Accident Analysis and Prevention 82 (2015) 192–198

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Support vector machine in crash prediction at the level of traffic analysis zones: Assessing the spatial proximity effects

Ni Dong, Helai Huang*, Liang Zheng

Urban Transport Research Center, School of Traffic and Transportation Engineering, Central South University, Changsha, Hunan, 410075 PR China

ARTICLE INFO

Article history:
Received 15 December 2014
Received in revised form 22 May 2015
Accepted 26 May 2015
Available online 16 June 2015

Keywords:
Support vector machine
Spatial weight features
CAR model
Correlation-based feature selector

ABSTRACT

In zone-level crash prediction, accounting for spatial dependence has become an extensively studied topic. This study proposes Support Vector Machine (SVM) model to address complex, large and multi-dimensional spatial data in crash prediction. Correlation-based Feature Selector (CFS) was applied to evaluate candidate factors possibly related to zonal crash frequency in handling high-dimension spatial data. To demonstrate the proposed approaches and to compare them with the Bayesian spatial model with conditional autoregressive prior (i.e., CAR), a dataset in Hillsborough county of Florida was employed. The results showed that SVM models accounting for spatial proximity outperform the non-spatial model in terms of model fitting and predictive performance, which indicates the reasonableness of considering cross-zonal spatial correlations. The best model predictive capability, relatively, is associated with the model considering proximity of the centroid distance by choosing the RBF kernel and setting the 10% of the whole dataset as the testing data, which further exhibits SVM models' capacity for addressing comparatively complex spatial data in regional crash prediction modeling. Moreover, SVM models exhibit the better goodness-of-fit compared with CAR models when utilizing the whole dataset as the samples. A sensitivity analysis of the centroid-distance-based spatial SVM models was conducted to capture the impacts of explanatory variables on the mean predicted probabilities for crash occurrence. While the results conform to the coefficient estimation in the CAR models, which supports the employment of the SVM model as an alternative in regional safety modeling.

1. Introduction

Crash prediction model (CPM) is an essential tool in traffic safety analysis. Numerous applications have been developed to evaluate safety level of various types of road entities and to examine effect of safety countermeasures. Recently, traffic crashes are aggregated by a certain spatial scale and researchers usually seek to relate safety to zone-level factors. One of the main objectives of macro-level crash prediction analysis is to explain observed cross-sectional variations in safety using zone-level covariates at different spatial scales (e.g., states, counties, traffic analysis zones, and census wards) (Washington et al., 2006; Quddus 2008; Huang et al., 2010). These macro-level CPMs may aid transportation agencies in more effectively incorporating safety consideration into transportation planning and management (Abdel-Aty et al., 2011; Huang et al., 2013; Xu and Huang, 2015).

In zonal crash prediction, accounting for spatial dependency has become an extensively studied topic. Previous studies (i.e., Quddus 2008; Huang and Abdel-Aty 2010; Siddiqui et al., 2012; Xu et al., 2014; Zeng and Huang, 2014a) found that traffic crashes exhibit extensive spatial dependency across neighboring zones. Research commonly seeks to address the issue of unmeasured spatial correlations using spatial econometric methods among the neighboring spatial units for two reasons: (a) the collection of crash data observations associated with the spatial units does not accurately reflect the nature of the underlying process that generates the sample data, which might induce measurement errors (Anselin, 2001); (b) the spatial dimensions of socio-demographic, economic or regional activities may truly represent an important aspect in model development and may help to improve the accuracy and robustness of crash prediction and avoid underestimation of standard errors for model parameters.

A key challenge associated with the consideration of spatial dependence effects in CPMs is to address the massive amounts of multi-dimensional spatial data found in crash prediction analyses. Specifically, to gain a more precise estimation of the variability in

* Corresponding author.
E-mail addresses: dongni722@foxmail.com (N. Dong), huanghelai@csu.edu.cn (H. Huang), zhengliang@csu.edu.cn (L. Zheng).

http://dx.doi.org/10.1016/j.aap.2015.05.018
0001-4575/

parameters by considering more complex spatial proximity structures, researchers have proposed a comprehensive investigation of different spatially neighboring structures for both road-segment-level and area-wide analyses (i.e., Aguero-Valverde and Jovanis, 2010; Wang et al., 2012). As in the study by Dong et al. (2014), CPMs accounting for spatial correlation perform better than non-spatial model and also model merely considering 0–1 first order adjacency-based proximity structure. A prevalent approach employs Bayesian spatial model with conditional autoregressive prior (i.e., CAR) to address the issue of unmeasured spatial dependences. But it has been claimed to suffer from selected limitations and fail to address complex and highly nonlinear data (the curse of dimensionality) (Karlaftis and Vlahogianni, 2011; Zeng and Huang, 2014b).

Support Vector Machine (SVM), a relatively new modeling technique, is theoretically supposed to be useful and has been employed in several studies (Yu and Abdel-Aty, 2013; Li et al., 2012, 2008; Zhang and Xie, 2008). Yu and Abdel-Aty (2013) constructed SVM models to compare with Bayesian logistic regression model in real-time crash risk evaluation. The better model predictive capability associated with SVM models implies the existence of nonlinear relationship between the dependent variables and explanatory variables which could not be captured by the logistic regression models. Li et al. (2008) investigated the potential of using an SVM model to evaluate safety performance functions for motor vehicle crashes and found that SVM models provide better goodness-of-fit than negative binomial models. It was argued that SVM model has a great ability to address classification problems while producing fewer over-fitting problems and better generalization abilities. The strength of SVM probably comes from its basis on structural risk minimization, which provides a trade-off between hypothesis space complexity and the quality of fitting the training data (Vapnik, 1998). Byvatov et al. (2003) also found that SVMs are able to efficiently address a substantial number of features due to the exploitation of kernel functions, especially for high-dimension data.

