Available online at www.sciencedirect.com
ScienceDirect
Journal of Hazardous Materials 147 (2007) 60–66
Journal of Hazardous Materials
www.elsevier.com/locate/jhazmat

Machine learning techniques applied to the determination of road suitability for the transportation of dangerous substances

J.M. Matías [a], J. Taboada [b], C. Ordóñez [b,*], P.G. Nieto [c]

[a] *Statistics Department, Vigo University, Vigo, Spain*
[b] *Natural Resources and Environmental Engineering Department, Vigo University, Vigo, Spain*
[c] *Applied Mathematics Department, Oviedo University, Oviedo, Spain*

Received 25 May 2006; received in revised form 14 December 2006; accepted 15 December 2006
Available online 21 December 2006

Abstract

This article describes a methodology to model the degree of remedial action required to make short stretches of a roadway suitable for dangerous goods transport (DGT), particularly pollutant substances, using different variables associated with the characteristics of each segment. Thirty-one factors determining the impact of an accident on a particular stretch of road were identified and subdivided into two major groups: accident probability factors and accident severity factors. Given the number of factors determining the state of a particular road segment, the only viable statistical methods for implementing the model were machine learning techniques, such as multilayer perceptron networks (MLPs), classification trees (CARTs) and support vector machines (SVMs). The results produced by these techniques on a test sample were more favourable than those produced by traditional discriminant analysis, irrespective of whether dimensionality reduction techniques were applied. The best results were obtained using SVMs specifically adapted to ordinal data. This technique takes advantage of the ordinal information contained in the data without penalising the computational load. Furthermore, the technique permits the estimation of the utility function that is latent in expert knowledge.

Keywords: Pollutant substances; Ordinal data; Machine learning; Support vector machines; Transportation

1. Introduction

The safety and efficiency of road transport is considered a strategic objective in countries like Spain, in which the proportion of goods transported by road is about 92%; 9% of road-transported goods, moreover, are classified as dangerous goods. The transportation of dangerous goods by road implies a risk for both humans and the environment, in that an accident may cause extensive material damage and may even endanger lives. For this reason, there is a growing interest among both public and private entities (e.g. insurance companies) in studies to assess the risks associated with dangerous goods transportation (DGT). Authors such as Glickman and Erkut [1] and Cassini [2] determined risk in terms of traffic volume and population density implied by a road accident involving the release of a dangerous substance. Other authors, such as Erkut and Verter [3], Lovett et al. [4] and Fabiano et al. [5] took a different approach, and endeavoured to reduce risk by selecting alternative, lower-risk routes. Huang et al. [6] integrated GIS and genetic algorithms to evaluate the risk of hazardous materials transportation and to plan safer alternative routes. Purdy [7] analysed the risk of transporting hazardous materials by road or rail in Great Britain and concluded that the inclusion of motorist and rail passenger populations significantly affected the calculated risk levels and that the safe routing of materials with large hazard ranges may be more easily achieved by road.

Of particular interest are reports published by the US Department of Transportation, such as their Guidelines for Applying Criteria to Designate Routes for Transporting Hazardous Materials (US Department of Transportation) [8], which provide an interesting overview of advances in terms of the assessment of the level of risk associated with dangerous goods transport. More recently, Martínez-Alegría et al. [9] proposed a conceptual model for identifying the stretches of roads within a network with the greatest accident risk. These authors took into account factors such as probability of occurrence, accident type and the product transported, as also the vulnerability of the

* Corresponding author at: ETSI de Minas de Vigo, Campus Lagoas-Marcosende, 36310 Vigo, Spain. Tel.: +34 986 814052; fax: +34 986 811924.
E-mail address: cgalan@uvigo.es (C. Ordóñez).

0304-3894/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2006.12.042

environmental and population elements exposed to each kind of hazardous substance transported. In many cases the stretches of road were over 100 km long since, given the information used to calculate accident probability, a more detailed analysis was not possible. This study can be viewed as a continuation of the study by Martinez-Alegría et al. [9]. Its aim is not to establish the level of risk for a road, which has already been studied, but rather to determine in detail – based on an analysis of 31 factors relating to the road and its physical surroundings – whether or not a stretch of road is suitable for transporting pollutant substances (data for the variables were collected at 100-m intervals).

