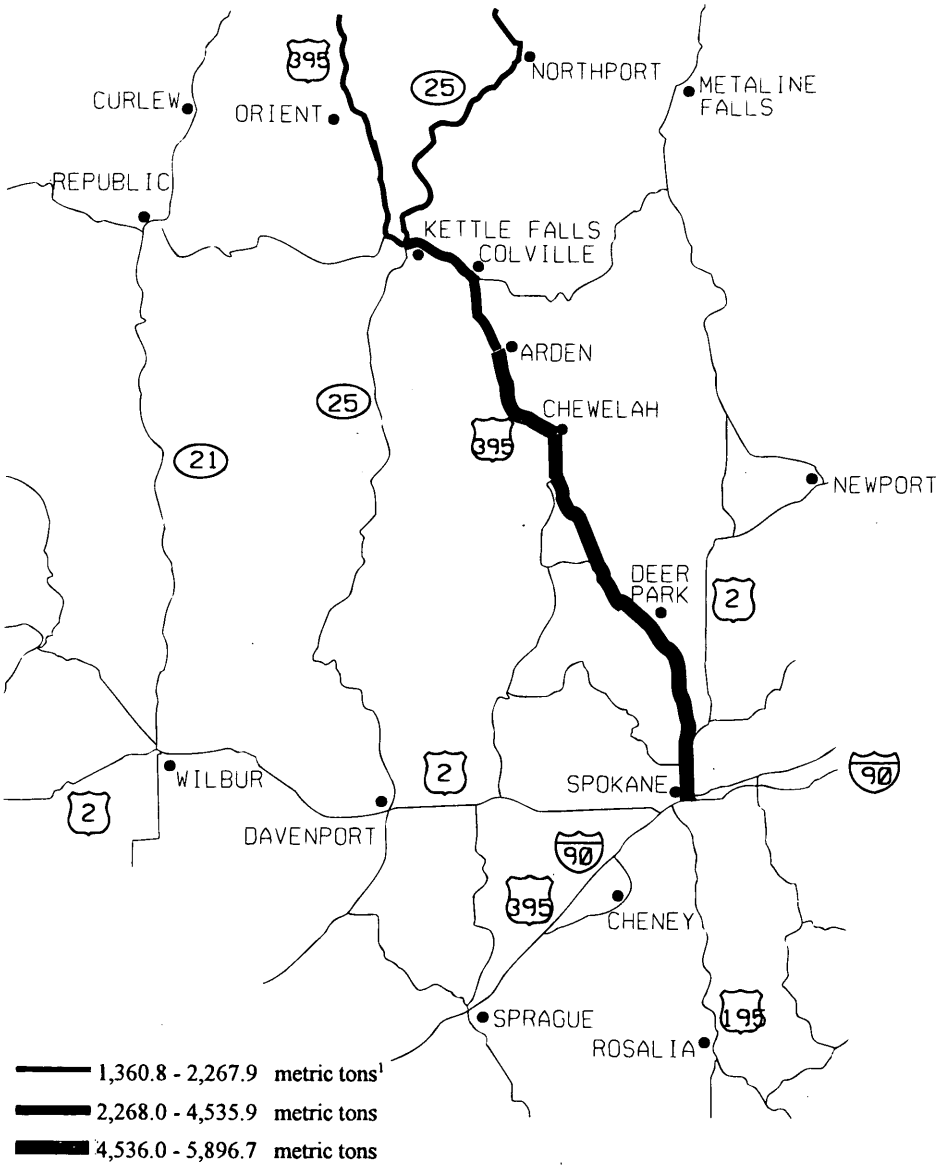


FIGURE 5 Total number of trucks carrying wood-related cargo (daily truck volumes).

truck driver interview record is assigned a specific Standard Industrial Classification (SIC) code consistent with the commodity being transported. Using the SIC code to link driver interview records with available industrial data bases will provide valuable information on both current and expected future freight truck movements on Washington highways. For example, highway corridors heavily traveled by trucks carrying commodities that are projected to grow rapidly as a component of the state's economy can expect future truck traffic increases, while corridors dominated by the shipment of commodities projected to decline should expect decreases in future truck traffic.

Tables generated from the Paradox data base will provide much of the detailed output describing key freight truck movements within the state of Washington. However, GIS-T is also used to provide graphical output highlighting characteristics of freight truck movements associated with specific segments of Washington's state highway network. Intergraph's PC-based MGE/MicroStation was selected as the GIS software for this study. Intergraph software is also being used by the WSDOT Geographic Services Division and WSDOT Planning Department. The digitized highway network, as well as substantial technical support, was provided to the research team by the WSDOT Geographic Services Division. The use of consistent software and digitized base maps will ensure the ability to easily interface with other WSDOT data bases for planning applications.

To interface with the Intergraph software, the truck driver interview data base from Paradox must be imported into Oracle. Freight truck characteristic data in Oracle is then interfaced with the tar-



[1] 1 metric ton = 2,204.6 pounds

FIGURE 6 Gross weight of total daily truck traffic.

geted digitized highway segments in MicroStation through MGE. The possible queries then become numerous.

REASONS FOR USING GIS-T

The bottom line purpose of the Washington State Freight Truck Origin and Destination Study is to provide information useful to the development of WSDOT, MPOs, and other regional transportation plans within the state of Washington. With this end purpose in mind, the use of geographic information systems provides two major advantages. First, a graphical presentation using GIS-T illustrates research findings in a form often more easily understood than the alternative of tabular output. Second, GIS-T enables a direct graphical interface with complementary planning data. For example, highway segments supporting the highest average daily cargo tonnage can be identified and directly compared with data bases documenting pavement conditions for those segments. This allows analytical questions to be posed and answered.

Graphical output complements but does not replace tabular output as a way to communicate and analyze freight truck origin and destination study results. Graphical presentation is most appropriate for geographical comparison of aggregate characteristics, such as total daily cargo tonnage on defined highway segments or key routes used by trucks traveling to a specific destination. However, tabular data is more appropriate for disaggregated information, such as detailed distribution of commodities carried by trucks traveling on I-90 to Puget Sound region ocean ports. Consequently, a combi-