Given this new line of research activity, to the best of our knowledge, little to no research has specifically worked on a fairly thorough treatment of SVM in zonal crash prediction accounting for spatial proximity effects. This motivates our interests to fill the gap by utilizing SVM model to explore the spatial proximity effects in crash prediction.

The major challenge associated with the SVM model lies in the optimal input feature subset especially in complex and highly multivariate prediction models because the choice of feature subset influences the appropriate kernel parameters and vice versa (Huang and Wang, 2006). Recent research has postulated that feature selection becomes necessary for machine-learning tasks when working with high-dimension data (Yu and Liu, 2003). Correlation-based Feature Selector (CFS) has been developed for selecting a list of candidate variables in SVMs, which may improve model fitness as well as predictive performance (Hall, 1999). Use of CFS method has greatly expanded the potential applications of the SVM model.

The objective of this study is to explore the possibility of using SVM models and CFS method for macro-level crash frequency analysis with comparatively complex spatial data structure. SVM models with radial-basis function (RBF) kernel and linear kernel are developed. Using a dataset of Hillsborough county of Florida, the model fitness and predictive performance are compared with the CAR models. Moreover, since the SVM is unable to contain a specified function to identify the effects of explanatory variables, a comprehensive sensitivity analysis is carried out to capture the impacts of explanatory variables on the mean predicted probabilities of crash occurrence.

2. Methodology

SVM model can be used to relate various zone-level risk factors to crash occurrence, while accounting for possible spatial proximity among adjacent zones. The spatial weight features are introduced to reflect the overall spatial proximity relationships of the traffic analysis zones (TAZs), which are considered as input vectors into a SVM model in improving the predictive accuracy in this study. For comparison purposes, we also develop CAR model based on the same dataset. They are briefly described in this section, followed by the presentation of the goodness of fit measures for model comparison.

3. SVM model

For this study, the ν-SVM is employed, which has been proposed by Schölkopf et al. (2000). Specifically, the data is separated into a training set and a testing set. The ν-SVM model produces a learning model based on the training set and subsequently makes predictions on the testing set. The ν-SVM model learns the relations between the TAZs-level crash frequency and explanatory variables based on the training dataset.

Assume the training input is defined as vectors $\mathbf{x}(i) \in R^{\text{In}}$ for $i = 1, \ldots, N$, which represents the full set of zone-level contributing factors of each TAZ including road and traffic characteristics, trip production/attraction, and demographic and socioeconomic, and the training output is defined as $\mathbf{y}(i) \in R^1$ for $i = 1, \ldots, N$, which represents the crash frequency that occurred in the TAZ. The ν-SVM maps $\mathbf{x}(i)$ into a feature space $R^h > \text{In}$ with higher dimension using a function $\Phi(\mathbf{x}(i))$ to linearize the nonlinear relation between $\mathbf{x}(i)$ and $\mathbf{y}(i)$. The estimation function of $\mathbf{y}(i)$ is

$$\hat{y} = f(\mathbf{x}) = w^T \Phi(\mathbf{x}) + b$$

where $w \in R^h$ and $b \in R^1$ are coefficients. Schölkopf et al. (2000) showed that the coefficients can be determined by solving the following optimization problem:

$$\text{MinZ}(w, \varepsilon, \xi_i, \xi_i^*) = \frac{1}{2} w^T w + C\{\nu\varepsilon + \frac{1}{N}\sum_{i=1}^{N}(\xi_i + \xi_i^*)\}$$

subject to

$$w^T \Phi(\mathbf{x}(i)) + b - \mathbf{y}(i) \le \varepsilon + \xi_i \forall i = 1, \ldots, N$$

$$\mathbf{y}(i) - w^T \Phi(\mathbf{x}(i)) - b \le \varepsilon + \xi_i^* \forall i = 1, \ldots, N$$

$$\xi_i, \xi_i^* \ge 0 \quad \forall i = 1, \ldots, N$$

$$\varepsilon \ge 0$$

where ξ_i, ξ_i^* are slack variables, C is a regularization parameter, and ν is a second parameter. For each $\mathbf{x}(i)$ the allowable error is ε. Slack variables ξ_i, ξ_i^* capture the errors above ε and are penalized in the objective function via a regularization constant C.

Therefore, the estimated function of $\mathbf{y}(i)$ becomes

$$\hat{\mathbf{y}} = f(\mathbf{x}) = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)\Phi(\mathbf{x}(i))^T \Phi(\mathbf{x}) + b$$
$$= \sum_{i=1}^{N}(\alpha_i^* - \alpha_i) \times K(\mathbf{x}(i), \mathbf{x}(j)) + b$$

where $K(\mathbf{x}(i), \mathbf{x}(j)) = \Phi(\mathbf{x}(i))^T \Phi(\mathbf{x}(j))$ is the kernel function, α_i and α_i^* ares. In this study, the RBF kernel and linear kernel were considered:

$$\text{Radial-basisfunctionkernel}: \quad K(\mathbf{x}(i),\mathbf{x}(j)) = \exp(-\gamma|\mathbf{x}(i)-\mathbf{x}(j)|^2)$$

$$\text{Linear kernel}: \quad K(\mathbf{x}(i),\mathbf{x}(j)) = \mathbf{x}(\mathrm{i})^T \times \mathbf{x}(j)$$

where γ is a parameter. With the radial basis function as the kernel function, the ν-SVM has three parameters (C, ν, γ) that need to be determined. There is always a globally optimal solution to w and b with the input of three parameters (C, ν, γ) (Burges, 2007).