The layout of the document is as follows: firstly, we present our model for evaluating the risk associated with an accident involving the transportation of pollutant substances. Next, we describe ordinal support vector machines (SVMs), a variation on the SVMs [10] obtained by considering a different loss function that penalises erroneous orderings. We then apply the ordinal SVMs to the problem of estimating our risk model, and compare the results to those obtained using other statistical classification techniques. Finally, we draw our conclusions on the work described.

2. Definition of the model

2.1. Impact factors

The initial model of the risk associated with an accident involving dangerous goods transportation along a particular stretch of roadway was constructed by combining elements of the Martínez-Alegría et al. [9] conceptual model with specific factors that, in the opinion of experts, affect risk on particular stretches of roadway.

Thirty-one impact factors were identified and subsequently subdivided into two main groups, namely accident probability factors and accident severity factors, discussed in turn below.

2.1.1. Accident probability factors

A total of 21 factors were considered as affecting the probability of the occurrence of an accident. These factors, which reflect the specific features of a stretch of roadway, are classified in six groups, as follows:

(a) *Design*: Road width, lane width, existence of slow lanes, types of feeder roads, protective barriers, and quality of drainage ditches and culverts. Taken as representing the lowest level of risk was a value of 7 m for road width, and a value of 1.5 m for lane width (values based on the Spanish road network). Lower values represent an increased risk, which results primarily from a reduction in the distance between vehicles and the reduced possibilities of avoiding an accident. The fact that a road has a slow lane means that faster vehicles are not necessarily affected by slower vehicles ahead (a major cause of accidents involving heavy vehicles). Feeder road types were graded in terms of a range of values, with the lowest risk associated with acceleration lanes, and the highest risk associated with direct intersections between feeder and main roads. Since protective barriers in good condition prevent animals or pedestrians from straying onto a road and causing an accident, higher values were assigned for adequate protective barriers in good condition, and lower values were assigned to protective barriers in poor condition and/or barriers that failed to fulfil their function; the lowest value was assigned when no protective barriers existed. The existence of suitably sized culverts and ditches determines the rainwater drainage capacity of a stretch of roadway, and prevents films of water accumulating on the road surface, with the resulting aquaplaning risk implied by loss of adherence.

(b) *Construction morphology*: Road condition, slope, altitude, exposure to sun and exposure to winds. Compared to wider curves, tighter curves are more likely to cause a loss of grip by a vehicle's tyres as a consequence of centrifugal forces. Slope is particularly likely to affect accident rates for vehicles travelling downhill. Long and gradual slopes place greater demands on the braking systems of heavy vehicles, with a greater likelihood of brake failure. Higher altitudes imply a harsher climate, and greater likelihood of ice and snow. Exposure to sun affects the probability of an accident, in that asphalt that does not receive direct sunlight is more likely to remain wet or to develop icy patches (which again affects tyre grip). Our area of study was located in the northern hemisphere, at latitude 40°; shadier areas were located between 330°NW and 30°NE and sunnier areas between 240°SE and 300°SW. Finally, greater exposure to wind also increases accident risk. Stretches of roadway that are exposed to strong side winds, particularly when these stretches alternate with sheltered stretches, are high accident risk areas. The risk associated with such exposed areas is determined by orthogonal orientation to prevailing winds with a west-to-east component. This factor, moreover, is aggravated by construction infrastructures, given that exposure to winds is greater on entry to and exit from bridges, viaducts and tunnels.

(c) *Signalling and signposting*: Painted road signs and lateral signs and signals. The type and condition of signalling and signposting on a road is determined above all by the field of vision in different weather conditions (rain, snow, fog, etc.), by light conditions at twilight, and by reflectivity at night. Elements that considerably reduce the risk of accidents include the existence of overhead neon-lit panels containing frequently updated information, painted road signs with good reflectivity, and legible lateral signs and signals. Moreover, road surfaces painted with anti-slip paint will also reduce accident rates. Higher values were assigned to those stretches of road without any signal and the lowest value is assigned to those stretches well signalled and painted.