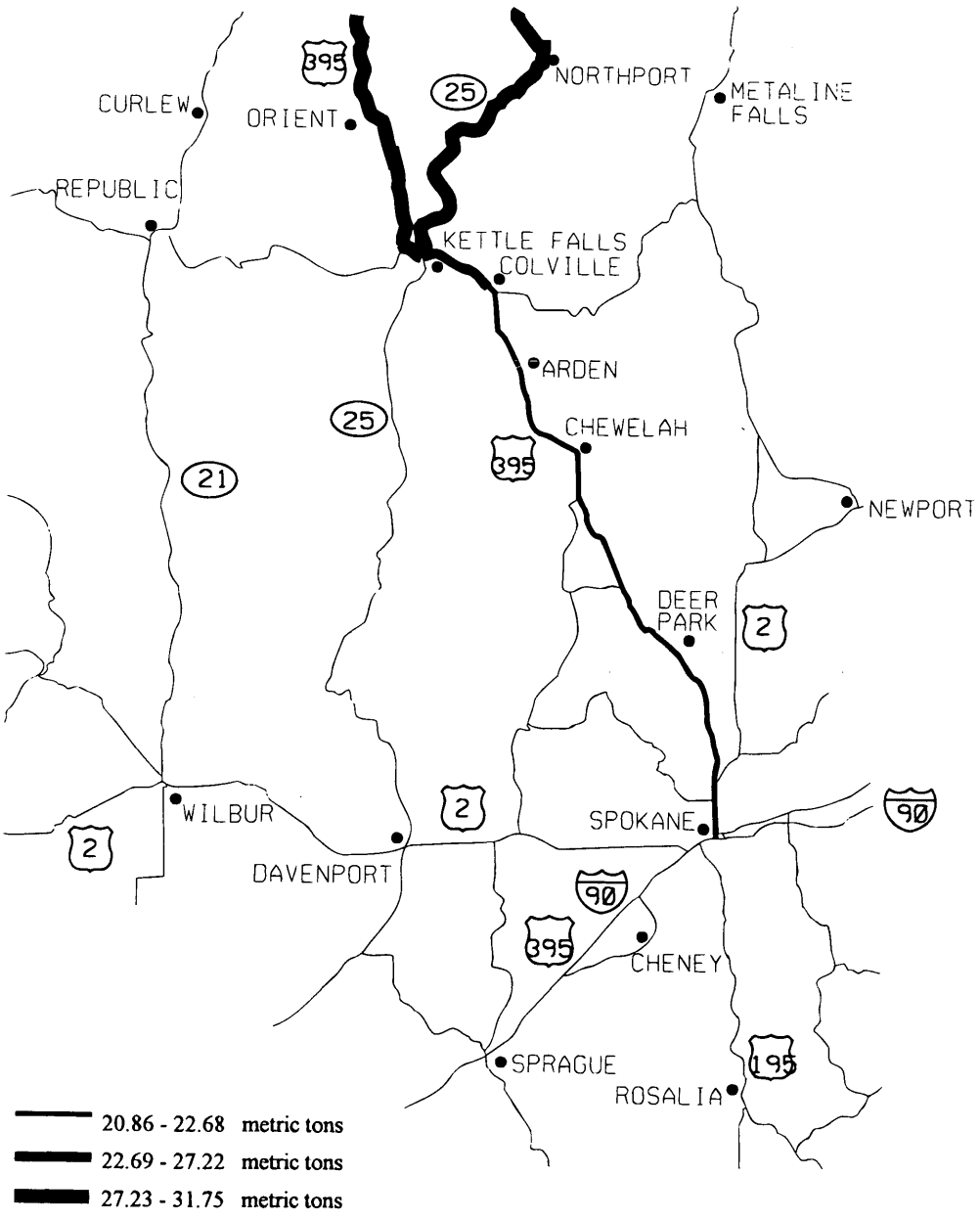


FIGURE 7 Median weight of freight truck traffic.

nation of graphical output and data tables is needed to effectively communicate findings from the freight truck origin and destination study.

CASE STUDY: SR-395 FROM CANADA TO SPOKANE

SR-395 from Canada to Spokane, Washington, is a critical transportation link, that support's the northeastern Washington economy and international trade with Canada (Figure 4). Because of the economic significance of this corridor, the WSDOT is undertaking a study of freight and goods movements on SR-395 from Canada to Spokane. GIS-T enables policy makers and WSDOT program administrators to analyze and compare key characteristics of truck movements along specific segments of this high-priority corridor. Several examples illustrate how GIS-T is being used to model freight and goods movements in the state of Washington.

Freight trucks from Canada account for 44 percent of the vehicles carrying cargo within the SR-395 study area (see Table 1). Wood chips, lumber products, and fertilizer are the primary products shipped from Canadian origins to markets in the United States. Wood-related products also are among the primary commodities transported by Washington state trucks traveling southbound on this

highway segment. Sawmills in Kettle Falls, Colville, and Arden are the primary generators of wood-related traffic from northeastern Washington.

GIS-T provides a way to compare the total volume of trucks carrying wood-related products on specific segments of SR-395 from Canada to Spokane (see Figure 5). Approximately 30 southbound trucks carrying wood products were recorded crossing the border at each of northeastern Washington's two major U.S. border crossings on SR-395 and SR-25. Approximately half the wood products trucks originating in Canada were chip trucks with loads terminating at a wood co-generation power plant located about 30 miles south in Kettle Falls. Most of the remaining Canadian-origin trucks carrying wood cargo continued on SR-395 to Spokane and destinations beyond. Additional SR-395 wood products traffic originated from sawmills in Kettle Falls, Colville, and Arden. Most of these trucks reported destinations in Spokane and points beyond. Consequently, the total volume of wood-related truck traffic was greatest on highway segments located on the southern end of the study area.

The aggregate daily gross weight of freight trucks traveling the SR-395 study area also was greatest on the southernmost portions of the corridor (see Figure 6). The pattern of increasing aggregate gross weight reflects the fact that most of the major generators of freight truck traffic in the SR-395 study area are located north of Chewelah. Total truck counts and aggregate gross weight are smallest on the highway segments between the Canadian border and Kettle Falls. Truck counts and aggregate gross weight increase steadily on highway segments located between Kettle Falls and Arden. Among highway segments in the study area, total truck counts and aggregate gross weight are highest between Arden and Spokane.

Within the study area, the road segments with the lowest daily truck traffic count support trucks with the heaviest median weight. The two highway segments connecting the Canadian border with Kettle Falls each carry trucks with a median weight over 27.2 metric tons (60,000 lbs). This reflects a high concentration of wood chip trucks that typically are loaded near the maximum legal limit for an eight-axle vehicle in the state of Washington. The relatively low truck count segment between Kettle Falls and Colville also carries a high proportion of heavier trucks, most of which originate in Canada. For road segments south of Colville, the median truck weight falls in the range of 20.9 and 21.8 metric tons (46,000 and 48,000 lbs., respectively) (See Figure 7).

CONCLUSION

Statewide, metropolitan, and regional transportation planning increasingly requires systems that can manage and analyze large volumes of data pertaining to freight and goods movements. GIS-T can play an important role in this process. A graphical presentation using GIS-T illustrates research findings in a form that is often more easily understood and communicated than the alternative of tabular data. GIS-T enables a direct graphical interface with information from other planning data bases, thus providing an environment for analytical investigations.

The statewide freight truck origin and destination study initiated by the WSDOT in April 1993 will provide substantial benefits for future transportation planning. The study will help the state of Washington comply with ISTEA planning requirements, and it will contribute to the Washington State Transportation Policy Plan and Statewide Transportation System Plan. The study also will be used by MPOs and regional transportation planning organizations to evaluate freight and goods mobility needs for their updated plans.

ACKNOWLEDGMENTS

Funding provided by the Washington State Department of Transportation, Washington State University Eastern Washington Intermodal Transportation Project, and the Intermodal Surface Transportation Efficiency Act of 1991.

Publication of this paper sponsored by Committee on Transportation Data and Information Systems.

A Framework for Integrating GIS-T with KBES: A Pavement Management System Example

WAYNE A. SARASUA AND XUDONG JIA

This paper provides a framework for integrating the spatial data manipulation strengths of a geographic information system (GIS) with the interactive problem-solving capabilities of a knowledge-based expert system (KBES) through emulation of the knowledge of human experts. While the integration of a GIS with KBES has practical applications throughout the transportation profession, this study uses an Intermodal Surface Transportation Efficiency Act (ISTEA) pavement management system example to illustrate this integration. The pavement management process involves spatially-indexed information, human expertise, heuristic knowledge, and multiobjective decision making. These characteristics make a pavement management system ideally suited for implementation in an integrated GIS/KBES environment. In this application, the GIS provides spatial data as context to a KBES that makes use of the National Aeronautics and Space Administration's CLIPS rule-based expert system shell. The KBES retrieves information from the GIS as needed to produce an outcome. As the KBES works, the knowledge base is updated for future processes. In this way, the KBES is able to learn from the previous applications of the system. Once processed by the KBES, the results can be passed back to the GIS for further analysis and display.

The integration of knowledge-based expert system (KBES) technology with geographic information systems (GISs) can address some of the difficulties associated with a GIS. Many of the transportation areas where GIS technology has been or will be applied, such as pavement management, involve very dynamic, iterative processes with no "right" answers. Engineering judgment, fiscal realities, and other irreducible factors preclude the development of "black box" solutions. Because of this, the potential for integrating expert systems with GIS is promising. Furthermore, because of high turnover among engineers in state and municipal transportation agencies, expertise can be scarce, compounding the need for integration.

There are a number of areas in which knowledge-based systems could be applied. A GIS could be provided with an intelligent user interface to guide an inexperienced user through the most efficient use of the system. Better database search techniques and querying capabilities could make the search of large geographic databases more efficient by using heuristic search methods, that is, search methods based on judgmental rules, which eliminate major portions of the database from consideration as early as possible. Learning capability could allow results of computationally expensive queries to be added to the knowledge base to process frequent queries faster. Finally, intelligent graphical output capabilities could produce high-quality maps and graphs. While the above areas of integration are general and can be applied to almost any type of transportation GIS application, this study illustrates this integration through an Intermodal Surface Transportation Efficiency Act (ISTEA) pavement management system example that the Georgia Department of Transportation (GDOT) is implementing.