4. SVM model with spatial weight feature

Spatial weight feature is consisted of spatial weights reflecting spatial association of two TAZs, which is specified with input vector into a SVM model. In the present study, four types of spatial weight features are constructed by a treatment of many aspects of spatial dependence including a different choice of spatial weight: based on 0–1 first order adjacency, common boundary length, geometry centroid distance and crash-weighted centroid distance. These spatial weight features are briefly depicted below.

As shown in Fig. 1, w_{ij} denotes the spatial weight between TAZ i $(i = 1, \ldots, N)$ and TAZ j $(j = 1, \ldots, N)$, which reflects the spatial proximity relationships among the TAZs. While swf_j is consisted of the spatial weights w_{ij} between TAZ i and TAZ j. The spatial weight features $(swf_1, swf_2, \ldots, swf_N)$ could be applied to separately represent the spatial weight feature for TAZ 1–N. The spatial weight features for the four cases in Fig. 2(a–d) are denoted as respectively. The centroid-distance based spatial weight feature is tested using an inverse squared distance decay function of the geometric/crash-weighted centroid distance of TAZs.

a) *0–1 first-order, adjacency-based spatial weight feature* (*0-1swf*): If TAZ i and TAZ j share a common border, they are considered neighbors, and as such, w_{ij} equals 1; otherwise, it equals 0.
b) *Common boundary-length-based spatial weight feature* (*CBL swf*): This method refers to the common boundary length of TAZ i and TAZ j.
c) *Geometric centroid-distance-based spatial weight feature* (*GCD swf*): The variable d_{ij} is the distance from the TAZ i' geometric centroid to the TAZ j' geometric centroid.
d) *Crash-weighted centroid-distance-based spatial weight feature* (*CCD swf*): The crash-weighted centroid is defined as the mean center of crash locations in each TAZ, and $d_{ij}{}'$ is the distance from the crash-weighted centroid of TAZ i to that of TAZ j. The weighted centroid is generally called the mean center because it is the average X coordinates and Y coordinates of all crashes in each zone, which reflects the zonal crash clustering effect or spatial orientation of crashes.

5. CAR model

Bayesian spatial models with CAR priors are developed with four types of neighboring structures from TAZs, which depend on the 0–1 first order adjacency, the common boundary length, the geometry centroid distance, and the weighted centroid distance.

$$\begin{array}{c} \\ \text{TAZ 1} \\ \text{TAZ 2} \\ \vdots \\ \text{TAZ } N \end{array} \begin{array}{c} \begin{array}{c}\text{swf}_1\\ \text{TAZ 1}\end{array} \\ \begin{bmatrix} 1 \\ w_{21} \\ \vdots \\ w_{N1} \end{bmatrix} \end{array} \begin{array}{c} \begin{array}{c}\text{swf}_2\\ \text{TAZ 2}\end{array} \\ \begin{bmatrix} w_{12} \\ 1 \\ \vdots \\ w_{N2} \end{bmatrix} \end{array} \cdots \begin{array}{c} \begin{array}{c}\text{swf}_N\\ \text{TAZ } N\end{array} \\ \begin{bmatrix} w_{1N} \\ w_{2N} \\ \vdots \\ 1 \end{bmatrix} \end{array}$$

Fig. 1. Spatial weight features in SVM model.

$$\begin{array}{c} \\ \text{TAZ 1} \\ \text{TAZ 2} \\ \vdots \\ \text{TAZ } N \end{array} \begin{array}{c} \begin{array}{c}\text{swf}_1\\ \text{TAZ 1}\end{array} \\ \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \end{array} \begin{array}{c} \begin{array}{c}\text{swf}_2\\ \text{TAZ 2}\end{array} \\ \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \end{array} \cdots \begin{array}{c} \begin{array}{c}\text{swf}_N\\ \text{TAZ } N\end{array} \\ \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \end{array}$$

(a)

$$\begin{array}{c} \\ \text{TAZ 1} \\ \text{TAZ 2} \\ \vdots \\ \text{TAZ } N \end{array} \begin{array}{c} \begin{array}{c}\text{swf}_1\\ \text{TAZ 1}\end{array} \\ \begin{bmatrix} 1 \\ l_{21} \\ \vdots \\ l_{N1} \end{bmatrix} \end{array} \begin{array}{c} \begin{array}{c}\text{swf}_2\\ \text{TAZ 2}\end{array} \\ \begin{bmatrix} l_{12} \\ 1 \\ \vdots \\ l_{N2} \end{bmatrix} \end{array} \cdots \begin{array}{c} \begin{array}{c}\text{swf}_N\\ \text{TAZ } N\end{array} \\ \begin{bmatrix} l_{1N} \\ l_{2N} \\ \vdots \\ 1 \end{bmatrix} \end{array}$$

(b)

$$\begin{array}{c} \\ \text{TAZ 1} \\ \text{TAZ 2} \\ \vdots \\ \text{TAZ } N \end{array} \begin{array}{c} \begin{array}{c}\text{swf}_1\\ \text{TAZ 1}\end{array} \\ \begin{bmatrix} 1 \\ d_{21} \\ \vdots \\ d_{N1} \end{bmatrix} \end{array} \begin{array}{c} \begin{array}{c}\text{swf}_2\\ \text{TAZ 2}\end{array} \\ \begin{bmatrix} d_{12} \\ 1 \\ \vdots \\ d_{N2} \end{bmatrix} \end{array} \cdots \begin{array}{c} \begin{array}{c}\text{swf}_N\\ \text{TAZ } N\end{array} \\ \begin{bmatrix} d_{1N} \\ d_{2N} \\ \vdots \\ 1 \end{bmatrix} \end{array}$$