(d) *Type of road works*: Existence of specific kinds of constructions on a stretch of roadway, such as earthworks, embankments, tunnels, viaducts, retaining walls, etc. The maximum value is applied if traffic flow is affected by roadworks and/or if there is only one-way traffic. This value diminishes to the minimum value (which is assigned when there are no roadworks), as roadwork bearing on traffic

movement is reduced. Medium values are assigned when the roadworks only affect the hard shoulder.

(e) *Visibility threshold*: Described in terms of five numeric intervals (0–100 m, 100–200 m, 200–500 m, 500–1000 m, and over 1000 m). The field of vision has a direct bearing on the possibility of an accident, as it directly affects a driver's reaction time in riskier vehicle manoeuvres.

(f) *Condition of the road*: The condition of the asphalt (drainage capacity, irregularities, defects and potholes) has a bearing on the level of accident risk. If a road surface is well maintained, accident risk from skids, for example, is reduced.

2.1.2. Accident severity factors

Severity factors comprise the inherent damage associated with the physical and chemical characteristics of the pollutant substances being transported, the dangers associated with the type of accident, and vulnerability factors associated with the environment and humans. The first two factors were excluded in this work since they had no bearing on the comparative study that we conducted. For the third factor, we considered 10 factors grouped into the following 3 categories:

(a) *Land use*: A distinction was made between irrigated land and non-irrigated land, between forested land and wasteland. Moreover, the presence of building and residential areas, industrial or mining areas, infrastructures and vulnerable population elements were also taken into account. Maximum values were assigned to cultivated irrigated land, to woods, and to high concentrations of dwellings in the vicinity. Minimum values were assigned to uncultivated non-irrigated land, to wasteland, and when there were no inhabited houses in the vicinity.

(b) *Natural land morphology*: Stability of the slopes or embankments, and natural slope of the land. The stability of excavations and embankments is likely to be endangered by an accident, which might even render a roadway unusable. The natural slope of the terrain in which a road is located has a multiplying effect on underlying risk. A fuel spillage on a steep slope, for example, is likely to spread beyond the road and into other vulnerable elements such as rivers or aquifers.

(c) *Surface and subterranean hydrology*: Lithological units containing aquifers located in the catchment area of spillage are highly vulnerable, given their importance as sources and reserves of water for different uses. The vulnerability of lithological units was evaluated on the basis of permeability and the type of permeability mechanism (intra-granular cracks or fissures), as these characteristics to a large extent determine a contaminated aquifer's capacity for recovery. The vulnerability of water sources was determined principally by their horizontal distance from the road. This value was weighted, moreover, according to the runoff gradient from the centre of the road to the water source.

Our model, consequently, takes the following form:

$$R = f(X_1, \ldots, X_d) \tag{1}$$

where X_i, $i = 1, \ldots, d$ are the 31 factors described above and where R is an ordinal variable that measures the degree of suitability of the road for transporting pollutant substances. The variables X_i, $i = 1, \ldots, 31$ were coded on ordinal scales of 0–10, with 0 representing the lowest level of hazard and 10 the highest level of hazard.

R, in turn, was coded on an ordinal scale of 1–3, which indicated the remedial work to be carried out on the roadway so equip it for pollutant substances transportation (the higher the rating, the less the number of remedial actions required).

Given the dimensions of the problem, the only viable statistical methods for implementing a non-linear risk model are machine learning techniques such as classification trees (CARTs) [11], multilayer perceptrons (MLPs) [12] and support vector machines [13–15]. The estimation of the model (1) can be viewed as a classification problem supervised by an expert. However, the above coding for R introduces a ranking for the different stretches of road in terms of their suitability for transporting pollutant substances; this is ignored by classification techniques by minimising the classification error rate criterion. By definition, under this criterion, two classification rules are equivalent if they result in the same error rate. However, the classification approach fails to take account of any possible violations in the order of the examples.

For this reason, we used SVMs for ordinal data (ordinal SVMs), following the approach developed by Herbrich et al. [16], with a view to taking advantage of the ordinal nature of the information contained in the data, while minimising the computational load. Furthermore, the SVM approach permits the utility function that is latent in expert knowledge to be estimated. The basic concepts of ordinal SVMs are described in the next section.