School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Ga. 30332-0355.

PREVIOUS WORK IN GIS/KBES INTEGRATION

Very little research has been done on the integration between KBES and a vector-based GIS. Evans (*1*) describes an expert GIS that combines rule-based reasoning with vector-based spatial data representation and analysis, but the implementation is experimental. The Australian Army is using a vector GIS to display results from a soil moisture-soil strength model (*2*). While this is a production application of both KBES and GIS technology, it is not an integrated application.

There have been a number of attempts at integrating KBES with a raster-based GIS. Researchers at the University of California at Santa Barbara carried out research and development on a knowledge-based geographic information system (KBGIS) (*3*). The objective of the KBGIS system was to respond intelligently to user queries on large spatial databases stored in a raster GIS. Several other research efforts focused on using a raster GIS in conjunction with a KBES for digital image processing (*4–6*).

The most beneficial GIS/KBES platform from a transportation standpoint is vector based, because the vast majority of existing transportation-related GIS applications are in this format. The primary reasons for this are that a vector GIS can precisely model transportation facilities, such as roads, and attribute linkages are easily accomplished. Only preliminary efforts have been made in using KBES in conjunction with a transportation-related GIS. One area that has had limited success is pavement management. The state of Wisconsin has hard-coded decision trees into its GIS to help prioritize and recommend pavement rehabilitation activities, but this effort does not make use of a KBES (*7*). Researchers at the University of Ljubljana in Slovenia are working on an expert system based on GIS technology (*8*). The main goal of this system is the optimization of road maintenance from both technical and managerial standpoints.

POTENTIAL ROLES OF A GIS/KBES IN THE GDOT ISTEA PAVEMENT MANAGEMENT SYSTEM

In 1991, Congress passed ISTEA, which requires state transportation agencies to develop and maintain six management systems (*9*). One of these systems deals with pavement management. A primary goal of GDOT's ISTEA pavement management system (PMS) is to

serve as a decision support tool that will help GDOT engineers and decision makers to determine when and where to spend pavement funds to enhance safety, preserve existing infrastructure, and serve commerce and the motoring public.

The specific objectives of the PMS that need to be achieved to meet the primary goal are as follows:

- Maintain a complete and up-to-date road inventory including physical features of the pavement, pavement history, pavement condition, and traffic information such as volume, vehicle classification, and load data;
- Develop and integrate systematic procedures for performing network-level analysis for projecting both short- and long-term pavement conditions across the network;
- Develop and implement prioritization schemes for investment in current and near-term projects;
- Routinely maintain and upgrade all of the components of the PMS; and
- Develop an efficient means by which the PMS interacts with other ISTEA management systems.

These objectives need to be achieved for all Federal-aid highways on the National Highway System and all roads that are maintained by GDOT.

The following sections explain how a GIS integrated with KBES technology may be incorporated into an ISTEA PMS for most aspects of the pavement management process. For each aspect we begin by describing the tasks involved in maintaining a stand-alone PMS without a GIS element, and then briefly discuss the effects of adding a GIS element. Further discussion is provided on how the addition of KBES technology may enhance completion of the task.

Data Collection and Maintenance

The first objective of the PMS is to maintain a complete and up-to-date road inventory. One type of data that needs to be collected is subjective pavement ratings that represent the condition of roadway segments. In a conventional PMS, subjective pavement ratings are coded into a database on a segment-by-segment basis and the results are displayed or printed in tabular form. By adding GIS technology with its visual display capabilities, segments could be color-coded by various attributes, which would greatly facilitate the process of data entry and editing. Omissions in the input of data would be immediately apparent from segments in the roadway showing no data. Errors in measurement or coding would also be readily apparent. Adding an intelligent user interface to the data entry process, based on KBES technology, could help to guide the user through the input process. Pavement ratings are entirely subjective and can vary greatly for a pavement depending on the judgment of the person rating the pavement. A KBES that guides the person making the rating can help to reduce the variability of these ratings. Of course this requires that the person collecting pavement ratings has a laptop computer in the field to enter the data and respond to questions of the system. Note that once the data are entered into the laptop, a simple program can be developed to transfer this data directly into the GIS database.

Preliminary Analysis and Interpretation

In a traditional PMS, the highway engineer transfers some of the tabular information to a base map by hand as a first step in understanding the data. For example, the engineer might construct a map showing the severity of rutting or block cracking, or create a map indicating the overall performance index. As in the situation for data editing, the GIS-PMS can integrate the database attributes describing the pavement condition with a map display of the road network; it can then create any number of illustrative visual displays of the status of the road system. For example, it would be possible to highlight all segments with block cracking greater than Condition 6.

Adding a knowledge-based system can help to standardize and automate the process of identifying problem areas. Consider a section of rutted pavement. A possible rule in the knowledge-based system may be as follows:

IF
Rutting is moderate and pavement age is less than 5 years old.
THEN
Mix is too soft.

This rule illustrates how an integrated KBES could be used to identify problem areas. The KBES could be taken one step further, and also recommend strategies to remedy the problem. Additional information pertaining to this will be presented later.

Performing Network-Level Analysis

While visual representations of the segment-by-segment status of the roadway are a valuable addition to the pavement management process, it is necessary to add analytical capabilities to assess the current status of the system as a whole, compare it with previous periods, and make predictions about the future. Doing this requires statistical and mathematical procedures, which a conventional database management system could be programmed to perform. A GIS would add the benefit of spatial querying to selectively isolate geographic regions for more detailed assessment. A network-level assessment can be reported in the form of graphical products that provide greater visual impact than tabular reports. These graphical products can be easily understood by management, politicians, and citizens' groups, helping to clarify issues and obtain needed support.

Incorporating a KBES into this task has many possibilities, since there is a great deal of subjectivity involved with network-level assessment. A KBES could be used to help interpret the output and guide the user through the process of using this output to develop the assessment. An additional possible KBES domain is automated map design, in which the system relies on heuristics to aid in the positioning of labels of spatial features.

Determination of Prioritization Strategies for Allocating Resources

Another use of the GIS-PMS is to develop and implement prioritization strategies for investment in current and near-term projects. This is where a KBES could again come into play. The determination of strategies could be based on a series of decision rules that match deficiency ratings with appropriate actions. Using these rules, a KBES could be used to identify a list of potential projects. Furthermore, a KBES could also be used to prioritize the allocation of resources to these projects.

A FRAMEWORK FOR GIS/KBES INTEGRATION

In the previous sections we have described a number of possible specific KBES applications in a GIS-PMS. The next issue is how a

KBES could be integrated with a GIS-PMS or even a standard GIS. Ideally, a KBES built into the core of the GIS would be most efficient, because the KBES would have direct access to the GIS database. Unfortunately, because of the proprietary nature of commercially available GIS products, this approach is not feasible. The intent in this study is to identify a framework that can be implemented on a wide variety of GIS platforms. The next question is how to interface with an existing GIS. This is a software question that depends on the type of GIS being used. Different GIS programs have varying capabilities for expansion. For example, ESRI's ARC/INFO product includes a large set of generic GIS tools and has extremely powerful customization capabilities that make it an attractive candidate for GIS/KBES integration.

Figure 1 presents a conceptual framework of how a GIS/KBES could be integrated. This figure shows a linkage between the GIS and the KBES. The user interface resides within the GIS. When the KBES is needed to solve a heuristic problem, the GIS would provide context (facts) to the KBES. The inference mechanism would then process the context using rules stored in the knowledge base. The results of the process can be passed back to the GIS for display. The updated context can be used in future processes which would use the KBES more efficiently. Based on this framework, the following sections present criteria that were used in the design of the pavement GIS/KBES discussed later.

The construction of a GIS/KBES requires a systematic approach to design. Design criteria, including the following points, must be rigorously applied and the rationale for each standard must be explicitly defined for each component of the GIS/KBES.