(c)

$$\begin{array}{c} \\ \text{TAZ 1} \\ \text{TAZ 2} \\ \vdots \\ \text{TAZ } N \end{array} \begin{array}{c} \begin{array}{c}\text{swf}_1\\ \text{TAZ 1}\end{array} \\ \begin{bmatrix} 1 \\ d_{21}{}' \\ \vdots \\ d_{N1}{}' \end{bmatrix} \end{array} \begin{array}{c} \begin{array}{c}\text{swf}_2\\ \text{TAZ 2}\end{array} \\ \begin{bmatrix} d_{12}{}' \\ 1 \\ \vdots \\ d_{N2}{}' \end{bmatrix} \end{array} \cdots \begin{array}{c} \begin{array}{c}\text{swf}_N\\ \text{TAZ } N\end{array} \\ \begin{bmatrix} d_{1N}{}' \\ d_{2N}{}' \\ \vdots \\ 1 \end{bmatrix} \end{array}$$

(d)

Fig. 2. Different spatial weight features in the SVM model: (a) 0–1 *swf*; (b) CBL *swf*; (c) GCD *swf*; (d) CCD *swf*.

In this study, the basic model structure developed by Besag (1974) is employed,

$$Y_i \sim \text{Poisson}(e_i\theta_i)$$

where for TAZ i $(i = 1, \ldots I)$, Y_i is the number of crashes, θ_i the crash risk, and e_i the crash exposure. The exposure is reflected by the average daily vehicle-miles traveled (DVMT) in an individual TAZ. The log risk is modeled as:

$$\log(\theta_i) = \boldsymbol{\alpha} + x_i'\boldsymbol{\beta} + \delta_i + \vartheta_i$$

where $\boldsymbol{x}_i$ denotes a vector of explanatory variables, or covariates, $\boldsymbol{\beta}$ a vector of fixed effect parameters, δ_i the random effect to account for unstructured over-dispersion error, which is specified via an ordinary exchangeable normal prior,

$$\delta_i \sim N\left(\frac{0,1}{\tau_h}\right)$$

where τ_h is the precision (reciprocal of the variance) of δ_i. Non-informative priors are assigned to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with Normal distribution (0, 1000).

A pair of areas is considered neighboring if they are adjacent. ϑ_i is the spatial correlation term reflecting a shared border, which is specified with a CAR prior as suggested by Besag (1974):

$$\vartheta_i \sim N\left(\frac{\overline{\vartheta}_i, 1}{\tau_i}\right)$$

where

$$\overline{\vartheta}_i = \frac{1}{\sum_{i\neq j}\omega_{ij}}\sum_{i\neq j}\vartheta_i\omega_{ij}$$

and

$$\tau_i = \frac{\tau_c}{\sum_{i\neq j}\omega_{ij}}$$

in which ω_{ij} is entries on the proximity matrix and generally reflects the spatial association of two TAZs, which can be specified by four types of spatial proximity structures as suggested by Dong et al. (2014). τ_c is the precision parameter in the CAR prior, which is assumed to be a prior gamma (0.5, 0.0005).

6. Model evaluation

To evaluate the overall model fitting and predictive performance, the Mean Absolute Deviance (MAD) and Mean Squared Prediction Error (MSPE) are used. The measures of effectiveness (MOEs) can be described as:

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

$$\text{MSPE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

where $\hat{y}_i$ is the predicted crash frequency of TAZ and n is the total number of TAZs. Models associated with lower values of MAD and MSPE fit better to the data.

With regards to the SVM model, the parameter accuracy which is the percent of correct predictions is used to compare the prediction performance of models. The SVM model is fitted using randomly separated fitting and testing datasets with a ratio of 4:1. The fitting data is using to fit the model while the testing dataset is used to evaluate the prediction performance of the model. Specifically, the parameter accuracy is defined as:

$$\text{Accuracy} = \frac{\text{the number of correct predictions}}{I} \times 100\%$$

where I is the number of observations in the testing dataset.

7. Data

The data used to evaluate the models were collected at 738 TAZs located in Hillsborough county in the state of Florida. Four datasets were collected from the Florida Department of Transportation (FDOT) and from the U.S. Census Bureau for a period of three years (2005–2007), including road and traffic characteristics, trip survey data, and demographic and socioeconomic data. The crash data were obtained from the FDOT Crash Analysis Reporting System, which contained a total of 57,694 crashes including fatalities, severe injuries, slightly injuries and Property Damage Only (PDO).

Road and traffic-related data were collected from two sources: FDOT' Roadway Characteristics Inventory and the GIS map of Hillsborough. These included Daily Vehicle Miles Traveled (DVMT), total roadway length, road segment length per TAZ with 25/35/45/55/65 mph speed limits, and intersections. Moreover, a number of trip productions and attractions per day per TAZ were also investigated, which were collected from the Intermodal Systems Development Unit of District 7 of the FDOT. These data sets contained total trip productions/attraction, home-based work productions/attraction, and non-home-base work productions/attraction. The number of demographical and socioeconomic factors was examined including the geographical area of each TAZ, population, and income, which were downloaded from the United States Census report.

ArcGIS 10.0 (ESRI) was used to join layers in the format of a shapefile for the crash data, trip survey data, road network data (95,487 road segments and 9651 intersections), and TAZ-based demographic and socioeconomic data. The 0–1 first-order neighboring matrix was obtained using "Find Adjacent and Neighboring Polygons" in the GIS toolbox. The common boundary lengths of the adjacent TAZs were calculated using the "intersect" command. The "mean center" and "point distance" commands in the spatial analysis toolbox were used to calculate the distances between the geometric centroids and the crash-weighted centroids.