2.2. Ordinal support vector machines

Assume a sample of independent observations $\{(X_i, Y_i)\}_{i=1}^n$ where $X_i \in \Omega \subset R^d$, $Y_i \in \Theta$ are random variables and where $\Theta = \{r_1, \ldots, r_c\}$ is a set of ordered ranks $r_i > r_j$ if $i > j$, such that $r_c \succ r_{c-1} \succ \cdots \succ r_1$ where $\succ$ is a preference relation with strict order properties (irreflexive, asymmetric and transitive).

Assume likewise, that the ranks r_i assigned by the expert are the result of a latent utility function $U : \Omega \to \mathsf{R}$, in such a way that, given a point $\mathbf{x} \in \Omega$, the expert assigns rank via:

$$y(\mathbf{x}) = r_j \Leftrightarrow U(\mathbf{x}) \in [\theta(r_{j-1}), \theta(r_j)], \tag{2}$$

where $\theta(r_i) \in R$, $j = 1, \ldots, c$ are the values used implicitly by the expert.

Rather than a classical loss function $\ell_{0-1}(y, \hat{y}) = 1_{\{\hat{y} \neq y\}}$ that just penalises classification errors, we define the following loss function [16] with a view to penalising violations in the order produced by an ordering rule $g : \Omega \to \Theta$ with $\hat{y} = g(\mathbf{x})$:

$$\ell_{\text{pref}}(y_i, y_j, \hat{y}_i, \hat{y}_j) = \begin{cases} 1, & \text{if } y_i \prec y_j \text{ and not } \hat{y}_i \prec \hat{y}_j \\ 1, & \text{if } y_j \prec y_i \text{ and not } \hat{y}_j \prec \hat{y}_i \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

In this framework, the problem of determining the best ordering rule for the points in Ω can be viewed as a classification

problem for the space $\Xi \subset \Omega \times \Omega$ containing all the different pairs of points in Ω, with the label $z \in \{-1, +1\}$ defined as (with $i \neq j$):

$$z_{ij} = z(\mathbf{x}_i, \mathbf{x}_j) = \left\{ \begin{array}{ll} +1, & \text{if } U(\mathbf{x}_i) > U(\mathbf{x}_j) \\ -1, & \text{if } U(\mathbf{x}_j) > U(\mathbf{x}_i) \end{array} \right| = \text{sign}(U(\mathbf{x}_i) - U(\mathbf{x}_j)) \quad (4)$$

The sample data are now: $\{(\mathbf{x}_i, \mathbf{x}_j, z_{ij}), i \neq j\}_{i,j=1}^{n}$.

In this context, if the expert's utility function were to apply a linear model $U(\mathbf{x}) = \mathbf{w}_e^T \mathbf{x}$, then using (4):

$$z_{ij} = \text{sign}(\mathbf{w}_e^T \mathbf{x}_i - \mathbf{w}_e^T \mathbf{x}_j) = \text{sign}(\mathbf{w}_e^T (\mathbf{x}_i - \mathbf{x}_j)) \quad (5)$$

If we resolve this classification problem following a soft-margin approach [10,13] and using the maximum margin hyperplane, the problem is formulated as:

$$\min_{\mathbf{w},\xi} \left\{ \frac{1}{2} ||\mathbf{w}||^2 + C \sum_{i \neq j, i, j=1}^{n} \xi_{ij} \right\} \quad (6)$$

$$\mathbf{w}^T (\mathbf{x}_i - \mathbf{x}_j) \geq 1 - \xi_{ij}; \quad i, j = 1, \ldots, n, \quad i \neq j \quad (7)$$

Bearing in mind that the points of the sample are now difference vectors $\mathbf{v}_{ij} = \mathbf{x}_i - \mathbf{x}_j$, $i \neq j$, the solution takes the form:

$$\hat{\mathbf{w}} = \sum_{\text{s.v.}} \alpha_{ij} z_{ij} \mathbf{v}_{ij} = \sum_{\text{s.v.}} \alpha_{ij} z_{ij} (\mathbf{x}_i - \mathbf{x}_j) \quad (8)$$

where the values α_{ij} are obtained from the resolution of the dual problem in (6) and (7).