- The inference engine must be able to address the class of general problems representative of any GIS while retaining a high degree of domain independence for other specific applications.
- The knowledge base must adequately reflect the complexity of highly specialized information while remaining internally consistent and logical.
- Conclusions and explanations derived by the rules of the knowledge base must be reproducible and supportable.

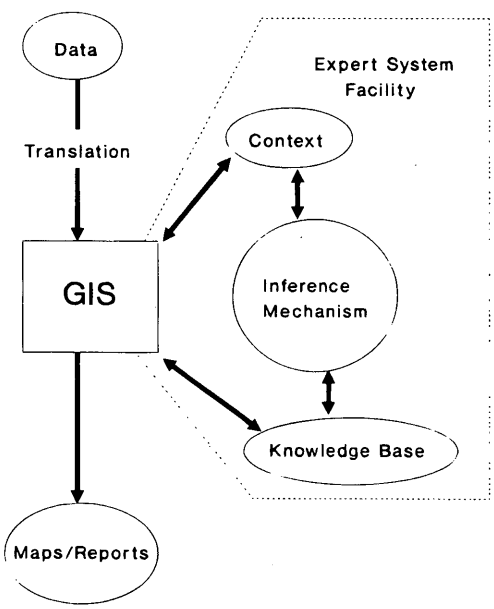


FIGURE 1 GIS/KBES conceptual framework.

- The link to the GIS and other databases must retrieve the exact information necessary to address an inquiry.
- The output must be in formats useful to the end users and decision makers, and sufficiently flexible to accommodate various needs.
- The user environment must be comfortable and encourage productivity while providing adequate power and capability to serve both experienced and novice users.
- The system must be amenable to updating and expansion in an open-ended, incremental fashion as new knowledge critical to the transportation application becomes available.

Standards for ensuring that these basic requirements are met once the system is working should be identified prior to the actual development of the system.

A number of inference algorithms were reviewed for incorporation into the GIS/KBES. These include forward chaining, backward chaining, and hybrid forward chaining with local backward-chaining inferencing. Forward chaining was chosen for this implementation because a wide range of possible scenarios can be explored in an efficient manner starting from the basic data. Thus, the system is free to draw any reasonable conclusion from the data, rather than seek out a particular conclusion or diagnosis, which is not as efficient (especially if it goes down the wrong path). Furthermore, conventional pavement management activities most closely resembles a forward-chaining decision process.

The performance of an expert system is most closely related to the content of the knowledge base. Thus, it is important that knowledge is stored in an internally consistent and logical manner.

The format of the GIS data is also important to system performance; however, like the data, it may be out of the hands of the system designer or knowledge engineer. It is important that the interaction between the GIS and the KBES be as seamless as possible and transparent to the user. The switching of different user interfaces is not very efficient. Because of the high-level rule structure of the GIS/KBES, it is preferable that GIS data be accessible to the KBES via fairly high-level calls at an operating system level. GIS query should comprise a functional description of the mapped data (e.g., asphalt overlay, no shoulder) rather than a structural description of the GIS organization (e.g., Columns 5–7 and 12). Otherwise, database information must be built into the KBES and any GIS update, change, or expansion will require a major effort.

DESIGN OF A PROTOTYPE ISTEA PAVEMENT MANAGEMENT SYSTEM GIS/KBES

The system architecture of the prototype ISTEA pavement management system GIS/KBES that evolved using the design criteria presented in the previous section is shown in Figure 2. The prototype GIS/KBES-based PMS illustrated here is designed to identify and interpret pavement distresses, evaluate the current problems of the pavements, and generate treatment or maintenance plans at the project or network level. It can be expanded to include other KBES domains such as map design, and intelligent querying to use the GIS database more efficiently. Figure 2 identifies several components that are divided into four areas. The areas are (*a*) external data, (*b*) the GIS subsystem, (*c*) the KBES subsystem, and (*d*) reporting of results. The KBES interacts with the GIS through the GIS user interface.

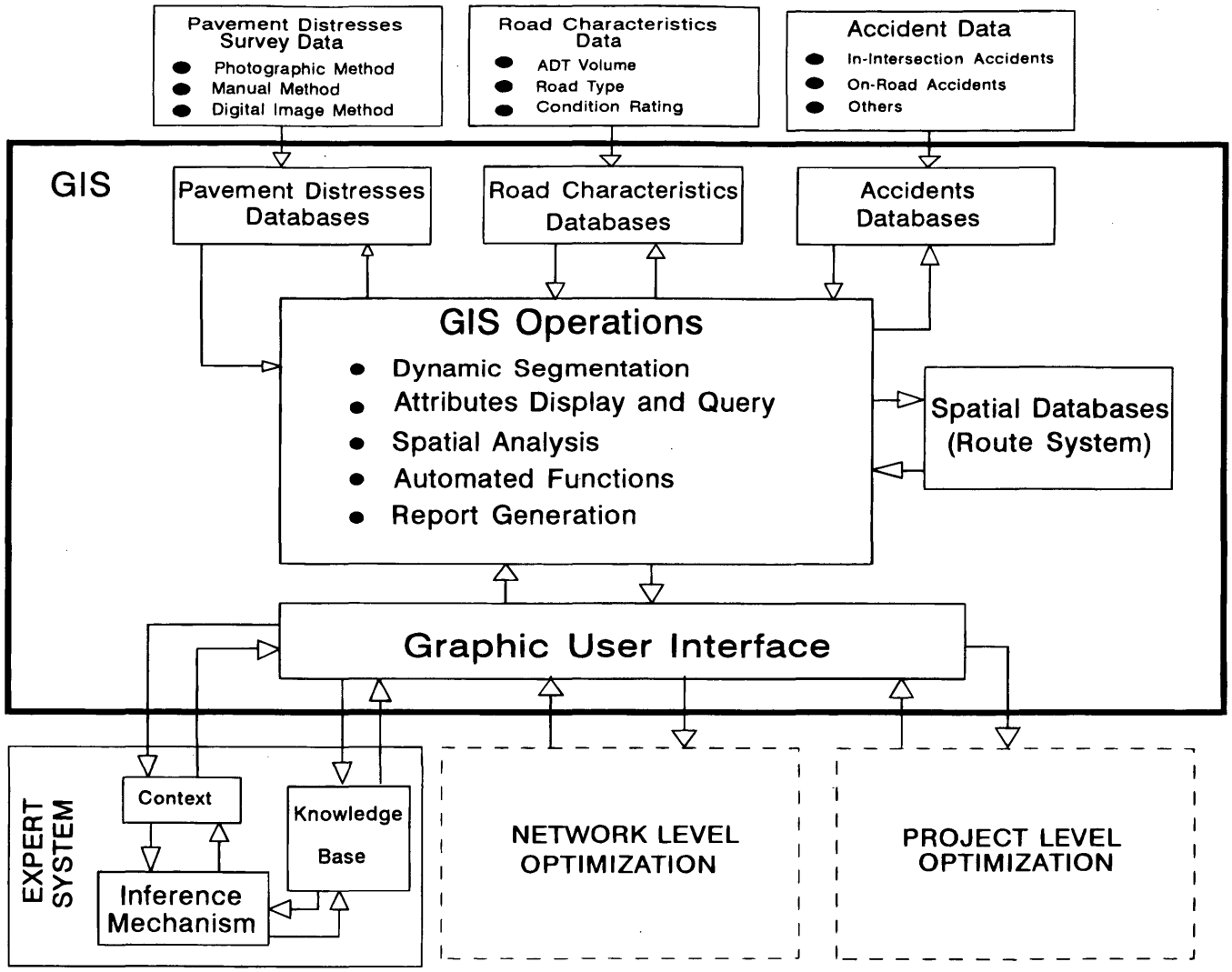


FIGURE 2 Architecture of a GIS/KBES PMS.

External Data

The external data shown in Figure 2 include pavement distress data, road characteristics data, and accident data. A series of translation programs were developed for importing these data into GIS attribute databases. The database of pavement distress attributes includes the severities and intensities of load cracking, block cracking, rutting, ravelling, and reflective cracking which were collected through biannual distress field surveys. Table 1 shows the attributes included in the pavement distress database. The road characteristics database contains average daily traffic volumes, skid factors, road types, and an overall pavement rating. The accident database currently contains only on-road accidents. It will be expanded to include accidents at intersections and the attributes associated with them. These three databases can be linked individually or in combination to the GIS route system that will be discussed in the following section. This linkage is crucial to the identification and interpretation of pavement problems. With this linkage, the pavement distresses can be graphically displayed or spatially queried on a segment-by-segment basis.