8. CFS-based selection of feature subsets

A CFS method suggested by Hall (1999) has been estimated to do the feature selection work in SVM model. The CFS method is a simple filter algorithm for machine learning that ranks the feature subsets according to a correlation-based heuristic evaluation function. The bias of the evaluation function is toward subsets that contain features that are highly correlated with (predictive of) the class and uncorrelated with each other. The CFS computes a heuristic measure of the "merit" of a feature subset from pair-wise feature correlations and a test theory derived formula. This procedure is conducted in Matlab (The MathWorks, Inc., 2009). Specifically, the heuristic measure of the "merit" is defined as:

$$Ms = k\frac{\overline{r_{fp}}}{\sqrt{k + k(k-1)\overline{r_{pp}}}}$$

where Ms is the heuristic "merit" of a feature subset S containing k features, $\overline{r_{fp}}$ is the mean feature-predict correlation ($f \in S$), and $\overline{r_{pp}}$ is the average feature–feature inter-correlation between inter-features.

This paper presents a measure of correlation between variables, i.e., the maximal information coefficient (MIC), which captures a wide range of associations not limited to specific function types (e.g., linear, exponential, or periodic) or even to all functional relations (Reshef et al., 2011). The MIC is calculated using the R software package in preparation for the CFS selection process.

The final attribute feature selection results are selected for model development, as well as their descriptive statistics, are shown in Table 1. Once the attribute feature subset is decided, the feature subset of SVM models with the four types of spatial weight features proposed are evaluated using the CFS technique. The

Table 2
Feature selection results of SVM models.

Feature subset	Merit
[Selected attribute features]	0.403
[Selected attribute features, 0–1 *swf*]	0.412
[Selected attribute features, CBL *swf*]	0.423
[Selected attribute features, GCD *swf*]	0.431
[Selected attribute features, CCD *swf*]	0.432

Table 1
Descriptions of the variables for modeling (per TAZ in 2005–2007).

Variables	Definition	Mean	S.D.	Min.	Max.
Total crash	Total number of crashes	78.18	72.89	0.00	481.00
DVMT	Daily vehicle miles traveled (in thousands)	95.07	110.24	0.06	788.77
hSeglen	Road segment length with speed limit ≥35 mph	0.45	0.75	0.00	14.47
TOTALP	Total productions	5016.01	4033.40	0.00	28638.00
TOTALA	Total attractions	5393.28	6464.14	0.00	79717.00
NHBWP	Non home-based work productions	296.06	356.76	0.00	3633.00
HU	Area of housing units (in acre)	51.61	50.77	0.00	363.53
MH_INC	Median household income (in thousands)	40.14	20.24	0.00	115.66

results of feature selection are given in Table 2. The subset search results according to the measure of the "merit" clearly indicate that consideration of spatial weigh features in SVM model is helpful on the accuracy of prediction model.

9. Model implementation and result analysis

The CAR model can be efficiently estimated by the freeware WinBUGS package and the car.normal function was used to specify four types of spatial proximity structures. The SVM model was built with the toolbox named LIBSVM in Matlab software (Chang and Lin, 2007). LIBSVM provided a grid search algorithm to determine three parameters (C, v, γ) for SVM model with RBF kernel.

The prediction results in Table 3 show that SVM with RBF kernel models generally outperform the linear SVM models, indicating the better performance of the RBF kernel in terms of the present dataset analysis than the linear kernel. Furthermore, it can be noted that choosing the RBF SVM models and setting 10% of the whole dataset as the testing data produce the better goodness-of-fit than all other cases, which would be employed to carry out the comparison task in the following sub-sections.

The model comparison results by MAD, MSPE and Accuracy clearly indicate that Model 1 (MAD = 0.82, MSPE = 45.23 and Accuracy = 58.4%) unsurprisingly underperforms when compared to the other four spatial models. which implies the positive effect of the consideration of spatial correlations between adjacent zones in improving the prediction accuracy. That is also to say, incorporating the spatial data into the input vectors of SVM models benefits the crash prediction, which was confirmed by the previous results of the CFS feature selection.

In comparison, the spatial weight features in Model 2 merely consider an equal weight for neighboring TAZs, whereas Models 3–5 introduce various ways to measure the exact proximity of zones through common boundary lengths or the distance between centroids. Clearly, the latter three models are more reasonable than Model 2 in terms of how the models are fitted to the specific data structures. This is obviously confirmed by the higher prediction accuracy values from Models 3–5. More interesting, it can also be seen that a gradual increase of prediction accuracy is identified as the complexity of the spatial weight features increased, which implies the advantage of measuring the exact proximity among the zones.

Moreover, Model 3 considers the proximity of the neighboring zones by weighing their common boundary length. Comparatively, the distance-based proximity in Models 4–5 not only account for the proximity of the neighboring zones but also involve the non-neighboring zones. Therefore, the outperformance of the centroid distance-based spatial models (c.f., Table 3) possibly results from that the full consideration of all possible spatial correlations for all zones (i.e., Models 4–5) occupies the better spatial correlation interpretation for the present dataset. Even more, the consideration of crash-weighted centroids in Model 5 would help discover the orientation of crashes or crash hotspots compared with the geometric centroid in Model 4. Unfortunately, its potential advantage is not adequately reflected by the fitting and predicting performance indexed by MAD, MSPE and Accuracy.