In the most realistic case of a non-linear utility function, we can use the kernel trick (references cited above; this consists of transforming the data in a space with a higher dimensionality through a transformation $\boldsymbol{\phi} : \Omega \to \Phi$ such that $k(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle$ is a positive definite function and so we consider the linear functions in this space:

$$u(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}). \quad (9)$$

The solution in (8) is converted in this case into:

$$\hat{\mathbf{w}} = \sum_{s.v.} \alpha_{ij} z_{ij} (\boldsymbol{\phi}(\mathbf{x}_i) - \boldsymbol{\phi}(\mathbf{x}_j)), \quad (10)$$

and so the resulting optimum hyperplane is:

$$\begin{aligned} f_{\hat{\mathbf{w}}}(\mathbf{x}, \mathbf{x}') &= \hat{\mathbf{w}}^T (\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{x}')) \\ &= \sum_{\text{s.v.}} \alpha_{ij} z_{ij} (\boldsymbol{\phi}(\mathbf{x}_i) - \boldsymbol{\phi}(\mathbf{x}_j))^T (\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{x}')) \\ &= \sum_{\text{s.v.}} \alpha_{ij} z_{ij} (k(\mathbf{x}_i, \mathbf{x}) - k(\mathbf{x}_i, \mathbf{x}') - k(\mathbf{x}_j, \mathbf{x}) + k(\mathbf{x}_j, \mathbf{x}')) \end{aligned} \quad (11)$$

an expression which means we can avoid determining and calculating the transformation $\boldsymbol{\phi}$.

Consequently, the estimated utility function is:

$$\begin{aligned} \hat{U}(\mathbf{x}) &= \hat{\mathbf{w}}^T \boldsymbol{\phi}(\mathbf{x}) = \sum_{\text{s.v.}} \alpha_{ij} z_{ij} (\boldsymbol{\phi}(\mathbf{x}_i) - \boldsymbol{\phi}(\mathbf{x}_j))^T \boldsymbol{\phi}(\mathbf{x}) \\ &= \sum_{\text{s.v.}} \alpha_{ij} z_{ij} (k(\mathbf{x}_i, \mathbf{x}) - k(\mathbf{x}_j, \mathbf{x})) \end{aligned} \quad (12)$$

Finally, to estimate the frontiers $\theta(r_j)$, $j = 1, \ldots, c$ for the intervals of the utility function that the expert implicitly uses in order to determine the ranks r_j, all we need to do is bear in mind that the pairs $(\mathbf{x}_i, \mathbf{x}_j)$ that verify $\xi_{ij} = 0$ have been classified correctly.

Therefore, if we choose a subset of pairs with ranks differing by just one unit:

$$A(s) = \{(\mathbf{x}_i, \mathbf{x}_j) : \ \xi_{ij} = 0, \ y_i = r_s, \ y_j = r_{s+1}\}, \quad (13)$$

the frontiers can be estimated through the mid-point of the closest points that differ by just one unit in their ranks, in other words:

$$\hat{\theta}(r_s) = \tfrac{1}{2}(U(\mathbf{x}^{(1)}; \mathbf{w}) + U(\mathbf{x}^{(2)}; \mathbf{w})), \quad (14)$$

with:

$$(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \arg \min_{(\mathbf{x}, \mathbf{x}') \in A(s)} \{U(\mathbf{x}'; \mathbf{w}) - U(\mathbf{x}; \mathbf{w})\}. \quad (15)$$

With these frontiers, the prediction of the rank that corresponds to a new point $\mathbf{x}$, is obtained using (2).

The main problem of the approach previously presented resides in its implementation due to the computational complexity of the problem (6) and (7) and in the necessary selection of the model, that in the SVM supposes the selection of parameter C and the corresponding ones to the family chosen for kernel. In order to overcome this serious difficulty and to apply the previous methodology, in this work we propose to estimate the kernel using the covariogram:

$$\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle = k((\mathbf{x}, \mathbf{x}') = \text{Cov}(Y(\mathbf{x}), Y(\mathbf{x}')) = \mathbb{C}(\mathbf{x} - \mathbf{x}') \quad (16)$$

More specifically, an isotropic covariogram was used:

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{C}(\mathbf{x}, \mathbf{x}') = \mathbb{C}(0) - \gamma(\mathbf{x}, \mathbf{x}') \quad (17)$$

where the variogram γ was estimated using the values of the dependent variable. This can be obtained merely by defining the feature vectors as $\boldsymbol{\phi}(\mathbf{x}) = Y(\mathbf{x})$.