GIS Subsystem

The GIS subsystem of the prototype GIS/KBES-based PMS provides the user interface for all activities of the integrated system and is responsible for GIS-related activities such as spatial and attribute data storage, display, and analysis.

The system uses ESRI's ARC/INFO software as the GIS platform operating on a Sun SPARC workstation. The system can be ported to other hardware environments that support ARC/INFO with minimal effort. There were two reasons for the selection of ARC/INFO. First, ARC/INFO has been adopted by GDOT as its primary GIS package. Second, ARC/INFO has capabilities that are vital for GIS/KBES implementation, including the following:

- It can be customized to automate the various GIS processes that the system will use, and can access external programs such as an external KBES shell.
- Its dynamic segmentation capabilities provide the fundamental tools for linking attributes such as pavement distresses to corresponding routes or segments using a mile-point linear referencing system.

TABLE 1 Attributes of the Pavement Distress Database

Field name	Field Type	Field Width	# of Dec	Remarks
Trip_ID	Integer	8		Trip ID of Pavement Distress Survey
Trip_Year	Integer	12		Trip Year of Pavement Distress Survey
Survey_No	Integer	11		Pavement Distress Survey Number
Route_Id	Character	10		Route ID
Route_Number	Integer	4		Route Number
Route_Suffix	Character	2		Route Suffix
County_FIPS	Integer	3		County FIPS
MP_From	Float	5	2	Beginning (or From) Mile Post of a Segment
MP_To	Float	5	2	Ending (or To) Mile Post a Segment
Rutting_Outside	Integer	12		Rut Depth of Outside Wheelpath
Rutting_Inside	Integer	12		Rut Depth of Inside Wheelpath
Percent_LC1	Integer	12		% of Load Cracks (Severity Level 1)
Precent_LC2	Integer	12		% of Load Cracks (Severity Level 2)
Percent_LC3	Integer	12		% of Load Cracks (Severity Level 3)
Percent_LC4	Integer	12		% of Load Cracks (Severity Level 4)
Percent_BC	Integer	12		% of Block Cracking Occurrences
Severity_BC	Integer	12		Severity Level of Block Cracks
Number_RC	Integer	12		Number of Reflective Cracks
Length_RC	Integer	12		Length of Reflective Cracks
Severity_RC	Integer	12		Severity Level of Reflective Cracks
Percent_Ravel	Integer	12		Percent of the Length of Raveling
Severity_Ravel	Integer	12		Severity Level of Raveling
Percent_ED	Integer	12		Percent of the Length of Edge Distress
Severity_ED	Integer	12		Severity Level of Edge Distress
Percent_Bleed	Integer	12		Percent of the Length of Bleeding
Severity_Bleed	Integer	12		Severity Level of Bleeding
Percent_Corrug	Integer	12		Percent of the Length of Corrugations
Severity_Corrug	Integer	12		Severity Level of Corrugations
Total_Patches	Integer	12		Total Number of Patches and Potholes
Percent_Loss_S	Integer	12		Percent of the Length of Section Loss
Severity_Loss_S	Integer	12		Severity Level of Section Loss
Cross_Slope_L	Integer	12		Cross Slope in the Leftside Road
Cross_Slope_R	Integer	12		Cross Slope in the Rightside Road

• It has network overlay capabilities that allow the user to integrate data from various sources (e.g., the pavement condition data obtained from a distress survey can be overlaid with road characteristics and accident data).

GIS Route System

The route system serves as the spatial database of the GIS. It is a base map for which database attributes can be integrated and displayed. The route system used in this demonstration project contains eighteen counties in northwest Georgia. The route system was created using maps that were digitized from rectified low-level aerial photographs.

The prototype GIS/KBES-based PMS was designed largely to be independent of the route system. Thus, expansion of the route system can be integrated into the PMS with little modification. Figure 3 shows the route system in Gilmer County.

GIS Operations

GIS operations include the graphic display and spatial query of route attributes, spatial overlay of these attributes, and the generation of reports. Figure 4 shows the user interface for creating thematic maps with the system. It includes the main menu on the left, which activates and controls the functions that implement the GIS operations. The main menu has four scroll lists that allows the user to (*a*) display and analyze road characteristics, (*b*) display and analyze pavement distress, (*c*) perform expert system analysis at both the project and network levels, (*d*) create a prioritized list of projects that includes identification of maintenance and rehabilitation treatments and the associated costs.

The road characteristics and pavement survey scroll lists include attribute items associated with the route system. One example of these attributes, or events, is the pavement condition evaluation system (PACES) rating, a subjective pavement distress rating scheme developed and used by the Georgia Department of Transportation. Once an event such as PACES rating is double-clicked or selected, a menu for the event is activated (top of Figure 4). The menu has a set of input fields, a spatial query builder, and a route-based querying and display facility.

The input fields require the user to specify the district, county, and route system to be analyzed. A help function is built into the menu. If a user misspells the county name, the help function can list all of the county names within the specified district for the user to choose from.

The query builder is a tool for creating conditional queries interactively. It provides a way of understanding the pavement problems at the network level. Once a user determines the query condition, the PMS will generate a thematic map based on the selected condition. Figure 4 shows an example of the use of the conditional query builder. In this figure, the user identified the query condition as "PACES rating < 65." The system will produce a thematic map that presents this information. One aspect of the system that has not yet been implemented is intelligent querying. In this scenario, information input into the query builder would be put into context. The KBES would use this information plus other information already in context to create a query that produces the desired map. This would have the benefit of querying the database more efficiently. Thus, it may be possible to eliminate large portions of the spatial database from consideration before the query is performed.

The route-based query and display facility provides a way of understanding pavement performance at the route level. It has two methods for choosing a route on which the PACES rating event can be displayed. These two alternatives are to select a route (*a*) by specifying its route number, or (*b*) by clicking on the route. Once a route is selected, the PACES rating event can be spatially linked and displayed on the route by the ARC/INFO dynamic segmentation functions. Figure 5 illustrates the route-based query and display

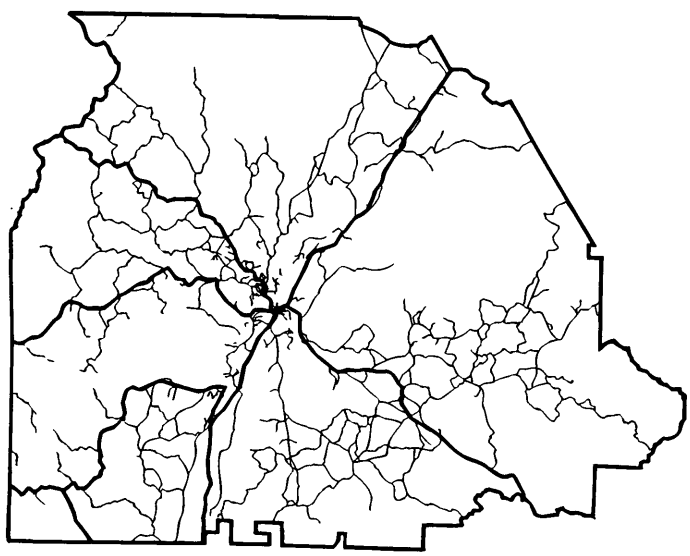


FIGURE 3 The route system of Gilmer County, 1992.