Table 4 presents the comparison results of various CAR models, based on which the potential advantage of SVM models is also proved by comparing their fitting results (indicated by MAD and MSPE) for the whole dataset (c.f., Tables 3 and 4). Note that their difference in model specification resides of the best goodness-of-fit in the way in which measurement of cross-zone spatial proximity is account for, where the SVM model considers it in weighting the centroid-distance, while the CAR model considers the common boundary length. This implies that the SVM models are regarded as more flexible compared to the CAR models when modeling complex datasets with possible nonlinearities data. It is again worth noting that, for the specific datasets in the present case study, the SVM models provide a more reasonable spatial crash prediction modeling at TAZ-level, which also supports its advantage in efficiently addressing complex spatial data (Byvatov et al., 2003).

Table 3
Performance comparisons of the SVM models.

		SVM with RBF			SVM with linear		
		MAD	MSPE	Accuracy (%)	MAD	MSPE	Accuracy (%)
Model 1	Training data	0.62	35.89	60.2	0.79	42.56	58.23
	Testing data 1	0.88	48.23	56.2	0.90	53.26	54.31
	Testing data 2	0.82	45.23	58.4	0.89	52.66	56.31
	The whole data	0.67	38.56	–	0.81	46.66	–
Model 2	Training data	0.45	29.11	68.9	0.56	32.12	67.43
	Testing data 1	0.52	32.25	65.3	0.62	35.12	63.78
	Testing data 2	0.51	31.23	66.2	0.61	35.08	63.81
	The whole data	0.47	31.02	–	0.58	34.03	–
Model 3	Training data	0.33	21.23	78.5	0.34	23.16	76.55
	Testing data 1	0.41	26.13	73.2	0.56	29.12	71.56
	Testing data 2	0.37	25.03	74.5	0.55	29.01	72.24
	The whole data	0.38	24.22	–	0.42	29.55	–
Model 4	Training data	0.12	15.13	82.1	0.23	16.23	79.34
	Testing data 1	0.22	17.15	80.9	0.33	19.25	78.45
	Testing data2	0.21	16.55	81.9	0.32	19.13	78.67
	The whole data	0.18	15.24	–	0.27	19.45	–
Model 5	Training data	0.13	15.12	82.5	0.21	17.23	81.23
	Testing data 1	0.21	16.03	81.3	0.31	20.12	80.06
	Testing data 2	0.20	16.24	81.6	0.30	19.14	81.06
	The whole data	0.18	15.77	–	0.29	19.21	–

Note: Model 1: non-spatial SVM model; Model 2: SVM model with 0–1 *swf*; Model 3: SVM model with CBL *swf*; Model 4: SVM model with GCD *swf*; Model 5: SVM model with CCD *swf*; training data: 80% of the whole dataset; testing data 1: 20% of the whole dataset; testing data 2: 10% of the whole dataset.

Table 4
Model comparisons of the CAR models.

Variable	Model 6		Model 7		Model 8		Model 9	
	Mean	95% BCI	Mean	95% BCI	Mean	95% BCI	Mean	95% BCI
DVMT	0.361	(0.358, 0.486)	0.382	(0.351, 0.473)	0.345	(0.324, 0.489)	0.377	(0.343, 0.467)
hSeglen	0.112	(0.009, 0.171)	0.137	(0.021, 0.257)	0.126	(0.016, 0.248)	0.138	(0.013, 0.234)
TOTALP	0.169	(−0.025, 0.347)	0.227	(−0.159, 0.238)	0.146	(−0.078, 0.328)	0.156	(−0.068, 0.326)
TOTALA	0.178	(0.054, 0.198)	0.164	(0.061, 0.179)	0.123	(0.059, 0.163)	0.169	(0.062, 0.179)
NHBWP	−0.143	(−0.185, −0.029)	−0.089	(−0.171, −0.014)	−0.123	(−0.186, −0.022)	−0.121	(−0.178, −0.022)
HU	0.173	(0.106, 0.242)	0.164	(0.099, 0.261)	0.159	(0.078, 0.233)	0.162	(0.091, 0.243)
MH_INC	−0.162	(−0.232, −0.112)	−0.156	(−0.246, −0.086)	−0.154	(−0.362, −0.133)	−0.167	(−0.251, −0.124)
MAD	2.45	–	1.23	–	1.78	–	1.73	–
MSPE	68.56	–	52.13	–	58.36	–	58.45	–

Note: Model 6: CAR model with 0–1 spatial proximity structure; Model 7: CAR model with CBL spatial proximity structure; Model 8: CAR model with GCD spatial proximity structure; Model 9: CAR model with GCD spatial proximity structure.

10. Sensitivity analysis

Sensitivity analysis based on the distance-based spatial SVM was conducted in this study to explore the effects of various explanatory variables. As suggested by Fish and Blodgett (2003), each input variable of the SVM model was changed by a user-defined amount (i.e., standard deviation) while maintaining all other variables as constant. The impacts of each input variable on the mean predicted probabilities for crash occurrence were estimated by one unit change of each input variable. Results of sensitivity analysis are depicted in Fig. 3.