Estimating k using the variogram enables the association structure contained in the data to be incorporated in the geometry of the feature space (where the problem is resolved).

3. Case study

With a view to construct the knowledge base represented by the model in (1), 28.6 km of roadway located between the Spanish regions of Castilla-León and Galicia were selected for modelling. This road runs through a mountainous region lying between 495 and 1105 m above sea-level (the deepest part of the valley and the highest part of the mountain pass, respectively). This road, which was constructed over 20 years ago, has many bends.

To study whether this road was suitable for pollutant substances transportation, the factors X_i, $i = 1, \ldots, 31$ (defined in the previous section) were analysed over intervals of 100 m. Also analysed was R, referring to the remedial measures necessary to make the stretch of road suitable for transportation of pollutants. Obtained as a result were 286 items of the form $(X_1, \ldots,$

X_{31}, R). Of these, $n_{train} = 150$ were used for the estimation of the model and $n_{test} = 136$ were used as a test sample to evaluate the behavior of the different techniques.

The evaluations of X and R are based on the knowledge and experience of human experts (a civil engineer and a Spanish Civil Protection Service advanced technician), and so contain an element of subjectivity. Nonetheless, some of the parameters were evaluated in an objective manner, such as, for example, the existence of slow lanes and protective barriers. Such subjectivity can be reduced by making comparisons with other experts, using techniques such as the well-known Delphi method. In practice, however, such a comparison is not feasible for the kind of problem described here (in terms of the time and the human/financial resources required).

Table 1 illustrates one of the data sheets used to record the values for the parameters. It shows the values assigned for five stretches of roadway, for both the parameters and the response R. Table 2, which shows the results obtained using ordinal SVMs, also includes, for comparison purposes, the results obtained using linear discriminant analysis, neural networks, multilayer perceptron, support vector machines for classification (multiclass), and classification trees. The error percentages for the different techniques were, respectively: 25.34% (linear discriminant analysis), 15.06% (CARTs), 14.52% (MLPs), 14.49% (SVMs) and 13.19% (ordinal SVMs).

As can be observed, the machine learning techniques produced much more satisfactory results than linear discriminant analysis (with or without a reduction in dimensionality). The results for the MLPs and SVMs were similar, although with a different error structure resulting from their different configurations (radial in the case of the SVM, and projection in the case of MLP). Finally, the ordinal SVMs produced the best results of all in percentage terms, although there were no significant statistical differences between this technique and the other techniques,

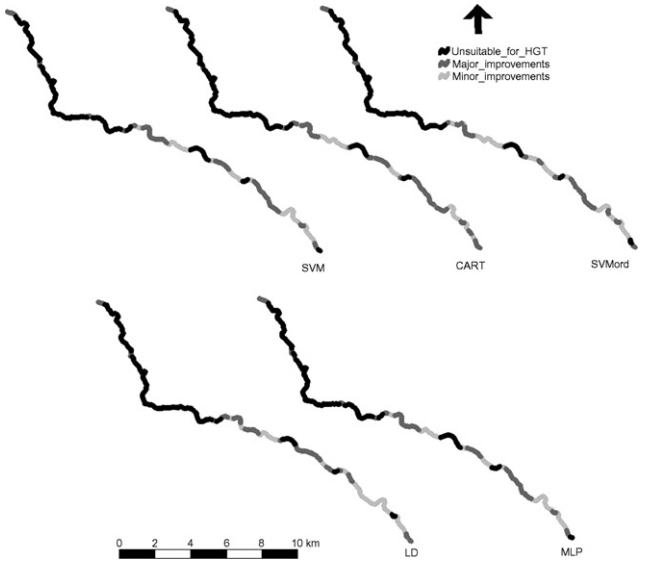

Fig. 1. Map of a stretch of roadway indicating the remedial actions necessary in order to make the stretch suitable for pollutant substances transportation, in accordance with the different techniques used. Progressively darker colours indicate the need for more drastic remedial actions.