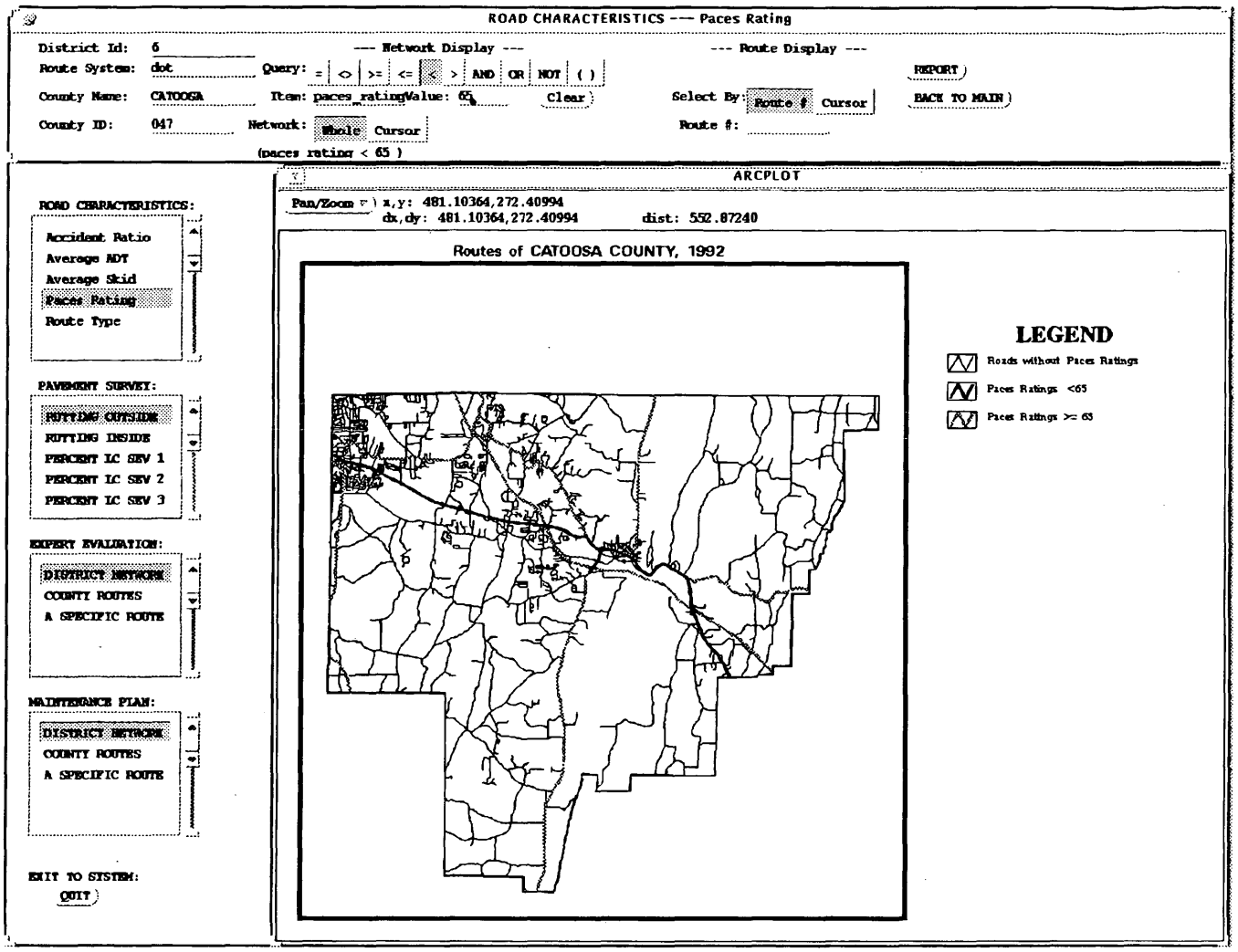


FIGURE 4 Sample thematic map creation.

facility. In this figure, the selected route is isolated in the right box for clarification. A strip map of the route is displayed with the PACES rating for each segment, and a window is displayed that includes a list of the pavement segments and their corresponding PACES ratings in tabular form. The system includes heuristics to display PACES information in a readable manner. This capability is useful when a route has too many segments to be displayed.

Knowledge-Based Expert System Design

The KBES subsystem uses NASA's CLIPS ruled-based expert system shell for the evaluation of pavement distresses and the determination of segment treatments. The CLIPS shell, written in C, can be embedded into other systems. The embedded feature allows for integration of the KBES with the GIS at the operating system level.

The KBES subsystem consists of three main components: the knowledge base, the inference mechanism, and context. The knowledge base contains control knowledge and domain knowledge. The inference mechanism uses the control knowledge to dictate how the domain knowledge is used. The knowledge is represented as a set of rules, which were developed through interviews with GDOT engineers and the FHWA district pavement engineer. Figure 6 shows an example of an inference network, which illustrates the process by which the inference mechanism uses control knowledge and domain knowledge to infer the solution of a pavement problem, similar to the way an experienced pavement engineer does. The context is developed automatically using inputs from the GIS including the event data and their associated route or network spatial information.

The interface between the GIS and the KBES is through a set of programs written in ARC/INFO's ARC macro language (AML) and C. The interface is activated by the pavement evaluation list, which has three levels of pavement evaluation: the route level, the county level, and the district level. The route level concentrates on an individual route, evaluating it through the KBES. The county and district levels consider a set of routes selected either by the spatial query builder or by the cursor and evaluate these routes through the KBES.

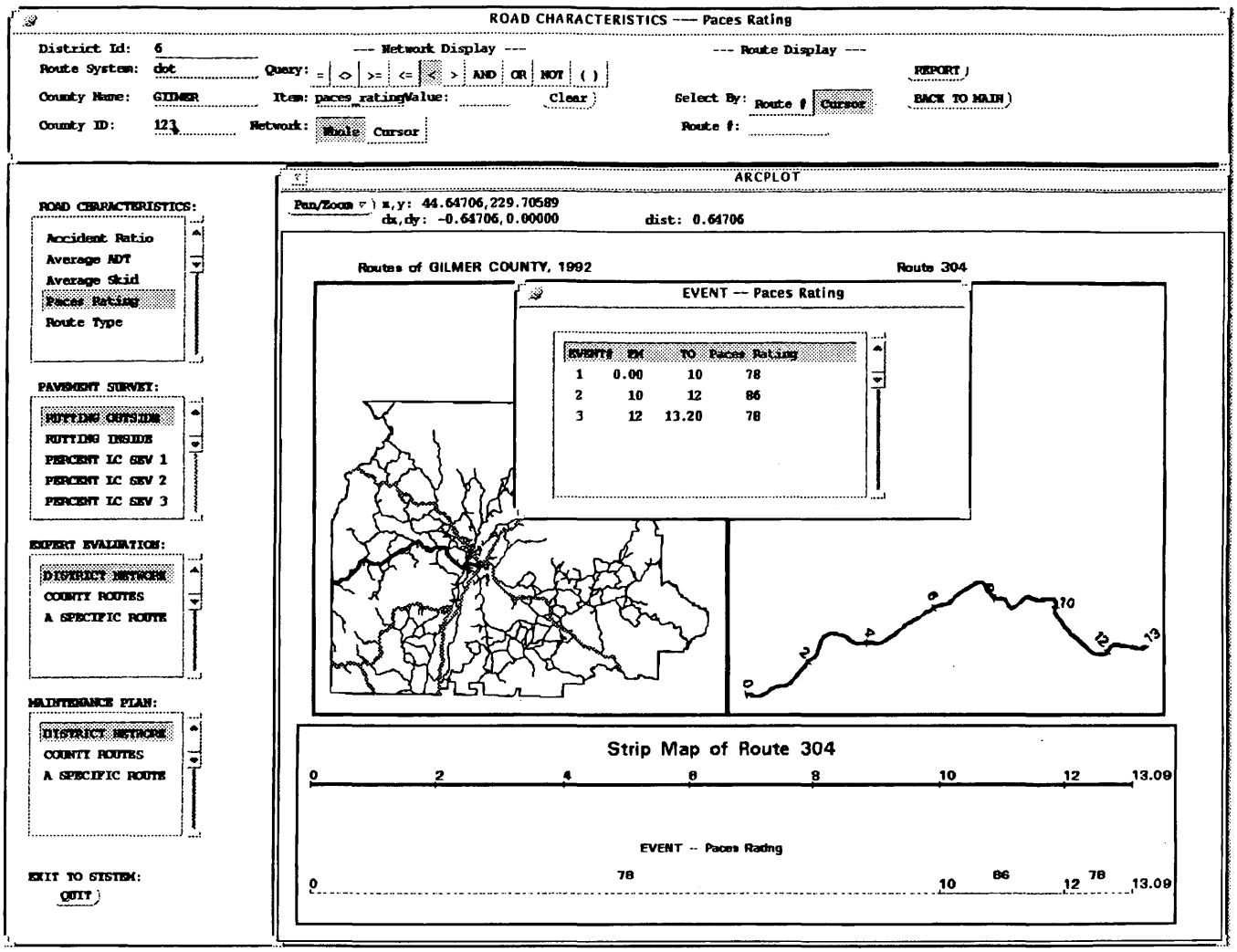


FIGURE 5 Sample route-based query.