DVMT is found to have the positive effects on crash risk. Similar results are also reported in the study of Huang et al. (2010). With respect to the road length, we divided all the roads into two categories according to the speed limit of 35 mph. The road lengths with speed limit equal to or greater than 35 mph was are positively related to the crash rate. This finding is consistent with the well-known fact that higher speed leads to a substantial increase of

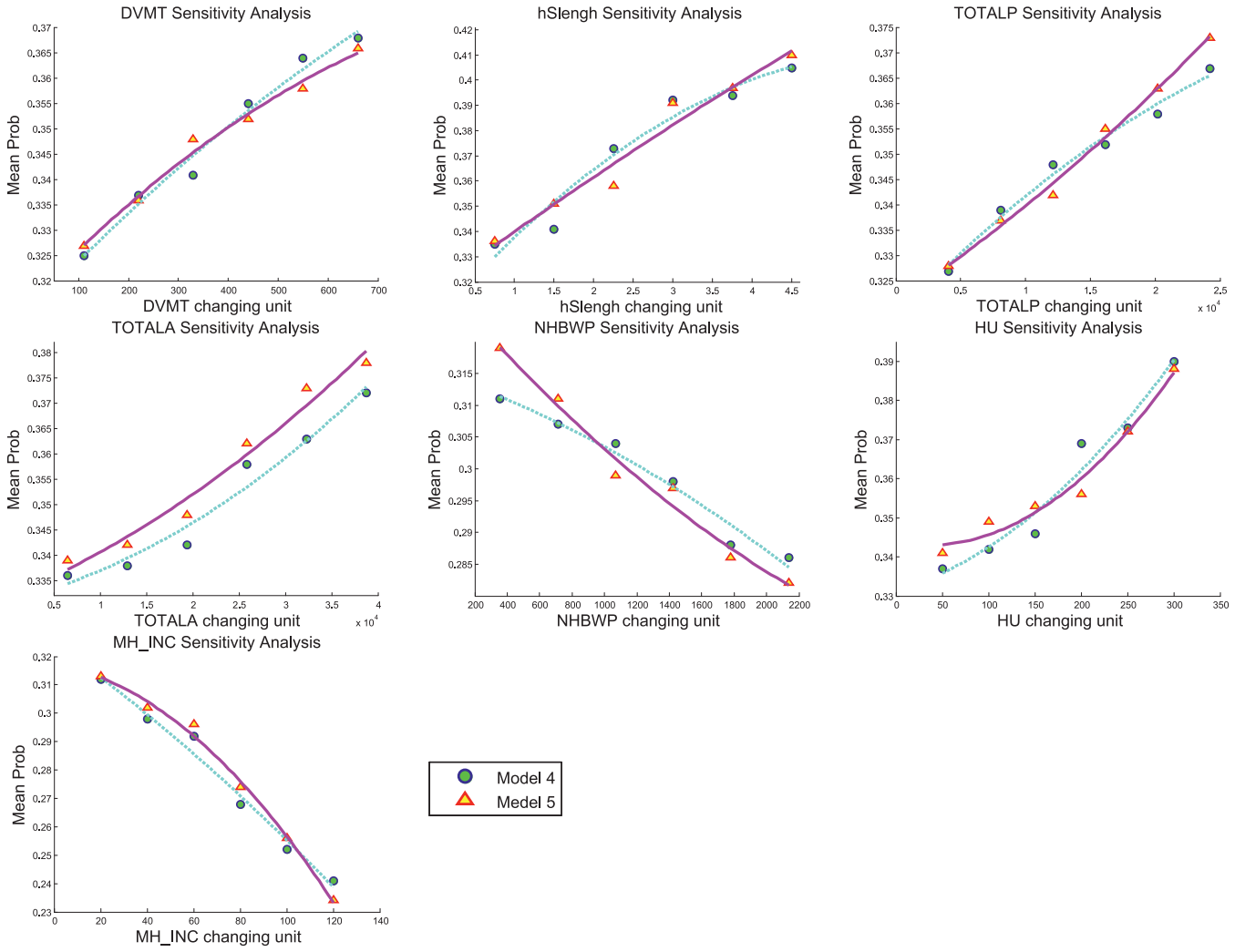

Fig. 3. Sensitivity analysis for the centroid-distance-based spatial SVM models regarding the impacts of variables on crash frequency.

crash risk, especially of severe crashes (Siddiqui et al., 2012). Trip data are well-known as the key determinants in the transportation planning process, which can reflect the population to certain extent (Huang et al., 2013). We used total trip productions, total trip attractions and non-home-base trip productions as the explanatory variables in the models. The total trip productions and attractions are found positively associated with higher crash rate. As expected, more total trip equal more crashes in general. While non-home-base trip productions have a negative effect related to crash risk. Regarding the socio-economic factors, median household income is negative, suggesting that TAZs with a lower median household income are relatively associated with a worse safety situation. The association between median household income and crash risk has been well recognized in previous road safety analysis (Noland and Quddus, 2004; Xu et al., 2014). In addition, area of housing units is found to have positive effects on crash rate, implying more residents in an area have more activities that could results in more traffic crashes.

11. Conclusion

This research proposes SVM to evaluate the crash risk at the TAZ-level by accounting for cross-zonal spatial correlations. Specifically, four types of spatial weight features were conducted based on 0-1 first order adjacency, common boundary-length, geometry centroid-distance and crash-weighted centroid-distance, respectively. Hillsborough data has been selected for model development and comparison with the CAR model. CFS method was employed to select critical variables which contribute to crash occurrence.

The results showed that the SVM models with spatial weight features outperform the non-spatial model in model fitting and predictive performance, which clearly states that the consideration of spatial correlations among adjacent zones is reasonable. The best model predictive capability, relatively, is associated with the model considering proximity of the centroid distance by choosing the RBF kernel and setting the 10% of the whole dataset as the testing data, which further exhibits SVM models' capacity for addressing comparatively complex spatial data in regional crash prediction modeling. Moreover, the goodness-of-fit of SVM models outperforms that of those CAR model with the whole dataset.

Using the sensitivity analysis, the centroid-distance-based spatial SVM models can be used to evaluate the impacts of explanatory variables on crash occurrence. The results of sensitivity analysis are consistent with those of the CAR model and conform to previous studies, which further imply SVM model is an alternative method for regional crash prediction modeling.