Table 1
Parameters used to evaluate the degree of roadway suitability for pollutant substances transportation, with examples of the values assigned to each parameter for five stretches of road

Probability factors																								
Design						Morphology					Signalling		Type of roadwork				Visibility	Condition						
Road width	Lanes width	Slow lanes	Feeder roads	Protective barriers	Ditches	Drainage	Road condition	Slope	Altimetry	Aspect	Exposure to wind	Painted	Lateral	Structures	Embankment	Earthwork	Tunnel		Asphalt	State of conservation			
10	4	0	8	6	6	6	9	9	0	3	0	4	10	10	9	9	10	2	7	5				
10	8	0	8	6	10	6	10	7	0	3	0	4	9	10	7	10	10	4	7	5				
10	6	0	10	8	10	8	2	7	0	3	0	4	10	10	8	8	10	0	9	7				
10	9	0	0	8	10	10	8	7	0	3	0	7	9	8	10	8	10	4	9	7				
10	7	0	10	8	0	0	6	7	0	3	0	7	9	10	8	10	10	2	7	7				
Severity factors																								
Land use											Morphology			Aquifers	Surface hydrology			Road suitability			Utility			
Irrigated/unirrigated land				Forested land		Urban		Indust./extract.		Infrastructures		Natural slope		Stability			River		Surface runoff					
7				10		2		10		10		8		10	6		10		8		3			7.35
10				10		10		10		10		7		7	6		10		7		1			−1.49
10				4		10		10		10		4		5	6		10		5		1			−3.42
10				2		10		10		10		4		5	6		10		10		1			1.37
10				5		0		10		10		4		9	4		10		0		2			4.86

The last column records the utility function.

Table 2
Stretches of roadway for the test sample, classified according to the different models as requiring minor (L) or mayor (M) improvements, or as being unsuitable (H) for pollutant substances transportation

	Model	Linear discriminant analysis			Classification trees			MLP			SVM			Ordinal SVM		
		L	M	H	L	M	H	L	M	H	L	M	H	L	M	H
Expert	L	17	8	0	22	3	0	19	6	0	20	5	0	24	1	0
	M	18	13	3	11	23	0	8	26	0	7	21	4	10	24	0
	H	0	6	71	0	6	71	0	6	71	0	3	74	0	6	71

The expert's classification is given in the rows.

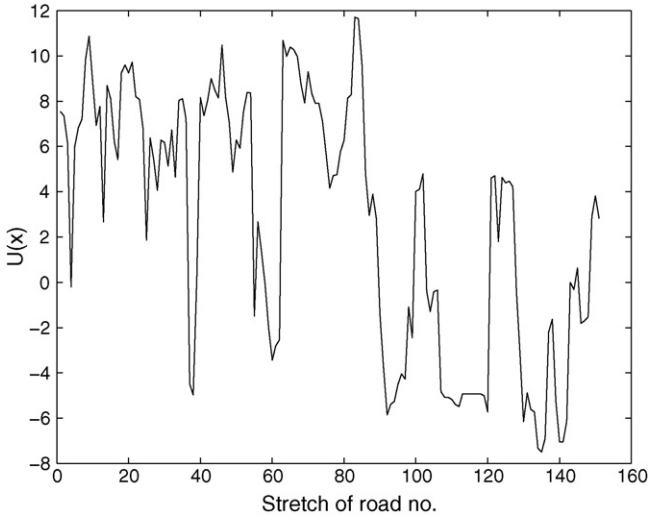

Fig. 2. Utility function for the stretches of road in the training sample (numbered sequentially).

with the exception of linear discriminant analysis. The low level of ordinal SVM error obtained would indicate that the non-linear modelling of the ordinal SVMs is capable of reproducing expert criteria with great accuracy.

Fig. 1 depicts the map of interventions in the road according to the different techniques used. The estimation of the intervals $[\theta(r_{j-1}), \theta(r_j)]$ for the ranks $r_j = 1$–3 (see Eq. (2)) were, respectively: $(-\infty, -5.73]$, $[-5.73, -0.72]$, $[-0.72, \infty]$).