GIS/KBES Integration

The process of integrating the KBES with the GIS-based PMS at the route level is shown in Figure 7. The figure illustrates three steps: network overlay of pavement distresses and road characteristics, the passage of control and information to and from the user interface, and the spatial display and query of segment treatment and the generation of the treatment report.

Network Overlay

Before using the KBES, the various attributes to be used in the analysis need to be overlaid. The overlay process, activated when a user selects a route with the cursor or the route number, can combine the pavement distress and other road characteristics to produce a new event database. Figure 7 illustrates the process of an event overlay. In the figure, the GIS overlays ADT with load cracking to create the new event database, which is then passed to the KBES.

Passage of Control and Information to and from the GIS

The GIS user interface calls the KBES at the operating system level using the Expert-Route button or Expert-Section button shown in Figure 8. As the KBES works, spatial and attribute data are passed from the GIS as needed to complete the process.

The Expert-Route button passes the overlaid event database for a specified route to the KBES, whereas the Expert-Section button passes a part of the overlaid event database for an individual segment. The Expert-Section button also opens a window listing the event data for all of the segments of a specified route. The window provides a visual tool for selecting segments to be passed to the KBES for analysis.

As the KBES works, system status and the segment treatment decisions are stored. This information allows the user to review how the KBES came up with a particular treatment. Status information, such as rules fired and associated explanations, can also be dynamically displayed in a window.

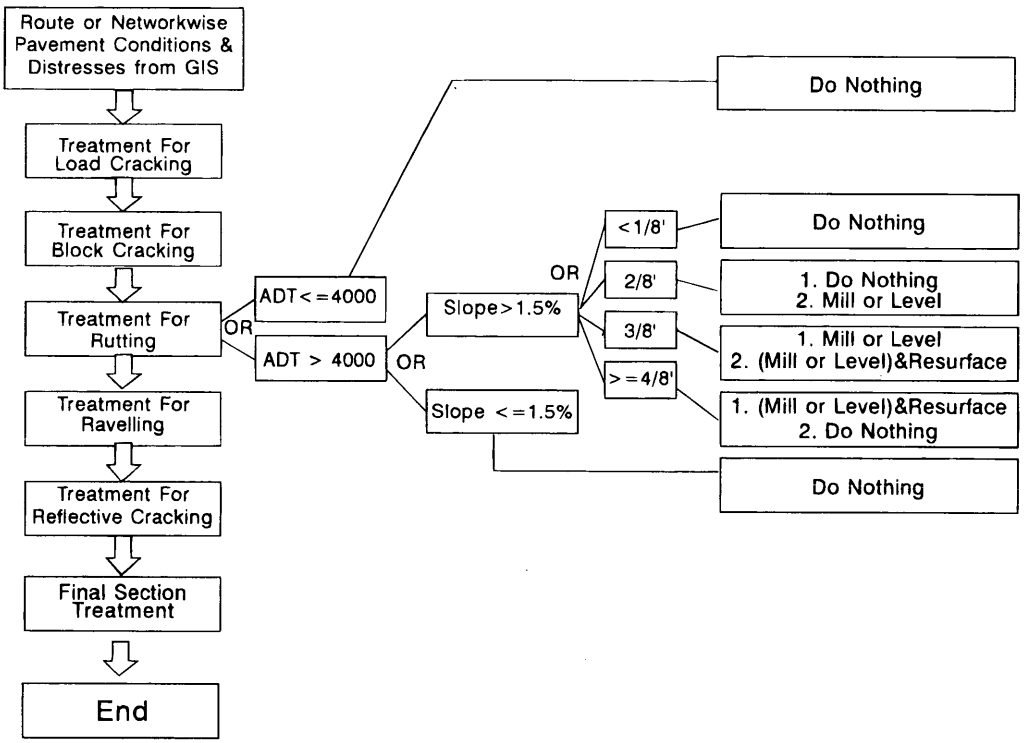


FIGURE 6 Sample inference network (rutting).

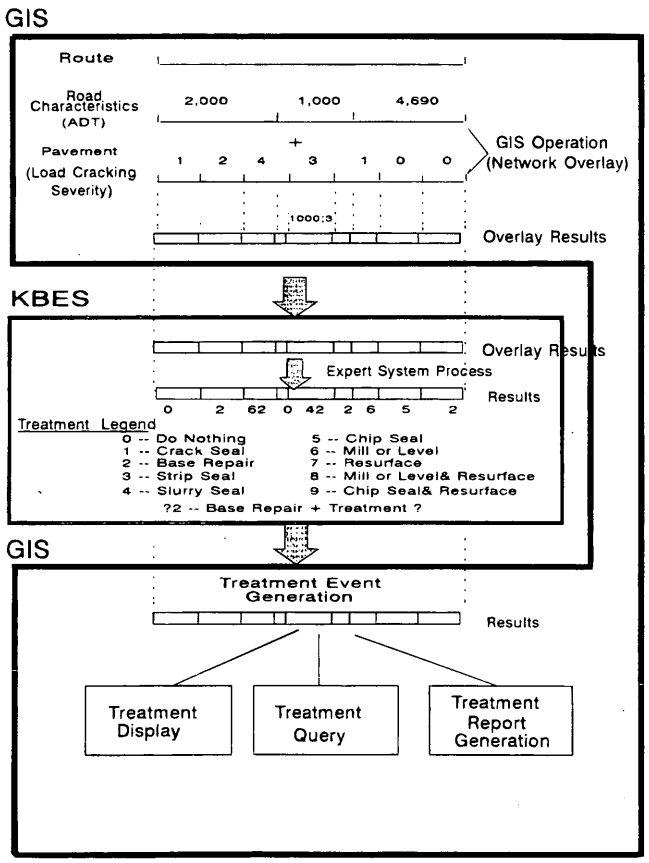


FIGURE 7 Interaction between GIS and KBES.