Compared to traditional regional crash prediction models, SVM model proposed provides a new perspective at TAZ-level by taking into account varied types of spatial proximity across the adjacent zones. Although the application of SVM models offered positive results, it is suggested to evaluate this kind of model using other datasets to validate the results obtained in further research. Besides, considering the fact that the performance of the SVM model highly depends on the learning procedure which contains function mapping and parameter selection, further efforts are need to tune the scale parameter's value and kernel functions selection for improving model performance.

Acknowledgements

This work was jointly supported by (1) the Natural Science Foundation of China (No. 71371192), (2) the Research Fund for Fok Ying Tong Education Foundation of Hong Kong (142005), and (3) Fundamental Research Funds for the Central Universities of CSU (No. 2014zzts039). The authors would like to thank Dr. Mohamed Abdel-Aty at the University of Central Florida and the Florida Department of Transportation for providing the data.

References

Abdel-Aty, M., Siddiqui, C., Huang, H., Wang, X., 2011. Integrating trip and roadway characteristics in managing safety at traffic analysis zones. Transp. Res. Rec. 2213, 20–28.
Aguero-Valverde, J., Jovanis, P.P., 2010. Spatial correlation in multilevel crash frequency models effects of different neighboring structures. Presented at the 89th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.
Anselin, L., 2001. Spatial Econometrics. Basil Blackwell, Oxford.
Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. 36, 192–236.
Burges, C.J., 2007. A tutorial on support vector machines for pattern recognition research. Microsoft.com/-cburges/papers/SVMTutorial.pdf.
Byvatov, E., Fechner, U., Sadowski, J., Schneider, G., 2003. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. J. Chem. Inf. Comput. Sci. 43, 1882–1889.
Chang, C.-C., Lin, C.-J., 2007. LIBSVM: A Library for Support Vector Machines. www.csie.ntu.edu.tw/_cjlin/libsvm (accessed 16.04.07.).
Dong, N., Huang, H., Xu, P., Ding, Z., Wang, D., 2014. Evaluating spatial proximity structures in crash prediction models at the level of traffic analysis zones. Transp. Res. Rec. 2432, 46–52.
Fish, K.E., Blodgett, J.G., 2003. A visual method for determining variable importance in an artificial neural network model: an empirical benchmark study. J. Target. Meas. Anal. Market. 11, 244–254.
Hall, M.A., 1999. Correlation-Based Feature Selection For Machine Learning. Doctoral dissertation. The University of Waikato.
Huang, H., Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. Accid. Anal. Prev. 42 (6), 1556–1565.
Huang, H., Abdel-Aty, M., Darwiche, A.L., 2010. County-level crash risk analysis in Florida: Bayesian spatial modeling. Transp. Res. Rec. 2148, 27–37.
Huang, H., Xu, P., Abdel-Aty, M., 2013. Transportation Safety Planning: A Spatial Analysis Approach. Presented at the 92th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.
Huang, C.L., Wang, C.J., 2006. A GA-based feature selection and parameters optimization for support vector machines. Expert Syst. Appl. 31, 231–240.
Karlaftis, M.G., Vlahogianni, E.I., 2011. Statistical methods versus neural networks in transportation research: differences, similarities and some insights. Transp. Res. Part C: Emerg. Technol. 19, 387–399.
Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accid. Anal. Prev. 40, 1611–1618.
Li, Z., Liu, P., Wang, W., Xu, C., 2012. Using support vector machine models for crash injury severity analysis. Accid. Anal. Prev. 45, 478–486.
Noland, R.B., Quddus, M.A., 2004. A spatial disaggregate analysis of road casualties in England. Accid. Anal. Prev. 36, 973–984.
Quddus, M.A., 2008. Modeling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. Accid. Anal. Prev. 40, 1486–1497.
Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., Sabeti, P.C., 2011. Detecting novel associations in large data sets. Science 334, 1518–1524.
Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L., 2000. New support vector algorithms. Neural Comput. 12, 1207–1245.
Siddiqui, C., Abdel-Aty, M., Choi, K., 2012. Macroscopic spatial analysis of pedestrian and bicycle crashes. Accid. Anal. Prev. 45, 382–391.
Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York, NY.
Wang, C., Quddus, M.A., Ryley, T., Enoch, M., Davison, L., 2012. Spatial models in transport: a review and assessment of methodological issues. Presented at 91th Annual Meeting of the Transportation Research Board, Washington, D.C., USA.
Washington, S.P., Van Schalkwyk, I., Mitra, S., Meyer, M., Dumbaugh, E., Zoll, M., 2006. Incorporating Safety Into Long-range Transportation Planning. NCHRP Report 546. Transportation Research Board of the National Academics, Washington, D.C.
Xu, P., Huang, H., Dong, N., Abdel-Aty, M., 2014. Sensitivity analysis in the context of regional safety modeling: identifying and assessing the MAUP effects. Accid. Anal. Prev. 70, 110–120.
Xu, P., Huang, H., 2015. Modeling crash spatial heterogeneity: random parameter versus geographically weighting. Accid. Anal. Prev. 75, 16–25.
Yu, L., Liu, H., 2003. Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. 5, 1205–1224.
Yu, R., Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash risk evaluation. Accid. Anal. Prev. 51, 252–259.
Zhang, Y., Xie, Y., 2008. Forecasting of short-term freeway volume with v-support vector machines. Transp. Res. Rec. 2024, 92–99.
Zeng, Q., Huang, H., 2014a. Bayesian spatial joint modeling of traffic crashes on an urban road network. Accid. Anal. Prev. 67, 105–112.
Zeng, Q., Huang, H., 2014b. A stable and optimized neural network model for crash injury severity prediction. Accid. Anal. Prev. 73, 351–358.