Fig. 2 shows the utility function for each of the 150 stretches of roadway represented in the training sample. This function provides a continuous justification for the expert's opinion in evaluating the stretches of road. Even though the information provided by the expert is discrete in nature ($\{1,2,3\}$), the utility function enables the true intrinsic quality to be assessed for stretches scored by the expert as having the same quality. Thus, as can be seen in Table 1, even when the expert classifies the second to fourth stretches as requiring significant remedial work, the last of these stretches is in better condition than the other two.

4. Conclusions

In this paper we have constructed a model for assessing a road's suitability for pollutant substances transportation that incorporates expert knowledge. The model can be applied on a large scale to other roads without the direct intervention of an expert, although expert supervision would be necessary. To estimate the model, a support vector machine approach applied to ordinal data was compared to linear discriminant analysis and other machine learning classification techniques.

The ordinal SVMs performed similarly to other machine learning techniques, and without increasing the computational burden to any significant extent. Moreover, the ordinal SVMs provided an estimation of both the expert latent utility function and the decision rule used to determine the level of risk for each stretch of roadway. The positive results would demonstrate the benefits of tackling such problems as ordinal regression problems rather than as mere classification problems, which focus on classification errors and fail to penalise inversion in the order of the examples.

References

[1] T. Glickman, E. Erkut, The Tradeoffs Associated with Risk-Conscious Routing of Trains with Hazardous Freight, Research Report 96-2, Department of Finances and Management Science, University of Alberta, Canada, 1996.
[2] P. Cassini, Road transportation of dangerous goods: quantitative risk assessment and route comparison, J. Hazard. Mater. 61 (1998) 133–138.
[3] E. Erkut, V. Verter, A framework for hazardous materials transport risk assessment, Risk Anal. 15 (1995) 589–601.
[4] A. Lovett, J.P. Parfitt, J.S. Brainard, Using GIS in risk analysis: a case study of hazardous waste transport, Risk Anal. 17 (1997) 625–633.
[5] B. Fabiano, F. Currò, E. Palazzi, R. Pastorino, A framework for risk assessment and decision-making strategies in dangerous good transportation, J. Hazard. Mater. 93 (2002) 1–15.
[6] B. Huang, R. Long, Y. Seng, GIS and genetic algorithms for route planning with security considerations, Int. J. Geog. Inform. Sci. 18 (2004) 769–787.
[7] G. Purdy, Risk analysis of the transportation of dangerous goods by road and rail, J. Hazard. Mater. 33 (1993) 229–259.
[8] US Department of Transportation (USDOT), Guidelines for Applying Criteria to Designate Routes for Transportation Hazardous Materials, Federal Highway Administration (FHWA-SA-94-083), Washington, DC, USA, 1994.
[9] R. Martínez-Alegría, C. Ordóñez, J. Taboada, A conceptual model for analyzing the risks involved in the transportation of hazardous goods: implementation in a geographic information system, Human Ecol. Risk Assess. 9 (2003) 857–873.
[10] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, New York, 1998.
[11] L. Breiman, J.H. Friedman, R.A. Olsen, C.J. Stone, Classification and Regression Trees, Chapman & Hall, 1984.
[12] S. Haykin, Neural Networks. A Comprehensive Foundation, Prentice Hall, Englewood Cliffs, New Jersey, 1999.

[13] B. Schölkopf, A.J. Smola, Learning with Kernels, The MIT Press, Cambridge, MA, 2002.
[14] J.M. Matías, A. Vaamonde, J. Taboada, W. González-Manteiga, Support vector machines and gradient boosting for graphical estimation of a slate deposit, J. Stochastic Environ. Res. Risk Assess. 18 (2004) 309–323.
[15] J.M. Matías, A. Saavedra, J. Taboada, C. Ordóñez, Neural networks for modelling the risks involved in the transportation of hazardous goods, Human Ecol. Risk Assess. 12 (2006) 174–191.
[16] R. Herbrich, T. Graepel, K. Obermayer, Large margin rank boundaries for ordinal regression, in: Advances in Large Margin Classifiers, MIT Press, 2000, pp. 115–132.