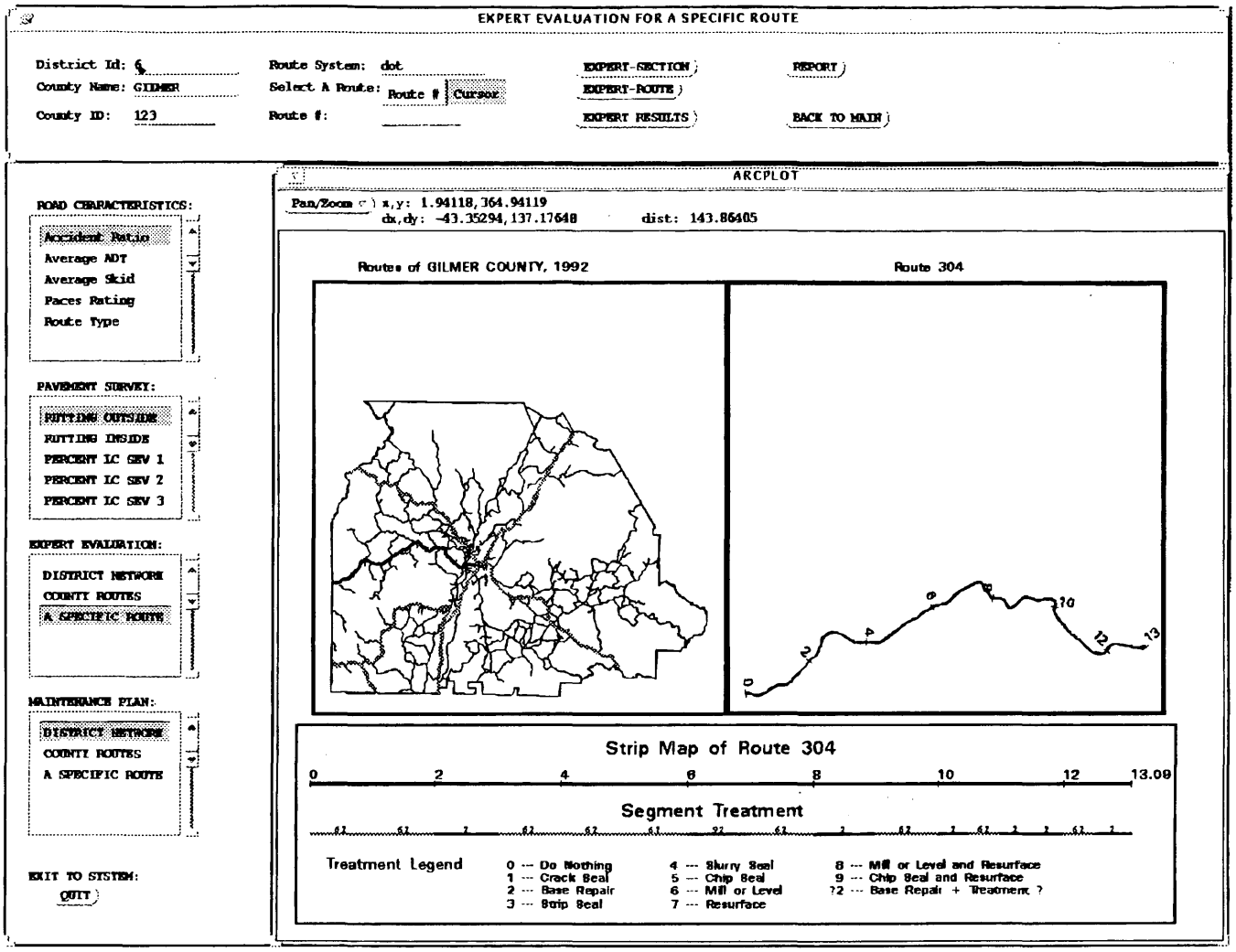


FIGURE 8 GIS/KBES route evaluation.

Segment Treatment Query, Display and Report Generation

After the segment treatment results are passed from the KBES to the GIS, the user interface can be used to spatially query and graphically display the recommended segment treatments, and to generate a treatment report. The Expert Results button invokes a strip map of the specified route and arranges the segment treatments graphically along the route (see Figure 8). The Report button activates the report generation function. Figure 9 shows a sample report.

The segment-based treatment decisions are the key to the determination of maintenance plans at the route level. Given these treatment decisions, their treatment costs, and other constraint factors, projects can be identified and prioritized into an overall maintenance plan.

FUTURE ENHANCEMENTS

The prototype GIS/KBES-based PMS presented in this paper provides users with a list of suggested roadway treatments based on the condition of the road. A future enhancement to the system includes incorporating network-level analysis and performance prediction capabilities, as required by ISTEA. From a research standpoint, there are a number of enhancements that can be made to demonstrate areas where the KBES would be beneficial. These areas include intelligent spatial querying and map design.

Further verification and validation of the prototype system based on criteria presented in this paper will be done once the system is in full implementation in GDOT.

CONCLUSION

The potential for the integration of GISs with KBESs promises to be of great value for the development of a better GIS. KBESs offer possibilities for making GISs more computationally efficient and user-friendly using expert knowledge and high-level reasoning procedures. Specific applications in transportation, such as pavement management, can benefit from KBES technology. Furthermore, many of the transportation areas in which GIS technology has been

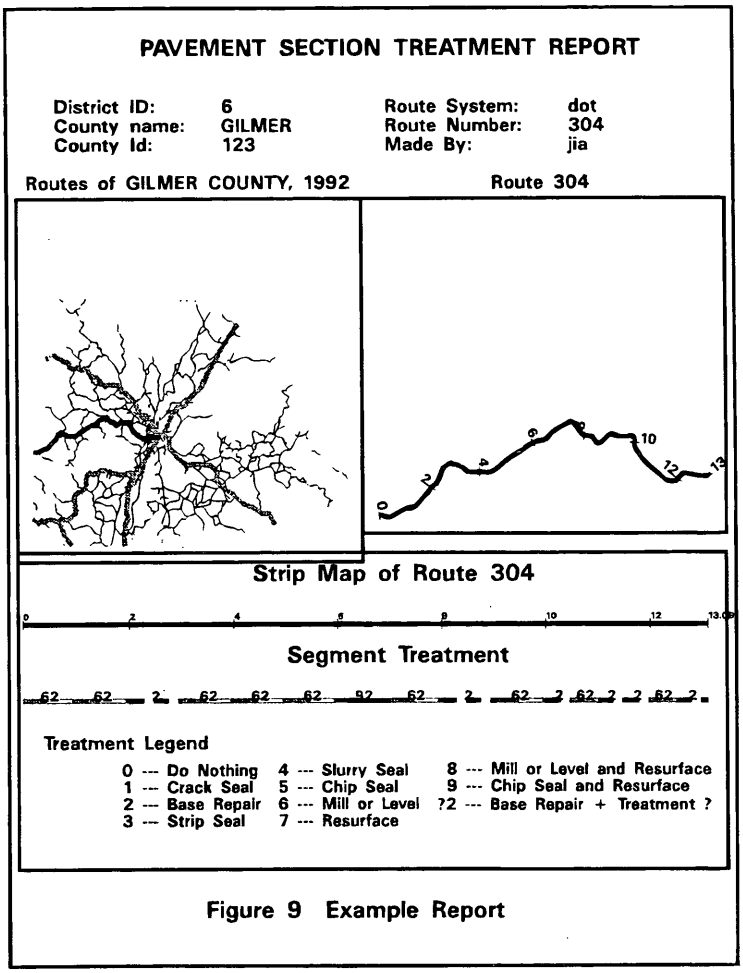


FIGURE 9 Sample report.

or will be applied involve very dynamic, iterative processes with no right answers. Numerous factors preclude the development of black box solutions; therefore, the potential for integrating expert systems with GIS is promising.

ACKNOWLEDGMENTS

The authors wish to thank Sonya Parker of the GDOT and Walter Boyd of the FHWA for their assistance on this project.

REFERENCES

1. Evans, T. A. Development and Application of Expert Geographic Information System. *Journal of Computing in Civil Engineering,* July 1993, pp. 339–353.
2. Davis, R., S. Cuddy, P. Laut, J. Guoodspeed, P. Whigham. Integrated GIS and Model for Assisting the Managers of an Army Training Area. *Proc., Symposium on Watershed Planning and Analysis in Action,* Durango, Colo., ASCE, Boston Society of Civil Engineers, Boston, Mass., 1990. p. 211–220.
3. Pazner, M. I. Geographic Knowledge Base Design and Implementation. Dissertation. University of California, Santa Barbara, March 1986.
4. Usery, E. L. Knowledge-Based GIS Techniques Applied to Geological Engineering. *Photogrammetric Engineering and Remote Sensing,* Nov. 1988, pp. 1623–1628.
5. Van Cleynenbreugel, J. Delineating Road Structures on Satellite Imagery by a GIS-Guided Technique. *Photogrammetric Engineering and Remote Sensing,* June 1990, pp. 893–898.
6. Johnson, K. J. Kanonier. Knowledge-Based Land-Use Classification. *International Geoscience and Remote Sensing Symposium Digest,* Vol. 3, 1991, pp. 1847–1850.
7. Wisconsin Department of Transportation. *Pavement Management Decision Support Using A Geographic Information System.* Report FHWA-DP-90-085-006. Federal Highway Administration, Washington, D.C., May 1990.
8. Kastelic, T., M. Zura, and D. Fajfar. Expert System for Road Evaluation and Maintenance: RESCEM. University of Ljubjana, Slovenia, undated.
9. Interim Final Rule. *Federal Register.* Department of Transportation, FHWA, Dec. 1, 1993.

Publication of this paper sponsored by Committee on Artificial Intelligence.